Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The CWI is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

CWI Tract 36

Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87)

edited by
A.H.P. van der Burgh
R.M.M. Mattheij

Contributions from the Netherlands

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

ISBN 90 6196 318 4

PREFACE

Through ICIAM 87 four major applied mathematics organisations, GAMM, IMA, SIAM and SMAI, from Germany, Great Britain, North America and France respectively took the initiative of organizing a conference on industrial and applied mathematics with a really international character. When the first ideas were being shaped the organizers became gradually convinced that this event would gain interest and importance when other organisations would be invited to join as associate members. In The Netherlands, being a smaller country with no special organisation for applied mathematicians as such, a Netherlands Recommendation Committee for ICIAM (NRCI) was formed for this purpose, in order to foster the conference locally.

From the large number of reactions to the call for contributions, it became clear that ICIAM 87 promised to become a major event indeed. Therefore the NRCI decided to invite the Dutch contributors to prepare a manuscript and have it published in this volume, which hopefully presents an interesting and fairly representative account of present day research activities in applied and industrial mathematics.

The publication of this volume is motivated not only by the feeling that efforts being put in preparing a conference contribution should be honoured by more than a short talk (the responce of which necessarily being limited) but also by the idea that it would be good to show the liveliness of research in industrial and applied mathematics in The Netherlands, both at universities and at industrial laboratories. Of course the latter fact might be similarly true elsewhere. However for smaller countries it is nearly impossible to organize activities like those at ICIAM on a national basis. Perhaps this is one of the reasons why this conference has attracted such a tremendous interest. It is certainly also a reason to hope that ICIAM will become a regularly organised event.

Finally it is a pleasure to thank the contributors who responded so positively to our request to prepare their paper on such a short notice. We also wish to thank the referees for their valuable advices and their understanding for very tight time constraints. Last but not least we acknowledge the Wiskundig Genootschap for their financial support.

April 1987,

A.H.P. van der Burgh, R.M.M. Mattheij.

TABLE OF CONTENTS

Applied Mathematical Analysis

Scientific Computing

Contol Theory and Signal Processing

Computational Geometry

Applied Probability and Statistics

Mathematics of Natural Sciences

Software and Hardware Aspects

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

The State Space Method in Problems of Analysis

H. Bart
Econometric Institute
Erasmus University
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

I. Gohberg
Department of Mathematical Sciences
The Raymond and Beverly Sackler
Faculty of Exact Sciences
Tel-Aviv University
Ramat-Aviv, Israel
and
M.A. Kaashoek
Department of Mathematics and Computer Science
Vrije Universiteit
P.O. Box 7161, 1007 MC Amsterdam, The Netherlands

ABSTRACT

A review is given of applications of the state space method from systems theory in analysis. The applications concern minimal factorization, Wiener-Hopf integral equations and Szegö limit formulas.

0. INTRODUCTION

Any rational $n \times n$ matrix function $W(\lambda)$, which is analytic at infinity, can be written in the form:

$$W(\lambda) = D + C(\lambda I - A)^{-1}B. \tag{0.1}$$

Here A is a square matrix of which the order m may be different from n. Furthermore, B, C and D are matrices of sizes $m \times n$, $n \times m$ and $n \times n$, respectively. Expressions of the type (0.1) are called *realizations*.

The idea of realization originated in mathematical systems theory and has led to a new approach which is now known as the state space method (see the books [16] and [17]). The representation (0.1) allows one to reduce analytic problems for rational matrix functions to linear algebra problems involving only the four matrices appearing in the realization. This connection is also a natural source of new questions about finite and infinite dimensional operators.

In this paper we review three applications of the state space method in analysis. The problems we shall deal with concern minimal factorization, Wiener-Hopf integral equations and Szegö limit formulas.

A few remarks about notation and terminology: The $n \times n$ identity matrix is denoted by I_n, or simply I. Whenever this is convenient, matrices are identified with linear operators. The null space of a matrix M is denoted by Ker M, the range by Im M. We use M^* for the adjoint (or conjugate transpose) of M and $\bar{\lambda}$ for the complex conjugate of a complex number λ. The symbol $\oplus$ denotes a possibly non-orthogonal direct sum and span V stands for the linear hull of a set V.

1. MINIMAL FACTORIZATION

In this section we consider rational $n \times n$ matrix functions. *We shall always assume that these functions are proper (i.e., analytic at ∞) and have the value I_n at ∞.* Up to a simple Möbius transformation and normalization (to I_n) at ∞, this amounts to requiring that these functions are regular, i.e., they have a determinant that does not vanish identically.

Let $W(\lambda)$ be a rational nxn matrix function. The McMillan degree of $W(\lambda)$, written $\delta(W)$, is the number of poles of $W(\lambda)$ counted according to pole multiplicity. By definition, the pole multiplicity of a pole λ_0 of $W(\lambda)$ is the rank of the block Hankel matrix

$$\begin{bmatrix} W_{-1} & W_{-2} & \cdot & \cdot & \cdot & W_{-p} \\ W_{-2} & W_{-3} & & & W_{-p} & 0 \\ \vdots & & & \cdot^{\cdot^{\cdot}} & & \vdots \\ W_{-p} & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix},$$

where $(\lambda - \lambda_0)^{-1}W_{-1} + \dots + (\lambda - \lambda_0)^{-p}W_{-p}$ is the principal part of the Laurent expansion of $W(\lambda)$ at λ_0.

The McMillan degree is sublogarithmic: If $W(\lambda) = W_1(\lambda)W_2(\lambda)$, then

$$(1.1) \qquad \delta(W) \leqq \delta(W_1) + \delta(W_2).$$

Of special importance are factorizations $W(\lambda) = W_1(\lambda)W_2(\lambda)$ such that in (1.1) equality holds (no "pole/zero cancellations"). These are called minimal factorizations.

PROBLEM 1.1. Describe all possible minimal factorizations of a given rational nxn matrix function $W(\lambda)$.

To deal with this problem, we apply the state space method and write $W(\lambda)$ in the form (0.1) with $D = I_n$, i.e.

$$(1.2) \qquad W(\lambda) = I_n + C(\lambda I_m - A)^{-1}B,$$

where A is an mxm matrix, B is an mxn matrix and C is an nxm matrix. We shall assume that the order m of the matrix A is taken as small as possible, so (1.2) is what is called a minimal realization of $W(\lambda)$. As a matter of fact this amounts to requiring that m is equal to the McMillan degree $\delta(W)$ of $W(\lambda)$. The well-known state space isomorphism theorem asserts that minimal realizations of a given function $W(\lambda)$ are unique up to similarity transformations.

THEOREM 1.2 ([6]). *Let the rational* $n\times n$ *matrix function* $W(\lambda)$ *be given by the minimal realization*

$$W(\lambda) = I_n + C(\lambda I_m - A)^{-1}B,$$

and put $A^\times = A - BC$. *Then there is a one-to-one correspondence between the minimal factorizations* $W(\lambda) = W_1(\lambda)W_2(\lambda)$ *at one hand and the pairs* M, $M^\times$ *of subspaces of* $\mathbb{C}^m$ *satisfying*

$$A[M] \subset M, \quad A^\times[M^\times] \subset M^\times, \quad \mathbb{C}^m = M \oplus M^\times$$

at the other. The correspondence is given by the expressions

$$\begin{aligned} W_1(\lambda) &= I_n + C(\lambda I_m - A)^{-1}(I_m - \Pi)B, \\ W_2(\lambda) &= I_n + C\Pi(\lambda I_m - A)^{-1}B, \\ W_1(\lambda)^{-1} &= I_n - C(I_m - \Pi)(\lambda I_m - A^\times)^{-1}B, \\ W_2(\lambda)^{-1} &= I_n - C(\lambda I_m - A^\times)^{-1}\Pi B, \end{aligned}$$

where Π *is the projection of* $\mathbb{C}^m$ *onto* $M^\times$ *along* M.

We conclude this section with two examples.

EXAMPLE 1.3. Let

$$W(\lambda) = \begin{bmatrix} 1 & \lambda^{-2} \\ 0 & 1 \end{bmatrix}.$$

Then (1.2) is a minimal realization for $W(\lambda)$, provided that one takes

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Clearly $BC = 0$, and so $A^\times = A - BC = A$. The matrix A has only one non-trivial invariant subspace. It follows that $W(\lambda)$ is *irreducible*, i.e., $W(\lambda)$ does not allow for any non-trivial minimal factorization. A similar argument works when in the definition of $W(\lambda)$ the entry λ^{-2} is replaced by λ^{-m}, for any positive integer m.

EXAMPLE 1.4. Let

$$(1.3)\qquad W(\lambda) = \begin{bmatrix} 1 + \dfrac{3}{\lambda^2 + 1} & \dfrac{\lambda^2 + 4}{(\lambda^2 + 1)(\lambda + i)} \\ \dfrac{-1}{\lambda^2 + 1} & 1 - \dfrac{1}{(\lambda^2 + 1)(\lambda + i)} \end{bmatrix}.$$

Then (1.2) is a minimal realization for $W(\lambda)$, provided that one takes

$$A = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & -i \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} -3 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Define M and $M^\times$ by

$$M = \operatorname{span}\left\{\begin{bmatrix} 1 \\ -i \\ 0 \end{bmatrix}\right\}, \quad M^\times = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 2i \\ -i \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right\}.$$

Then M is A-invariant, $M^\times$ is $A^\times$-invariant and $\mathbb{C}^m = M \oplus M^\times$. Here, as usual, $A^\times = A - BC$. The projection Π of $\mathbb{C}^m$ onto $M^\times$ along M is given by

$$\Pi = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3}i & 0 \\ \frac{2}{3}i & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

One can now apply Theorem 1.2 to compute the following non-trivial minimal factorization of $W(\lambda)$:

$$W(\lambda) = \begin{bmatrix} 1 - \dfrac{i}{\lambda - i} & 0 \\ \dfrac{i}{3(\lambda - i)} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 + \dfrac{i}{\lambda + i} & \dfrac{\lambda + 2i}{(\lambda + i)^2} \\ \dfrac{-i}{3(\lambda + i)} & 1 - \dfrac{i}{3(\lambda + i)^2} \end{bmatrix}.$$

References: For details and additional material, see [1], [6], [8], [14] and [23].

2. THE WIENER-HOPF INTEGRAL EQUATION

Consider the (vector-valued) Wiener-Hopf integral equation

$$(2.1)\qquad \phi(t) - \int_0^\infty k(t-s)\phi(s)ds = f(t), \qquad t \geqq 0,$$

with $k \in L_1^{n\times n}(-\infty,\infty)$ and $f, \phi \in L_p^n[0,\infty)$. Here n and p are fixed, $1 \leqq p \leqq \infty$. We say that (2.1) is uniquely solvable $\left(\text{in } L_p^n[0,\infty)\right)$ if for every f in $L_p^n[0,\infty)$, the equation (2.1) has a unique solution ϕ in $L_p^n[0,\infty)$.

PROBLEM 2.1. Give necessary and sufficient conditions for (2.1) to be uniquely solvable and describe the solution.

The problem may be reformulated in terms of the symbol of the equation (2.1), that is the function

$$W(\lambda) = I_n - \int_{-\infty}^{\infty} e^{i\lambda t}k(t)dt.$$

Note that $W(\lambda)$ is a well-defined continuous nxn matrix function on the extended real line. The value of $W(\lambda)$ at infinity is I_n. A (right) canonical Wiener-Hopf factorization of $W(\lambda)$ with respect to the (extended) real line is a factorization

$$(2.2)\qquad W(\lambda) = W_-(\lambda)W_+(\lambda), \qquad \lambda \in \mathbb{R}$$

involving Wiener-Hopf factors having the following properties:

(i) $W_-(\lambda)$ is continuous on the closed lower half plane $\bar{\mathbb{C}}_\infty^-$ (including the point ∞) and analytic on the open lower half plane,

(ii) $W_+(\lambda)$ is continuous on the closed upper half plane $\mathbb{C}_\infty^+$ (including the point ∞) and analytic on the open upper half plane,

(iii) $\det W_-(\lambda) \neq 0,\ \lambda \in \mathbb{C}_\infty^-$, $\qquad \det W_+(\lambda) \neq 0,\ \lambda \in \mathbb{C}_\infty^+$.

It will always be assumed that $W_-(\lambda)$ and $W_+(\lambda)$ are normalized to I_n at ∞. With this extra assumption canonical Wiener-Hopf factorization is unique, provided it exists. We are now ready to present the reformulation of Problem 2.1.

PROBLEM 2.2. *Give necessary and sufficient conditions for the symbol* $W(\lambda)$ *of* (2.1) *to admit canonical Wiener-Hopf factorization and describe the Wiener-Hopf factors.*

The equivalence of Problems 1 and 2 was established by I. Gohberg and M.G. Krein [12].

In the case when $W(\lambda)$ is rational, we can apply the state space method to get the solution of Problems 2.1 and 2.2 explicitly in terms of a realization. Our first theorem concerns Problem 2.2.

THEOREM 2.3 ([1, Section 4.4]). *Let*

$$W(\lambda) = I_n + C(\lambda I_m - A)^{-1}B \tag{2.3}$$

be a realization of $W(\lambda)$ *such that* A *has no eigenvalues on the real line, and put* $A^\times = A - BC$. *Then* $W(\lambda)$ *admits (right) canonical Wiener-Hopf factorization with respect to the real line if and only if*

(1) $A^\times$ *has no eigenvalues on the real line,*

(2) $\mathbb{C}^m = M \oplus M^\times$,

where M *is the (direct) sum of the generalized eigenspaces of* A *corresponding to the eigenvalues of* A *located in the upper half plane and* $M^\times$ *is the (direct) sum of the generalized eigenspaces of* $A^\times$ *corresponding to the eigenvalues of* $A^\times$ *located in the lower half plane.*

In that case the canonical Wiener-Hopf factorization has the form (2.2) *with*

$$\begin{aligned}
W_-(\lambda) &= I_n + C(\lambda I_m - A)^{-1}(I_m - \Pi)B,\\
W_+(\lambda) &= I_n + C\Pi(\lambda I_m - A)^{-1}B,\\
W_-(\lambda)^{-1} &= I_n - C(I_m - \Pi)(\lambda I_m - A^\times)^{-1}B,\\
W_+(\lambda)^{-1} &= I_n - C(\lambda I_m - A^\times)^{-1}\Pi B.
\end{aligned}$$

Here Π *is the projection of* $\mathbb{C}^m$ *onto* M *along* $M^\times$.

Example 1.4 may be seen as an illustration of Theorem 2.3. Indeed, the subspaces M and $M^\times$ featuring there are of the type described in Theorem 2.3 and accordingly the factorization obtained in Example 1.4 is a canonical Wiener-Hopf factorization with respect to the real line.

Theorem 2.3 deals with *right* canonical Wiener-Hopf factorization. An analogous result holds for *left* factorization, where the order of the factors $W_-(\lambda)$ and $W_+(\lambda)$ is interchanged. There are also similar results for contours in the Riemann sphere other than the (extended) real line (cf. the end of Section 3 below).

Next we turn to Problem 2.1. Again there is an explicit solution when the symbol $W(\lambda)$ of the Wiener-Hopf integral equation (2.1) is rational.

THEOREM 2.4 ([1, Section 4.5]). *Assume that the symbol* $W(\lambda)$ *of the Wiener-Hopf integral equation* (2.1) *is rational, and let* (2.3) *be a realization of* $W(\lambda)$ *such that* A *has no poles on the real line. Then* (2.1) *is uniquely solvable in* $L_p^n[0,\infty)$ *if and only if conditions* (1) *and* (2) *of Theorem* 2.3 *are satisfied. In that case the unique solution of* (2.1) *is given by*

$$\phi(t) = f(t) - \int_0^\infty \gamma(t,s)f(s)ds, \quad t \geq 0, \tag{2.4}$$

$$\gamma(t,s) = \begin{cases} -iCe^{-itA^\times}\Pi e^{isA^\times}B, & s < t,\\ iCe^{-itA^\times}(I_m-\Pi)e^{isA^\times}B, & s > t, \end{cases} \tag{2.5}$$

where Π *is the projection of* $\mathbb{C}^m$ *onto* M *along* $M^\times$.

Condition (1) in Theorem 2.3 is equivalent to requiring that $\det W(\lambda) \neq 0$ for λ on the real line. If the symbol $W(\lambda)$ of the equation (2.1) is rational, there does exist a realization (2.3) of $W(\lambda)$ such that A has no eigenvalues on the real line. Indeed, $W(\lambda)$ has no poles on the real line, and for instance any minimal realization will do.

EXAMPLE 2.5. Consider the Wiener-Hopf integral equation (2.1), where

$$k(t) = \begin{bmatrix} -\frac{3}{2}e^t & \frac{3}{4}ie^t \\ \frac{1}{2}e^t & -\frac{1}{4}ie^t \end{bmatrix}, \qquad t < 0,$$

$$k(t) = \begin{bmatrix} -\frac{3}{2}e^{-t} & \frac{7}{4}ie^{-t} + \frac{3}{2}ite^{-t} \\ \frac{1}{2}e^{-t} & -\frac{1}{4}ie^{-t} - \frac{1}{2}ite^{-t} \end{bmatrix}, \qquad t > 0.$$

Then the symbol $W(\lambda)$ is given by (1.3). So (2.3) is a (minimal) realization for $W(\lambda)$, provided that A, B and C are as in Example 1.4. From Theorem 2.4 and the remark made after Theorem 2.3 it is now clear that the integral equation is uniquely solvable in $L_p^2[0,\infty)$. Further, the solution is given by (2.4) with $\gamma(t,s)$ as in (2.5). Note that

$$A^\times = A - BC = \begin{bmatrix} 0 & -1 & 0 \\ 4 & 0 & 0 \\ -1 & 0 & -i \end{bmatrix},$$

and hence

$$e^{-itA^\times} = \begin{bmatrix} \frac{1}{2}e^{2t} + \frac{1}{2}e^{-2t} & \frac{1}{4}ie^{2t} - \frac{1}{4}ie^{-2t} & 0 \\ -ie^{2t} + ie^{-2t} & \frac{1}{2}e^{2t} + \frac{1}{2}e^{-2t} & 0 \\ \frac{1}{6}ie^{2t} - \frac{1}{2}ie^{-2t} + \frac{1}{3}ie^{-t} & -\frac{1}{12}e^{2t} - \frac{1}{4}e^{-2t} + \frac{1}{3}e^{-t} & e^{-t} \end{bmatrix}.$$

A similar expression holds of course for $e^{isA^\times}$: just replace t by $-s$. Since the projection Π was already described in Example 1.4, all ingredients for calculating $\gamma(t,s)$ explicitly are available. We leave the final matrix computations to the reader.

References: For details and additional material (also on non-canonical Wiener-Hopf factorization, singular integral equations, the Riemann-Hilbert boundary value problem and the infinite dimensional case), see [1] - [5] and [9].

Consider the situation of Theorem 2.3, and assume in addition that for each real λ the matrix $W(\lambda)$ is positive definite, i.e.,

$$x^*W(\lambda)x > 0, \quad -\infty < \lambda < \infty; \quad 0 \neq x \in \mathbb{C}^n.$$

It is well-known that under these conditions the matrix function $W(\lambda)$ admits canonical Wiener-Hopf factorization. Let us explain how in this context the state space method works.

First observe that $W(\lambda) = W(\bar{\lambda})^*$. So besides the realization (2.3), we have

$$(2.6) \qquad W(\lambda) = I_n + B^*(\lambda I_m - A^*)^{-1}C^*.$$

Assume now that (2.3) is minimal. Then (2.6) is a minimal realization too, and the state space isomorphism theorem guarantees the existence of a unique matrix H such that

$$(2.7) \qquad A^* = HAH^{-1}, \quad C^* = HB, \quad B^* = CH^{-1}.$$

Taking adjoints and using the uniqueness of H, one sees that H is selfadjoint (cf. [7] and [10]). The first identity in (2.7) means that A is H-selfadjoint, i.e., A is selfadjoint with respect to the (possibly) indefinite inner product x^*Hy induced by H.

Next we follow A.C.M. Ran [19]. Observe that (2.7) implies that $A^\times$ is H-selfadjoint too. So

$$(2.8) \qquad A^*H = HA, \quad (A^\times)^*H = HA^\times.$$

Here, as usual, $A^\times = A - BC$. Since $\det W(\lambda) \neq 0$ for all real λ, we have that $A^\times$ has no eigenvalues on the real line. Define M and $M^\times$ as in Theorem 2.3. The identities (2.8) yield

$$(2.9) \qquad H[M] = M^\perp, \quad H[M^\times] = (M^\times)^\perp,$$

where $M^\perp$ and $(M^\times)^\perp$ are the orthogonal complements of M and $M^\times$, respectively. In other words M and $M^\times$ are H-neutral. Take $x \in M \cap M^\times$. Then $Ax \in M$, and so $x^*HAx = 0$. Similarly $x^*HA^*x = 0$. This leads to $x^*HBCx = 0$. Now $HB = C^*$, and we get $x^*C^*Cx = 0$. Thus $Cx = 0$ and $Ax = A^\times x \in M \cap M^\times$. Hence $M \cap M^\times$ is A-invariant and $M \cap M^\times \subset \operatorname{Ker} C$. Since (2.3) is minimal, the largest A-invariant subspace contained in Ker C is (0). So $M \cap M^\times = (0)$.

From (2.9) it is clear that $\dim M = \dim M^\perp$ and $\dim M^\times = \dim(M^\times)^\perp$. But then $\dim M = \dim M^\times = \frac{1}{2}m$. Together with $M \cap M^\times = (0)$, this gives $\mathbb{C}^m = M \oplus M^\times$. So our hypotheses imply that condition (2) of Theorem 2.3 is satisfied ("automatic matching"). As a result, we obtain a canonical Wiener-Hopf factorization involving factors $W_-(\lambda\ )$ and $W_+(\lambda)$ as described in the theorem. Analysis of the factors reveals that the factorization is symmetric, i.e., $W_-(\lambda) = W_+(\bar{\lambda})^*$.

References: For additional material (also on symmetric non-canonical Wiener-Hopf factorization), see [9] and [13].

Another case of "automatic matching" (but then in an infinite dimensional context) appears in linear transport theory (see [1, Ch. 6], [15] and [18]).

3. SZEGÖ LIMIT FORMULAS

Let $W(\lambda)$ be a continuous nxn matrix function on the unit circle $|\lambda| = 1$, and introduce

$$D_k(W) = \det([W_{i-j}]_{i,j=1}^k),$$

where $\ldots, W_{-2}, W_{-1}, W_0, W_1, W_2, \ldots$ are the (matrix) Fourier coefficients of $W(\lambda)$. Under additional assumptions on $W(\lambda)$, the strong

Szegö limit theorem holds true, that is

$$(3.1)\qquad \Sigma_2(W) = \lim_{k\to\infty} \frac{D_k(W)}{\Sigma_1(W)^{k+1}}$$

exists. Here

$$(3.2)\qquad \Sigma_1(W) = \exp[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det W(e^{it})dt].$$

The additional assumptions alluded to include

$$(3.3)\qquad \det W(e^{it}) \neq 0, \quad -\pi \leqq t \leqq \pi,$$

$$(3.4)\qquad \arg\det W(e^{it})\Big|_{t=-\pi}^{\pi} = 0,$$

i.e., the winding number of the curve $W(e^{it})$, $-\pi \leqq t \leqq \pi$ with respect to the origin is zero. Under these circumstances, the right hand side of (3.2) is well-defined.

PROBLEM 3.1. *Describe, for a rational* n×n *matrix function* $W(\lambda)$ *having no poles on the unit circle, the Szegö constants* $\Sigma_1(W)$ *and* $\Sigma_2(W)$ *in terms of a realization.*

We shall assume that $W(\lambda)$ is normalized to I_n at the origin. Thus $W(\lambda)$ can be written as

$$(3.5)\qquad W(\lambda) = I_n + \lambda C(I - \lambda A)^{-1}B,$$

where A has no eigenvalues on the unit circle. Condition (3.3) then amounts to requiring that $A^\times = A - BC$ has no eigenvalues on the unit circle too.

THEOREM 3.2 ([11]). *Let the rational* n×n *matrix function* $W(\lambda)$ *be given by* (3.5), *and suppose that neither* A *nor* $A^\times$ *has eigenvalues on the unit circle. Assume, in addition, that* (3.4) *is satisfied. Then the limit in the right hand side of* (3.1) *exists, and the Szegö constants* $\Sigma_1(W)$ *and* $\Sigma_2(W)$ *are given by*

$$\Sigma_1(W) = \frac{\det(I_m - P^\times + A^\times P^\times)}{\det(I_m - P + AP)},$$

$$\Sigma_2(W) = \det\big(PP^\times + (I_m - P)(I_m - P^\times)\big),$$

where the Riesz projections P *and* $P^\times$ *are defined by*

$$P = I_m - \frac{1}{2\pi i} \int_{|\lambda|=1} (\lambda I_m - A)^{-1} d\lambda,$$

$$P^\times = I_m - \frac{1}{2\pi i} \int_{|\lambda|=1} (\lambda I_m - A^\times)^{-1} d\lambda.$$

Here as usual $A^\times = A - BC$.

Put $V = PP^\times + (I_m - P)(I_m - P^\times)$, where P and $P^\times$ are as in Theorem 3.2. Then $\Sigma_2(W) = \det V$. Note that $\det V \neq 0$ if and only if

$$\mathbb{C}^m = \operatorname{Ker} P \dotplus \operatorname{Im} P^\times, \quad \mathbb{C}^m = \operatorname{Ker} P^\times \dotplus \operatorname{Im} P.$$

Hence $\Sigma_2(W) \neq 0$ if and only if $W(\lambda)$ admits both left and right canonical Wiener-Hopf factorization with respect to the unit circle (cf. Section 2). In that case (3.1) implies another Szegö limit formula, namely

$$\text{(3.6)} \qquad \lim_{k\to\infty} D_k(W)^{\frac{1}{k}} = \Sigma_1(W).$$

By way of illustration, we present a simple example.

EXAMPLE 3.3. Consider the (scalar) rational function

$$W(\lambda) = \frac{\lambda^2 + \frac{5}{2}\lambda + 1}{\lambda^2 - \frac{5}{2}\lambda + 1}.$$

This function can be written in the form (3.5) with

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} -\frac{5}{3} & \frac{20}{3} \end{bmatrix}.$$

We now have

$$A^{\times} = A - BC = \begin{bmatrix} \frac{13}{6} & \frac{-20}{3} \\ \frac{5}{3} & \frac{-14}{3} \end{bmatrix}$$

and

$$P = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^{\times} = \begin{bmatrix} \frac{-16}{9} & \frac{40}{9} \\ \frac{-10}{9} & \frac{25}{9} \end{bmatrix},$$

where P and $P^{\times}$ are as in Theorem 3.2. It follows that

$$\Sigma_1(W) = -1, \quad \Sigma_2(W) = \frac{25}{9}.$$

Since $\Sigma_2(W) \neq 0$, the identity (3.6) holds true.

References: For details and additional material (also on the Kac-Achiezer formulas, the continual analogues of the Szegö formulas), see [11], [20] - [22] and [24] - [26].

REFERENCES

[1] H. BART, I. GOHBERG and M.A. KAASHOEK, Minimal Factorization of Matrix and Operator Functions, Operator Theory: Advances and Applications, Vol. 1, Birkhäuser Verlag, Basel etc., 1979.

[2] H. BART, I. GOHBERG and M.A. KAASHOEK, Wiener-Hopf integral equations, Toeplitz matrices and linear systems, in Toeplitz Centennial, Operator Theory: Advances and Applications, Vol. 4 (I. Gohberg, Ed.), Birkhäuser Verlag, Basel etc. 1982, pp. 85-135.

[3] H. BART, I. GOHBERG and M.A. KAASHOEK, Convolution equations and linear systems, Integral Equations and Operator Theory 5: 283-340 (1982).

[4] H. BART, I. GOHBERG and M.A. KAASHOEK, Fredholm theory of Wiener-Hopf equations in terms of realization of their symbols, Integral Equations and Operator Theory 8: 590-613 (1985).

[5] H. BART, I. GOHBERG and M.A. KAASHOEK, *Wiener-Hopf factorization, inverse Fourier transforms and exponentially dichotomous operators*, J. Functional Analysis 68: 1-42 (1986).

[6] H. BART, I. GOHBERG, M.A. KAASHOEK and P. VAN DOOREN, *Factorizations of transfer fuctions*, Siam J. Control Optimization 18: 675-696 (1980).

[7] R.W. BROCKETT and R.A. SKOOG, *A new perturbation theory for the synthesis of nonlinear networks*, Mathematical Aspects of Electrical Networks Analysis, IAM-AMS Proceedings V3: 17-33 (1971).

[8] N. COHEN, *On minimal factorizations of rational matrix functions*, Integral Equations and Operator Theory 6: 647-671 (1983).

[9] *Constructive Methods of Wiener-Hopf Factorization*, Operator Theory: Advances and Applications, Vol. 21 (I. Gohberg and M.A. Kaashoek, Eds.), Birkhäuser Verlag, Basel etc., 1986.

[10] P.A. FUHRMANN, *On symmetric rational transfer functions*, Lin. Alg. Appl. 50: 167-250 (1983).

[11] I. GOHBERG, M.A. KAASHOEK and F. VAN SCHAGEN, *Szegö-Kac-Achiezer formulas in terms of realizations of the symbol*, Wiskundig Seminarium der Vrije Universiteit, Rapport no. 310, Amsterdam, 1986; to appear in J. Functional Analysis.

[12] I. GOHBERG and M.G. KREIN, *Systems of integral equations on a half line with kernels depending on the difference of arguments*, Uspehi Mat. Nauk 13, no. 2 (80): 3-72 (1958); English transl. Math. Soc. Transl. (2) 14: 217-287 (1960).

[13] I. GOHBERG, P. LANCASTER and L. RODMAN, *Matrices and Indefinite Scalar Products*, Operator Theory: Advances and Applications, Vol. 8, Birkhäuser Verlag, Basel etc., 1983.

[14] I. GOHBERG, P. LANCASTER and L. RODMAN, *Invariant Subspaces of Matrices with Applications*, Canadian Mathematical Society series of monographs and advanced texts, Wiley-Interscience, New York etc., 1986.

[15] W. GREENBERG, C.V.M. VAN DER MEE and V. PROTOPOPESCU, *Boundary Value Problems in Abstract Kinetic Theory*, Operator Theory: Advances and Applications, Vol. 23, Birkhäuser Verlag, Basel etc., 1986.

[16] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

[17] R.E. KALMAN, P.L. FALB and M.A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[18] H.G. KAPER, C.G. LEKKERKERKER and J. HEJTMANEK, *Spectral Methods in Linear Transport Theory*, Operator Theory: Advances and Applications, Vol. 5, Birkhäuser Verlag, Basel etc., 1982.

[19] A.C.M. RAN, *Minimal factorization of selfadjoint rational matrix functions*, Integral Equations and Operator Theory 5: 850-869 (1982).

[20] G. SZEGÖ, *Ein Grenzwertsatz über der Toeplitzen Determinanten einer reellen positiven Function*, Math. Ann. 76: 490-503 (1915).

[21] G. SZEGÖ, *Beiträge zur Theorie der Toeplitzen Formen*, Math. Z. 6: 167-202 (1920).

[22] G. SZEGÖ, *On certain Hermitian forms associated with the Fourier series of a positive function*, Commun. du Sèminaire Math. de l'Univ. de Lund, Festschrift Marcel Riesz, Lund, 1952, 228-238.

[23] P. VAN DOOREN, *Factorization of a rational matrix: the singular case*, Integral Equations and Operator Theory 7: 704-741 (1984).

[24] H. WIDOM, *Asymptotic behavior of block Toeplitz matrices and determinants*, Adv. in Math. 13: 284-322 (1974).

[25] H. WIDOM, *On the limit of block Toeplitz determinants*, Proc. Amer. Math. Soc. 50: 167-173 (1975).

[26] H. WIDOM, *Asymptotic behavior of block Toeplitz matrices and determinants*, II, Adv. in Math. 21: 1-29 (1976).

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Discretizations Conserving Energy and Other Constants of the Motion

F.P.H. van Beckum and E. van Groesen
Department of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

Various evolution equations from mathematical physics conserve one or more integrals (constants of the motion; e.g. the energy) and have solutions in the form of steadily propagating waves (e.g. solitairy waves).
In spatial discretizations these properties are generally lost. However, observing that the properties are a consequence of a certain variational structure (Poisson structure) of the evolution equation, we derive discretizations in such a way that they inherit this structure. Consequently the constants of the motion and the existence of steadily propagating waves are conserved.
Calculations are shown for the Korteweg-de Vries equation as an example.

Introduction

Many numerical methods for evolutionary partial differential equations treat the discretization in space and that in time as two separate and consecutive operations. We will focus in this paper on the spatial discretization. Generally speaking this leads to a system of ordinary differential equations in time. Specifically, let the partial differential equation be given as

$$(1) \qquad \partial_t u = Ku$$

where K is some (non-)linear mapping in an infinite dimensional space V, and u(t) ∈ V. A spatial discretization of (1) will be based on approximating the elements of V, i.e. on a projection $u \rightarrow \sum_1^N \hat{u}_k s_k$ of V, onto a N-dimensional subspace $B = [s_1, \ldots, s_N]$, through which the problem can be reformulated in B, or, equivalently, in the space $\hat{V}$ of N-dimensional vectors $\hat{u}$ representing elements of B. With an appropiate choice of a mapping $\hat{K} : \hat{V} \rightarrow \hat{V}$ the system of ordinary differential equations

$$(2) \qquad d_t \hat{u} = \hat{K}\hat{u}$$

may be, in a certain sense, an approximation to (1) provided, of course, that requirements concerning numerical consistency, stability, etcetera, are fulfilled.

In this paper we consider a class of evolution equations (1) that has a specific structure. As a consequence special properties will hold, such as (a) the existence of constants of the motion and (b) a simple characterization of steady states (travelling waves). The aim is to exploit the specific structure to find spatial discretizations that satisfy the additional requirement that (2) conserves as much as possible these properties, thereby anticipating that conservation of these properties does not decrease the numerical consistency for arbitrary solutions, while improving specific features such as travelling waves.
We will outline and demonstrate these ideas for a class of partial differential equations of the form

$$(3) \qquad \partial_t u = \partial_x \frac{\delta P}{\delta u} (u)$$

where u: $[-\pi,\pi] \rightarrow \mathbb{R}$ is periodic, and P is a translation invariant functional on V. (The extension to vector valued functions u is obvious.) As a particular example we will take the Korteweg-de Vries equation:

$$(4) \qquad u_t + u_{xxx} + 6uu_x = 0$$

which is of the form (3) with

(5) $$P(u) = \frac{1}{2}\int u_x{}^2\, dx - \int u^3 dx.$$

The translation invariance that is present in (3) is fundamental for many of the specific properties mentioned. The aim to preserve this invariance almost naturally leads us to consider discretizations that are based on trigonometric approximation of functions. Although a formulation in the spectral plane could equally well be used, we present a translation invariant discretization in the physical plane, i.e. based on nodal values, facilitating a possible comparison with finite difference methods.

Remark 1. We assume that the functional P is such that (3) admits only smooth solutions, i.e. we consider dispersive wave equations, excluding equations with shock wave solutions.
Remark 2. We will be very concise here; a more elaborate treatment is given in Van Groesen & Van Beckum 1987.
Remark 3. We are not going into time integration methods; for integration of systems of ODEs with conservation properties see e.g. Baumgarte 1973, Shampine 1984, Feng 1986 and Gear 1986.

Features of the continuous equation

In order to specify the particular properties of the continuous equation (3), we will first show that it admits (at least) three constants of the motion, and examine the underlying structure that leads to their conservation. The three conserved quantities are (with a name that is justified for some specific applications):

(6) the total mass $M(u) = \int u,$

(7) the momentum $C(u) = \frac{1}{2}\int u^2,$

(8) the energy $P(u).$

The conservation of M is a consequence of the fact that the operator ∂_x is an anti-symmetric mapping (on periodic functions) and that $\partial_x 1 = 0$. Indeed, writing $<\ ,\ >$ for the usual L_2-innerproduct:

$$\partial_t M(u) = \int u_t = < 1, \partial_x \frac{\delta P}{\delta u} > = -< \partial_x 1, \frac{\delta P}{\delta u} > = 0.$$

For the conservation of P the essentialities are both the anti-symmetry of ∂_x and the appearance of the variational derivative $\frac{\delta P}{\delta u}$ in the right-handside of (3):

$$\partial_t P(u) = < \frac{\delta P}{\delta u}, u_t > = < \frac{\delta P}{\delta u}, \partial_x \frac{\delta P}{\delta u} > = 0.$$

The conservation of C is a consequence of the translation invariance of P. More precisely, with T_ε the shift operator in V, defined by

(9) $\qquad (T_\varepsilon u)(x) = u(x+\varepsilon) \qquad$ for $u \in V$

we suppose that

(10) $\qquad P(T_\varepsilon u) = P(u) \qquad$ for all $u \in V$ and all $\varepsilon \in \mathbb{R}$.

For density functionals, like (5), this simply means that the density does not depend explicitely on x. By differentiating (10) with respect to ε at $\varepsilon = 0$, it follows that

$$< \frac{\delta P}{\delta u}, \partial_x u > = 0 \qquad \text{for all } u \in V.$$

This leads to the conservation of C:

$$\partial_t C(u) = < u, \partial_t u > = < u, \partial_x \frac{\delta P}{\delta u} > = - < \frac{\delta P}{\delta u}, \partial_x u > = 0.$$

Remark 1. For linear equations, with $\frac{\delta P}{\delta u} = Lu$ for some symmetric operator L, any quadratic density functional (u,Qu) for which the symmetric operator Q commutes with $\partial_x L$ is conserved.

Remark 2. The structure, based on the form of the differential equation (3) and the translation invariance of the functional P, can be recognized as a Poisson structure. In this respect M and C are Casimir functionals, i.e. conserved functionals that only regard the kinematics of the system (independent of the specific choice of the functional P). See Van Groesen & Van Beckum 1987.

The existence of three conserved quantities (which are independent in interesting cases) restricts the dynamics to a space of codimension 3, which, in an infinite dimensional space V, is only a very limited gain in this sense. However, a restricted, but usually interesting, class of solutions is completely determined by these conserved quantities and can be obtained by looking for critical points of one of the functionals on level sets of the others. For instance, for a rather large class of functionals P, the constrained minimization problem

$$\text{(11)}\qquad \min \{P(u) \mid M(u) = m,\ C(u) = c\}$$

has a solution, ϕ say. From the translation invariance of M, C and P it then follows that $T_\varepsilon \phi$ is also a solution for any $\varepsilon \in \mathbb{R}$. These functions $T_\varepsilon \phi$ satisfy the equation

$$\text{(12)}\qquad \frac{\delta P}{\delta u}(\phi) = \lambda \frac{\delta C}{\delta u}(\phi) + \mu = \lambda\phi + \mu,$$

where λ and μ are real Lagrange multipliers, arising as a consequence of the constraints. A translation of a specific solution ϕ uniform in time with velocity $-\lambda$,

$$\text{(13)}\qquad U(x,t) := T_{\lambda t}\phi(x) = \phi(x+\lambda t)$$

then satisfies

$$\partial_t U(x,t) = \lambda\, \partial_x \phi(x+\lambda t) = \partial_x \frac{\delta P}{\delta u}(\phi(x+\lambda t))$$

i.e. it is a travelling wave solution of (3), propagating undisturbed in shape with velocity $-\lambda$.

Remark. From general variational theory it is known that for the minimum value of P in (11), considered as a function of c at constant m, the derivative with respect to c is equal to the multiplier $\lambda = \lambda(c)$.

Summarizing, we can say that the presence of translation invariance in equation (3) provides the conserved quantity C, with the aid of which travelling wave solutions of (3) can be found from the variational characterization (11).
Specifically, for the Korteweg-de Vries equation, ϕ is a solution of the eigenvalue problem

$$-\phi_{xxx} - 6\phi\phi_x = \lambda\phi_x;$$

in the periodic case it is known as a cnoidal wave; see e.g. Whitham.

Discretization

Like in finite difference methods, we will describe a state by function values at a given set of grid points x_j over the interval $[-\pi,\pi]$. The concept of translation invariance, however, requires some specification of the behaviour in between these discrete values, i.e. we have to interpolate with a given set of continuous functions. Trigonometric interpolation will turn out to be most natural. We start however with a general definition of discretization.
Given a subspace B of V, spanned by a finite number of basefunctions $B = [s_1,\dots,s_N]$, we conceive a discretization as a projection of V on B: $u \to \sum_1^N \hat{u}_k s_k$, or equivalently, as the mapping $u \to \hat{u}$ of V on $\hat{V}$, where $\hat{V}$ is the N-dimensional Euclidian space of vectors $(\hat{u}_1,\dots,\hat{u}_N)$. If discretization is done by collocation, e.g. with splines or trigonometric functions, $\{\hat{u}_k\}$ is to be solved from $u(x_j) = \sum_k \hat{u}_k s_k(x_j)$. In the special case that s_k is chosen to satisfy $s_k(x_j) = \delta_{kj}$ (Kronecker delta), $\hat{u}_k$ coincides with the function value $u(x_k)$.

Remark. In a certain sense finite difference methods can be included in this treatment; although in that case the introduction of a subspace B is ambiguous, the mapping $V \to \hat{V}$ still applies.

Guided by the aim to discretize (3) in such a way that the desired properties are conserved, we will reformulate the previous section within the framework of the space $\hat{V}$.

In order to arrive at an equation of which functionals like M and P are conserved, it is required to take a discretization for ∂_x, D say, operating in $\hat{V}$, with the properties: (1) $D\hat{1} = 0$ (where $\hat{1}$ is the discretization of $u = 1$), and (2) D is antisymmetric. Moreover, the functional derivative $\frac{\delta P}{\delta u}$ has to be discretised in such a way that it is again a variational derivative of some function $\hat{P}$ of $\hat{u}$ i.e. a gradient with respect to the innerproduct $< \, , \, >$ in $\hat{V}$ defined by restriction of the L_2-innerproduct of the infinite dimensional space V. The resulting equation then reads:

$$(14) \qquad d_t\hat{u} = D\,\frac{\delta\hat{P}}{\delta\hat{u}} \quad ,$$

and it is easily verified that for (14) the functions

$$(15) \qquad \hat{M}(\hat{u}) := < \hat{u}, \hat{1} >$$

and

$$(16) \qquad \hat{P}(\hat{u})$$

are conserved. Indeed:

$$d_t\hat{M}(\hat{u}) = < d_t\hat{u}, \hat{1} > = < D\frac{\delta\hat{P}}{\delta\hat{u}}\,(\hat{u}),\ 1 > = -< \frac{\delta\hat{P}}{\delta\hat{u}}(\hat{u}),\ D\hat{1} > = 0$$

and

$$d_t\hat{P}(\hat{u}) = < \frac{\delta\hat{P}}{\delta\hat{u}}(\hat{u}),\ d_t\hat{u} > = < \frac{\delta\hat{P}}{\delta\hat{u}}(\hat{u}),\ D\frac{\delta\hat{P}}{\delta\hat{u}}(\hat{u}) > = 0$$

Note that, up to now, we have "derived" the equation (14) only by analogy with equation (3), and no arguments about numerical consistency have been invoked. Of course, it is anticipated that if $\hat{P}$ is chosen in a proper way, i.e. related to P for the given discretization of functions u, then (14) will be a numerically consistent approximation of (4). Before entering into the arguments of conserving a functional $\hat{C}$ and consequently the existence of discrete travelling waves for (14), let us specify the ideas so far in an example.

Example. Let us see how various discretizations work out on the simple linear evolution equation

$$(17) \qquad \partial_t u = \partial_x u$$

where, as throughout this paper, periodicity on $-\pi \le x \le \pi$ is assumed. Every function

$$u(x,t) := \psi(x+t),$$

with ψ differentiable, is a solution to (17), i.e. every initial state ψ just moves with constant speed (= -1), conserving all translation invariant functionals. In particular for (7) we find

$$d_t C = \int u\, \partial_t u = \int u\, \partial_x u = \frac{1}{2} \int \partial_x u^2 = 0$$

an explicit proof that C is conserved. The functional P coincides with C in this simple case.

a. Finite differences, with forward differencing on a uniform grid: $x_j = j\Delta x$, $\Delta x = 2\pi/N$, and $\hat{u}_j(t)$ is meant to approximate $u(x_j,t)$. As discretization of (17), (7) and (8) we take, naively:

$$(18) \qquad d_t \hat{u}_j = (\hat{u}_{j+1} - \hat{u}_j) \;/\; \Delta x$$

$$(19) \qquad \hat{C}(t) = \frac{\Delta x}{2} \Sigma\, \hat{u}_j^{\,2} = \hat{P}(t).$$

Conservation of $\hat{C}$ would require

$$0 = d_t \hat{C} = \Delta x\, \Sigma\, \hat{u}_j\, d_t \hat{u}_j = \Sigma\, \hat{u}_j \hat{u}_{j+1} - \Sigma\, \hat{u}_j^{\,2}$$

which is not true for most functions u, not even for u = sin ax. The reason of failure is that the discretization of ∂_x, i.e. the mapping $\hat{u}_j \rightarrow (\hat{u}_{j+1} - \hat{u}_j) \;/\; \Delta x$, is not an antisymmetric operator.

b. Finite differences, with central differencing:

(20) $$d_t \hat{u}_j = (\hat{u}_{j+1} - \hat{u}_{j-1}) \;/\; 2\Delta x$$

and $\hat{C}$, $\hat{P}$ as (19). In this case we find

$$d_t \hat{C} = \frac{1}{2} \Sigma\, \hat{u}_j \hat{u}_{j+1} - \frac{1}{2} \Sigma\, \hat{u}_j \hat{u}_{j-1} = 0.$$

So $\hat{C}$ and $\hat{P}$ are conserved. Indeed, central differencing is an antisymmetric operator, and $\hat{P}$ fits in the variational formulation (14). However, the agreement between solutions to (20) and solutions to (17) is rather restricted. A complete set of solutions to (20) is:

$$\{\hat{u}^{(k)}(t) \mid k = 1,2,\ldots,2\pi/\Delta x\}$$

$$\text{with } \hat{u}_j^{(k)}(t) = e^{ik(j\Delta x + \lambda t)} \text{ and } \lambda = \frac{\sin k\Delta x}{k\Delta x}.$$

From this we see that every eigensolution $\hat{u}^{(k)}$ travels with its own speed and none of them with the correct speed -1. In this case the reason of failure is the fact that expression (19) is not translation invariant: if $\hat{w}$ is the discrete representation of u after a translation: $\hat{w}_j := u(x_j+\varepsilon)$, then in general $\Sigma\, \hat{w}_j^{\,2}$ is unequal to $\Sigma\, \hat{u}_j^{\,2}$.

c. Spectral method: truncated Fourier series.

We choose $B = [e^{-inx}, .., 1, e^{ix}, .., e^{inx}]$, and for $u \in B$ we write $u(x) = \sum_{-n}^{n} \hat{u}_k e^{ikx}$. So in this case $\hat{u}$ is not a vector of function values but a vector of Fourier coefficients with respect to the basis $\{e^{ikx}\}$. For real u we have

(22) $$\hat{u}_k = \hat{u}_{-k}$$

For $\hat{P}$ and $\hat{C}$ we take

(23) $$\hat{P} = \hat{C} = \frac{1}{2} \int (\Sigma\, \hat{u}_k e^{ikx})^2 = \pi \sum_{-n}^{n} \hat{u}_k^{\,2}$$

Translation invariance of $\hat{P}$ is an immediate consequence of Parseval's identity. In the Euclidean space $\hat{V}$ we define as discretization of ∂_x:

(24) $$D = \mathrm{diag}\ (-in, .., 0, i, .., in)$$

so that D corresponds with ∂_x exactly for $u \in B$. D is antisymmetric, so $\hat{P}$ will be conserved. Equation (14) becomes:

(25) $$d_t\hat{u} = D\ \hat{u}$$

and an explicit proof of conservation is:

$$d_t\hat{P} = 2\pi \sum_{-n}^{n} \hat{u}_k d_t \hat{u}_k = 2\pi i \sum_{-n}^{n} k\hat{u}_k^2 = 0 \qquad \text{by (22).}$$

The solution of (25) is $\hat{u}(t) = e^{tD}\ \hat{u}(0)$, and for every function $u \in B$ we have:

$$u(x,t) = \sum e^{ikx}\ \hat{u}_k(t) = \sum e^{ikx}\ e^{tD}\ \hat{u}_k(0) =$$

$$= \sum e^{ikx}\ e^{ikt}\ \hat{u}_k(0) = u(x+t,0),$$

so every function in B travels with the correct speed -1.

Translation invariance. Let us now return to the translation invariance of $\hat{P}$ as condition for conservation of $\hat{C}$. The example, though too simple to distinguish between $\hat{C}$ and $\hat{P}$, has already shown the importance of translation invariance.

Translation invariance of $\hat{P}$ is to be understood as follows. Let $\hat{u} \in \hat{V}$; then $u := \sum \hat{u}_k s_k \in B$; perform an arbitrary translation $u \to T_\varepsilon u$ and let the projection of $T_\varepsilon u$ on B be denoted by $w = \sum \hat{w}_k s_k$; then

(26) $$\hat{P}(\hat{w}) = \hat{P}(\hat{u})$$

must hold for every $\varepsilon \in \mathbb{R}$. Since the original P is assumed to be translation invariant, one can hope to transfer this property to $\hat{P}$ by defining:

$$(27) \qquad \hat{P}(\hat{u}) := P(\sum \hat{u}_k s_k).$$

With this definition equality (26) implies

$$P(w) = P(\sum \hat{w}_k s_k) = \hat{P}(\hat{w}) = \hat{P}(\hat{u}) = P(\sum \hat{u}_k s_k) = P(u) = P(T_\varepsilon u)$$

Comparing left and right end we see that usually this equality will be satisfied only if $w = T_\varepsilon u$, i.e. if $T_\varepsilon u \in B$ for every ε.
Since by continuity and consistency arguments (27) is the only way to define $\hat{P}$ in $\hat{V}$, we arrive at the conclusion that $\hat{P}$ inherits the translation invariance from P provided that the subspace B is translation invariant.

Translation invariance is a rather severe requirement for B; e.g. spline functions do not meet this condition. The case of degree 1 (piecewise linear interpolation) is illustrated in figure 1, showing (a) a set of values $\hat{u}_k = u(x_k)$, (b) $u = \sum \hat{u}_k s_k \in B$, (c) $T_\varepsilon u$, (d) the set of values $\hat{w}_k$, (e) $w = \sum \hat{w}_k s_k$. Clearly figure e is different from figure c.

The translation invariance almost naturally leads to approximation with trigonometric functions, i.e. to spectral-like methods. Before presenting a specific example, let us first look at the travelling wave solutions of the discretised equation.

A consequence of the invariance of $\hat{P}$ will be that exact travelling wave solutions do exist for the discretised equation (14) and that these solutions -just as in the continuous case- are translations of the solutions of the minimization problem

$$(28) \qquad \min_{\hat{u} \in \hat{V}} \{ \hat{P}(\hat{u}) \mid \hat{M}(\hat{u}) = m, \ \hat{C}(\hat{u}) = c \}$$

which problem is nothing but the discretization of (11). Denoting a mi-

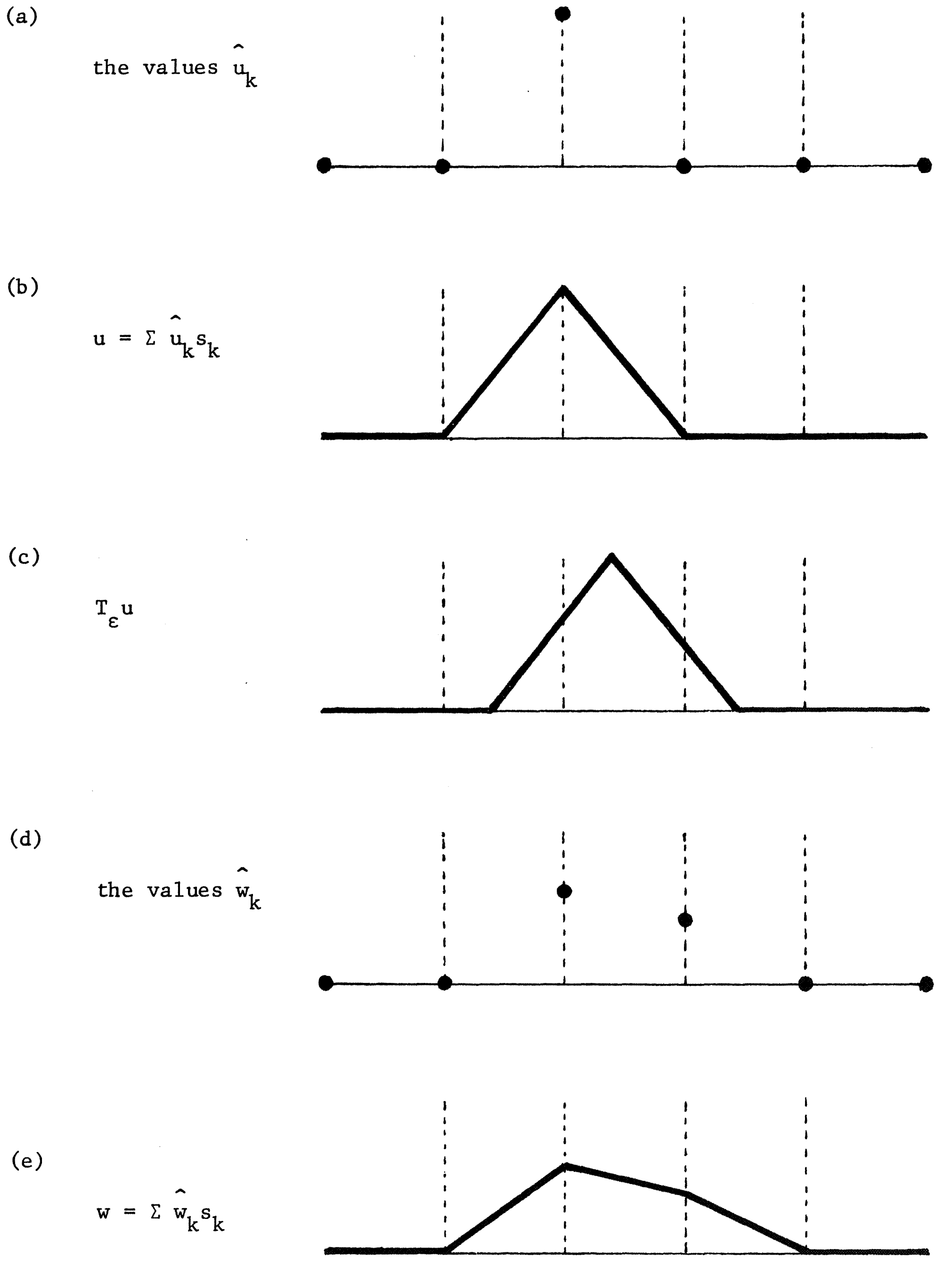

(a)
the values $\hat{u}_k$
(b)
$u = \Sigma \hat{u}_k s_k$
(c)
$T_\varepsilon u$
(d)
the values $\hat{w}_k$
(e)
$w = \Sigma \hat{w}_k s_k$

Figure 1.

nimizing vector by $\hat{\phi}$ we now have

$$(29) \qquad D \text{ grad } \hat{P}(\hat{\phi}) = \hat{\lambda} \, D \text{ grad } \hat{C}(\hat{\phi})$$

Again the multiplier $\hat{\lambda}$ is related to the travelling wave speed $-\hat{\lambda}$, and for a family of solutions of (28), to be denoted by $\hat{\phi}(\hat{C})$, it holds

$$(30) \qquad \hat{\lambda} = \hat{\lambda}(\hat{C}) = \frac{d\hat{P}(\hat{\phi}(\hat{C}))}{d\hat{C}}.$$

Discretization of the Korteweg-de Vries equation.

Having examined the general properties of translation invariant discretizations, we will work out a numerical example regarding the KdV-equation. As was explained before, and indicated already by the simple example, the spectral method is the most promising. On the other hand we have chosen to work with function values as representing a function, like in common finite difference methods.
To combine these ideas we consider again:

$$B = [\, e^{-inx}, \, .., \, e^{inx}]$$

but now we choose another basis: the functions

$$\psi_k(x) = \frac{1}{2n+1} \sum_{p=-n}^{n} e^{ip(x-x_k)} \qquad \text{with} \quad x_k = k\Delta x, \quad \Delta x = 2\pi/(2n+1)$$
$$-n \le k \le n$$

satisfy $\psi_k(x_j) = \delta_{kj}$ (Kronecker delta), and are orthogonal in the L_2-innerproduct:

$$\int \psi_k(x) \, \psi_j(x) \, dx = \Delta x \, \delta_{kj}$$

This implies that the representation $\hat{u}$ of a function u is given by $\hat{u}_j = u(x_j)$, and that the functional $\hat{C}$, see (7), becomes:

$$(31) \qquad \hat{C} = \frac{\Delta x}{2} \sum_{-n}^{n} \hat{u}_j^{\,2}.$$

The discretization A of the differential operator ∂_x is an antisymmetric matrix, operating in $\hat{V}$, and is found to be given by

$$A_{j,j+k} = - \frac{(-1)^k}{2 \sin \frac{k\Delta x}{2}} , \qquad -n \le j,j+k \le n, \quad k \ne 0.$$

In contrast to (24) we now find a full matrix, the entries of which are equal along diagonals (Toeplitz-form). Another property of numerical interest is that the absolute values of the entries are increasing towards the main diagonal. The matrix A is exact for functions in B, i.e. if $u(x) = \sum \hat{u}_k\psi_k(x)$ and $\hat{v} = A\hat{u}$, then $v(x) = \sum \hat{v}_k\psi_k(x)$ is the derivative of u for all x. Consequently $e^{\varepsilon A}$ is an exact shift in B. In fact the exponential, seen as an infinite power series in A, reduces to a polynomial in A of degree 2n: as A is of finite order, higher powers are linearly dependent on lower ones. A concise illustration is found in the simplest case, n = 1, where

$$A = \frac{1}{\sqrt{3}} \begin{vmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{vmatrix} , \quad A^2 = \frac{1}{3} \begin{vmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{vmatrix} ,$$

$A^3 = -A$, $A^4 = -A^2$, etcetera. The exponential becomes

$$\begin{aligned} e^{\varepsilon A} &= 1 + \varepsilon A + \frac{\varepsilon^2}{2} A^2 + \frac{\varepsilon^3}{3!} A^3 + \dots \\ &= 1 + (\varepsilon - \frac{\varepsilon^3}{3!} + \dots) A + (\frac{\varepsilon^2}{2!} - \frac{\varepsilon^4}{4!} + \dots) A^2 \\ &= 1 + (\sin \varepsilon) A + (1 - \cos \varepsilon) A^2, \end{aligned}$$

a quadratic polynomial. In particular, for $\varepsilon = \Delta x = 2\pi/3$ we have

$$e^{\frac{2}{3}\pi A} = 1 + \frac{1}{2}\sqrt{3}\, A + \frac{3}{2} A^2 = \begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix}$$

which represents an exact shift over Δx.

To continue with the dynamics of the KdV-equation, the functional P, defined by (5), will be discretised according to (27):

$$\hat{P}(\hat{u}) = \frac{1}{2}\int \Big(\sum_{j,k} A_{jk}\hat{u}_k\psi_j(x)\Big)^2 - \int \Big(\sum_j \hat{u}_j\psi_j(x)\Big)^3$$

$$= \frac{\Delta x}{2}\sum_j (A\hat{u})_j^{\,2} - \sum_{jkm} \hat{u}_j\hat{u}_k\hat{u}_m\int\psi_j\psi_k\psi_m .$$

The last integral is invariant under permutation of subscripts and under shifting of all subscripts together ($j,k,m \rightarrow j+1,\ k+1,\ m+1$), so it is sufficient to calculate

$$Q_{k,j} = \frac{1}{\Delta x}\int\psi_0\psi_k\psi_j =$$

$$= \frac{1}{\sin \frac{j-k}{2}\Delta x}\left((-1)^k \frac{\sin^2 n\frac{j}{2}\Delta x}{\sin \frac{j}{2}\Delta x} - (-1)^j \frac{\sin^2 n\frac{k}{2}\Delta x}{\sin \frac{k}{2}\Delta x}\right)$$

$$Q_{0,j} = \left(\frac{\sin n\frac{j}{2}\Delta x}{\sin \frac{j}{2}\Delta x}\right)^2$$

$$Q_{0,0} = \frac{3n^2 + 3n + 1}{4n^2 + 4n + 1}$$

Defining $\hat{q} = \hat{q}(\hat{u})$ by:

$$\text{(32)} \qquad \hat{q}(\hat{u})_m = \sum_{k,j} \hat{u}_{k+m}\, Q_{k,j}\, \hat{u}_{j+m}$$

we find for (14) the following system of ordinary differential equations:

$$\text{(33)} \qquad d_t\hat{u} = A\,(\,-\,A^2\hat{u} - 3\hat{q}(\hat{u}))$$

The righthandside reflects interaction between all the components $\hat{u}_k$, as is usual in spectral methods.

Compared with finite difference methods we see that we miss the benefit of sparce matrices, but on the other hand equation (33) has the advantage that it admits travelling wave solutions and that it conserves $\hat{C}$, $\hat{M}$ and $\hat{P}$, while these quantities are exactly equal to C, M and P for u ∊ B. For the sake of completeness we remind that $\hat{C}$ is given by (30), and $\hat{M}$ and $\hat{P}$ are:

$$(34) \qquad \hat{M} = \Delta x \sum_{-n}^{n} \hat{u}_j , \qquad \hat{P} = \frac{\Delta x}{2} \sum_{-n}^{n} \{(A\hat{u})_j{}^2 - \hat{u}_j \hat{q}_j(\hat{u})\}.$$

Numerical results.

We have shown how to discretise certain evolution equations, in particular the periodic KdV-problem, on a given number of grid points, and how to discretise certain quantities in such a way that they will be conserved in time. To illustrate the conservation numerically it will be sufficient to use a standard fourth order Runge-Kutta time integration.
The computer runs have been made using only five grid points for the spatial discretization. Of course, the approximation of solutions will become better with grid refinement. But to show the numerical consistency is not our main object. On the contrary, we want to demonstrate that conservation is not a matter of grid spacing: in spite of the spatial approximation error the quantities $\hat{P}$, $\hat{C}$ and $\hat{M}$ are found to be conserved exactly, if only the time integration could be performed exactly.

First we have solved minimization problem (28) for m = 0 and various values of C.
The minimizing values of $\hat{P}$ are depicted in the $\hat{C}$-$\hat{P}$-diagram in figure 2. The slope $\hat{\lambda}$ of the curve -see (30)- is a decreasing function of $\hat{C}$. (For the linear equation $u_t + u_{xxx} = 0$ the graph would be a straight line with slope 1.)
Since $\hat{\lambda}$ is equal to the velocity of the travelling wave solution $\hat{\phi}$, we see that these waves slow down with increasing $\hat{C}$. At the maximum value of $\hat{P}$, occuring for $\hat{C} = \pi/2$, the wave speed is zero; this solution will not appear physically, because such high values of $\hat{C}$ lie beyond the model assumptions for the KdV-equation. Nevertheless the mathematical treatment still applies, theoretically as well as numerically.

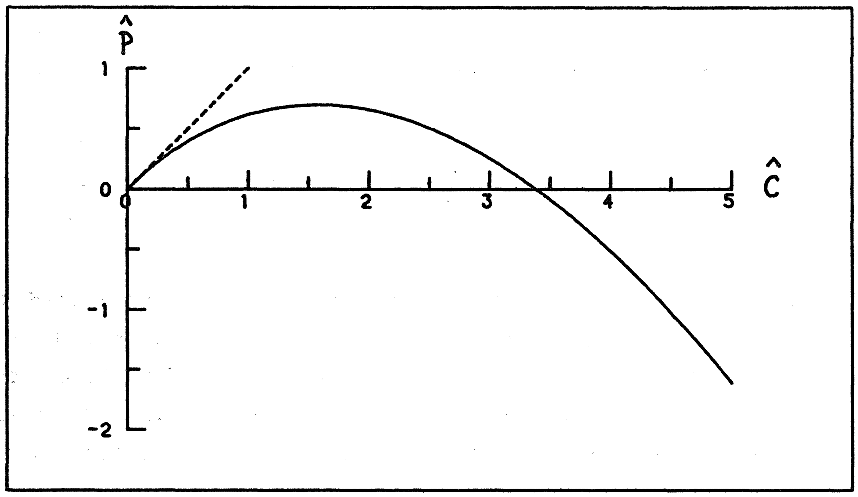


Figure 2. $\hat{C}$-$\hat{P}$-diagram for the discretised Korteweg-de Vries equation

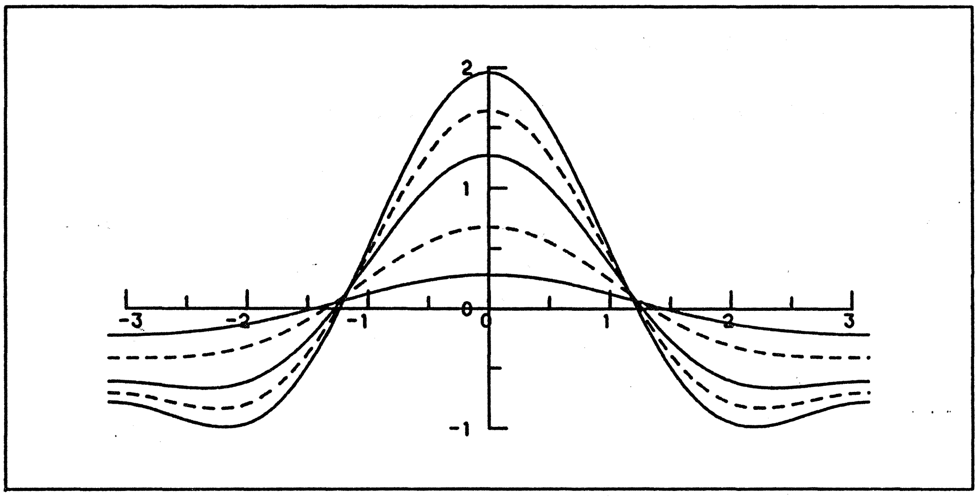


Figure 3. 5-point-approximation of cnoidal waves for the KdV equation, for $\hat{C}$ = 0.1, 0.5, $\pi/2$, 2.5 and 3.5.

Solutions to (28), interpolated with trigonometric functions, are shown in figure 3 for several values of $\hat{C}$. For $\hat{C} \to 0$ the "discrete cnoidal wave" approximates the harmonic wave cos x, which is the exact solution to the linearised equation $u_t + u_{xxx} = 0$.
For increasing $\hat{C}$ the deviation from the continuous cnoidal wave becomes apparent, obviously due to the small number of grid points. Yet we know theoretically that every solution $\hat{\phi}$ will be propagated as a travelling wave solution to (33).
To find this property confirmed we have run a fourth order Runge-Kutta integration of the ODE system (33), and compared it with a uniform shift based on the velocity $\hat{\lambda}$ predicted by (29). The errors after 10 periods, run with 240 time steps per period, are shown in the table. Runs with smaller time steps have produced even smaller errors. The machine accuracy is 10^{-8}. The table reflects the universal tendency that solutions with steeper time derivatives have greater errors. It is clear that the errors in $\hat{M}$, $\hat{C}$ and $\hat{P}$ are only due to the time integration and machine round off. This confirms the theoretical result that in the discretised problem the quantities $\hat{M}$, $\hat{C}$ and $\hat{P}$ would be conserved and that discrete cnoidal waves would exist provided that the time integration could be done exactly.

Table of errors after 10 periods

(fourth order Runge-Kutta with 240 time steps per period)

$\hat{C}$	$\Delta\hat{M}$	$\Delta\hat{C}$	$\Delta\hat{P}$	$\Delta\hat{u}$
0.01	4.0 (-8)	4.2 (-8)	4.0 (-8)	1.6 (-6)
0.1	3.5 (-7)	3.7 (-8)	6.7 (-8)	1.3 (-5)
0.5	2.5 (-6)	7.2 (-7)	7.2 (-7)	7.5 (-5)
$\pi/2^*$	0	2.4 (-7)	4.7 (-7)	1.6 (-5)
2.5	3.8 (-6)	3.3 (-6)	1.1 (-5)	4.2 (-4)
3.5	1.2 (-5)	3.5 (-5)	7.5 (-5)	2.6 (-3)

*) Since no period can be defined (wave speed is zero), the time interval has been taken from the case $\hat{C} = 0.5$.

Concluding we can say that for a whole class of evolution equations of a specific form we have given a procedure to find spatial discretizations that preserve conservation properties exactly, even in the case of strong nonlinearity. For long time computations, e.g. to study wave interactions, our numerical results motivate further investigations into time integration processes that can maintain these conservation properties over a long time.

References

Baumgarte, J., 1973, Asymptotische Stabilisierung von Integralen bei gewöhnlichen Differentialgleichungen 1. Ordnung, ZAMM 53 p. 701.

Feng, K., 1986, Symplectic geometry and numerical methods in fluid dynamics, Tenth Int. Conf. Num. Meth. Fluid Dyn., Beijing, Lecture Notes in Physics 264, p. 1, Springer-Verlag.

Gear, C.W., 1986, Maintaining solution invariants in the numerical solution of ODEs, SIAM J. Sci. Stat. Comput., vol. 7, p. 734.

Groesen, E. van, & F.P.H. van Beckum, 1987, Model consistent discretizations of Poisson systems with applications to fluid dynamics. Part I: Spatial discretizations of 1D translation invariant systems. Memorandum no. 611 Dept. Appl. Math., University of Twente.

Shampine, L.F., 1984, Conservation laws and the numerical solution of ODEs, Sandia Albuquerque Report SAND84-1241.

Whitham, G.B., 1974, Linear and nonlinear waves, John Wiley & Sons, New York.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

On the Periodic Wind-Induced Vibrations of an Oscillator with Two Degrees of Freedom

C.G.A. van der Beek and A.H.P. van der Burgh
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands

ABSTRACT

In this paper the dynamics of an oscillator with two degrees of freedom in a steady flow is studied. Principles from the theory of galloping are used to derive the equations of motion.
The normal forms for the equations of motion for a number of interesting cases are presented and the existence of periodic solutions and their stability is established. Formulas, which may be used to calculate amplitudes and periods in an approximative way, are presented.

1. INTRODUCTION

Overhead transmission lines on which ice has accreted may have cross-sectional shapes that are aerodynamically unstable to transverse disturbances in a wind-field. The evolution from this unstable equilibrium position may result in galloping: a large amplitude oscillation with a low frequency (< 1 Hz). The very complicated phenomenon of galloping of overhead transmission lines which involves the aeroelastic interaction of longitudinal, transversal and torsional oscillations of a continuous system is far from being understood.

For an interesting survey paper on wind-induced vibrations of overhead power transmission lines the reader is referred to [1]. In the present paper a simple oscillator, which has some relation with this galloping problem is introduced. At this stage the oscillator is a theoretical model, that is, no experimental prototype has been developed yet: however, it is believed that the ideas presented here, may be used for the actual development of an experimental model. The oscillator, sketched in Fig. 2.1, consists of a rigid circular cylinder with a small ridge and a number of springs mounted in a frame. The oscillator is constructed in such a way that the cylinder-spring system has two degrees of freedom, i.e. oscillations in the direction and perpendicular to the direction of the air flow; both modes of vibrations are decoupled in the absence of an air flow. A more detailed description of the oscillator is given in section 2. The oscillator may be considered as an extension of the one degree of freedom oscillator introduced and studied theoretically and experimentally in a wind-tunnel in [2] and [3], where only oscillations perpendicular to the direction of the air flow are possible.

The remainder of the paper is organized as follows.
In section 2 the equations of motion of the oscillator in a uniform windfield are derived. The assumption is made that the aerodynamic forces are quasi-steady which implies that they can be derived from static force measurements (in steady flow); for the study of galloping there is no disagreement in the literature (e.g. [2] and [3]) on the use of a quasi-steady theory. The mathematical modeling of the oscillator is analogous to the modeling of a swinging-spring oscillator introduced in [4]. This oscillator consists of a cylinder, hung from springs, such that spring and pendulum oscillations may be carried out. However, in the equations of motion of the oscillator as considered here a new parameter representing the position of the ice accretion (ridge) on the cylinder is introduced. Finally, by truncating the equations of motion, the so-called model equations are obtained.
In section 3 the theory of normal forms is presented and used for the study of periodic solutions. The main result is a theorem which states existence of periodic solutions with a period close to the period of the solutions of the unperturbed model equations, which are equations describing the free oscillations (absence of the external aerodynamical forces) of the oscillator. This theorem is an extension of theorems to be found in [5,6,7] where the existence of periodic solutions, with a period that equals the period of the solutions of the unperturbed system (or the period of the time-periodic vectorfield f in $\dot{x} = \varepsilon f(t,x,\varepsilon)$), is established.
In section 4 periodic solutions for the model equations in a number of interesting cases are established and criteria for the stability of these periodic solutions are given (based on the theorem mentioned above).
The paper ends with some concluding remarks. With respect to the calculations of the normal forms use has been made of the computer-algebra system Macsyma [9] which has the capability of manipulating algebraic expressions

involving unevaluated variables; in fact, theorem 3.1 has been implemented on the computer.

2. THE OSCILLATOR AND THE EQUATIONS OF MOTION

In this section the oscillator, as sketched in Fig. 2.1 and 2.2 is described in more detail. In addition, the equations of motion will be derived. The oscillator, as sketched in Fig. 2.1, consists of a rigid circular cylinder with a small ridge (representing the ice-accretion) and six springs providing linear elasticity. The cylinder is rigidly attached with two sup-

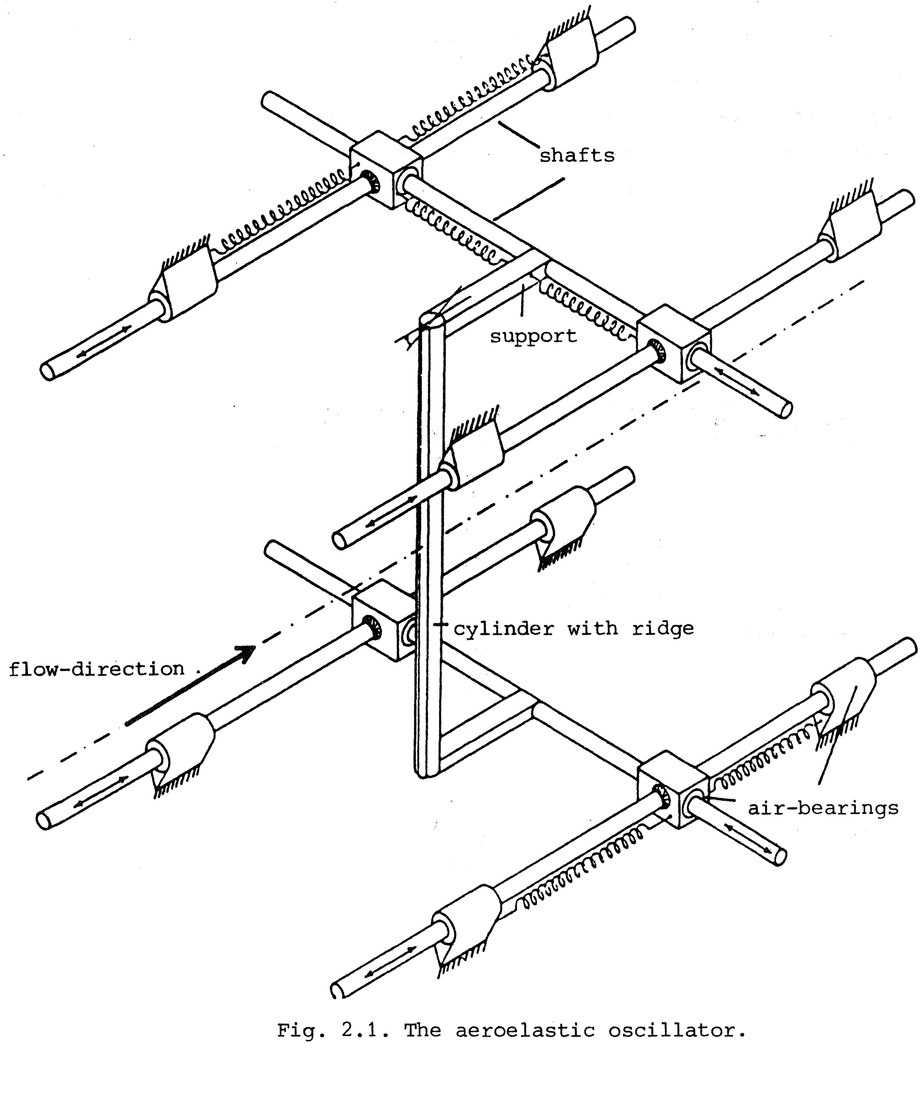


Fig. 2.1. The aeroelastic oscillator.

ports to two shafts that are supported by four air-bearings (indicated with a rectangular parallelepiped in the sketch). These four air-bearings are subsequently rigidly attached to a system of four shafts that are supported by eight air-bearings fixed to a frame indicated with slant lines in the sketch. The two springs fixed to the supports of the cylinder and the air-bearings provide restoring forces in the direction perpendicular to the flow directions whereas the other four springs provide restoring forces in the direction of the flow.

Figure 2.2 is a sketch of the view from above; the position of the cylinder, remaining always perpendicular to the picture-plane, can be defined by the two coordinates x and y as indicated. The origin of the (x,y) coordinate-system is the equilibrium position of the centre of the cylinder in absence of aerodynamic forces. Although the cross-section is not circular, due to the ridge, it has still an axis of symmetry; α_s (static angle of attack) is the angle between the direction of the wind-field and the axis of symmetry

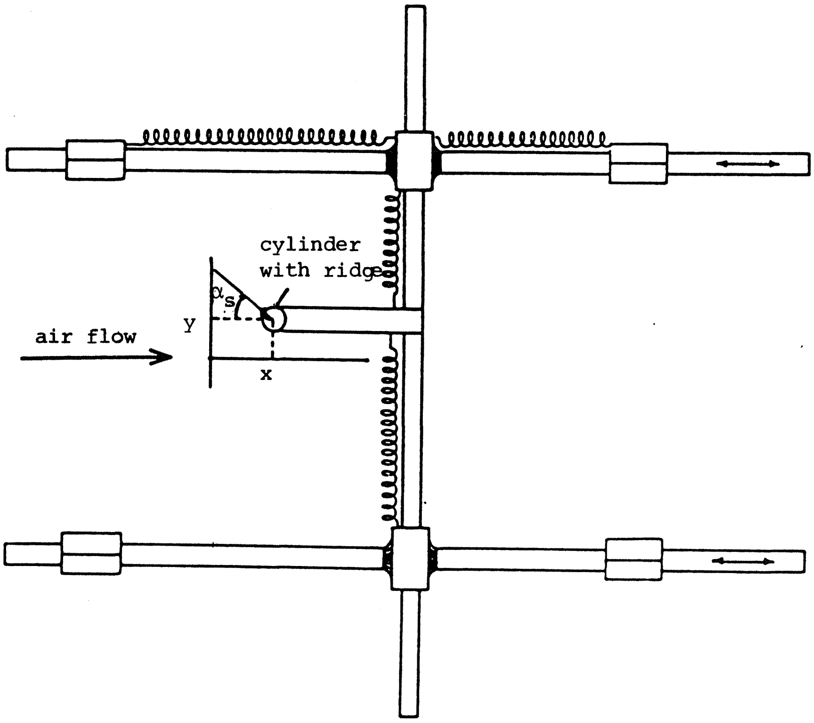


Fig. 2.2. The aeroelastic oscillator; view from above.

of the cross-section. To be more precise α_s is the angle obtained by rotating $\underline{e}_{\alpha_s}$ (unit vector along the axis of symmetry pointing from the ridge to the centre of the cylinder) to $\underline{e}_x$ (unit vector in the direction of the windfield), positive in anti-clockwise direction.

It may be clear now that the oscillator has the property that the two degrees of freedom are mechanically uncoupled. However, as will be shown in what follows there will be an aerodynamic coupling when the oscillator is interacting with a wind-field.

If one puts this oscillator in a uniform wind-field with wind-velocity $\underline{v}_\infty, \underline{v}_\infty = v_\infty \underline{e}_x$ $(v_\infty > 0)$, forces will be generated on the cylinder. These forces, the drag $D\underline{e}_D$ and the lift $L\underline{e}_L$ are sketched in Fig. 2.3. $\underline{e}_D$ and $\underline{e}_L$ are unit-vectors (D and L are the magnitudes of the drag respectively lift force) such that $\underline{e}_D$ has the same direction as the virtual wind velocity $\underline{v}_s, \underline{v}_s := \underline{v}_\infty - (\dot{x}\underline{e}_x + \dot{y}\underline{e}_y)$, and $\underline{e}_L$ is obtained by rotating $\underline{e}_D$ over an angle $\pi/2$ in anti-clockwise direction. ($\dot{x}$ and $\dot{y}$ are the velocity components of the cylinder).

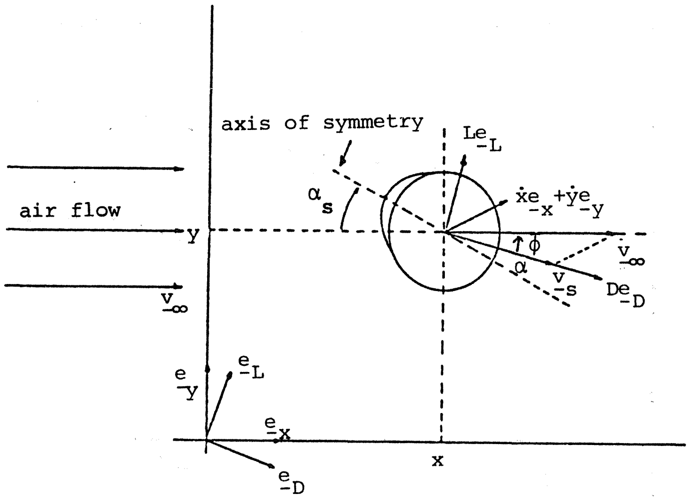


Fig. 2.3. Wind-velocities and aerodynamic forces acting on the cross-section.

It is easy to verify that the equations of motion for the cylinder become:

$$\begin{aligned} m_x\ddot{x} + s_x x &= D\cos\phi - L\sin\phi \\ m_y\ddot{y} + s_y y &= D\sin\phi + L\cos\phi \end{aligned} \qquad (2.1)$$

with m_x (m_y) the total mass of that part of the oscillator that can move in x-direction (y-direction), s_x (s_y) the total stiffness of the springs in x-direction (y-direction) and ϕ the angle between $\underline{e}_x$ and $\underline{e}_D$, positive in anti-clockwise direction. It will be assumed that the mass of the shafts and air-bearings is small compared with the mass of the cylinder which implies that $m_x = m_y := m$.

The magnitude of the aerodynamic forces D and L may be expressed in terms of aerodynamic coefficients as:

$$\begin{aligned} D &= \frac{1}{2}\rho d c^D(\alpha) v_s^2 \\ L &= \frac{1}{2}\rho d c^L(\alpha) v_s^2, \end{aligned} \qquad (2.2)$$

where ρ is the density of air, d is the cylinder diameter, $v_s = |\underline{v}_s|$ and α the angle between $\underline{v}_s$ and the axis of symmetry of the cylinder (see Fig. 2.3). $c^D(\alpha)$ and $c^L(\alpha)$ (functions of α) are the quasi-static drag and lift coefficients which may be obtained from wind-tunnel experiments; typical results obtained from measurements in a wind-tunnel are sketched in Fig. 2.4.

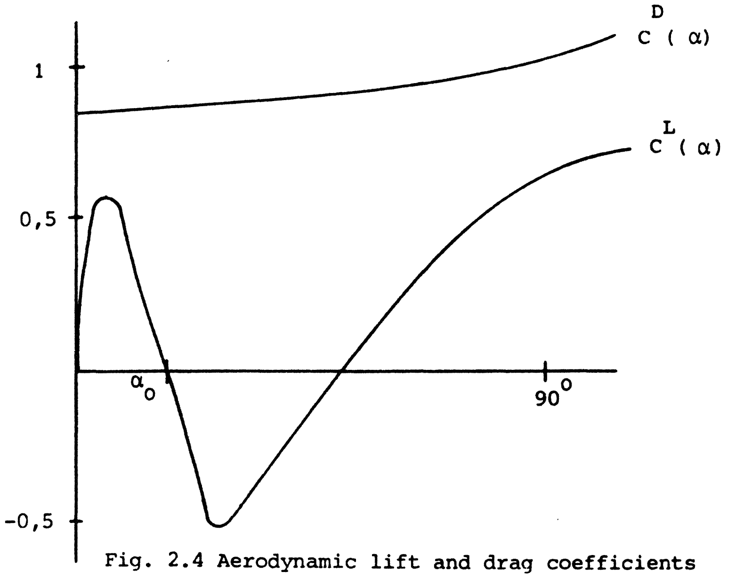


Fig. 2.4 Aerodynamic lift and drag coefficients

According to the den Hartog criterion galloping may occur if

$$\left[c^D(\alpha) + \frac{\partial}{\partial\alpha} c^L(\alpha)\right]_{|\alpha_s} < 0 \qquad \text{(linear instability of the equilibrium position).}$$

Since $c^D(\alpha)$ is always positive attention is only paid to the case where $\frac{\partial}{\partial\alpha} c^L(\alpha)\,|_{\alpha_s} < 0$.

Hence the drag and lift coefficients curves are approximated by:

$$\begin{aligned} c^D(\alpha) &= c_0^D, && \text{where } c_0^D > 0 \text{ a constant and} \\ c^L(\alpha) &= c_1^L(\alpha - \alpha_0) + c_3^L(\alpha - \alpha_0)^3, && \text{where } c_1^L < 0 \text{ and } c_3^L > 0 \text{ are constants.} \end{aligned} \tag{2.3}$$

Note that:

(.) $v_s^2 = v_\infty^2 - 2\dot{x}v_\infty + \dot{x}^2 + \dot{y}^2$,

(.) $\cos\phi = \dfrac{v_\infty - \dot{x}}{v_s}$, $\sin\phi = -\dfrac{\dot{y}}{v_s}$ and $\tan\phi = \dfrac{-\dot{y}}{v_\infty - \dot{x}}$,

(.) $\alpha = \phi + \alpha_s$ (in figure 2.3 the angle ϕ is negative),

By using these expressions one can show that the equations of motion become (the terms of fourth and higher order are neglected):

$$\begin{aligned} \ddot{x} + \omega_1^2 x = {} & \frac{\rho d c_0^D v_\infty^2}{2m} - \frac{2\rho d c_0^D v_\infty}{2m}\dot{x} + \frac{\rho d v_\infty}{2m}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right)\dot{y} + \\ & + \frac{\rho d c_0^D}{2m}\dot{x}^2 - \frac{\rho d}{2m}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right)\dot{x}\dot{y} + \frac{\rho d}{2m}\left(\frac{c_0^D}{2} - c_1^L + 3\bar{\alpha}_s^2 c_3^L\right)\dot{y}^2 \\ & + \frac{\rho d}{2mv_\infty}\left[\frac{c_1^L}{2}\bar{\alpha}_s + 3c_3^L\bar{\alpha}_s + \frac{c_3^L}{2}\bar{\alpha}_s^3\right]\dot{y}^3, \\ \ddot{y} + \omega_2^2 y = {} & \frac{\rho d v_\infty^2}{2m}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right) - \frac{2\rho d v_\infty}{2m}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right)\dot{x} + \\ & - \frac{\rho d v_\infty}{2m}\left(c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L\right)\dot{y} + \frac{\rho d}{2m}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right)\dot{x}^2 + \\ & + \frac{\rho d}{2m}\left[3\bar{\alpha}_s c_3^L + \frac{c_3^L}{2}\bar{\alpha}_s^3 + \frac{c_1^L}{2}\bar{\alpha}_s\right]\dot{y}^2 + \\ & + \frac{\rho d}{2m}\left(c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L\right)\dot{x}\dot{y} + \frac{\rho d}{2mv_\infty}\left[-\frac{c_0^D}{2} - \frac{c_1^L}{6} - \left(1 + \frac{1}{2}\bar{\alpha}_s^2\right)c_3^L\right]\dot{y}^3, \end{aligned} \tag{2.4}$$

where $\omega_1 = \sqrt{\dfrac{s_x}{m}}$, $\omega_2 = \sqrt{\dfrac{s_y}{m}}$, $\bar{\alpha}_s = \alpha_s - \alpha_0$ and where use has been made of:

$$(.)\ \sqrt{1+z} \;= 1 + \frac{z}{2} - \frac{z^2}{8} + \frac{z^3}{16} - \ldots, \quad z \in \mathbb{R},\ |z| < 1,$$

$$(.)\ \arctan(z) = z - \frac{z^3}{3} + \frac{z^5}{5} - \ldots, \qquad z \in \mathbb{R},\ |z| < 1.$$

It is assumed that this system of equations describes the motion of the cylinder; in what follows an analysis of this system of model equations will be carried out in a rigorous way, that is all elaborations will be motivated.

Introduction of the dimensionless variables X, Y, τ defined by $X = \omega_2 x/v_\infty$, $Y = \omega_2 y/v_\infty$ and $\tau = \omega_2 t$ yields the following system:

$$\begin{aligned}
\ddot{X} + \Omega^2 X &= \varepsilon \Bigg[c_0^D - 2c_0^D \dot{X} + \left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right) \dot{Y} + c_0^D \dot{X}^2 - \left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right) \dot{X}\dot{Y} + \\
&\quad + \left(\frac{c_0^D}{2} - c_1^L - 3\bar{\alpha}_s^2 c_3^L \right) \dot{Y}^2 + \left(\frac{c_1^L}{2} \bar{\alpha}_s + 3c_3^L \bar{\alpha}_s + \frac{c_3^L}{2} \bar{\alpha}_s^3 \right) \dot{Y}^3 \Bigg], \\
\ddot{Y} + Y &= \varepsilon \Bigg[\left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right) - 2 \left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right) \dot{X} - \left(c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L \right) \dot{Y} + \\
&\quad + \left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right) \dot{X}^2 + \left[3\bar{\alpha}_s c_3^L + \frac{c_3^L}{2} \bar{\alpha}_s^3 + \frac{c_1^L}{2} \bar{\alpha}_s \right] \dot{Y}^2 + \\
&\quad + \left(c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L \right) \dot{X}\dot{Y} - \left[\frac{c_0^D}{2} + \frac{c_1^L}{6} + \left(1 + \frac{1}{2} \bar{\alpha}_s^2 \right) c_3^L \right] \dot{Y}^3 \Bigg],
\end{aligned} \tag{2.5}$$

where $\Omega^2 = \omega_1^2/\omega_2^2$, $\varepsilon = \rho d v_\infty / 2m\omega_2$ and a dot now stands for differentiation with respect to τ.

A system of four first order equations may be obtained by setting $X = x_1 + \varepsilon c_0^D/\Omega^2$, $\dot{X} = x_2$, $Y = x_3 + \varepsilon \left(c_1^L \bar{\alpha}_s + c_3^L \bar{\alpha}_s^3 \right)$, $\dot{Y} = x_4$ from which follows that:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} x_2 \\ -\Omega^2 x_1 \\ x_4 \\ -x_3 \end{pmatrix} + \varepsilon \begin{pmatrix} 0 \\ a_2 x_2 + a_4 x_4 + a_{22} x_2^2 + a_{24} x_2 x_4 + a_{44} x_4^2 + a_{444} x_4^3 \\ 0 \\ b_2 x_2 + b_4 x_4 + b_{22} x_2^2 + b_{24} x_2 x_4 + b_{44} x^2 + b_{444} x_4^3 \end{pmatrix}, \tag{2.6}$$

where:

$$\begin{aligned}
a_2 &= -2c_0^D < 0 & \qquad b_2 &= -2\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right) \\
a_4 &= \left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right) & \qquad b_4 &= -\left(c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L\right) > 0 \\
a_{22} &= c_0^D > 0 & \qquad b_{22} &= c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3 \\
a_{24} &= -\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right) & \qquad b_{24} &= c_0^D + c_1^L + 3\bar{\alpha}_s^2 c_3^L < 0 \\
a_{44} &= c_0^D/2 - c_1^L + 3\bar{\alpha}_s^2 c_3^L > 0 & \qquad b_{44} &= 3\bar{\alpha}_s c_3^L + c_3^L/2\bar{\alpha}_s^3 + c_1^L\bar{\alpha}_s/2 \\
a_{444} &= \frac{c_1^L\bar{\alpha}_s}{2} + 3c_3^L\bar{\alpha}_s + c_3^L\,\frac{\bar{\alpha}_s^3}{2} & \qquad b_{444} &= -\left[\frac{c_0^D}{2} + \frac{c_1^L}{6} + \left(1 + \frac{1}{2}\,\bar{\alpha}_s^2\right) c_3^L\right] < 0.
\end{aligned}$$

The equations (2.6) will be studied for the case that the coefficients a_2, a_{22}, a_{44}, b_4, b_{24} and b_{444} have a sign as indicated above. As follows from measurements of the aerodynamic coefficients in a wind-tunnel these signs are relevant for the description of the galloping phenomenon.

3. PERIODIC SOLUTIONS AND THE THEORY OF NORMAL FORMS

In this section normal forms will be used for the study of periodic solutions for a class of differential equations, of which the system of model equations derived in the previous section is a prototype. The study concerns existence, stability and location of the periodic solutions in the phase space as well as the calculation of amplitudes and periods in an approximative way. The results are formulated in two theorems. For the proof of these theorems the reader is referred to [8] where also the relation with the (extensive) literature on normal forms is pointed out.

The model equations derived in section 2 are a special case of:

$$\dot{\underline{x}} = A_0\underline{x} + \varepsilon\underline{f}(\underline{x},\varepsilon), \ \underline{x} \in D, \qquad (3.1)$$

where:

(.) $\underline{0} \in D \subset \mathbb{R}^{2n}$ $(n \in \mathbb{N})$ open and bounded,

(.) $\varepsilon \in (-\varepsilon_0, \varepsilon_0) \subset \mathbb{R}$, $0 < \varepsilon_0 \ll 1$,

(.) $A_0 = \begin{pmatrix} A_0^1 & & \emptyset \\ & A_0^2 & \\ & & \ddots & \\ \emptyset & & & A_0^n \end{pmatrix}$, with $A_0^i = \begin{pmatrix} 0 & 1 \\ -\omega_i^2 & 0 \end{pmatrix}$ and $\omega_i > 0$,

(.) $\underline{f} \in C^2\ (D\times(-\varepsilon_0,\varepsilon_0), \mathbb{R}^{2n})$, $\underline{f}(\underline{0},\varepsilon) = \underline{0}$ for all $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$.

Furthermore, it is assumed that

(.) $\frac{\omega_i}{\omega_j} \in \mathbb{Q}$ for all $i,j \in \{1,\dots,n\}$,

(.) the domain D has the property that if $\underline{x} \in D$ then also $e^{\phi A_0}\underline{x} \in D$ for all $\phi \in \mathbb{R}$, where:

$$e^{\phi A_0} = \sum_{n=0}^{\infty} \phi^n \frac{A_0^n}{n!} = \begin{pmatrix} e^{\phi A_0^1} & & \emptyset \\ & \ddots & \\ \emptyset & & e^{\phi A_0^n} \end{pmatrix},\ e^{\phi A_0^i} = \begin{pmatrix} \cos(\omega_i\phi) & \omega_i^{-1}\sin(\omega_i\phi) \\ -\omega_i\sin(\omega_i\phi) & \cos(\omega_i\phi) \end{pmatrix}.$$

Note that the condition $\frac{\omega_i}{\omega_j} \in \mathbb{Q}$ for all $i,j \in \{1,\dots,n\}$ implies that there exists a $T > 0$, independent of ε, such that $\phi \rightarrow e^{\phi A_0}$ is T-periodic; T_0 is the primitive period of $\phi \rightarrow e^{\phi A_0}$.

To introduce and motivate the use of the normal forms used here consider the differential equation:

$$\dot{\underline{\xi}} = A_0\underline{\xi} + \varepsilon\underline{f}^0(\underline{\xi}),\ \text{with } \underline{\xi} \in D, \tag{3.2}$$

and suppose that $\underline{f}^0$ satisfies the following condition:

$$\underline{f}^0\left(e^{\phi A_0}\underline{\xi}\right) = e^{\phi A_0}\underline{f}^0(\underline{\xi}) \text{ for all } \phi \in \mathbb{R} \text{ and } \underline{\xi} \in D. \tag{3.3}$$

By using the transformation $\underline{\xi} = e^{tA_0}\underline{y}$ the following differential equation for $\underline{y}$ holds:

$$\dot{\underline{y}} = \varepsilon\underline{f}^0(\underline{y}). \tag{3.4}$$

Critical points of (3.4) induce periodic solutions of (3.2): $\underline{y}_0 \in D \setminus \{0\}$ and $\underline{f}^0(\underline{y}_0) = \underline{0}$ implies that $t \rightarrow e^{tA_0}\underline{y}_0$ is a non-trivial T_0-periodic solution of (3.2). In what follows property (3.3) is used to define a normal

form for system (3.1) and a (near-identity) transformation will be introduced to put system (3.1) in normal form.
This normal form will then be used to give a theorem on the existence of periodic solutions for system (3.1).
More explicitly one needs:

Definition: A vectorfield $\underline{f}^0 \in C^0(D, \mathbb{R}^{2n})$ is called a normal vectorfield (with respect to A_0) if it is invariant under the flow produced by A_0, i.e.:

$$\underline{f}^0\,(e^{\phi A_0}\underline{x}) = e^{\phi A_0}\underline{f}^0(\underline{x}) \text{ for all } \underline{x} \in D, \phi \in \mathbb{R}$$

and

Definition: System (3.1) is in first order normal form when $\underline{f}(\underline{x},0)$ is a normal vectorfield (with respect to A_0) for all $\underline{x} \in D$.

It may be clear that in general system (3.1) is not in first order normal form. In order to establish this, one may use a near-identity transformation of the form ($\underline{P}$ will be specified in theorem 3.1):

$$\underline{x} = \underline{\xi} + \varepsilon\underline{P}(\underline{\xi}). \qquad (3.5)$$

Substituting (3.5) in (3.1) yields, for ε small enough, the following differential equation for $\underline{\xi}$:

$$\dot{\underline{\xi}} = A_0\underline{\xi} + \varepsilon\left(A_0\underline{P}(\underline{\xi}) - D\underline{P}(\underline{\xi})A_0\underline{\xi} + \underline{f}(\underline{\xi},0)\right) + O(\varepsilon^2); \qquad (3.6)$$

here is $\underline{\xi} \to D\underline{P}(\underline{\xi})$ the Jacobian matrix of $\underline{P}$ with respect to $\underline{\xi}$.
In order to have system (3.6) in first order normal form one has to apply a near-identity transformation as given in the following theorem (see [8]):

Theorem 3.1: Consider system (3.1), i.e.

$$\dot{\underline{x}} = A_0\underline{x} + \varepsilon\underline{f}(\underline{x},\varepsilon), \quad \underline{x} \in D, \qquad (3.7)$$

where

(.) $D \subset \mathbb{R}^{2n}$ open and bounded, $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$, $0 < \varepsilon_0 \ll 1$,

(.) $\underline{f} \in C^2\ (D\times(-\varepsilon_0,\varepsilon_0), \mathbb{R}^{2n})$, $\underline{f}(\underline{0},\varepsilon) = \underline{0}$ for all $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$,

(.) $\frac{\omega_i}{\omega_j} \in \mathbb{Q}$ for all $i,j \in \{1,\dots,n\}$ and D has the property that $\underline{x} \in D$ implies $e^{tA_0}\underline{x} \in D$ for all $\phi \in \mathbb{R}$.

With

$$\underline{f}^0(\underline{x}) \equiv \frac{1}{T_0}\int_0^{T_0} e^{-\phi A_0}\underline{f}\left(e^{\phi A_0}\underline{x},0\right) d\phi$$

and

$$\underline{P}(\underline{x}) \equiv \frac{1}{T_0}\int_0^{T_0} \phi \left(e^{-\phi A_0}\underline{f}\ (e^{\phi A_0}\underline{x},0) - \underline{f}^0(\underline{x})\right) d\phi,$$

the following holds:

(1) $\underline{f}^0,\underline{P} \in C^2(D, \mathbb{R}^{2n})$, $\underline{f}^0(\underline{0}) = \underline{P}(\underline{0}) = \underline{0}$,

(2) $\underline{f}^0$ is a normal vectorfield,

(3) application of $\underline{x} = \underline{\xi} + \varepsilon\underline{P}(\underline{\xi})$ yields, for ε small enough

$$\dot{\underline{\xi}} = A_0\underline{\xi} + \varepsilon\underline{f}^0(\underline{\xi}) + \varepsilon^2\underline{g}(\xi,\varepsilon),\ \underline{\xi} \in \tilde{\Omega},\ |\varepsilon| < \tilde{\varepsilon}_0,$$

where $\underline{g} \in C^1\ (\tilde{D}\times(-\tilde{\varepsilon}_0,\tilde{\varepsilon}_0), \mathbb{R}^{2n})$;
here is $0 < \tilde{\varepsilon}_0 \leqslant \varepsilon_0$ and $\tilde{D}$ a domain close to D (that is $\tilde{D} \to D$ for $\varepsilon_0 \to 0$). ■

Remark 1: It is not difficult to extend theorem 3.1 to the case where $\frac{\omega_i}{\omega_j} = \frac{n_i}{n_j} + \varepsilon\delta_i$, with $n_i,n_j \in \mathbb{N}$ and $\delta_i \in \mathbb{R}$.

Remark 2: Theorem 3.1 concerns an algorithm for the calculation of the first order normal form of (3.7). In a straightforward way this algorithm can be extended to calculate higher order normal forms.

In order to formulate a theorem on the existence and stability of periodic solutions of system (3.1) two transformations are needed.

The first and most obvious one is (3.5) which puts by virtue of theorem 3.1 system (3.1) into first order normal form:

$$\dot{\underline{\xi}} = A_0\underline{\xi} + \varepsilon\underline{f}^0(\underline{\xi}) + \varepsilon^2\underline{g}(\underline{\xi},\varepsilon), \; \underline{\xi} \in \tilde{D}, \; \varepsilon \in (-\tilde{\varepsilon}_0,\tilde{\varepsilon}_0).$$

The independent variable and the invariance property (3.3) of the normalised vectorfield play a part in the second transformation. With:

$$\underline{\xi}(t) = e^{t(1-\varepsilon\eta_\varepsilon)A_0}\underline{y}(t) \tag{3.8}$$

for some $\eta_\varepsilon \in \mathbb{R}$ (η_ε depends on ε and will be specified in theorem 3.2), it follows that $t \to \underline{y}(t)$ satisfies:

$$\begin{aligned}\dot{\underline{y}}(t) &= \varepsilon\left(\eta_\varepsilon A_0\underline{y} + \underline{f}^0(\underline{y})\right) + \varepsilon^2\tilde{\underline{g}}(t,\underline{y},\eta_\varepsilon,\varepsilon) = \\ &= \varepsilon\underline{F}(\underline{y},\eta_\varepsilon) + \varepsilon^2\tilde{\underline{g}}(t,\underline{y},\eta_\varepsilon,\varepsilon),\end{aligned} \tag{3.9}$$

where

(.) $\underline{F}(\underline{y},\eta_\varepsilon) \equiv \eta_\varepsilon A_0\underline{y} + \underline{f}^0(\underline{y})$; $\underline{F}$ is a normal vectorfield,

(.) $\tilde{\underline{g}}(t,\underline{y},\eta_\varepsilon,\varepsilon) \equiv e^{-t(1-\varepsilon\eta_\varepsilon)A_0}\underline{g}(e^{t(1-\varepsilon\eta_\varepsilon)A_0}\underline{y},\varepsilon)$.

As aforementioned the transformation $\underline{y}(t) = e^{-A_0t}\underline{\xi}(t)$ may be used to establish periodic solutions with period T_0. In general, however, when periodic solutions exist of system (3.1) their period depends on ε. Assuming that the period is smooth with respect to ε one may write $T_\varepsilon = \frac{T_0}{1-\varepsilon\eta_\varepsilon}$ where η_ε depends continuous on ε and $\lim\limits_{\varepsilon\to 0}\varepsilon\eta_\varepsilon = 0$. This motivates the introduction of the second transformation.

Definition: $\underline{G} : \tilde{D}\times\mathbb{R}\times(-\tilde{\varepsilon}_0,\tilde{\varepsilon}_0) \to \mathbb{R}^{2n}$,

$$\underline{G}(\underline{y}^0_\varepsilon,\eta_\varepsilon,\varepsilon) = \int_0^{T_\varepsilon}\left[\underline{F}(\underline{y}_\varepsilon(t,\underline{y}^0),\eta_\varepsilon) + \varepsilon\tilde{\underline{g}}(t,\underline{y}_\varepsilon(t,\underline{y}^0_\varepsilon),\eta_\varepsilon,\varepsilon)\right]dt,$$

where $t \to \underline{y}_\varepsilon(t,\underline{y}^0_\varepsilon)$ is the solution of (3.9) satisfying $\underline{y}_\varepsilon(0,\underline{y}^0_\varepsilon) = \underline{y}^0_\varepsilon$.

Observe that:

(1) $\underline{G} \in C^1\ (\tilde{D}\times\mathbb{R}\times(-\tilde{\varepsilon}_0,\tilde{\varepsilon}_0),\mathbb{R}^{2n})$,

(2) for $\varepsilon \neq 0$ $t \to \underline{y}_\varepsilon(t,\underline{y}_\varepsilon^0)$ is a periodic solution of (3.9) with period T if and only if $G(\underline{y}_\varepsilon^0,\eta_\varepsilon,\varepsilon) = \underline{0}$.

With the previous transformations the problem of finding periodic solutions of 3.1 has been reduced to the problem of finding critical points of G. Now, by using continuity arguments and the implicit function theorem the following theorem may be proved (see [8]):

Theorem 3.2: Suppose that there exists a $(\underline{y}_0^0,\eta_0) \in \tilde{D}\times\mathbb{R}$, $\underline{y}_0^0 \neq \underline{0}$ and a $i \in \{1,\ldots,2n\}$ such that:

(i) $\underline{F}(\underline{y}_0^0,\eta_0) = \underline{0}$,

(ii) The Jacobian matrix of $\underline{F}$ with respect to the variables $\eta,y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_{2n}$ in the point $(\underline{y}_0^0,\eta_0)$ is regular. Then, for ε sufficiently small ($\varepsilon \neq 0$), there is exactly one critical point of $\underline{G}$, say $(\underline{y}_\varepsilon^0,\eta_\varepsilon)$, in a $O(\varepsilon)$ neighbourhood of $(\underline{y}_0^0,\eta_0)$. This implies the existence of a non-trivial unique periodic solution $\underline{x}(t,\underline{x}_\varepsilon^0)$ of (3.1) with period $T_\varepsilon = \dfrac{T_0}{1-\varepsilon\eta_\varepsilon}$ satisfying $\underline{x}(0,\underline{x}_\varepsilon^0) = \underline{x}_\varepsilon^0$; $\underline{x}_\varepsilon^0 := \underline{y}_\varepsilon^0 + \varepsilon\underline{P}(\underline{y}_\varepsilon^0)$. Furthermore, the periodic solution is stable if the Jacobian matrix $\varepsilon\dfrac{\delta(F_1,\ldots,F_{i-1},F_{i+1},\ldots,F_{2n})}{\delta(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_{2n})}$ in the point $(\underline{y}_0^0,\eta_0)$ is stable and unstable if the Jacobian is unstable

■

Remark 3: The condition (i) in theorem 3.2 is a necessary condition i.e. if $\underline{F}(\underline{y},\underline{\eta}) \neq \underline{0}$ for all $\underline{y} \in \tilde{D}$ and $\eta \in \mathbb{R}$ then there does not exist a periodic solution of 3.1 with smooth period.

4. PERIODIC SOLUTIONS OF THE MODEL EQUATIONS

In this section periodic solutions of the model equations will be established on the basis of the theory of section 3. In this theory it is assumed that the unperturbed system has periodic solutions with period T_0. This implies that $\Omega \in Q$ (in the model equations (2.6) $\omega_1 = \Omega$, $\omega_2 = 1$). However, it is not difficult to show that the results which will be obtained also apply for the case that $\Omega = \frac{n}{m} + \delta\varepsilon^2$ with $n,m \in \mathbb{N}$ and $\delta \in \mathbb{R}$. The model equations (2.6) may be written in the form:

$$\dot{\underline{x}} = A_0\underline{x} + \varepsilon\underline{f}(\underline{x}), \quad \underline{x} \in \mathbb{R}^4, \quad \varepsilon > 0. \tag{4.1}$$

In order to analyse this system one has to calculate:

(1) The first order normal form of (4.1) with respect to A_0 (Theorem 3.1):

$$\dot{\underline{\xi}} = A_0\underline{\xi} + \varepsilon\underline{f}^0(\underline{\xi}) + O(\varepsilon^2), \tag{4.2}$$

(2) The zeroes $(\underline{y}_0^0, \eta_0) \in \mathbb{R}^4 \times \mathbb{R}$ of the vector function $\underline{F}$:

$$\underline{F}(\underline{y},\eta) = \eta A_0\underline{y} + \underline{f}^0(\underline{y}). \tag{4.3}$$

Subsequently, Theorem (3.2) may be applied; firstly, one may decide whether the critical points, if any, correspond to periodic solutions of (4.1) and secondly, one may investigate their stability.

An interesting parameter in the equations of motion is $\bar{\alpha}_s$, which determines the position of the ridge. Interesting here means that a number of terms in f i.e. $a_4, a_{24}, a_{444}, b_2, b_{22}$ and b_{44} vanish when one sets $\bar{\alpha}_s = 0$. In what follows the case $\bar{\alpha}_s = O(\varepsilon)$, which includes $\bar{\alpha}_s = 0$, will be studied and some results will be presented when $\bar{\alpha}_s = O(1)$.

The case $\bar{\alpha}_s = O(\varepsilon)$

As already pointed out the normal form depends on A_0 and hence on Ω. One may distinguish two cases: $\Omega \neq 2$ and $\Omega = 2$.

(i) $\Omega \neq 2$:

The first order normal form of system (2.6) becomes:

$$\begin{aligned}
\dot{\xi}_1 &= \quad \xi_2 + \varepsilon\frac{a_2}{2}\xi_1 + O(\varepsilon^2),\\
\dot{\xi}_2 &= -\Omega^2\xi_1 + \varepsilon\frac{a_2}{2}\xi_2 + O(\varepsilon^2),\\
\dot{\xi}_3 &= \quad \xi_4 + \varepsilon\left[\frac{b_4}{2}\xi_3 + \frac{3b_{444}}{8}(\xi_3^2 + \xi_4^2)\xi_3\right] + O(\varepsilon^2),
\end{aligned} \tag{4.4}$$

$$\dot{\xi}_4 = -\xi_3 + \varepsilon\left[\frac{b_4}{2}\xi_4 + \frac{3b_{444}}{8}(\xi_3^2+\xi_4^2)\xi_4\right] + O(\varepsilon^2).$$

As critical points of the vector function F one finds:

(1) $\underline{y}_0^0 = (0,0,0,0)^T$ and η_0 arbitrary,

(2) $\underline{y}_0^0 = (0,0,\rho\cos\phi,\rho\sin\phi)^T$ and $\eta_0 = 0$,

where $\rho = \sqrt{-\frac{4}{3}\frac{b_4}{b_{444}}}$ and $\phi \in [0,2\pi]$ arbitrary.

The critical point (1) corresponds with the trivial periodic solution $\underline{x}(\tau) \equiv \underline{0}$; further analysis show that there are no non-trivial periodic solutions of (4.1) in the neighbourhood of the origin (the origin is unstable). Although (2) suggests that there is an infinite number of periodic solutions one can show that they all correspond to one and the same periodic solution which is *stable*. As an approximation for the periodic solution one has:

$\underline{x}(\tau) = e^{\tau A_0}\underline{y}_0^0 + O(\varepsilon)$ on time-scale $O(\frac{1}{\varepsilon})$ and with period $T_\varepsilon = 2\pi m + O(\varepsilon^2)$ where $\Omega = \frac{n}{m}$ and gcd(n,m) = 1.

Remark 4: One may observe that the first two equations up to order ε of system (4.4) are decoupled from the other two and can be solved exactly; both solutions are asymptotically stable, i.e. tend to zero if $t \to \infty$. The third and fourth equation represent, up to order ε, the Van der Pol equation in first order normal form.

(ii) $\Omega = 2$:

Now the first order normal form of system (2.6) becomes:

$$\begin{aligned}
\dot{\xi}_1 &= \xi_2 + \varepsilon\left[\frac{a_2}{2}\xi_1 + \frac{a_{44}}{4}\xi_3\xi_4\right] + O(\varepsilon^2),\\
\dot{\xi}_2 &= -4\xi_1 + \varepsilon\left[\frac{a_2}{2}\xi_2 + \frac{a_{44}}{4}(\xi_4^2 - \xi_3^2)\right] + O(\varepsilon^2),\\
\dot{\xi}_3 &= \xi_4 + \varepsilon\left[\frac{b_4}{2}\xi_3 + \frac{b_{24}}{4}(2\xi_1\xi_4 - \xi_2\xi_3) + \frac{3b_{444}}{8}(\xi_3^2+\xi_4^2)\xi_3\right] + O(\varepsilon^2),\\
\dot{\xi}_4 &= -\xi_3 + \varepsilon\left[\frac{b_4}{2}\xi_4 + \frac{b_{24}}{4}(2\xi_1\xi_3 + \xi_2\xi_4) + \frac{3b_{444}}{8}(\xi_3^2+\xi_4^2)\xi_4\right] + O(\varepsilon^2).
\end{aligned} \tag{4.5}$$

As zeroes of $\underline{F}$ one finds:

(1) $\underline{y}_0^0 = (0,0,0,0)^T$ and η_0 arbitrary,

(2) $\underline{y}_0^0 = (0,\rho_1,0,\rho_2)^T$ and $\eta_0 = 0$,

(3) $\underline{y}_0^0 = (y_1^0,y_2^0,0,y_4^0)^T$ and $\eta_0 = \hat{\eta}$,

(4) $\underline{y}_0^0 = (-y_1^0,y_2^0,0,y_4^0)^T$ and $\eta_0 = -\hat{\eta}$.

where

$$(.)\ \rho_1 = \frac{-a_{44}}{2a_2}\rho_2^2,\ \rho_2 = \sqrt{\frac{4a_2b_4}{a_{44}b_{24} - 3a_2b_{444}}},$$

$$(.)\ y_1^0 = -\frac{2\hat{\eta}}{b_{24}},\ y_2^0 = \frac{a_2}{b_{24}},\ y_4^0 = \sqrt{\frac{2(a_2+2b_4)}{-3b_{444}}},\ \hat{\eta} = \frac{1}{4}\sqrt{\frac{a_{44}b_{24}(a_2+2b_4)}{3b_{444}} - a_2^2}. \tag{4.6}$$

The critical point (1) again corresponds with the trivial solution ($\underline{x}(\tau) \equiv 0$) and there is no non-trivial periodic solution in the neighbourhood of the origin. For the other three critical points condition (ii) in theorem 3.2 holds if

$$a_2^2 \neq \frac{a_{44}b_{24}(a_2 + 2b_4)}{3b_{444}} \quad (\Leftrightarrow \hat{\eta} \neq 0).$$

One can thus distinguish:

(a) $a_2^2 \neq \dfrac{a_{44}b_{24}(a_2 + 2b_4)}{3b_{444}}$.

In this case one can apply theorem 3.2. One finds that system (4.1) has at least three different periodic solutions; these periodic solutions are stable and as approximation one finds:

$x(\tau) = e^{\tau A_0}y_0^0 + O(\varepsilon)$ on time-scale $O(\frac{1}{\varepsilon})$ and

with period $T_\varepsilon = 2\pi + \varepsilon\eta_0 + O(\varepsilon^2)$.

Finally, by virtue of remark 3 one may conclude that there are exactly 3 periodic solutions with smooth period (for ε sufficiently small).

(b) $a_2^2 = \dfrac{a_{44}b_{24}(a_2 + 2b_4)}{3b_{444}}$.

In this case the critical points (2), (3) and (4) coincide. One may therefore conclude, on the basis of remark 3, that system (4.1) has at most one periodic solution with smooth period.

Remark 5: Since in case (a) $\hat{\eta} \neq 0$ the periodic solutions have also different periods:

$$2\pi + O(\varepsilon^2) \quad \text{for} \quad \underline{y}_0^0 = (0,\rho_1,0,\rho_2)^T,$$

$$2\pi + \varepsilon\hat{\eta} + O(\varepsilon^2) \quad \text{for} \quad \underline{y}_0^0 = (y_1^0,y_2^0,0,y_3^0)^T \quad \text{and}$$

$$2\pi - \varepsilon\hat{\eta} + O(\varepsilon^2) \quad \text{for} \quad \underline{y}_0^0 = (-y_1^0,y_2^0,0,y_4^0)^T.$$

As is shown in section 1, the relation between the motion of the cylinder in the (x,y)-plane and a solution of 4.1 is given by:

$$x(t) = \frac{v_\infty}{\omega_2} x_1(\omega_2 t) + \frac{\varepsilon c_0^D v_\infty}{\omega_2^2\Omega^2}, \qquad \dot{x}(t) = v_\infty x_2(\omega_2 t)$$

$$y(t) = \frac{v_\infty}{\omega_2} x_3(\omega_2 t) + \frac{\varepsilon}{\omega_2\Omega^2}\left(c_1^L\bar{\alpha}_s + c_3^L\bar{\alpha}_s^3\right) v_\infty, \quad \dot{y}(t) = v_\infty x_4(\omega_2 t)$$

One can therefore conclude that in the case $\bar{\alpha}_s = O(\varepsilon)$ the cylinder can have exactly 4 periodic orbits (if $\hat{\eta} \neq 0$) in the (x,y)-plane; 1 in the case of $\Omega \neq 2$ and 3 if $\Omega = 2$.

The approximations for these orbits are listed in table 4.1 and they are sketched in Fig. 4.1 to Fig. 4.4.

TABLE 4.1

Resonance	*Periodic orbit (on time-scale $O(1/\varepsilon)$)*	*Period*	*Stability*	*Fig.*
$\Omega \neq 2$	$x(t) = O(\varepsilon)$ $y(t) = \frac{\rho v_\infty}{\omega_2} \sin(\omega_2 t) + O(\varepsilon)$	$\frac{2\pi m}{\omega_2} + O(\varepsilon^2)$	stable	4.1
$\Omega = 2$	$x(t) = \frac{\rho_1 v_\infty}{2\omega_2} \sin(2\omega_2 t) + O(\varepsilon)$ $y(t) = \frac{\rho_2 v_\infty}{\omega_2} \sin(\omega_2 t) + O(\varepsilon)$	$\frac{2\pi}{\omega_2} + O(\varepsilon^2)$	stable	4.2
$\Omega = 2$	$x(t) = \frac{\rho_3 v_\infty}{\omega_2} \sin(2\omega_2 t+\theta) + O(\varepsilon)$ $y(t) = \frac{y_4^0 v_\infty}{\omega_2} \sin(\omega_2 t) + O(\varepsilon)$	$\frac{2\pi+\varepsilon\hat{\eta}}{\omega_2} + O(\varepsilon^2)$	stable	4.3
$\Omega = 2$	$x(t) = \frac{\rho_3 v_\infty}{\omega_2} \sin(2\omega_2 t-\theta) + O(\varepsilon)$ $y(t) = \frac{y_4^0 v_\infty}{\omega_2} \sin(\omega_2 t) + O(\varepsilon)$	$\frac{2\pi-\varepsilon\hat{\eta}}{\omega_2} + O(\varepsilon^2)$	stable	4.4

Note: $\rho_3 = \sqrt{(y_1^0)^2 + (y_2^0)^2/4}$, $\theta \in (0,\frac{\pi}{2})$ satisfying $\mathrm{tg}\theta = \frac{y_2^0}{2y_1^0}$.

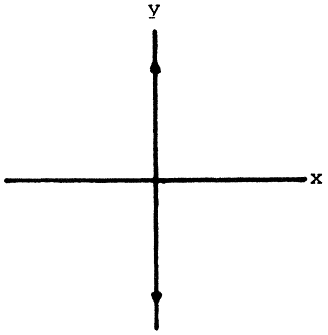


Fig. 4.1

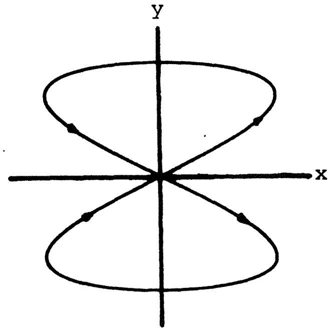


Fig. 4.2

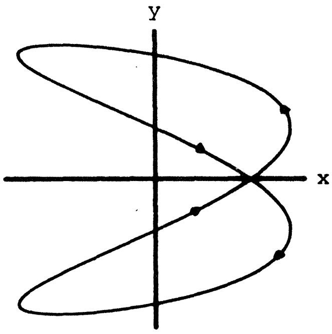


Fig. 4.3. $\theta = \frac{\pi}{4}$

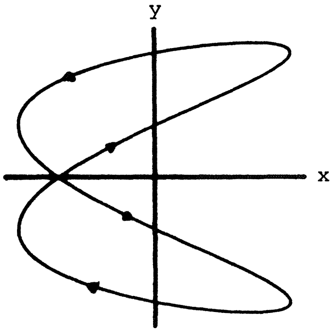


Fig. 4.4. $\theta = \frac{\pi}{4}$

The case $\bar{\alpha}_s = O(1)$

In this case all non-linear terms in the model equations in section 2 have to be taken into account. It turns out that if one compares the results one gets here with the case $\bar{\alpha}_s = O(\varepsilon)$ one should distinguish 2 cases:

(i) $\Omega \neq 1$ and $\Omega \neq \frac{1}{2}$:
Although the original differential equations ((4.1)) are different from the case $\bar{\alpha}_s = O(\varepsilon)$, they have the same first order normal form ((4.4) if $\Omega \neq 2$, (4.5) if $\Omega = 2$). Hence the same results apply.

(ii) $\Omega = 1$ or $\Omega = \frac{1}{2}$:

In these 2 cases the first order normal forms are different from those found for $\bar{\alpha}_s = O(\varepsilon)$. In fact, preliminary results show that they do not only give rise to resonance, which did not appear for these frequency ratio's in the case $\bar{\alpha}_s = O(\varepsilon)$, but also instability occurs and periodic solutions vanish as $\bar{\alpha}_s \to 0$. It would go too far to present all the possible periodic solutions here. In order to give the reader an idea of the kind of periodic solutions that may occur some typical periodic motions of the cylinder are sketched in Fig. 4.5 and Fig. 4.6. Both periodic orbits are unstable and vanish as $\bar{\alpha}_s \to 0$.

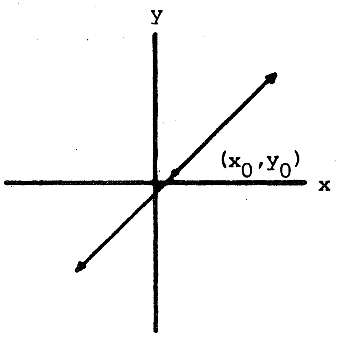


Fig. 4.5. $\Omega = 1$
(1:1 resonance)

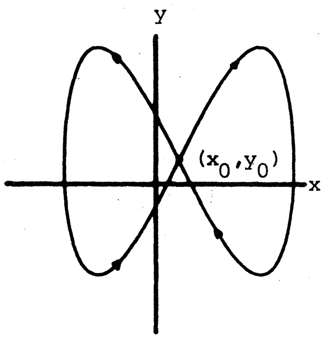


Fig. 4.6. $\Omega = \frac{1}{2}$
(1:2 resonance)

5. CONCLUDING REMARKS

In this work a non-Hamiltonian system of two non-linearly perturbed harmonic oscillators is studied with respect to the occurrence of periodic solutions. The mathematical framework e.g. the theory of normal forms and the theory of existence for periodic solutions has been developed and presented in a rather general way, that is, extensions to autonomous problems involving n ($n \in \mathbb{N}$, $n > 2$) degrees of freedom and general polynomial vectorfields can be carried out.

The non-linear coupling terms in the model equations as considered here are non-homogenous polynomials of the third degree. In the model equations the parameter $\bar{\alpha}_s$ determining the position of the ridge at the cylinder plays an important part. When $\bar{\alpha}_s$ is small and the frequency ratio $\Omega \neq 2$ the two degree of freedom motions are decoupled and the resulting oscillations involve one degree of freedom, perpendicular to the directions of the windfield only. When the frequency ratio $\Omega = 2$ (and $\bar{\alpha}_s$ is small) the harmonic oscillators are coupled by quadratic terms, which give rise to resonance. This resonance is induced by an interaction of (non-Hamiltonian) external forces. The type of solution corresponding to this resonance phenomenon is a limit-cycle in $\mathbb{R}^4$. As this type of resonance seems not to be known in the literature one may wonder whether a new type of resonance is involved in this problem.

Preliminary results show that in the case that $\bar{\alpha}_s$ is not anymore small as well in the 1:1 as in 1:2 frequency ratios resonance occurs, which disappears when $\alpha_s \rightarrow 0$. This is a bifurcation problem which has to be studied in more detail. In addition, it is of interest to introduce a detuning parameter δ defined by $\omega_1/\omega_2 = 2 + \varepsilon\delta$. It is known that this parameter is very appropriate to study near-resonance cases.

Finally one may say that since the general theory as presented here applies to problems involving a finite number of degrees of freedom, the way is open to develop and study extended model problems with three degrees of freedom including rotational oscillations of the cylinder induced by aerodynamic moments.

ACKNOWLEDGEMENT

The authors wish to thank Prof. J.W. Reyn and Ir. W.T. van Horssen for reading the manuscript and suggesting improvements.

The investigations were supported by the Netherlands Foundation for Mathematics SMC with financial aid from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

REFERENCES

[1] SIMPSON, A., 1982, *Wind-induced vibration of overhead power transmission lines*, Sci. Progr. Oxf. 68, 285-308.

[2] WAWZONEK, M.A. and PARKINSON, G.V., 1979, *Combined effects of galloping instability and vortex resonance*, Wind Eng. Proceedings of the 5th Conf., Fort Collins, Colorado, USA, Vol. 2, CERMAK, J.E. (ed.), Pergamon Press Oxford, 673-684.

[3] PARKINSON, G.V., 1974, *Mathematical models of flow-induced vibrations of bluff bodies*, IUTAM-IAHR Symp. on Flow-induced structural vibrations, Karlsruhe (1972), Springer-Verlag, Berlin.

[4] VAN DER BURGH, A.H.P., 1984, *On the galloping response of a simple aero-elastic oscillator*, Symp. on Flow-induced vibrations, Am. Soc. of Mech. Eng. 2, 37-53.

[5] ROSEAU, M., 1966, *Vibrations non linéaires et théorie de la stabilité*, Springer Tracts in Natural Philosophy 8, 108, Berlin.

[6] URABE, M., 1967, *Nonlinear autonomous oscillations*, Academic Press, New York, 106.

[7] HALE, J.K., 1980, *Ordinary Differential Equations*, R.E. Krieger Publ. Comp. Inc., USA, 194.

[8] VAN DER BEEK, C.G.A., 1987, *Normal forms and periodic solutions in the theory of non-linear oscillations. Existence and asymptotic theory*, Report to be published, Faculty of Mathematics and Informatics, Delft.

[9] MACSYMA, 1985, Reference Manual, Symbolics Inc., Version 11.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

The Interface Between Fresh and Salt Groundwater

J.R. Chan Hong
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands
and
D. Hilhorst
Laboratoire d'Analyse Numérique
C.N.R.S. et Université Paris-Sud
91405 Orsay, France

ABSTRACT.

We study a mathematical model for the interface between fresh and salt groundwater which consists of a Poisson equation in a strip for the stream function coupled to a time evolution equation for the moving interface. We first present a numerical study. The equation for the stream function is solved by means of a finite element method while an $S^{\alpha\beta}$ scheme is used to discretize the interface equation. We then prove a local existence and uniqueness result in a space of analytic functions; our proof also extends to the Rayleigh-Taylor instability in the case that the flow domain is a strip.

1. INTRODUCTION.

We consider a model which describes the two-dimensional motion of fresh and salt water through a horizontal aquifer. The fresh and salt water have different specific weights, denoted by γ_f and γ_s ($\gamma_f < \gamma_s$), respectively. As is common in hydrology (e.g. see Bear [2] and de Josselin de Jong [12]) it is assumed here that the fluids do not mix and are separated by a sharp interface. The difference in specific weight induces a flow and thus a displacement of the fluids and their interface. Our interest here is in the time evolution of this interface. Mathematically, this yields the following problem :

$$
\mathrm{P}\quad\begin{cases}
-\Delta\psi = \Gamma \dfrac{\partial}{\partial x}\,[H(u(x,t) - z)] & \text{in } \Omega \times \mathbb{R}^{+} \\
\psi = 0 & \text{on } \partial\Omega \times \mathbb{R}^{+} \\
u_t = \dfrac{d}{dx}\,[\psi(x,u(x,t),t] & \text{in } \mathbb{R} \times \mathbb{R}^{+} \\
u(x,0) = u_0(x) \qquad x \in \mathbb{R} &
\end{cases}
$$

where

$$\Omega = \{(x,z) \in \mathbb{R} \times (0,\Pi)\}\ ,\ \Gamma = \gamma_s - \gamma_f ,$$

and H is the Heaviside function : $H(s) = 1$ if $s > 0$ and $H(s) = 0$ if $s < 0$.

The function ψ is the stream function of the flow and u represents the height of the fresh-salt water interface, $0 \leqslant u \leqslant \Pi$.

We give a physical derivation of Problem P in Section 2. Using the Green function of the Laplace operator we then give an explicit formula for ψ and transform Problem P into a problem with a single integro-differential equation for u.

We describe a numerical method in Section 3. The equation for the stream function is solved by means of a finite element method. The equation for the interface, when considered apart, is hyperbolic ; the $S^{\alpha\beta}$ scheme introduced by Lerat and Peyret [14] is used for its discretization. A particularly interesting case is that where $u_0 = 0$ for small x, $u_0 = \Pi$ for large x, and $0 < u_0 < \Pi$ elsewhere. It is then essential to calculate as precisely as possible the x-coordinates $S_1(t)$ and $S_2(t)$ of the points where the interface reaches the bottom $z = 0$ and the top $z = \Pi$ of the aquifer. We do so by discretizing as well the differential equations for S_1 and S_2 and calculating u only between S_1 and S_2. Similar techniques have been used by DiBenedetto and Hoff [5] and Hoff [11] for the discretization of the porous media equation. We then show some numerical results. In particular it clearly appears that for large t u behaves as a rotating line. The contents of Sections 2 and 3 summarize a joint paper with C.J. van Duijn and J. van Kester [4].

Our analytical treatment of Problem P still leaves many questions unanswered. In Section 4 we restrict ourselves to the case where u_0 is bounded away from zero and from Π ; no hypothesis is made concerning the initial configuration of the salt and the fresh water : the salt water can either be below or above the fresh water. We prove the local existence and uniqueness in time of the solution in a space of analytic functions, closely following a method due to Bardos, Frisch, Sulem and Sulem [1] and Sulem and Sulem [15] ; the idea is to apply a Cauchy-Kowalewsky theorem in a scale of Banach spaces.

In Section 5 we extend results of [1] and [15] for the Kelvin-Helmholtz and the Rayleigh-Taylor instabilities to the case
that the spatial domain is a strip. For more details about the proofs of the results of Sections 4 and 5 we refer to Chan Hong [3].

By other methods, Duchon and Robert [6,7,8,9] obtain local and global existence and uniqueness results in the case that $\Omega = \mathbb{R}^2$.

2. THE PHYSICAL DERIVATION.

In this section we give a physical derivation of Problem P. We suppose that the interface Γ_u separating the fluids can be parametrized in the form $z = u(x)$. Then the specific weight throughout the flow domain is given by

$$\gamma(x,z) = (\gamma_s - \gamma_f)\, H(u(x) - z) + \gamma_f \quad \text{for } (x,z) \in \Omega.$$

A situation of particular interest is that where $u = 0$ for small x and $u = \Pi$ for large x (see Figure 1).

The motion of the fluid is governed by Darcy's law

$$(2.1) \qquad q + grad\ p + \gamma e_z = 0 \quad in\ \Omega$$

and the continuity equation (expressing the incompressibility of the fluid)

$$(2.2) \qquad div\ q = 0 \ .$$

In these equations q and p denote the velocity field and the pressure, respectively, and e_z the unit vector in the positive z direction. Equation (2.1) is written here in dimensionless variables. Finally we suppose that q satisfies the no-flow boundary condition

$$(2.3) \qquad q.\nu = 0 \qquad on\ \partial\Omega.$$

Suppose now that $q \in \{L^2(\Omega)\}^2$ and that u is a continuous function such that $\dfrac{\partial \gamma}{\partial x} \in H^{-1}(\Omega)$. One can deduce from (2.2) and (2.3) that there exists a function $\psi \in H^1_0(\Omega)$ such that $q =$ *curl* ψ and it follows from (2.1) and (2.3) that ψ is the unique solution of the boundary value problem P_ψ

$$(2.4) \qquad \left.\begin{array}{l} -\Delta\psi = \dfrac{\partial \gamma}{\partial x} = \Gamma \dfrac{\partial}{\partial x} \{H(u(x) - z)\}\ \ in\ H^{-1}(\Omega) \\ \psi = 0 \quad on \quad \partial\Omega \end{array}\right\} \ P_\psi$$

where $\Gamma = \gamma_s - \gamma_f$. Before we proceed to the physical derivation of the equation for the interface let us make two remarks.

(i) For $R > 0$, we define $\Omega_R = (-R,R) \times (0,\Pi)$. Then one can show that $\psi \in C^{0,\alpha}(\bar{\Omega}_R)$ for all $R > 0$ and all $\alpha \in (0,1)$. However, ψ is <u>not</u> continuously differentiable : if n is the normal unit vector to Γ_u, it follows from [4] that $\dfrac{\partial\psi}{\partial n}$ is discontinuous across Γ_u ;

(ii) in order to solve Problem P_ψ numerically, we shall solve in fact the corresponding boundary value problem with a homogeneous Dirichlet boundary condition in the bounded domain Ω_R. This procedure is justified by the following result. Let ψ be the solution of Problem P_ψ and ψ_R be the solution of the corresponding problem in Ω_R. Then, as $R \longrightarrow \infty$, ψ_R converges to ψ uniformly on compact subsets of Ω (see [4]).

Next we derive the interface equation. Let $u = u(x,t)$ denote the height of the fresh-salt interface at a certain time $t > 0$. Then the corresponding velocity field can be found by solving problem P_ψ. From this the displacement of the interface is calculated with the kinematic condition

$$q.n = \frac{u_t}{\sqrt{1 + u_x^2}} .$$

Using $q = curl\ \psi$, this equation becomes

$$(2.5) \qquad u_t = \frac{d}{dx} [\psi(x, u(x,t), t)] \qquad in\ \mathbb{R} \times \mathbb{R}^+ .$$

Problem P_ψ together with equation (2.5) and an initial condition for u gives Problem P.

In Section 3 we consider the case where $u_0 - \Pi H$ (where H is Heaviside function) has compact support. In view of the nature of the problem, we conjecture that the speed of propagation of the points $S_1(t)$ and $S_2(t)$ where the interface $u(x,t)$ reaches the bottom and the top of the aquifer, respectively, (see Figure 1), is finite. We define $S_1(t)$ and $S_2(t)$ by

$$S_1(t) = sup\ \{x \in \mathbb{R} \mid u(s,t) = 0 \quad for\ all\ s \leqslant x\}$$

and

$$S_2(t) = inf\ \{x \in \mathbb{R} \mid u(s,t) = \Pi \quad for\ all\ s \geqslant x\}$$

The differential equation for S_1 is found by observing that the speed at which S_1 travels in the (x,t) plane must be equal to the velocity of the salt water in the salt water toe. We have the formulas

$$(2.6) \qquad \dot{S}_1(t) = \lim_{x \downarrow S_1(t)} q_x(x,0,t) = - \lim_{x \downarrow S_1(t)} \frac{\partial \psi}{\partial z}(x,0,t)$$

$$(2.7) \qquad \dot{S}_2(t) = \lim_{x \uparrow S_2(t)} q_x(x,\Pi,t) = - \lim_{x \uparrow S_2(t)} \frac{\partial \psi}{\partial z}(x,\Pi,t)$$

We observe that equations (2.6) and (2.7) are not part of the original problem. However they will be used in the numerical algorithm.

Next we suppose that u is a smooth function such that

$u_x \in L^1(\mathbb{R})$ and $0 < u < \Pi$; in what follows we transform Problem P into an initial value problem with an integro-differential equation. Using the Green function of the Laplace operator we express ψ as

$$\psi(x,z,t) = -\frac{\Gamma}{4\Pi}\int \ln \frac{ch(x-y) - \cos\,(z - u(y,t))}{ch(x-y) - \cos\,(z + u(y,t))}\; u_y\,(y,t)\,dy$$

and deduce from (2.5) that u satisfies the integro-differential equation

$$u_t = -\frac{\Gamma}{4\Pi}\fint \frac{\partial}{\partial x}\left\{\ln \frac{ch(x-y) - \cos\,(u(x,t) - u(y,t))}{ch(x-y) - \cos\,(u(x,t) + u(y,t))}\right\}\; u_y\,(y,t)\,dy$$

For what follows it is handy to introduce the quantities

$$V(u,f)(x,t) = \fint \frac{sh(x-y)}{ch(x-y) - \cos(u(x,t) - u(y,t))}\, f(y,t)\,dy$$

$$W(u,f)(x,t) = \int \frac{sh(x-y)}{ch(x-y) - \cos(u(x,t) + u(y,t))}\, f(y,t)\,dy$$

$$X(u,f)(x,t) = \fint \frac{\sin\,(u(x,t) - u(y,t))}{ch(x-y) - \cos(u(x,t) - u(y,t))}\, f(y,t)\,dy$$

$$Y(u,f)(x,t) = \int \frac{\sin\,(u(x,t) + u(y,t))}{ch(x-y) - \cos(u(x,t) + u(y,t))}\, f(y,t)\,dy$$

$Q_1(u,f) = V(u,f) - W(u,f)$ and $Q_2(u,f) = X(u,f) - Y(u,f)$. Then the system (2.4),(2.5) can be written as the equation E,

$$\text{E} \qquad u_t = -\frac{\Gamma}{4\Pi}\{Q_1(u,u_x) + u_x\, Q_2(u,u_x)\}\ .$$

3. THE NUMERICAL METHOD.

In this section we describe a numerical algorithm for solving Problem P and show some numerical results. The algorithm is based on an explicit time integration scheme for the initial value problem

$$\begin{cases} u_t = \dfrac{d}{dx}\,[\psi(x,u(x,t),t)] & in\ (-R,R) \times \mathbb{R}^+ \\ u(x,0) = u_0(x) & in\ (-R,R) \end{cases}$$

3.1. *Discretization of the problem for* ψ .

Let $u^n(x)$ be the interface at time t^n. The corresponding stream function satisfies the problem

$$P^n_\psi \begin{cases} -\Delta\psi = \Gamma \dfrac{\partial}{\partial x}\{H(u^n(x) - z)\} & in\ \Omega_R \\ \psi = 0 & on\ \partial\Omega_R \end{cases}$$

Let $\mathcal{T}_h$ be a triangularization of $\bar{\Omega}_R$. Using the finite element method with piecewise linear basis functions, we obtain the following discretized problem

$$\text{Find } \psi_h \in V_h \text{ such that}$$

$$\int_{\Omega_R} grad\ \psi_h\ grad\ v_h = -\Gamma \int_{-R}^{R} (u^n_h)'(x)\ v_h(x,u^n_h(x))\,dx$$

$$\text{for all } v_h \in V_h$$

where

$$V_h = \{v_h \in \mathcal{C}(\bar{\Omega}_R) \mid \forall K \in \mathcal{T}_h \quad v_h \text{ is linear on } K \text{ and } v_h = 0 \text{ on } \partial\Omega_R\}$$

In [4] we present two variants for the triangularization of Ω_R. The first one consists in a fixed triangle distribution throughout Ω_R ; the other one is done with the help of an automatic mesh generator, allowing the mesh to vary at each time step in such a way that the discretized interface coincides with sides of triangles. In this way only values of ψ_h at mesh points are needed in the computations.

3.2. *Discretization of the interface equation.*

In order to discretize the interface equation, we use the $S^{\alpha\beta}$ explicit scheme of Lerat and Peyret [14] with α, β optimal. We consider in particular two cases :if $0 < u_0 < \Pi$, we compute u on the whole interval $(-R,R)$; in the case that

$u_0 = 0$ for small x, $u_0 = \Pi$ for large x, and $0 < u_0 < \Pi$ elsewhere, we use the extra equations (2.6) and (2.7) and calculate u only between S_1 and S_2.

a) **The $S^{\alpha\beta}$ scheme with α, β optimal.**

Let u_i^n be the approximation of $u(x_i^n, t^n)$ where the x_i^n's are the x-coordinates of the mesh points at time t^n. The function u_h^n that we have introduced above is obtained by linear interpolation. Further we use the notation $h_i^n = x_{i+1}^n - x_i^n$ and $\Delta t^n = t^{n+1} - t^n$. The $S^{\alpha\beta}$ scheme is given by

$$\tilde{u}_i^n = (1 - \beta) u_i^n + \beta\, u_{i+1}^n + \alpha \frac{\Delta t^n}{h_i^n}$$

$$\{\psi_h(x_{i+1}^n, u_{i+1}^n, t^n) - \psi_h(x_i^n, u_i^n, t^n)\}$$

$$u_i^{n+1} - u_i^n = \frac{\Delta t^n}{\alpha(h_i^n + h_{i-1}^n)} \{(\alpha - \beta)\psi_h(x_{i+1}^n, u_{i+1}^n, t^n) +$$

$$+ (2\beta - 1)\psi_h(x_i^n, u_i^n, t^n) +$$

$$(1 - \alpha - \beta)\psi_h(x_{i-1}^n, u_{i-1}^n, t^n) +$$

$$+ \tilde{\psi}_h(\tilde{x}_i^n, \tilde{u}_i^n, \tilde{t}^n) - \tilde{\psi}_h(\tilde{x}_{i-1}^n, \tilde{u}_{i-1}^n, \tilde{t}^n)\}$$

where $\tilde{x}_i^n = x_i^n + \beta\, h_i^n$, $\tilde{t}^n = t^n + \alpha(t^{n+1} - t^n)$, the predictor term $\tilde{u}_i^n$ is an approximation of $u(\tilde{x}_i^n, \tilde{t}^n)$ and $\psi_h \in V_h$, $\tilde{\psi}_h \in \tilde{V}_h$ are the solutions of

$$\int_{\Omega_R} grad\ \psi_h\ grad\ v_h = \Gamma \int_{-R}^{R} (u_h^n)'(x)\, v_h(x, u_h^n(x))\, dx$$

$$\textit{for all } v_h \in V_h$$

and of

$$\int_{\Omega_R} grad\ \tilde{\psi}_h\ grad\ \tilde{v}_h = \Gamma \int_{-R}^{R} (\tilde{u}_h^n)'(x)\, \tilde{v}_h(x, \tilde{u}_h^n(x))\, dx$$

$$\text{for all } \tilde{v}_h \in \tilde{V}_h ,$$

respectively, where $\tilde{u}_h^n$ and $\tilde{V}_h$ are the analogs of u_h^n and V_h. Our choice of the parameters $\alpha = 1 + \sqrt{5}/2$ and $\beta = 1/2$ is called optimal ; when applied to Burger's equation this choice of the parameters minimizes the dissipative effect of the scheme.

In order to insure the stability of the $S^{\alpha\beta}$scheme, we choose Δt^n such that it satisfies the CFL condition

$$C^n \ \Delta t^n \ / \ h^n_{max} \ \leqslant 1$$

where $h^n_{max} = \max_i \{h^n_i\}$ and where C^n is an approximation of the maximum of $\left|\dfrac{\partial \varphi_h}{\partial z}\right|$ on both sides of the discretized interface.

b) **Boundary conditions.**

In the case that $0 < u < \Pi$, numerical boundary conditions are necessary. Since the lines $x = \pm R$ are characteristics of the differential equation (2.5), we obtain the boundary values by approximating the equations on the characteristics

$$u_t \ (\pm R, t) \ = \ \varphi_x (\pm R, u(\pm R, t), t)$$

by means of a suitable scheme [4].

c. **Computation of $S_1$and S_2.**

Let S_1^n and $\tilde{S}_1^n$ be the approximated values of $S_1(t^n)$ and $S_1(\tilde{t}^n)$. We assume that $S_1^n > - R$ and define k_1^n and $\tilde{k}_1^n$ by

$$k_1^n = \min \{i, \ x_i^n > S_1^n\} \quad \text{and} \quad \tilde{k}_1^n = \min \{i, \ \tilde{x}_i^n > \tilde{S}_1^n\}$$

We discretize the equation (2.6) by the following analog of the second order Runge-Kutta scheme :

$$\tilde{S}^n_1 = S^n_1 - \alpha\,\Delta t^n \frac{\varphi_h\,(x^n_{k^n_1}, u^n_{k^n_1}, t^n)}{u^{\,n}_{k^n_1}},$$

$$S^{n+1}_1 = S^n_1 -$$

$$- \Delta t^n \left(\left(1 - \frac{1}{2\alpha}\right) \frac{\varphi_n\,(x^n_{k^n_1}, u^n_{k^n_1}, t^n)}{u^{\,n}_{k^n_1}} + \frac{1}{2\alpha} \frac{\tilde{\varphi}_h\,(\tilde{x}^n_{\tilde{k}^n_1}, \tilde{u}^n_{\tilde{k}^n_1}, \tilde{t}^n)}{\tilde{u}^n_{\tilde{k}^n_1}} \right)$$

and use similar formulas for the discretization of (2.7).

3.3. *Some numerical results.*

We choose $\Omega_R = (-3,3) \times (0,1)$, $\Gamma = 1$ and take as initial condition the function u_0

$$u_0 = \begin{cases} 0 & -R < x \leqslant -1/2 \\ 2x+1 & -1/2 < x \leqslant -1/6 \\ -x+1/2 & -1/6 < x \leqslant 1/6 \\ 2x & 1/6 < x \leqslant 1/2 \\ 1 & 1/2 < x < R \end{cases}$$

The computations of Figures 2 and 3 have been performed with the adaptive mesh. One observes that for large t the interface behaves as a rotating (nearly straight) line. It turns out that this rotating line is very close to the similarity solution computed by van Duijn and Hilhorst [10] in the case of an approximated model.

4. LOCAL EXISTENCE AND UNIQUENESS.

In this section we give a local existence and uniqueness result for Equation E together with an analytic initial function which is valid in both the cases where the constant Γ is positive and negative.

In order to get a feeling for the sort of difficulties that arise, we first linearize this problem around the constant $u = \Pi/2$. We obtain the initial value problem

$$(4.1)\quad \begin{cases} v_t = -\dfrac{\Gamma}{2\Pi} \displaystyle\int \dfrac{1}{sh(x-y)}\, v_y(y,t)\,dy & in\ \mathbb{R} \times \mathbb{R}^+ \\ v(x,0) = v_0(x) & in\ \mathbb{R} \end{cases}$$

Let $\hat{v}$ be the Fourier transform of v

$$\hat{v}(\xi,t) = \int e^{-2\Pi i x \xi}\, v(x,t)\,dx\ .$$

Then $\hat{v}$ satisfies the problem

$$\begin{cases} \hat{v}_t = -\Gamma\, \Pi\, \xi\ th\ (\Pi^2 \xi)\ \hat{v}(\xi,t) \\ \hat{v}(\xi,0) = \hat{v}_0(\xi) \end{cases}$$

whose solution is given by

$$\hat{v}(\xi,t) = e^{-\Gamma \Pi \xi\, th(\Pi^2 \xi)\, t}\ \hat{v}_0(\xi)$$

If $\Gamma > 0$, that is, if the salt water lies below the fresh water, and if $v_0 \in L^2(\mathbb{R})$, then $\hat{v}_0 \in L^2(\mathbb{R})$ and $\hat{v}(t) \in L^2(\mathbb{R})$ for all $t > 0$ so that Problem (4.1) is well-posed in $L^2(\mathbb{R})$. If on the other hand $\Gamma < 0$, that is if the fresh water lies below the salt water, then in general,

$$e^{-\Gamma \Pi \xi\, th(\Pi^2 \xi)\, t}\ \hat{v}_0(\xi) \notin L^2(\mathbb{R}) \quad for\ t > 0.$$

However, if v_0 is analytic in a strip of the complex plane, its Fourier transform decreases exponentially fast as $\xi \longrightarrow \pm\infty$ and the function

$$e^{-\Gamma \Pi \xi\, th(\Pi^2 \xi)\, t}\ \hat{v}_0(\xi)$$

remains in $L^2(\mathbb{R})$ during a time interval which is proportional to the width of the strip. Hence the motivation to use spaces of analytic functions for the study of Equation E.

Next we present our basic tools of study.

Definition 4.1. Let $\{B_s\}_{s>0}$ be a set of Banach spaces. $S = \bigcup_{s>0} B_s$ is a scale of Banach spaces if for all $0 < s' < s$ there holds$B_s \subset B_{s'}$ and $\|u\|_{s'} \leqslant \|u\|_s$ for all $u \in B_s$ where $\|.\|_s$ stands for the norm in B_s.

Fix $\alpha \in (0,1)$. For each $s > 0$ we consider the set of functions $\{u(z)\}$ analytic in the strip of the complex plane

$$b_s = \{z = x + i\sigma,\ x \in \mathbb{R},\ |\sigma| < s\} .$$

We define

$$\|u\|_s = |u|_s + \sup_{\substack{x + i\sigma \in b_s \\ y + i\sigma \in b_s}} \frac{|u(x + i\sigma) - u(y + i\sigma)|}{|x - y|^{\alpha}}$$

where

$$|u|_s = \sup_{x + i\sigma \in b_s} |u(x + i\sigma)|$$

and

$$\|u\|_{\mathcal{L}_s} = |u|_{L_s} + \\ + \sup_{|\sigma| < s} \sup_{d > 0} \frac{1}{d^{\alpha}} \int |u(x + d + i\sigma) - u(x + i\sigma)|dx$$

where

$$|u|_{L_s} = \sup_{|\sigma| < s} \int |u(x + i\sigma)|dx .$$

We will use the following notations, which are due to Bardos, Frisch, Sulem and Sulem [1] and Kano and Nishida [13] :

$$\mathcal{C}_s = \{u,\ u \text{ analytic in } b_s\ ,\ \|u\|_s < +\infty\}$$

$$\mathcal{L}_s = \{u,\ u \textit{ analytic in } b_s\ ,\ \|u\|_{\mathcal{L}_s} < +\infty\}$$

and

$$\mathcal{B}_s = \{u,\ u \textit{ analytic in } b_s\ ,\ |||u|||_s := \|u\|_s + \|u\|_{\mathcal{L}_s} < +\infty\}$$

and introduce the scale of Banach spaces

$$S = \bigcup_{s > 0} B_s$$

where

$$B_s = \{u \in \mathcal{C}_s \mid u_x \in \mathcal{B}_s \textit{ and } u_{xx} \in \mathcal{C}_s\}$$

is equipped with the norm

$$\|u\|_{B_s} = \max\ (\|u\|_s\ ,\ |||u_x|||_s\ ,\ \|u_{xx}\|_s)\ .$$

The following relations are very useful in what follows. For all $0 < s' < s$

$$\left\|\frac{\partial u}{\partial x}\right\|_{s'} \leqslant \frac{C}{s - s'}\ \|u\|_s \quad \text{for all } u \in \mathcal{C}_s \tag{4.2}$$

and

$$\left\|\frac{\partial u}{\partial x}\right\|_{\mathcal{L}_{s'}} \leqslant \frac{C}{s - s'}\ \|u\|_{\mathcal{L}_s} \quad \text{for all } u \in \mathcal{L}_s\ . \tag{4.3}$$

Our proof of the local existence and uniqueness of a solution of Equation E is based on the following Cauchy-Kowalewski theorem

<u>Theorem 4.2.</u> Let $S = \bigcup_{s > 0} B_s$ be a scale of Banach spaces and let s_0, R and η be positive constants, $u_0 \in B_{s_0}$ and $(u,t) \longrightarrow F(u,t)$ a continuous mapping of $\{u \in B_s\ ,\ \|u - u_0\|_{B_s} < R\} \times [-\eta,\eta]$ into $B_{s'}$, satisfying for any $0 < s' < s < s_0$ and any $t \in [-\eta,\eta]$

(i) $$\|F(u_1 , t) - F(u_2 t)\|_{B_{s'}} \leqslant C \frac{\|u_1 - u_2\|_{B_s}}{s - s'}$$

for all u_i , $i = 1,2$ such that $\|u_i - u_0\|_{B_s} < R$, where C is a constant which does not depend on t, u_1 ; u_2 , s and s' ;

(ii) $$\|F(u_0 , t)\|_{B_s} \leqslant \frac{C'}{s_0 - s}$$

where C' is a fixed constant.

Then there exists a positive number a and a unique function $u(t)$ such that, for every positive $s < s_0$ and $|t| < a(s_0 - s)$, u is a continuously differentiable function of t with values in B_s , $\|u - u_0\|_{B_s} < R$ and u satisfies

$$\begin{cases} u_t = F(u(t), t) & |t| < a(s_0 - s) \\ u(0) = u_0 \end{cases}$$

If, in addition to (i) and (ii) with t complex, F satisfies the following assumption : for $0 < s' < s < s_0$ and u holomorphic for $t \in \mathbb{C}$, $|t| < \eta$ valued in B_s with $\displaystyle\operatorname{Sup}_{|t| < \eta} \|u(t) - u_0\|_{B_s} < R$, $t \longrightarrow F(u(t),t)$ is a holomorphic function for $|t| < \eta$ valued in $B_{s'}$, then u is a holomorphic function of t with values in B_s .

In order to apply Theorem 4.2, we look for functions u and f which have an analytic continuation in the strips b_s satisfying $|\mathcal{I}m\, u_x|_s < K < 1$ and $|u - \frac{\Pi}{2}|_s < \gamma \frac{\Pi}{2}$ with $\gamma \in (0,1)$ so that the functions $Q_1 (u,f)$ and $Q_2 (u,f)$ can be analytically continued in those strips. Further we set

$$F(u,t) = - \frac{\Gamma}{4\Pi} \{Q_1 (u,u_x) + u_x\, Q_2 (u,u_x)\}$$

We have to obtain a priori estimates for $\|F(u_1 , t) - F(u_2 , t)\|_{B_{s'}}$ and for $\|F(u_0 , t)\|_{B_s}$ for suitable functions u_0 , u_1 and u_2 . Next we indicate the main steps needed to estimate $\|F(u_1 , t) - F(u_2 , t)\|_{B_{s'}}$, or in other words,

$\|F(u_1,t) - F(u_2,t)\|_s$, , $|||\frac{\partial}{\partial x}(F(u_1,t) - F(u_2,t)|||_s$, and $\|\frac{\partial^2}{\partial x^2}(F(u_1,t) - F(u_2,t))\|_s$,. These estimates will follow from estimating $\|F(u_1,t) - F(u_2,t)\|_s$, $|||F(u_1,t) - F(u_2,t)|||_s$ and $\|\frac{\partial}{\partial x}(F(u_1,t) - F(u_2,t)\|_s$ and using the relations (4.2) and (4.3).

Next we state results which permit to obtain estimates of $\|F(u_1,t) - F(u_2,t)\|_s$ and $|||F(u_1,t) - F(u_2,t)|||_s$.

<u>Proposition 4.3.</u> For u and $\tilde{u}$ in B_s with $|\mathcal{I}m\, u_x|_s < K < 1$, $|u - \frac{\Pi}{2}| < \gamma\frac{\Pi}{2}$, $\gamma \in (0,1)$, $max\,(\|u\|_s, \|u_x\|_s, \|u_{xx}\|_s) < R$ and similar conditions for $\tilde{u}$, $\tilde{u}_x$ and $\tilde{u}_{xx}$ and for f, $\tilde{f} \in \mathcal{B}_s$, we have the following estimates with i = 1,2 :

(i) if $\|f\|_s$, $\|\tilde{f}\|_s < R$, then

$$\|Q_i(u,f) - Q_i(\tilde{u},\tilde{f})\|_s \leqslant$$

$$\leqslant C(\gamma,R)\ \{\|u-\tilde{u}\|_s + \|u_x-\tilde{u}_x\|_s + \|u_{xx}-\tilde{u}_{xx}\|_s + \|f-\tilde{f}\|_s\}$$

(ii) if $\|f\|_{\mathcal{L}_s}$, $\|\tilde{f}\|_{\mathcal{L}_s} < R$, then

$$\|Q_i(u,f) - Q_i(\tilde{u},\tilde{f})\|_{\mathcal{L}_s} \leqslant$$

$$\leqslant C(\gamma,R)\ \{\|u-\tilde{u}\|_s + \|u_x-\tilde{u}_x\|_s + \|u_{xx}-\tilde{u}_{xx}\|_s + \|f-\tilde{f}\|_{\mathcal{L}_s}\}$$

where the constant $C(\gamma,R)$ is uniformly bounded in γ for $\gamma \in [0,\gamma_0]$ with $\gamma_0 < 1$.

There remains to estimate the term

$$\left\|\frac{\partial}{\partial x}(F(u_1,t) - F(u_2,t))\right\|_s .$$

To that purpose, one first derives the equalities

$$\frac{\partial}{\partial x} Q_1(u,f)(z) = Q_1\left(u, \frac{\partial}{\partial x}\left(\frac{f}{1+u_x^2}\right)\right)(z) + Q_2\left(u, \frac{\partial}{\partial x}\left(\frac{fu_x}{1+u_x^2}\right)\right)(z)$$

$$+ u_x(z)\, Q_2\left(u, \frac{\partial}{\partial x}\left(\frac{f}{1+u_x^2}\right)\right)(z) + u_x(z)\,(V+W)\left(u, \frac{\partial}{\partial x}\left(\frac{fu_x}{1+u_x^2}\right)\right)(z)$$

and

$$\frac{\partial}{\partial x} Q_2(u,f)(z) = Q_2\left(u, \frac{\partial}{\partial x}\left(\frac{f}{1+u_x^2}\right)\right)(z) + (V+W)\left(u, \frac{\partial}{\partial x}\left(\frac{fu_x}{1+u_x^2}\right)\right)(z)$$

$$- u_x(z)\, Q_1\left(u, \frac{\partial}{\partial x}\left(\frac{f}{1+u_x^2}\right)\right)(z) + u_x(z)\,(X+Y)\left(u, \frac{\partial}{\partial x}\left(\frac{fu_x}{1+u_x^2}\right)\right)(z)$$

for all $z \in b_s$.

Using these equalities, the estimates of Proposition 4.3 and similar estimates for *V, W, X and Y*, one can show the proposition

<u>Proposition 4.4.</u> For u and $\tilde{u}$ in B_s with $|\mathscr{Im}\, u_x| < K < 1$, $|u - \frac{\Pi}{2}|_s < \gamma \frac{\Pi}{2}$, $\gamma \in (0,1)$, $\|u\|_{B_s} < R$ and similar conditions for $\tilde{u}$ and for f and $\tilde{f}$ in $\mathscr{B}_s$ and f_x and $\tilde{f}_x$ in $\mathscr{C}_s$ with *max* $(|||f|||_s,\ \|f_x\|_s) < R$ and $\max(|||\tilde{f}|||_s, \|\tilde{f}_x\|_s) < R$, we have the following estimates with $i = 1,2$

$$\left\|\frac{\partial}{\partial x}\,(Q_i(u,f) - Q_i(\tilde{u},\tilde{f}))\right\|_s \leqslant$$

$$C(\gamma,R)\ \{\|u - \tilde{u}\|_{B_s} + |||f - \tilde{f}|||_s + \|f_x - \tilde{f}_x\|_s\}$$

where the constant $C(\gamma,R)$ is uniformly bounded in γ for $\gamma \in [0,\gamma_0]$ with $\gamma_0 < 1$.

Finally, one can deduce from Propositions 4.3 and 4.4 the following result :

Let u_0 be a function which can be analytically continued in b_{s_0} such that $u_0 \in B_{s_0}$ $|\mathcal{I}m\; u_0'|_{s_0} < K < 1$ and $|u_0 - \frac{\Pi}{2}|_{s_0} < r_0 \frac{\Pi}{2}$ with $r_0 \in (0,1)$ and let $R > 0$ be such that $R < min \left(K - |\mathcal{I}m\; u_0'|_{s_0}, r_0 \frac{\Pi}{2} - |u_0 - \frac{\Pi}{2}|_{s_0} \right)$. Then for any $0 < s' < s \leqslant s_0$ and any $t \in \mathbb{C}$

$$\|F(u_1, t) - F(u_2, t)\|_{B_{s'}} \leqslant C \frac{\|u_1 - u_2\|_{B_s}}{s - s'}$$

for all u_i, $i = 1,2$ such that $\|u_i - u_0\|_{B_s} \leqslant R$. Furthermore estimates similar to those of Propositions 4.3 and 4.4 allow to show that

$$\|F(u_0, t)\|_{B_s} \leqslant \frac{C'}{s_0 - s} .$$

Finally, we are able to state our main result.

<u>Theorem 4.5.</u> For any initial condition u_0 whose analytic continuation satisfies $u_0 \in B_{s_0}$ with $|\mathcal{I}m\; u_{0x}|_{s_0} < K < 1$ and $|u_0 - \frac{\Pi}{2}|_{s_0} < r_0 \frac{\Pi}{2}$ with $r_0 \in (0,1)$, there exists a constant a such that for $|t| < a(s_0 - s)$ Equation E has a unique solution u which is an analytic function of t with values in B_s.

5. FINITE TIME ANALYTICITY FOR THE RAYLEIGH-TAYLOR INSTABILITY IN A STRIP.

We consider the movement of two ideal incompressible fluids of constant densities ρ_1 and ρ_2 separated by a vortex sheet S_u. We suppose that the flow takes place in a strip $\Omega = \mathbb{R} \times (0,\Pi)$ and that the interface S_u can be parametrized in the form $z = u(x,t)$.

The momentum equation reads

$$\begin{cases} \dfrac{\partial}{\partial t}(\rho q) + \dfrac{\partial}{\partial x}(\rho q_x q) + \dfrac{\partial}{\partial z}(\rho q_z q) + grad\ p + \rho g = 0 \\ div\ q = 0 \end{cases}$$

where

ρ is the fluid density ; $\rho = \rho_1$ in $\{(x,z) \mid z > u(x,t)\}$ and $\rho = \rho_2$ in $\{(x,z) \mid z < u(x,t)\}$;

q_x and q_z are the x and z-components of the velocity q;

g is the gravity field.

One supposes here as well that the flow cannot cross the boundary :

$$q.\nu = 0 \quad on\ \partial\Omega$$

The vorticity density $\omega(x,t)$ on the interface is defined by

$$\int_\Omega \varphi\ curl\ q = \int \varphi(x,u(x,t))\ \omega(x,t)\,dx$$

for all $\varphi \in \mathscr{C}_0^\infty(\Omega)$.

Following the main lines of a derivation of Sulem and Sulem [15], Chan Hong [3] obtains the system

$$S \quad \begin{cases} u_t = \dfrac{1}{4\Pi}\{Q_1(u,\omega) + u_x\ Q_2(u,\omega)\} \\ \omega_t - 2\alpha\ B(u)\,\omega_t = 2S(u,\omega) \end{cases}$$

where

$$\alpha = \frac{\rho_2 - \rho_1}{\rho_2 + \rho_1}$$

$$B(u)\,\omega_t = \frac{1}{4\Pi}\,\{Q_2\,(u,\omega_t) - u_x\ Q_1\,(u,\omega_t)\}$$

and

$$S(u,\omega) = -\frac{\partial}{\partial x}\left\{-\frac{\omega\ Q_2\,(u,\omega)}{8\Pi} + \alpha\left(\frac{Q_1^2 + Q_2^2}{32\ \Pi^2} + \frac{\omega^2}{8\,(1+u_x^2)} + gu\right)\right\}$$

$$+ \frac{\alpha}{16\ \Pi^2}\,\frac{\partial}{\partial x}\,Q_1\ (u,(Q_1 + u_x\ Q_2)\,\omega)$$

Our result is quite similar to those that Sulem and Sulem [15] obtain in the case of other flow domains. Let $B_s^{(1)}$ be the Banach space

$$B_s^{(1)} = \{\omega \in \mathscr{B}_s \mid \omega_x \in \mathscr{C}_s\ \}$$

with the norm

$$\|\omega\|_{1,s} = |||\omega|||_s + \|\omega_x\|_s$$

and let $B_s^{(2)}$ be the Banach space

$$B_s^{(2)} = \{u \in \mathscr{C}_s \mid u_x \in \mathscr{B}_s \textit{ and } u_{xx} \in \mathscr{C}_s\}$$

with the norm

$$\|u\|_{2,s} = \|u\|_s + |||u_x|||_s + \|u_{xx}\|_s$$

Using an extension of Theorem 4.2 one can prove the following result [3] :

Theorem 5.1.
There exists a constant k_0 such that for any initial condition (u_0,ω_0) whose analytic continuation belongs to $B_{s_0}^{(2)} \times B_{s_0}^{(1)}$ and satisfies $|\mathscr{I}m\ u_{0x}|_{s_0} < K < 1$, $|u_0 - \frac{\Pi}{2}|_{s_0} < \gamma_0\,\frac{\Pi}{2}$ with $\gamma_0 \in (0,1)$ and $\|u_0 - \frac{\Pi}{2}\|_{2,s} \leqslant k_0$, there exists a constant a such that for all $|t| < a(s_0 - s)$ the system S has a unique solution (u,ω) which is an analytic function of t with values in $B_s^{(2)} \times B_s^{(1)}$.

Remark 5.2. In the case that $\alpha = 0$, which corresponds to the

Kelvin-Helmholtz instability, the result of Theorem 5.1 holds without any restriction on $\|u_0 - \frac{\Pi}{2}\|_{2,s}$.

REFERENCES.

[1] BARDOS C., FRISCH U., SULEM C. and P.L. SULEM, Finite time analyticity for the two and three dimensional Kelvin-Helmholtz instability, *Comm. Math. Phys.,* 80 (1981), 485-516.

[2] BEAR J., Dynamics of Fluids in Porous Media, Elsevier, New-York, 1975.

[3] CHAN HONG J.R., Sur quelques problèmes à frontière libre en hydrologie, Thèse de Doctorat, Ecole Centrale de Lyon, 1987.

[4] CHAN HONG J.R., VAN DUIJN C.J., HILHORST D. and J. VAN KESTER, The interface between fresh and salt groundwater : a numerical study, to appear.

[5] DIBENEDETTO E. and D. HOFF, An interface tracking algorithm for the porous media equation, *Trans. Amer. Math. Soc,* 284 (1984), 463-500.

[6] DUCHON J. and R. ROBERT, Estimation d'opérateurs intégraux du type de Cauchy dans les échelles d'Ovsjannikov et application, *Ann. Inst. Fourier,* 36 (1986), 83-95.

[7] DUCHON J. and R. ROBERT, Sur quelques problèmes à frontière libre analytique dans le plan, Séminaire Bony-Sjöstrand-Meyer, 1984-85, exposé n° X, Ecole Polytechnique.

[8] DUCHON J. and R. ROBERT, Perturbation quasi-différentielle d'un semi-groupe régularisant dans une échelle d'espaces de Banach, *C.R. Acad. Sci. Paris,* 301 (1985), 561-564.

[9] DUCHON J. and R. ROBERT, Solutions globales avec nappe tourbillonnaire pour les équations d'Euler dans le plan, *C. R. Acad. Sci. Paris,* 302 (1986), 183-186.

[10] VAN DUIJN C.J. and D. HILHORST, On a doubly degenerate parabolic equation in hydrology, to appear in J. Nonlinear Analysis T.M.A., 1987.

[11] HOFF D., A scheme for computing solutions and interface curves for a doubly degenerate parabolic equation, *SIAM J. Num. Anal.,* 22 (1985), 687-712.

[12] DE JOSSELIN DE JONG G., The simultaneous flow of fresh and salt water in aquifers of large horizontal extension determined by shear flow and vortex theory, in Proc. Euromech. 143, A. Verruit and F.B.J. Barends eds, Balkema Rotterdam, 1981.

[13] KANO T. and T. NISHIDA, Sur les ondes de surface de l'eau avec une justification mathématique des équations des ondes en eau peu profonde, *J. Math. Kyoto Univ.,* 19 (1979), 335-370.

[14] LERAT A. and R. PEYRET, Sur le choix de schémas aux

différences finies du second ordre fournissant des profils de choc sans oscillation, *C.R. Acad. Sci. Paris,* 277 (1973), 363-366.

[15] SULEM C. and P.L. SULEM, Finite time analyticity for the two and three dimensional Rayleigh-Taylor instability, *Trans. Amer. Math. Soc.,* 287 (1985) 127-160.

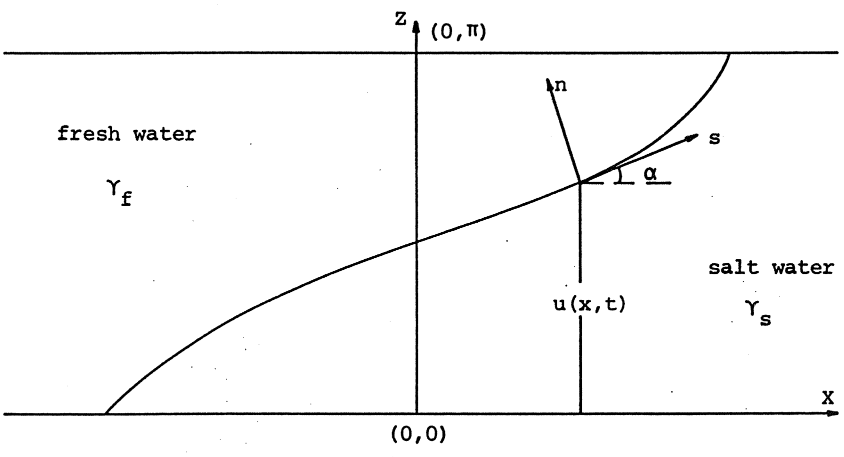


Figure 1. The distribution of fresh and salt groundwater in an aquifer.

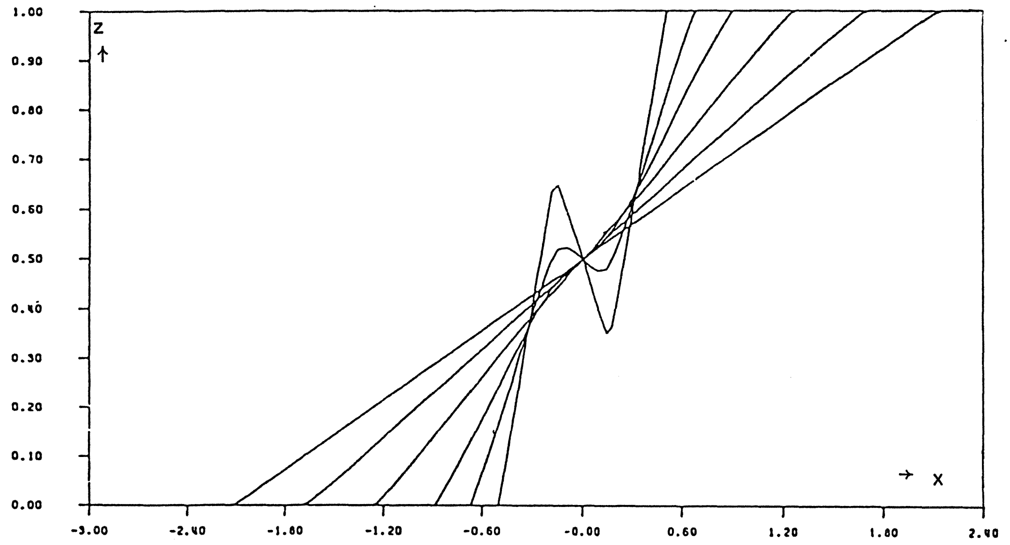


Figure 2. Convergence to a rotating line. The interface is computed at $t = 0.24$, $t = 0.6$, $t = 1.43$, $t = 2.77$, $t = 4.57$

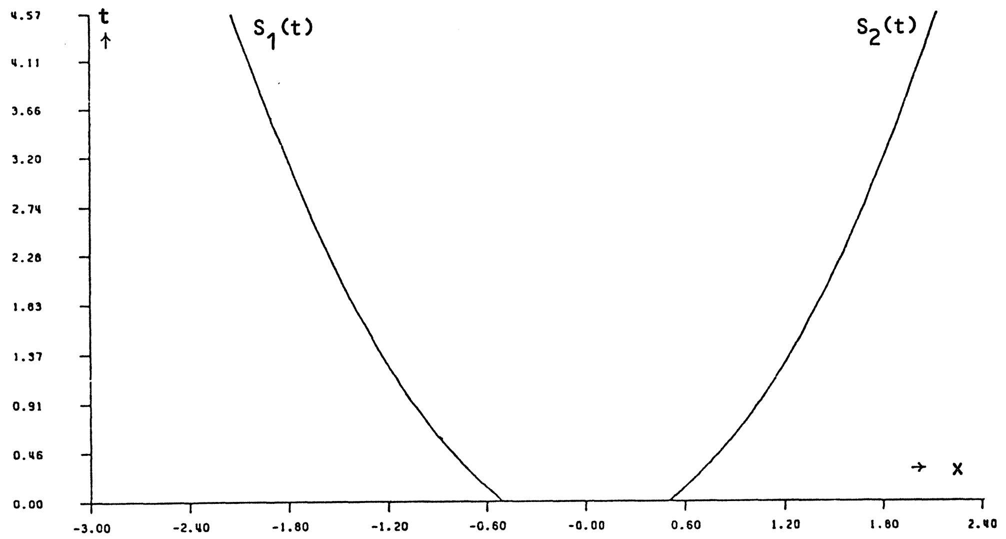


Figure 3. Evolution of the free boundaries S_1 and S_2 .

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Phase Portraits for Quadratic Systems with a Higher Order Singularity. I. Third and Fourth Order Points with Two Zero Eigenvalues

P. de Jager and J.W. Reyn
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands

ABSTRACT

This paper presents a classification of all possible phase portraits of quadratic systems of differential equations with a third or fourth order singular point with two zero eigenvalues. The singular points include the fourth order saddle node, the third order saddle point and the third order point, having an elliptic and a hyperbolic sector.

In a survey paper [4] on general properties of quadratic systems of differential equations in the plane Coppel states that what remains to be done is to determine all possible phase portraits of such systems, this being of great practical value. The present paper aims at giving a contribution in this direction. By a quadratic system is meant the system

$$\begin{aligned}\dot{x} &= a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 \equiv P(x,y) \quad ,\\ \dot{y} &= b_{00} + b_{10}x + b_{01}y + b_{20}x^2 + b_{11}xy + b_{02}y^2 \equiv Q(x,y) \quad ,\end{aligned} \tag{1}$$

for the functions $x = x(t)$, $y = y(t)$, where $\dot{} = \frac{d}{dt}$ and a_{ij}, $b_{ij} \in \mathbb{R}$. Since quadratic systems without singular points in the finite part of the plane have been classified by Gasull, Sheng Li-Ren and Llibre [6], we may assume that (1) has at least one singular point in the finite part of the plane, and we may shift the origin into this point: $a_{00} = b_{00} = 0$. Two limiting cases are then also classified: the linear case ($a_{20} = a_{11} = a_{02} = b_{20} = b_{11} = b_{02} = 0$) and the homogeneous case ($a_{10} = a_{01} = b_{10} = b_{01} = 0$) [5], [7]. If both linear and quadratic terms are present a lot of work remains to be done. If one or both eigenvalues in the singularity are zero, yet at least one linear term remains after transformation to the origin, a higher order singularity or multiple equilibrium point exists, which is considered in the present paper. Such higher order points were classified by Berlinskii [2], who makes use of other papers in Russian, which are not easily accessible. At present it is more convenient to use the classification of multiple equilibrium points for analytic systems as given by Andronov et al. in Chapter 9 of their book [1], of which an English translation is available. Also, the notion of order of a singular point (or multiplicity of an equilibrium point) as used in [2] may be improved somewhat by using that, implicitly present in the analysis given in [1]. The results of a renewed classification [8], however, agree with those given by Berlinskii {See also [3]} .

If the higher order singular point is in the origin, the quadratic system may be brought into a normal form by an affine transformation. If both eigenvalues are zero this normal forms reads [1, p. 346].

$$\begin{aligned}\dot{x} &= y + ax^2 + bxy + cy^2 \equiv P(x,y) \quad ,\\ \dot{y} &= dx^2 + exy + fy^2 \equiv Q(x,y) \quad .\end{aligned} \tag{2}$$

The order of the singular point in the origin may be defined as the maximum number of common zeros near the origin in the unfoldings of the functions $P(x,y)$ and $Q(x,y)$. This is the same as the maximum number of zeros of the unfolding of $\psi(x) \equiv Q\{x,\phi(x)\}$, where $\phi(x)$ is defined by $P\{x,\phi(x)\} = 0$, satisfying $\phi(0) = \phi'(0) = 0$; thus if $\psi(x) = a_n x^n + \ldots$, with $a_n \neq 0$, n is the order of the singular point. On the basis of Theorems 66 and 67, Chapter 9 of [1] there are for (2) the following cases:

(i) a fourth order saddle node, if $a \neq 0$, $d = e = 0$, $f \neq 0$; index 0,

(ii) a third order saddle point, if $ae < 0$, $d = 0$; index -1,

(iii) a third order point, having an elliptic and a hyperbolic sector, if $ae > 0$, $d = 0$; index 1,

(iv) a cusp point; the phase portrait is the union of two hyperbolic sectors and two separatrices, both tangent to the x asis, if $d \neq 0$, order 2; index 0.

In all cases div $\{P(0,0), Q(0,0)\} = 0$.

If only one eigenvalue in the singular point is zero, the normal form reads [1, p. 338]

$$\begin{aligned} \dot{x} &= ax^2 + bxy + cy^2 \equiv P(x,y) \quad , \\ \dot{y} &= y + dx^2 + exy + fy^2 \equiv Q(x,y) \quad . \end{aligned} \tag{3}$$

Similar to the previous case, the order of the point may now be defined as the maximum number of zeros in the unfolding of $\psi(x) \equiv P\{x,\phi(x)\}$, where $\phi(x)$ is defined by $Q\{x,\phi(x)\} = 0$, satisfying $\phi(0) = \phi'(0) = 0$. On the basis of Theorem 65, Chapter 9 of [1] there are for (3) the cases:

(i) a fourth order saddle node, if $a = b = 0$, $c \neq 0$, $d \neq 0$; index 0,

(ii) a third order saddle point, if $a = 0$, $bd > 0$; index -1,

(iii) a third order node, if $a = 0$, $bd < 0$; index 1,

(iv) a second order saddle node, if $a \neq 0$; index 0.

In all these cases div$\{(P(0,0), Q(0,0)\} \neq 0$ (= 1).

In the present paper the possible phase portraits are given for systems with a third order or a fourth order singular point having two zero eigenvalues. They include the fourth order saddle node, the third order saddle point and the third order point, having an elliptic and a hyperbolic sector. The phase portraits are characterized in the usual way: by the number, position and character of its singular points; by the position of its periodic solutions, if any; by the position of the separatrices, and by the behavior at infinity. Standard arguments will be used for the classification, such as local linearization in singular points, if possible, Dulac functions, integrating factors, continuity and index arguments. For the investigation of points at infinity a slightly different transformation than that given by Poincaré will be used, by putting $x = \rho(1-\rho)^{-1}\cos\theta$, $y = \rho(1-\rho)^{-1}\sin\theta$. System (2) then becomes, with $\rho(1-\rho)\frac{d}{dt} = \frac{d}{d\tau} = {}'$

$$\begin{aligned} \rho' &= \rho^2(1-\rho)^2 B_1(\theta) + \rho^3(1-\rho)C_1(\theta) \quad , \\ \theta' &= \rho(1-\rho)\, B_2(\theta) + \rho^2 C_2(\theta) \quad , \end{aligned} \tag{4}$$

, where $B_1(\theta) = \sin\theta\cos\theta$, $C_1(\theta) = a\cos^3\theta + (b+d)\sin\theta\cos^2\theta + (c+e)\sin^2\theta\cos\theta + f\sin^3\theta$,

, $B_2(\theta) = -\sin^2\theta$, $C_2(\theta) = d\cos^3\theta + (e-a)\sin\theta\cos^2\theta + (f-b)\sin^2\theta\cos\theta - c\sin^3\theta$.

Points at infinity are then represented on $\rho \equiv 1$; singular points appear in diametrically opposed pairs. If $C_2(\theta) \not\equiv 0$, $\rho \equiv 1$ consists of integral curves and possibly of singular points. In order to include such a curve into considerations using index theory, an extension of the usual Poincaré index of a planar vector field will be adopted by regarding $\rho \equiv 1$ as the limiting position of a closed curve near it [9]. If $C_2(\theta) \not\equiv 0$, it can then be deduced that the sum of the indices of the singular points on $\rho \le 1$ is equal to 1, where for the index of a singular point on $\rho \equiv 1$ only the vector field for $\rho \le 1$ is considered.

1. Quadratic systems with a fourth order saddle node and two zero eigenvalues.

If (0,0) is a fourth order saddle node with two zero eigenvalues, (2), with $d = e = 0$, $af \neq 0$ may be written, - if necessary by applying an affine transformation and/or a scaling of t -, in the form

$$\begin{aligned} \dot{x} &= y + x^2 + \lambda y^2 \quad , \\ \dot{y} &= y^2 \quad , \end{aligned} \tag{5}$$

, with $\lambda \in \mathbb{R}$.

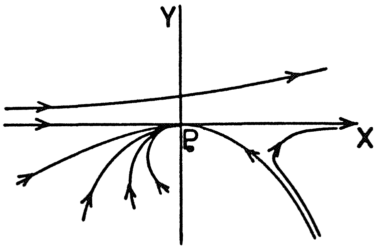


figure 1.

As illustrated in figure 1, the saddle node P_0 in (0,0) consists of two hyperbolic sectors, separated by the positive x axis, and a parabolic sector. One hyperbolic sector coincides with the half plane $y > 0$ and the other lies between the axis $x > 0$ and a separatrix tangent to this axis in P_0. The integral curves in the parabolic sector are tangent to the negative x axis in P_0.

Since there are no other singular points P_0 is the only possible candidate to be situated inside a periodic solution. However, such a periodic solution would have to intersect the x axis, which is a integral curve. Thus, uniqueness makes a periodic solution impossible.
From (4) follows, that the singular points at infinity are given by

$$C_2(\theta) \equiv (-\cos^2\theta + \sin\theta\cos\theta - \lambda\sin^2\theta)\sin\theta = 0,$$

so that there is a point P_1 at $\theta = 0(\pi)$, and, for $\lambda < \frac{1}{4}$, a point P_3 at $\theta = \operatorname{arc\,cot}\frac{1}{2}(1-\sqrt{1-4\lambda})$, and a point P_4 at $\theta = \operatorname{arc\,cot}\frac{1}{2}(1+\sqrt{1-4\lambda})$; these two points coincide for $\lambda = \frac{1}{4}$ in the point P_2 at $\theta = \operatorname{arc\,cot}\frac{1}{2}$. The character of these points may be found by local linearization and using Theorem 65, Chapter 9 of [1]. As a result follows, that P_1 and P_3 are nodes, P_4 is a saddle point and P_2 a saddle node. If is further observed that from (4) follows, that on P_0P_3 and on P_0P_4 (so also on P_0P_2) there is $\theta' < 0$, the continuity argument shows, that the phase portraits are as given in figure 2.

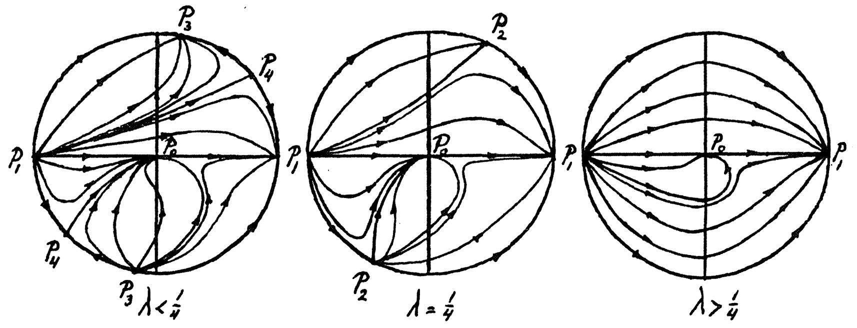


figure 2.

2. Quadratic systems with a third order saddle point and two zero eigenvalues.

If (0,0) is a third order saddle point with two zero eigenvalues, (2), with $d = 0$, $ae < 0$, may be written, if necessary by applying an affine transformation and/or a scaling of t, in the form

$$\begin{aligned} \dot{x} &= y + \lambda_1 x^2 + \lambda_2 xy + \lambda_3 y^2 \equiv P(x,y) \quad , \\ \dot{y} &= xy \qquad\qquad\qquad\quad\; \equiv Q(x,y) \quad , \end{aligned} \tag{6}$$

, with $\lambda_1 < 0$, $\lambda_2 \in \{0,1\}$, $\lambda_3 \in \mathbb{R}$.

As illustrated in figure 3, the saddle point in P_0: (0,0) consists of four hyperbolic sectors separated by four separatrices: the positive x axis and a curved separatrix tangent to it in P_0, and the negative x axis and a curved separatrix tangent to it in P_0; both curved separatrices lying in $y > 0$. System (6) has no limit cycles. In fact, it can be shown that for $\lambda_2 = 1$ the system has no periodic solutions, since $B(x,y) = y^{-2\lambda_1-1}$ is a Dulac function yielding

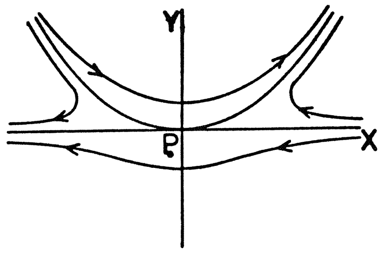


figure 3.

$$\frac{\partial}{\partial x}(BP) + \frac{\partial}{\partial y}(BQ) = \lambda_2 y^{-2\lambda_1}$$

, which is of constant sign in a half plane $y \gtrless 0$. A periodic solution must then cross the x axis, which is not possible since it consists of integral curves. For $\lambda_2 = 0$ the system is symmetric around the y axis, which excludes limit cycles, since all possible singular points are on the y axis and a limit cycle has to have a singular point in its interior.Another argumentation would be to use div(BP,BQ) $\equiv 0$, when applying Green's theorem to an annulus with a limit cycle as one of the boundary curves [9]. For $\lambda_2 = 0$, the Dulac function acts as an integrating factor to yield the solutions

$$y \equiv 0 \text{ and } \frac{2}{2\lambda_1 - 1} y + x^2 + \frac{\lambda_3}{\lambda_1 - 1} y^2 + \lambda_4 |y|^{2\lambda_1} = 0 , \quad \lambda_4 \in \mathbb{R} \quad ,$$

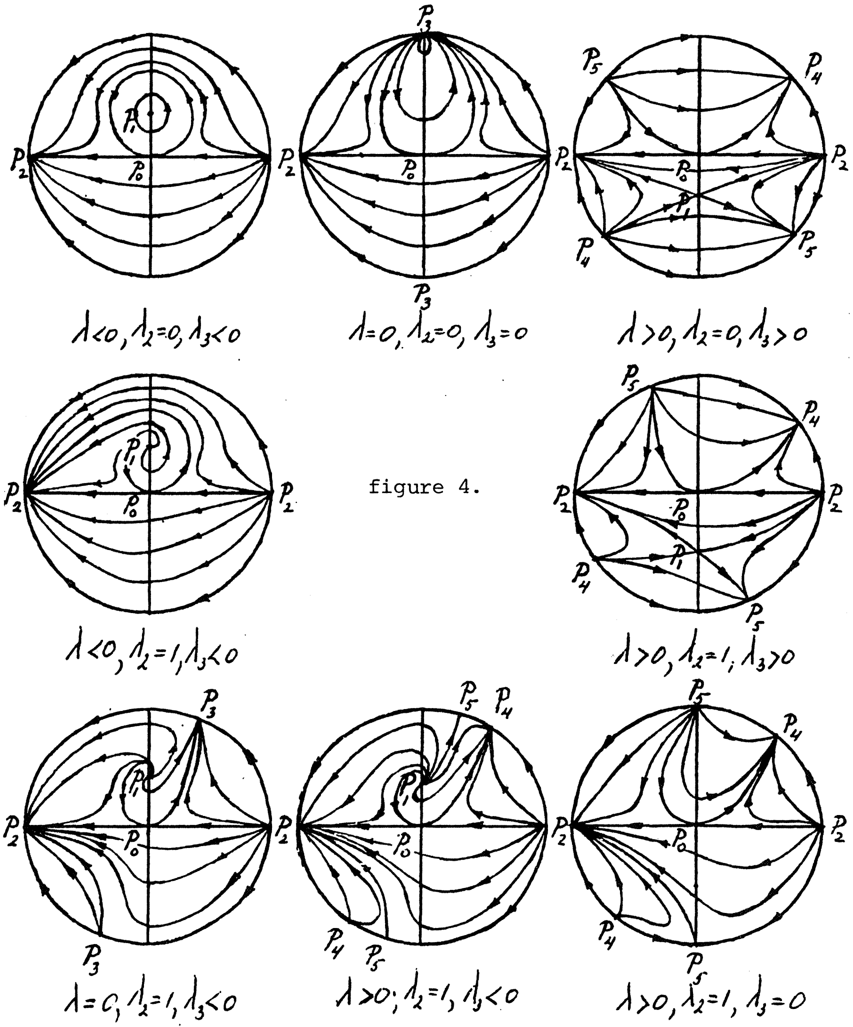
, which includes also the periodic solution of (6).

The other finite singular point P_1: $(0,-\lambda_3^{-1})$, (for $\lambda_3 \neq 0$) may be analyzed by local linearization and using the symmetry for $\lambda_2 = 0$.

From (4) follows, that the singular points at infinity are given by

$$C_2(\theta) \equiv [(1-\lambda_1)\cos^2\theta - \lambda_2 \sin\theta\cos\theta - \lambda_3 \sin^2\theta]\sin\theta = 0,$$

so that there is a point P_2 at $\theta = 0(\pi)$, and, for $\lambda \equiv \lambda_2^2 + 4\lambda_3(1-\lambda_1) > 0$, a point P_4 at $\theta = \text{arc cot} - \frac{1}{2}(\lambda_2 + \sqrt{\lambda})(\lambda_1 - 1)^{-1}$, and a point P_5 at $\theta = \text{arc cot} - \frac{1}{2}(\lambda_2 - \sqrt{\lambda})(\lambda_1 - 1)^{-1}$; these two points coincide for $\lambda = 0$ in the point P_3 at $\text{arc cot} - \frac{1}{2}\lambda_2(\lambda_1 - 1)^{-1}$. The character of these points may be found by local linearization and using Theorems 65, 66 and 67 of Chapter 9 of [1]. The character of the singular points is listed in Table 1 and the phase portraits are given in figure 4. Use should be made of the argument that $\theta' < 0$ on P_0P_4 and P_0P_5 (and thereby on P_0P_3).



$\lambda<0, \lambda_2=0, \lambda_3<0$

$\lambda=0, \lambda_2=0, \lambda_3=0$

$\lambda>0, \lambda_2=0, \lambda_3>0$

$\lambda<0, \lambda_2=1, \lambda_3<0$

figure 4.

$\lambda>0, \lambda_2=1; \lambda_3>0$

$\lambda=0, \lambda_2=1, \lambda_3<0$

$\lambda>0; \lambda_2=1, \lambda_3<0$

$\lambda>0, \lambda_2=1, \lambda_3=0$

λ	λ_2	λ_3	P_0	P_1	P_2	P_3	P_4	P_5
-	0	-	s	c	n			
0	0	0	s		n	eh		
+	0	+	s	s	n		n	n
+	1	+	s	s	n		n	n
+	1	0	s		n		n	sn
+	1	-	s	n,f	n		n	s
0	1	-	s	n,f	n	sn		
-	1	-	s	n,f	n			

Table 1

s = saddle point
n = node
sn = saddle node
f = focus
c = centre point
eh = point with an elliptic and hyperbolic sector.

3. Quadratic systems with a third order singular point with an elliptic and a hyperbolic sector.

If (0,0) is a third order singular point with an elliptic and hyperbolic sector; (2), with $d = 0$, $ae > 0$, may be written, if necessary by applying an affine transformation and / or a scaling of t, in the form

$$\begin{aligned} \dot{x} &= y + \lambda_1 x^2 + \lambda_2 xy + \lambda_3 y^2 \equiv P(x,y) \quad , \\ \dot{y} &= xy \qquad\qquad\qquad\qquad \equiv Q(x,y) \quad , \end{aligned} \tag{7}$$

, with $\lambda_1 > 0$, $\lambda_2 \in \{0,1\}$, $\lambda_3 \in \mathbb{R}$.

The hyperbolic sector is situated in $y > 0$ with the negative and positive x axis as separatrices. For $y < 0$ there is an elliptic sector, the extent of which near the x axis cannot be determined from local considerations. As a result, for $y < 0$, there is a parabolic sector near the x axis ($x < 0$ and $x > 0$).

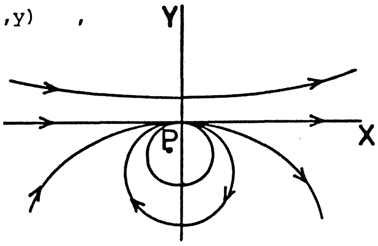


figure 5.

As for the previous case it can be shown that (7) has no limit cycles by using the same Dulac function $B(x,y) \equiv y^{-2\lambda_1 - 1}$. For $\lambda_2 = 0$, this Dulac function acts as an integrating factor to yield the solutions

for $\lambda_1 \neq \frac{1}{2}, 1$: $y \equiv 0$ and $\frac{2}{2\lambda_1 - 1} y + x^2 + \frac{\lambda_3}{\lambda_1 - 1} y^2 + \lambda_4 |y|^{2\lambda_1} = 0$, $\lambda_4 \in \mathbb{R}$,

for $\lambda_1 = \frac{1}{2}$: $y \equiv 0$ and $-2y \ln|y| + x^2 - 2\lambda_3 y^2 + \lambda_4 y = 0$, $\lambda_4 \in \mathbb{R}$,

for $\lambda_1 = 1$: $y \equiv 0$ and $2y + x^2 - 2\lambda_3 y^2 \ln|y| + \lambda_4 y^2 = 0$, $\lambda_4 \in \mathbb{R}$,

, which also include the periodic solutions of (7) for $\lambda_3 < 0$. [See also [10].]

The character of the other finite singular point P_1: $(0, -\lambda_3^{-1})$, (for $\lambda_3 \neq 0$) may be analyzed by local linearization and using the symmetry for $\lambda_2 = 0$.

In order to determine the phase portraits of (7) we consider the cases $\lambda_2 = 0$ and $\lambda_2 \neq 0$ separately.

Case $\lambda_2 = 0$. There follows

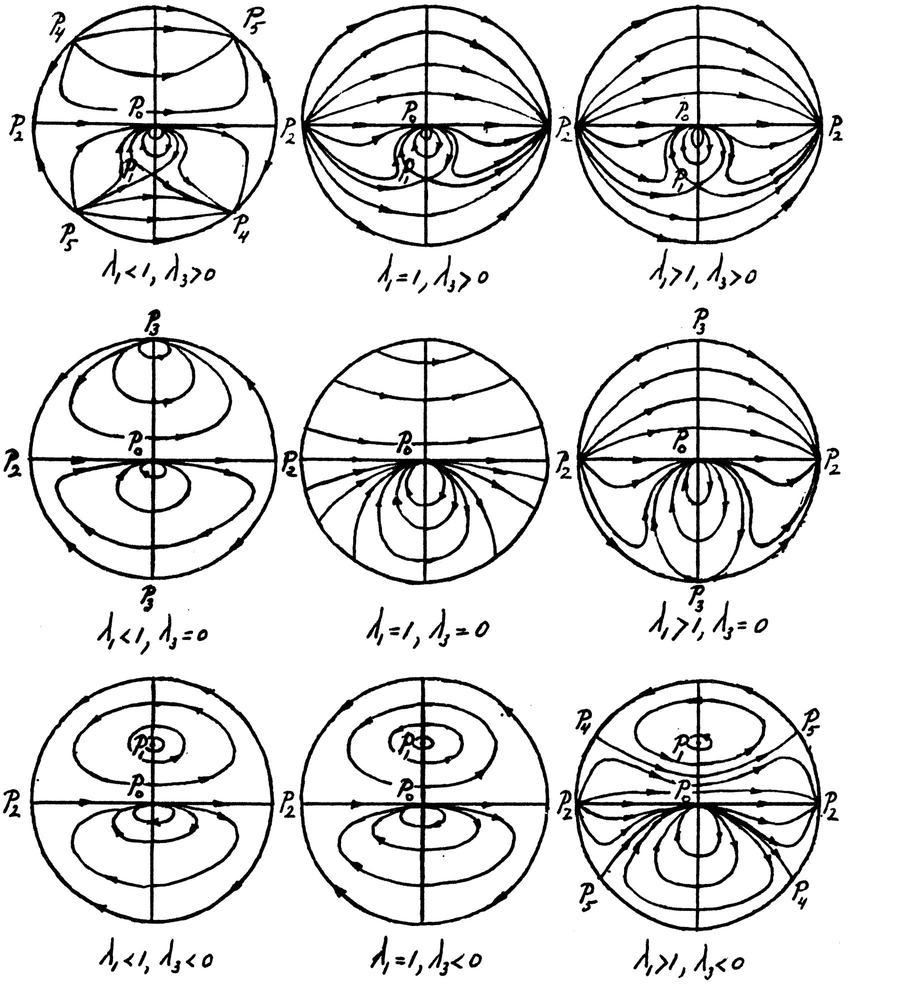
$$C_2(\theta) \equiv \left[(1-\lambda_1)\cos^2\theta - \lambda_3 \sin^2\theta\right]\sin\theta.$$

For $\lambda_1 = 1$, $\lambda_3 = 0$, all points on $\rho \equiv 1$ are singular, or, if the factor $(1-\rho)$ is divided out in (4), ordinary points. In the x,y plane the integral curves are conics through P_0. For $(\lambda_1,\lambda_3) \neq (1,0)$, $C_2(\theta) = 0$ shows that there is a singular point P_2 at $\theta = 0(\pi)$, and for $\lambda \equiv \lambda_3(1-\lambda_1)^{-1} > 0$, a point P_4 at $\theta = \text{arc cot} -\sqrt{\lambda}$, and a point P_5 at $\theta = \text{arc cot} \sqrt{\lambda}$; these points coincide for $\lambda = 0$ in point P_3 at $\theta = \frac{\pi}{2}(\frac{3\pi}{2})$. The character of these points may be found by local linearization and using Theorems 65, 66 and 67 of Chapter 9 of [1]. The character of the singular points is listed in Table 2 and the phase portraits are given in figure 6. It may be seen, that for $\lambda_3 < 0$, eq. (7) has periodic solutions and the phase portraits should also be found in the classification of quadratic systems with a centre as given by Vulpe [10]. It appears, however, that phase portrait 22 in [10] is incorrect and should be the same as phase portrait 23 (or as in fig. 6 for $\lambda_1 \leq 1$, $\lambda_3 = 0$).

Table 2

$\lambda_1 - 1$	λ_3	P_0	P_1	P_2	P_3	P_4	P_5
0	0	eh					
-	+	eh	s	s		n	n
0	+	eh	s	n*			
+	+	eh	s	n			
+	0	eh		n	s*		
+	-	eh	c	n		s	s
0	-	eh	c	s*			
-	-	eh	c	s			
-	0	eh		s	eh		

*third order point



$\lambda_1<1,\ \lambda_3>0$ $\lambda_1=1,\ \lambda_3>0$ $\lambda_1>1,\ \lambda_3>0$

$\lambda_1<1,\ \lambda_3=0$ $\lambda_1=1,\ \lambda_3=0$ $\lambda_1>1,\ \lambda_3=0$

$\lambda_1<1,\ \lambda_3<0$ $\lambda_1=1,\ \lambda_3<0$ $\lambda_1>1,\ \lambda_3<0$

figure 6 : $\lambda_2 = 0$

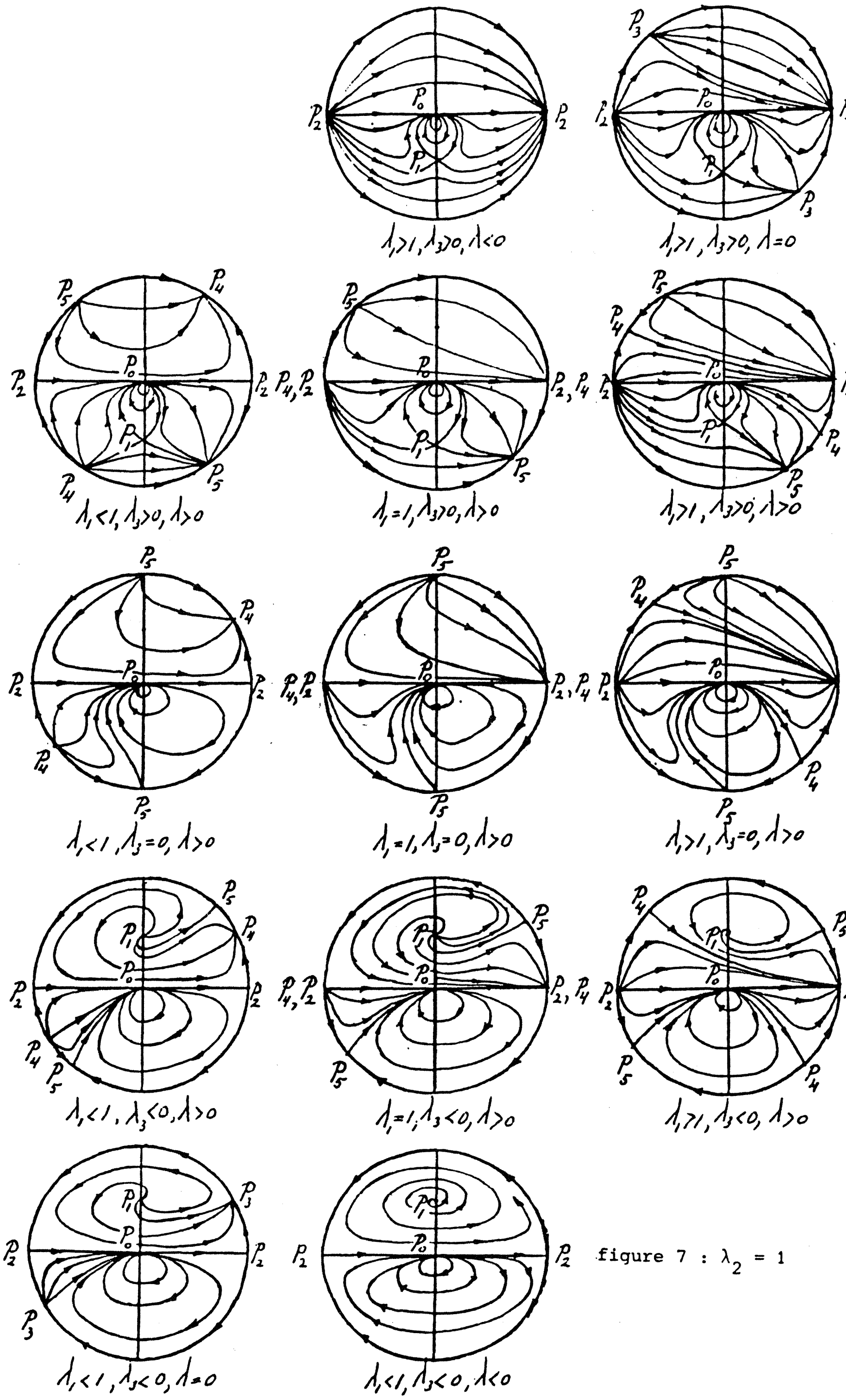
<u>Case $\lambda_2 = 1$.</u> From (4) follows, that the singular points at infinity are given by

$$C_2(\theta) = \left[(1-\lambda_1)\cos^2\theta - \sin\theta\cos\theta - \lambda_3\sin^2\theta\right]\sin\theta = 0,$$

so that there is a point P_2 at $\theta = 0(\pi)$, and, for $\lambda \equiv 1 + 4\lambda_3(1-\lambda_1) > 0$, a point P_4 at $\theta = \text{arc cot} - \frac{1}{2}(1+\sqrt{\lambda})(\lambda_1-1)^{-1}$ and a point P_5 at $\theta = \text{arc cot} - \frac{1}{2}(1-\sqrt{\lambda})(\lambda_1-1)^{-1}$; these two points coincide for $\lambda = 0$ at point P_3 at

$$\theta = \text{arc cot} - \frac{1}{2}(\lambda_1-1)^{-1}.$$

Moreover, for $\lambda_1 = 1$, the point P_4 coincides with P_2, and point P_5 is situated in $\theta = \text{arc cot} -\lambda_3$. The character of these points may be found by local



$\lambda_1 > 1, \lambda_3 > 0, \lambda < 0$

$\lambda_1 > 1, \lambda_3 > 0, \lambda = 0$

$\lambda_1 < 1, \lambda_3 > 0, \lambda > 0$

$\lambda_1 = 1, \lambda_3 > 0, \lambda > 0$

$\lambda_1 > 1, \lambda_3 > 0, \lambda > 0$

$\lambda_1 < 1, \lambda_3 = 0, \lambda > 0$

$\lambda_1 = 1, \lambda_3 = 0, \lambda > 0$

$\lambda_1 > 1, \lambda_3 = 0, \lambda > 0$

$\lambda_1 < 1, \lambda_3 < 0, \lambda > 0$

$\lambda_1 = 1, \lambda_3 < 0, \lambda > 0$

$\lambda_1 > 1, \lambda_3 < 0, \lambda > 0$

$\lambda_1 < 1, \lambda_3 < 0, \lambda = 0$

$\lambda_1 < 1, \lambda_3 < 0, \lambda < 0$

figure 7 : $\lambda_2 = 1$

linearization and using Theorem 65 of Chapter 9 of [1]. The character of the singular points is listed in Table 3 and the phase portraits are given in figure 7.

Table 3

λ_1-1	λ_3	λ	P_0	P_1	P_2	P_3	P_4	P_5
-	+	+	eh	s	s		n	n
-	0	+	eh		s		n	sn
-	-	+	eh	n,f	s		n	s
-	-	0	eh	n,f	s	sn		
-	-	-	eh	n,f	s			
0	+	+	eh	s	sn		sn	n
0	0	+	eh		sn		sn	sn
0	-	+	eh	n,f	sn		sn	s
+	+	-	eh	s	n			
+	+	0	eh	s	n	sn		
+	+	+	eh	s	n		s	n
+	0	+	eh		n		s	sn
+	-	+	eh	n,f	n		s	s

Acknowledgement.

The research presented in this paper was made possible by the Delftse Hogeschool Fonds through a stipendium awarded to the first author.

References.

1. Andronov, A.A., E.A. Leontovich, J.J. Gordon and A.G. Maier, 1973, Qualitative theory of second-order dynamic systems, Israel Program for Scientific Translation, Jeruzalem, Wiley, New York.

2. Berlinskii, A.N., On the behaviour of integral curves of a differential equation, Izv. Vyss. Ucebn. Zaved. Matematika, 15, no. 2, 3-18, 1960 (Russian), translated by the National Lending Library for Science and Technology, Boston Spa, Yorkshire, England, Russian Translating Programme RIS 5158, June 1969.

3. Chicone, C. and D.S. Shafer, Separatrix and limit cycles of quadratic systems and Dulac's theorem, Trans. of the American Mathematical Society, Vol. 278, no. 2, August 1983.

4. Coppel, W.A., A survey of quadratic systems, Journ. of Diff. Eq., 2, 293-304, 1966.

5. Date, Tsumotu, Classification and analysis of two-dimensional real homogeneous quadratic differential equation systems, Journ. of Diff. Eq., 32, 311-334, 1979.

6. Gasull, Armengol, Sheng Li-Ren and Jaume Llibre, Chordal quadratic systems, Universitat Autonoma de Barcelona, Bellaterra, Barcelona.

7. Newton, T.A., Two dimensional homogeneous quadratic differential systems, SIAM Review, 20, 120-138, 1978.

8. Private communication: W.T. v. Horssen, P. de Jager.

9. Reyn, J.W., Phase portraits of a quadratic system of differential equations occurring frequently in applications, Nieuw Archief voor Wiskunde, Series 4, Deel 5, no. 1,March 1987.

10. Vulpe, N.I., Affine-invariant conditions for the topological discrimination of quadratic systems with a center, Diff. Uravn., 19, no. 3, pp. 371-379, March 1983.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Analytical Approximations For Offshore Pipelaying Problems

S.W. Rienstra
Mathematics Consulting Department
University of Nijmegen
Toernooiveld, 6525 ED Nijmegen, The Netherlands

1. Abstract

The geometrically non-linear slender bar equation is solved for a number of problems involving suspended pipelines, related to the off-shore gas- and oil-pipe laying. The problems concern the use of a lay-barge with stinger, and the process of abandoning and recovery of a pipe. The usually stiff equation requires for a completely numerical solution considerable computer power, not always available on board. Therefore, the solutions are analytical (matched asymptotic expansions, and linear theory) to allow the results being evaluated on a small computer. It is shown that for the majority of the practical cases the two solution methods complement each other very well.

2. Introduction

Exploitation of gas- and oil-wells offshore requires the presence of pipelines along the sea floor for transport of the products. The laying of these pipelines is usually done by suspending the pipeline via a stinger from a lay-barge. On board the pipe is composed by welding pipe elements together at the welding ramp (figure 1). During the process of laying the pipe is bent by its own weight into a stretched S-curve, causing bending stresses in the pipe. If the water is (relatively) shallow, the pipe sufficiently stiff, and the weight (per unit of length) is sufficiently low, these stresses remain low enough without further precautions. However, in modern applications the pipes are laid in deep water, sometimes in a considerable current necessitating a heavy pipe (especially when it is a gas pipe), in a way that the bending stresses become so high that the pipe would buckle. In that case, a horizontal tension is applied by the ship, to stretch the bends and reduce the stresses. Furthermore, sometimes the pipe has to be abandoned and recovered by means of a cable (let down to and pulled up from the sea floor, figure 2, for example when a storm prohibits the continuation of laying). Also during this process a certain tension is to be applied at the cable to avoid buckling.

Since both the tension machines, stinger equipment etc., and the repairing of a buckled pipe are very expensive, it is necessary to calculate in advance, for a given configuration, the tension just sufficient to obtain a given maximum stress level. This problem will be discussed here.

With sea current, dynamics of the sea, nonlinearity of the steel elasticity, and variation of pipe weight and flexural rigidity being usually of minor importance, we consider the model of a linearly elastic, geometrically nonlinear suspended bar, loaded by its own weight and a horizontal tension. The equations and boundary conditions will be presented in the next section; the derivation may be found in [1]. The differential equation is of second order, with an unknown free pipe length and an unknown bottom reaction. Effectively, the problem is therefore of fourth order. Due to the nonlinear character of the problem only very few exact solutions are known. For example, if the flexural rigidity vanishes we obtain the catenary (some boundary conditions have to be given up), and if the pipe weight vanishes the equation becomes equivalent to the nonlinear pendulum equation allowing an implicit solution with elliptic functions (Kirchoff's analogy, [1]). For the present problem no exact solutions are known, and we will therefore consider approximations.

A very well-known approximation is based on the pipe being nearly horizontal (small deflection angle) allowing linearized equations with solutions of exponential or (if the horizontal tension vanishes) polynomial type (beam theory; [1,2]). A generalization of this approach is a linearization around a non-zero mean deflection angle, yielding solutions in terms of Airy functions [1]. These linearizations are uniform approximations: physically, the behaviour of the pipe is everywhere the same with basically the same equation valid. Another approach is based on a small (relative) flexural rigidity, giving the pipe a shape close to a catenary (important weight and tension, unimportant bending stiffness), except for the regions near the ends where bending stiffness is important, but weight is unimportant. Evidently, this approximation is not uniform, with physically different behaviour in boundary layers at the ends. It has become known in the literature as "stiffened catenary". The presence of these boundary layers was first noted by Plunkett [3] (for a related problem without free boundaries); however, his asymptotic solution is only correct to leading order. Matched asymptotic expansion solutions based on this boundary layer behaviour were also given by Dixon and Rutledge [4] (using Plunketts solution), van der Heyden [5] (for suspended cables), Konuk [6], and (in a more general setting) by Flaherty and O'Malley [7]. None of these authors, however, do consider the present *free* boundary problem.

In the following sections we will derive two approximate solutions (or in any case, reduce the problem to an algebraic equation): a small-angle linearization (beam), and a matched asymptotic expansion based on a small flexural rigidity (stiffened catenary). Special attention will be paid to the use of the first integral of the equation (free bending energy) to deal with the inherent problem of the unknown suspended pipe length. Furthermore, we will show that these two approximations appear to provide excellent (complementary) solutions for almost all the investigated practical cases, so that they are probably sufficient in a practical situation for on-board calculations with only a small computer available. At the same time, of course, they provide efficient starting values for completely numerical solutions which might otherwise suffer from the stiffness of the equation.

3. The Problems

Equilibrium of forces, together with application of the Bernoulli-Euler law, relating bending moment to radius of curvature, yields the following equation for $\psi(s)$, the angle between horizon and the tangent at the local coordinate s , in non-dimensional form

$$(\varepsilon/\mu)^2\psi_{ss} = \sin(\psi) - (\mu s - \lambda)\cos(\psi) \tag{1}$$

along the interval [0,1], and where $\varepsilon^2 = EIQ^2/H^3$, $\mu = LQ/H$, $\lambda = V/H$, with EI denoting the flexural rigidity, Q the pipe weight per unit length, H the horizontal tension, L the (unknown) free pipe length, and V the (unknown) bottom reaction force. The corresponding boundary conditions for the pipelay problem are given by

$$\psi(0) = 0,\ \psi_s(0) = 0,\ \psi_s(1) = -\mu/r, \tag{2.a}$$

$$d = d_{sh} + r\cos(\psi(1)) - r\cos(\phi), \tag{2.b}$$

and for the abandon/recovery problem

$$\psi(0) = 0,\ \psi_s(0) = 0,\ \psi_s(1) = 0, \tag{3.a}$$

$$d = d_{sh} + r\cos(\gamma) - r\cos(\phi) - (c - r\gamma + r\phi)\sin(\gamma), \quad \gamma = \arctan(\mu - \lambda). \tag{3.b}$$

Here is $r = RQ/H$, $d = DQ/H$, $d_{sh} = D_{sh}Q/H$, $c = CQ/H$, with R denoting the stinger radius, $D = L\int_0^1 \sin(\psi(s))ds$ the height of the pipe end, D_{sh} the height of the stinger hinge, ϕ the angle at the hinge, C the cable length (measured from the stinger hinge), and γ the cable angle. The cable is for simplicity taken with zero weight, but a nonzero weight can be included without much difficulty.

An important relation is the first integral of (1), expressing the elastic free bending energy density [2], and providing an explicit relation between d and $\psi(1)$:

$$\tfrac{1}{2}(\varepsilon/r)^2 = 1 - \cos(\psi(1)) - (\mu - \lambda)\sin(\psi(1)) + d. \tag{4}$$

In the following section we will present a stiffened catenary solution for small ε, and a beam solution for small $|\psi|$, of (1) with (2) and (1) with (3) , and using (where appropriate) (4). We note in passing that the present problems have no unique solution without an additional condition to minimize energy or pipe length; further research is in progress [8].

4. Solution

4.1 *Stiffened catenary* ($\varepsilon \to 0$)

The solution is built up from local asymptotic expansions in three regions: $\psi = h$ in $s = O(1)$, $\psi = f$ in $s = O(\varepsilon)$, and $\psi = g$ in $s = 1 + O(\varepsilon)$. Unknown constants are determined via matching. The complete solution is constructed by adding the three solutions and subtracting common terms:

$$\psi = \tilde{f} + h + \tilde{g}.$$

This whole matched asymptotic expansion procedure is relatively standard, and will not be repeated here. The only point to be noted, is that L and V, and therefore μ and λ, are unknown, so dependent on ε, and should therefore be expanded into powers of ε, like f,g and h. This will, however, not be carried through right from the start. It is more convenient to begin with assuming μ and λ fixed, or rather, known to any desired accuracy, and to postpone the actual calculation to a later stage.

If $s = O(1)$, we introduce $z = \mu s - \lambda$, and rewrite (1) into

$$h = \arctan(z) + \arcsin(\varepsilon^2 h_{zz}/(1+z^2)^{1/2}).$$

By successive substitution, or otherwise, we obtain easily

$$h = \arctan(z) - 2\varepsilon^2 z/(1+z^2)^{5/2} + O(\varepsilon^4). \tag{5}$$

Note that the leading order term is just the catenary.

If $s = O(\varepsilon)$, we introduce $t = \mu s/\varepsilon$ to obtain

$$f_{tt} = \sin(f) - (\varepsilon t - \lambda)\cos(f).$$

Expanding f and λ (which is determined at this stage) in an ε-power series, yields, after matching and application of the boundary conditions,

$$\lambda = \varepsilon/(1+\tfrac{3}{4}\varepsilon^2) + O(\varepsilon^5), \tag{6}$$

$$\tilde{f} = \varepsilon e^{-t} - \tfrac{1}{48}\varepsilon^3 e^{-t}[e^{-2t} + 4t^3 - 6t^2 + 6t + 147] + O(\varepsilon^5). \tag{7}$$

(We already skipped the terms common to f and h).
If $s = 1+O(\varepsilon)$, we introduce $\tau = \mu(s-1)/\varepsilon\varkappa$, where $\varkappa = (1+(\mu-\lambda)^2)^{-1/4}$, to obtain

$$g_{\tau\tau} = \varkappa^2 \sin(g) - \varkappa^2(\mu-\lambda+\varepsilon\varkappa\tau)\cos(g) .$$

Following the usual steps we arrive at

$$\tilde{g} = -\varepsilon\varkappa e^{\tau}(\varkappa^4+1/r)[1 + \tfrac{1}{4}\varepsilon(\mu-\lambda)\varkappa^5(\tau^2-\tau+1)] + O(\varepsilon^3) \tag{8}$$

for the pipelay problem; in case of the abandon/recovery problem the term $1/r$ is set to zero. Up to now we have applied the boundary conditions (2.a) and (3.a). The final step to be taken, to determine μ, is substitution of the results obtained so far into (2.b) and (3.b), and utilizing (4) to get rid of d. Rewritten in suitable form it becomes for the pipelay problem

$$\varkappa^2 = [A + (A^2 + 4r(\cos\alpha-(\mu-\lambda)\sin\alpha)\cos\alpha)^{1/2}] \,/\, 2r(\cos\alpha-(\mu-\lambda)\sin\alpha) \tag{9}$$

with $A = r\cos(\phi) - d_{sh} + \frac{1}{2}(\varepsilon/r)^2 - 1$, and $\alpha = \psi(1) - \arctan(\mu-\lambda)$. Since $\alpha = O(\varepsilon)$, the right-hand side of (9) is to leading order independent of ε, and the solution for μ is simply obtained by successive substitution of $\varkappa^2$, starting with $\alpha = 0$. The equation for the abandon/recovery problem, corresponding to (9), is

$$r\varkappa^2 - (c+r\phi-r\gamma)(\mu-\lambda)\varkappa^2 - \cos(\alpha)/\varkappa^2 = A\,, \tag{10}$$

but this equation cannot be written in a form allowing an explicit asymptotic solution, and therefore has to be solved numerically.

This solves the present stiffened catenary problems. One final remark to be made is that it is practically very useful to modify the boundary layer contributions $\bar{f}$ and $\bar{g}$ a little bit, by adding exponentially small terms, of the order of $\exp(-\mu/\varepsilon)$ and $\exp(-\mu/\varepsilon\varkappa)$, in a way that the coupling between $\bar{f}$ and $\bar{g}$ in each others domain is reduced, for example:
$\psi = \bar{f}(s) - \bar{f}(1)s + h(s) + \bar{g}(s) - \bar{g}(0)(1-s)$, and similarly for ψ_s. Asymptotically for $\varepsilon \to 0$, these terms have no meaning, of course, since they are smaller than any power of ε, but for any finite ε they appear to be very useful, and extend the region of validity to values of ε as high as 0.35 .

4.2 *Beam* ($|\psi| \ll 1$).

Linearization of equation (1) yields

$$(\varepsilon/\mu)^2\psi_{ss} = \psi - \mu s + \lambda, \tag{11}$$

with solution (satisfying $\psi(0)=\psi_s(0)=0$)

$$\psi = \lambda(\cosh(t)-1) - \varepsilon(\sinh(t)-t) \tag{12}$$

where $t = \mu s/\varepsilon$. From (12) we derive an expression for $d = \mu\int_0^1 \psi ds$ by direct integration (eq. (4) is not a first integral of (11) any more). Then, for given d, the solution $x=x_0$ of

$$\tfrac{1}{2}x\sinh(x) - \cosh(x) + 1 - (d/\varepsilon^2)\sinh(x)/x - (\sinh(x)/x - 1)/r = 0 \tag{13}$$

gives $\mu = \varepsilon x_0$, $\lambda = \mu/2 - d/\mu - \varepsilon^2/\mu r$, of course with $1/r = 0$ in case of the abandon/recovery problem. Finally, d is determined by solving equation (2.b) or (3.b). So for the beam problem we arrive at two coupled algebraic equations. We note, that we do *not* linearize (2.b) (which would imply $\cos(\psi(1)) = 1$), since in that case we would lose all information on the lift-off angle $\psi(1)$, which is of great practical importance as it determines the required length of the stinger.

5. Examples

We start with an example of the pipelay problem, figure 3. We plotted the curve of required tension versus water depth (i.e. D_{sh}) to obtain a prescribed minimum radius of curvature. The plot is given in dimensional form, since we scaled previously on H, which is now the varying quantity. We see the results from the beam theory, valid up to, say, 20°−25°, smoothly taken over by the stiffened catenary theory, suggesting correct results from both theories also in the transition region. This appeared to be typical for practically all the cases considered, and is, furthermore, confirmed by comparison with completely numerical results. For example, the results in the transition region near $D_{sh}=50$ appear to be indeed very accurate, in spite of the rather high value of $\varepsilon=0.32$, the boundary layer widths of 0.23 and 0.27, and the rather large maximum angle of about 24°, as is seen from the following comparison with a completely numerical solution at $D_{sh}=50$ and $H=110$:

	numerical	catenary	beam
$\psi(1)$	21.89°	21.82°	22.18°
V	30.05	30.63	30.30
L	145.80	146.9	141.8
min.radius	208.2	209.2	202.0

This is much more accurate than could be estimated theoretically: an error of $O(\varepsilon^3)$ for ψ, V and L in the catenary theory would predict 3% (here 0.3%), an error of $O(\varepsilon^2)$ for the minimum radius would give 10% (here 0.5%), and an error of $O(|\psi|^2)$ in the beam theory, giving 15%, is really less than 3%. This remarkably better performance than the a-priori estimates remains also for other examples, and does not seem to be accidental. Probably, the higher order corrections are numerically small or cancel each other.

In figure 4 we have an example of the abandon/recovery problem, where maximum bending stress ($\sim\psi_s$) is plottes versus cable length. A similar transition from catenary to beam is seen, with again overall accurate results.

6. Acknowledgments

We would like to thank C.J. Negenman for posing the problem (many years ago), and his advice and stimulation. The completely numerical results used for comparison were obtained by the use of a subroutine for nonlinear boundary value problems, developed by R.M. Mattheij and G. Staarink.

7. References

[1] FRISCH-FAY, R., *Flexible Bars.* London, Butterworths (1962).

[2] LANDAU, L.D. and LIFSHITZ, E.M., *Theory of Elasticity,* Pergamon, Oxford, 1970.

[3] PLUNKETT, R., *Static Bending Stresses in Catenaries and Drill Strings.* Journal of Engineering for Industry, Transactions of the ASME, Vol.89, no.1, p.31-36, 1967.

[4] DIXON, D.A. and RUTLEDGE, D.R., *Stiffened Catenary Calculations in Pipeline Laying Problem.* Journal of Engineering for Industry, Transactions of the ASME, Vol.90, no.1, p.153-160, 1968.

[5] VAN DER HEYDEN, A.M.A., *On the Influence of the Bending Stiffness in Cable Analysis.* Proceedings of the KNAW, B76, p.217-229, 1973.

[6] KONUK, I., *Higher Order Approximations in Stress Analysis of Submarine Pipelines.* ASME 80-Pet-72, presented at ETC&E, New Orleans, La., February 3-7, 1980.

[7] FLAHERTY, J.E. and O'MALLEY, R.E., *Singularly Perturbed Boundary Value Problems for Nonlinear Systems, Including a Challenging Problem for a Nonlinear Beam.* Lecture Notes in Mathematics 942, p.170-191, Springer Verlag, Berlin, 1982.

[8] MATTHEIJ, R.M.M. and RIENSTRA, S.W., *On an offshore pipelaying problem.* To appear.

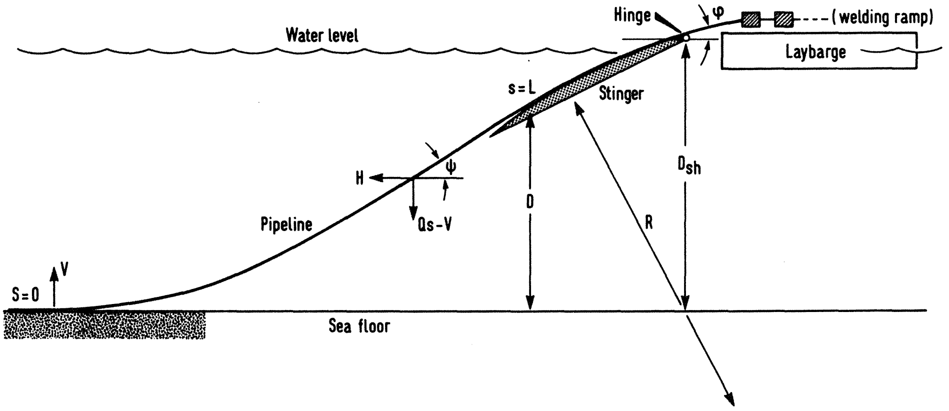


Figure 1. Sketch of the pipelay problem

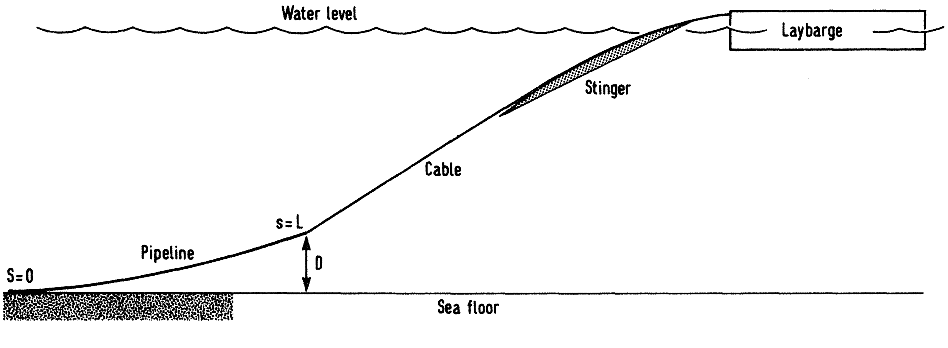


Figure 2. Sketch of the abandon / recovery problem

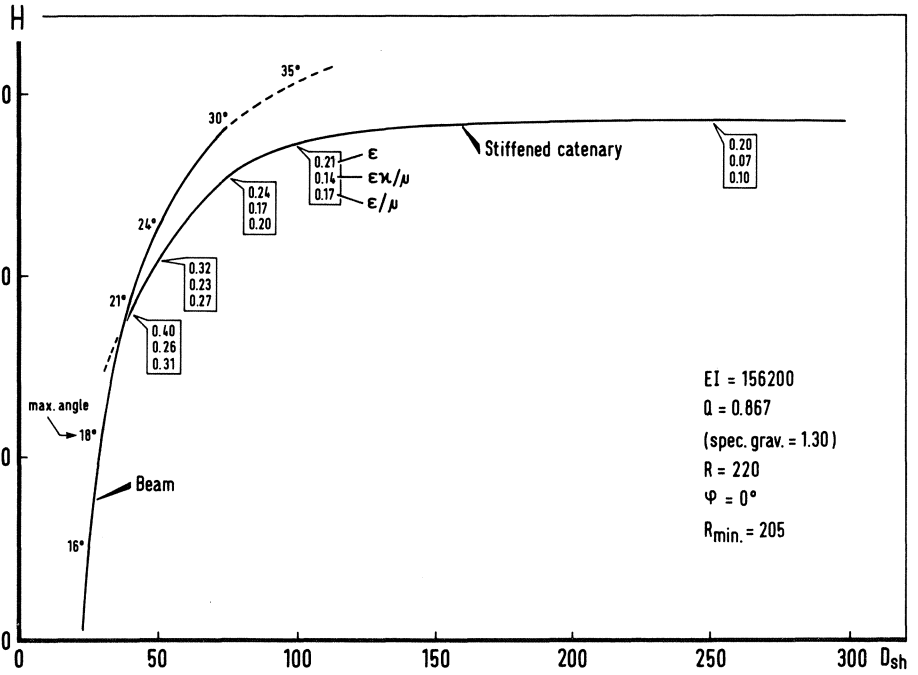


Figure 3. Tension / waterdepth diagram for constant minimum radius of curvature ($R_{\min}$)

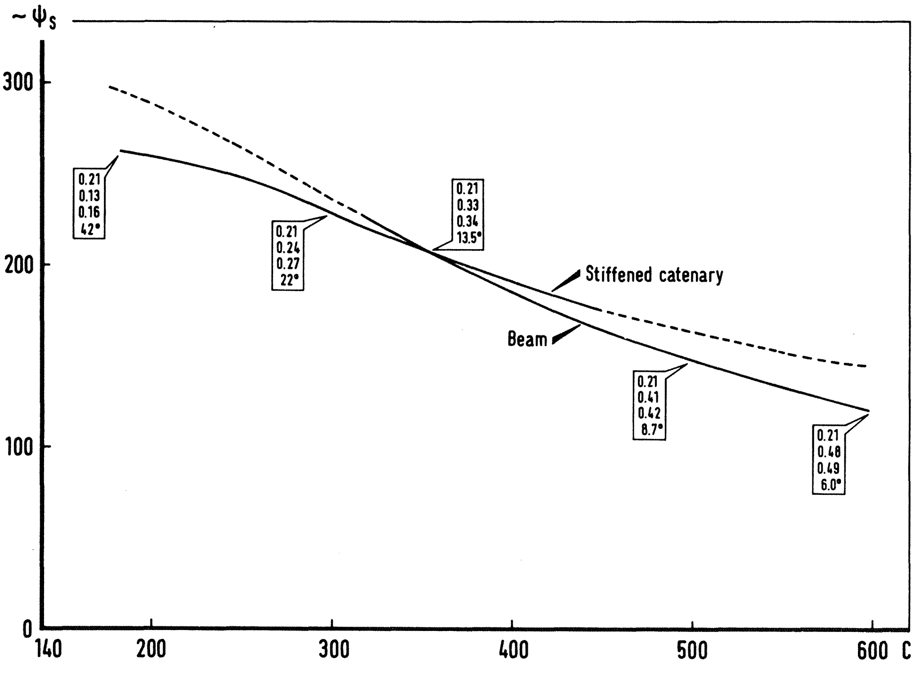


Figure 4. Maximum bending stress / cable length for pipe of fig. 3, with $D_{sh}=150$ *and* $H=140$. *Boxed numbers:* ε, εκ/μ, ε/μ, *maximum angle.*

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Some Results for a Semilinear Elliptic Problem with a Large Parameter

G. Sweers
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands

ABSTRACT

Consider the following eigenvalue problem

$$(P)\begin{cases} -\Delta u = \lambda f(u) & \text{in } \Omega \subset \mathbb{R}^n\text{, bounded,} \\ u = 0 & \text{on } \partial\Omega\text{, smooth,} \end{cases}$$

where f changes sign. In this note we will show results which can be found by using the so-called sweeping principle of Serrin, 1971. Especially we will give estimates for the boundary layer of positive solutions near a zero of f. For some f a solution u will have a free boundary. We show for such f that $f(u)=0$ except near $\partial\Omega$. Next to this we improve a result for existence of a solution.

1. INTRODUCTION

We are interested in pairs $(\lambda,u) \in \mathbb{R}_+ \times C^2(\Omega)$ satisfying (P) and $u > 0$ in Ω. First, note that a solution satisfies $f(\max u) \geq 0$. If $f \in C^1$, the strong maximum principle even shows $f(\max u) > 0$. Secondly, if ρ is a zero of f then $u \equiv \rho$ satisfies the differential equation for all λ. So one could expect the existence of a solution (λ,u), where λ is large and u is near a zero of f (with $f(\max u) \geq 0$) except for a boundary layer. Results for this problem were presented by Fife,1973 and by Clément et al., 1986. The results here are strongly related to this last paper.
Assume that there are two numbers $0 < \rho_1 < \rho_2$ such that

(F1) $\quad f(\rho_1) = f(\rho_2) = 0$ and $f > 0$ in (ρ_1,ρ_2), (F2) $f \in C^{\gamma}(-\infty,\rho_2] \cap C^1(-\infty,\rho_2)$ and there is $\delta > 0$ such that $f' \leq 0$ in $(\rho_2-\delta,\rho_2)$, where $\gamma \in (0,1)$.

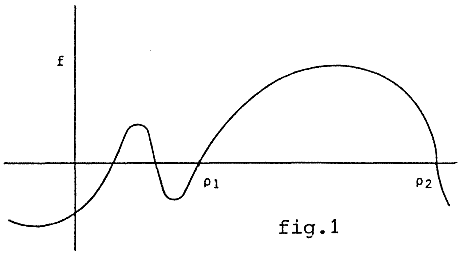


fig.1

In 1981 Hess showed, if $f(0) > 0$, that the following condition is sufficient for existence of a positive solution (λ,u) with max $u \in (\rho_1,\rho_2)$.

(F3) $$J(\rho) := \int_{\rho}^{\rho_2} f(s)ds > 0 \text{ for every } \rho \in [0,\rho_1].$$

In the first theorem, it will be proven that this condition is sufficient and necessary when $f \in C^1[0,\max u]$, even if $f(0) < 0$. In the second theorem we will show that the solutions, which are found in this way, are near ρ_2.

2. THEOREMS AND PROOFS

Before stating the first theorem we will shortly explain the sweeping principle of Serrin, 1971. A formulation can also be found in the paper by Clément et al., 1986.

Fix λ, let u be a solution of (P) and let $\{v(t) \in C(\bar{\Omega});\ t \in [0,1]\}$ be a continuous family of subsolutions, such that $v(0) < u$ in Ω and for all t $v(t) < u$ on $\partial\Omega$ as well as $v(t) < \rho_2$ in Ω. Then $v(t) < u$ in Ω for all $t \in [0,1]$. Since, if there exists $t^* \in [0,1]$ such that $v(t^*) \leq u$ and for some $x^* \in \Omega$ $v(t^*,x^*) = u(x^*)$, the strong maximum principle implies $v(t^*) \equiv u$, a contradiction.

THEOREM 1 :

Let f satisfy (F1)(F2)(F3) and let Ω satisfy a uniform interior sphere condition. Then there exists $c_1 > 0$, $c_2 \in (\rho_1,\rho_2)$ and $\lambda_0 > 0$ such that for all $\lambda > \lambda_0$ a positive solution $(\lambda,u(\lambda))$ of (P) exists with

(1) $$\min (c_1.d(x,\partial\Omega).\lambda^{\frac{1}{2}},\ c_2) < u(\lambda) \leq \rho_2.$$

Moreover every solution (λ,u) of (P) (not necessarily positive) with $\max u \in (\rho_1,\rho_2)$ satisfies $\int_{\rho}^{\max u} f(s)ds > 0$ for every $\rho \in [0,\rho_1]$.

PROOF:

Replace f by f^*, where f^* satisfies (F1) and

$f^*(u) = 1 \quad$ for $u < -1$,

$f^*(u) \leq f(u)$ for $0 \leq u \leq \rho_2$,

$f^* \in C^1(\mathbb{R})$,

$\int_{\rho}^{\rho_2} f^*(s)ds > 0$ for all $\rho < \rho_2$.

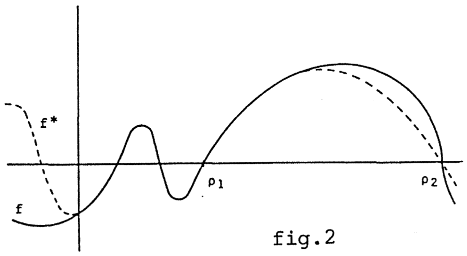


fig.2

Like Hess in 1981, one finds for μ large enough, a minimizer v of $I(v,\mu) = \frac{1}{2}\int_{\Omega}|\nabla v|^2dx - \mu\int_{\Omega}\int_0^v f^*(s)ds\,dx$ in the cone $\{v \in W^{1,2}(\Omega);\ v > -1$ in B, $v = -1$ on $\partial B\}$ with $\max v \in (\rho_1,\rho_2)$. (B denotes the unit ball). Gidas et al. showed in 1979 that v is radially symmetric and $v'(r) < 0$ for $r \in (0,1]$. Let $\theta \in (0,1)$ be the number such that $v(\theta) = 0$. Since Ω satisfies a uniform interior sphere condition, $\Omega = \cup\{B(x,\varepsilon);\ x \in \Omega(\varepsilon)\}$ for all $\varepsilon \in (0,\varepsilon_0)$, where ε_0 is some positive constant and $B(x,\varepsilon) = \{y \in \mathbb{R}^N;\ |x-y| < \varepsilon\}$, $\Omega(\varepsilon) = \{x \in \Omega;\ d(x,\partial\Omega) > \varepsilon\}$. Then $w(\lambda,x) := \sup\{v(\theta.\varepsilon^{-1}.|x-y|);\ y \in \Omega(\varepsilon)\}$, with $\lambda = \mu.(\theta/\varepsilon)^2$, is a subsolution of (P), with f replaced by f^*, for all $\lambda \geq \lambda_0 := \mu.(\theta/\varepsilon_0)^2$. Since $0 < w(\lambda) < \rho_2$ and $f^* \leq f$ on $[0,\rho_2]$, $w(\lambda)$ is also a subsolution of the original (P). Note that $W(\lambda) \equiv \rho_2$ is a supersolution of (P) for all λ. By an iteration scheme one shows the existence of a solution in between. By condition (F2) there exist two strictly increasing continuous functions f_1 and f_2 such that $f = f_1 - f_2$ on $[0,\rho_2]$ and $f_2(0) = 0$. Because of (F2) one may assume $f_1 \in C^1[0,\rho_2]$. Define T by $u = T(v)$, where u is the unique solution of

$$\begin{cases} -\Delta u + \lambda f_2(u) = \lambda f_1(v) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

See the paper of Brezis et al. from 1973. Define $W_n = T^n(W(\lambda))$ and $w_n = T^n(w(\lambda))$. $\{W_n\}$ and $\{w_n\}$ are sequences of respectively decreasing supersolutions and increasing subsolutions. Since $W_n > w_n$ in Ω the sequences converge to a solution of (P). Standard regularity theory shows that these solutions, or maybe just one solution , are $C^2(\Omega)$. The estimate (1) is valid since the solutions are between $w(\lambda)$ and $W(\lambda)$.

The last part will also be proven with a sweeping argument. Suppose there is a solution of (P) with $\max u \in (\rho_1,\rho_2)$ and $\int_{\rho^*}^{\max u} f(s)ds = 0$ for some $\rho^* \in [0,\rho_1]$.

Let $\bar{u}$ be the solution of

$$-\bar{u}'' = \lambda f(\bar{u}) \quad , t \in \mathbb{R} ,$$
$$\bar{u}(0) = \max u \quad ,$$
$$\bar{u}'(0) = 0.$$

Set $U(t,x_1,\dots,x_N) = \bar{u}(x_1-t)$ for $x \in \mathbb{R}^N$.

Note that $\max U = \max u$ and $\inf U \geq \rho^*$. Moreover there exists t^* and $x^* \in \overline{\Omega^*}$, with $\Omega^* = \Omega \cap \{x \in \mathbb{R}^N ; x_1 > t^*\}$, such that

$$U(t^*) \geq u \text{ in } \Omega^*,$$
$$U(t^*,x^*) = u(x^*) \text{ and } \nabla U(t^*,x^*) = \nabla u(x^*).$$

The strong maximum principle shows $U(t^*) \equiv u$, which is a contradiction. For a more detailed proof see the authors paper of 1986. □

THEOREM 2:

Let Ω satisfy an interior sphere condition and let f satisfy (F1) and (F2) with ρ_1 not necessarily positive. If $\rho_1 > 0$ then assume (F3) is also satisfied.

Suppose that $f(u) > c(\rho_2-u)^\alpha$ for $u \in (\rho_2-\delta,\rho_2)$, where $c,\alpha,\delta > 0$. Then there is $C > 0$ such that for any nonnegative $z \in C_0^\infty(\Omega)$, with $\max z \in (\rho_1,\rho_2)$, $\lambda(z) > \lambda_0$ exists for which the following holds.

Let (λ,u) be a solution of (P) with $z \leq u \leq \rho_2$ in Ω and $\lambda > \lambda(z)$.

1) If $0 < \alpha < 1$ then $u(x) \geq \min (C.\lambda^{\frac{1}{2}}.d(x,\partial\Omega),\rho_2)$.

2) If $\alpha = 1$, then $u(x) > \rho_2(1-\exp(-C.\lambda^{\frac{1}{2}}.d(x,\partial\Omega)))$, for $x \in \Omega$.

3) If $1 < \alpha$, then $u(x) > \rho_2(1-(1+C.\lambda^{\frac{1}{2}}.d(x,\partial\Omega))^{-p})$ for $x \in \Omega$, with $p = 2(\alpha-1)^{-1}$.

REMARK 1.

Case 1) shows that a solution near ρ_2 will have a free boundary within a distance of order $\lambda^{-\frac{1}{2}}$ from $\partial\Omega$.

REMARK 2.

For the cases ii) and iii) it was proven by Clément et al. in 1986, if $f \in C^{1,\gamma}[0,\rho_2]$ and $\partial\Omega \in C^3$, that there exists a unique solution $u \in [z,\rho_2]$ for every λ large enough.

The key to the proof of theorem 2 will be the following lemma.

LEMMA :

Let (λ,u) be a solution of (P) and let ν be the first eigenvalue of

$$\text{(L)} \qquad \begin{cases} -\Delta\psi = \nu.\psi & \text{in } B(0,1), \\ \psi = 0 & \text{on } \partial B(0,1). \end{cases}$$

If $f(u) > \sigma.(u-m)$ for $u \in [m,M]$ and $u > m$ on $B(y,(\sigma.\lambda/\nu)^{-\frac{1}{2}})$ then $u(y) > M$.

PROOF :

Let ψ be the associated eigenfunction of (L) with $\psi(0) = 1$. Define $v(t,x) = m + (t-m).\psi((\sigma.\lambda/\nu)^{\frac{1}{2}}.|x-y|)$ for $x \in B(y,(\sigma.\lambda/\nu)^{-\frac{1}{2}})$.

Then
$$\begin{aligned} -\Delta v(t) &= (t-m).(\sigma.\lambda/\nu).(-\Delta\psi) = \\ &= \lambda.\sigma.(t-m)\psi = \\ &= \lambda.\sigma.(v(t)-m) < \lambda f(v(t)) \qquad \text{for } t \in [m,M]. \end{aligned}$$

Since $v(t,x) = m < u(x)$ for $x \in \partial B(y,(\sigma\lambda/\nu)^{-\frac{1}{2}})$, $v(t)$ is a subsolution of (P) for all $t \in [m,M]$. And since $v(m,x) \equiv m < u(x)$ the sweeping principle shows $v(M,x) < u(x)$ in $B(y,(\sigma\lambda/\nu)^{-\frac{1}{2}})$. Hence $v(M,y) = M < u(y)$. □

PROOF OF THEOREM 2:

In the first step we will show that there exists $\lambda(z)$ such that if (λ,u) is a solution of (P) with $\lambda > \lambda(z)$ and $z < u$ then $u > w(\lambda)$, which is also defined in the proof of theorem 1. If $\rho_1 < 0$ then set $w(\lambda) \equiv 0$. If (F3) is satisfied there exists a radially symmetric solution (μ,v) of

$$\begin{cases} -\Delta v = \mu.f^*(v) & \text{in } B(0,1), \\ v = -1 & \text{on } \partial B(0,1) \end{cases} \quad \text{with } 0 \le \max v \in (\rho_1,\rho_2).$$

Let σ and ε_0 be as before, and set

$$w(\lambda,x) = \sup\,\{v((\lambda/\mu)^{\frac{1}{2}}|x-y|);\ y \in \Omega(\theta.(\lambda/\mu)^{-\frac{1}{2}})\},$$

which is a positive subsolution of (P) for $\lambda \ge \lambda_0 = \mu.(\theta/\varepsilon_0)^2$. If one can show $u(x) > v((\lambda/\mu)^{\frac{1}{2}}|x-y|)$ for some $y \in \Omega\ (\theta.(\lambda/\mu)^{-\frac{1}{2}})$ then by sweeping and the fact that $\Omega(\theta.(\lambda/\mu)^{-\frac{1}{2}})$ is connected by arc, this inequality holds for all $y \in \Omega(\theta.(\lambda/\mu)^{-\frac{1}{2}})$.

Define $m := \frac{1}{2}(\rho_1+\max z)$ and $M := v(0)$. Then there exists a ball $B(x^*,r)$, such that $B(x^*,r) \subset \{x \in \Omega;\ z(x) > m\}$, and a constant σ, such that $f(u) > \sigma(u-m)$ for $u \in [m,M]$. By the lemma one finds

$$u(x) > M \text{ for } x \in B(x^*, r-(\sigma.\lambda/\nu)^{-\frac{1}{2}}).$$

When $r-(\sigma.\lambda/\nu)^{-\frac{1}{2}} > \theta.(\lambda/\mu)^{-\frac{1}{2}}$ the first step is finished since

$$u(x) > M \geq v((\lambda/\mu)^{\frac{1}{2}}|x-x^*|) \text{ for } x \in B(x^*,\theta(\lambda/\mu)^{-\frac{1}{2}})$$

Hence set $\lambda(z) = \max(\lambda_0, r^{-2}((\nu/\sigma)^{\frac{1}{2}} + \mu^{\frac{1}{2}})^2)$.

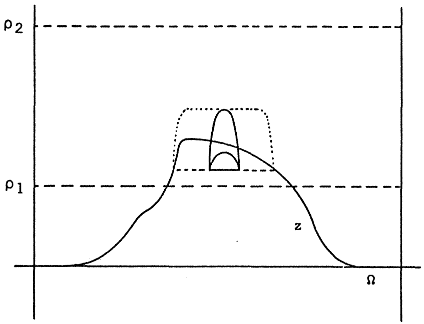


fig.3

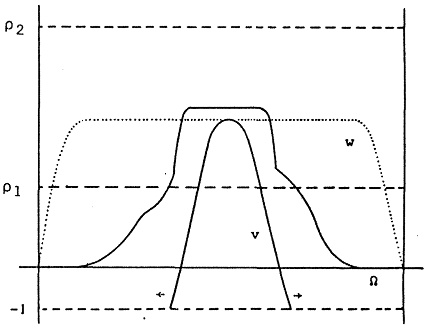


fig.4

In the second step we prove that a solution (λ,u), with $u \in (w(\lambda),\rho_2]$ and $\lambda > \lambda(z)$, satisfies the statement of the theorem. If $\rho_1 < 0$ set $M = 0$. We may assume that c is such that

$$f(u) > c(\rho_2-u)^{\alpha} \text{ for } u \in [M,\rho_2] .$$

Define $M_k = \rho_2 - 2^{-k} . (\rho_2-M)$

and $\sigma_k = c . 2^{-(k+1).(\alpha-1)} . (\rho_2-M)^{\alpha-1}$.

Then $f(u) > \sigma_k . (u-M_k)$ for $u \in [M_k,M_{k+1}]$.

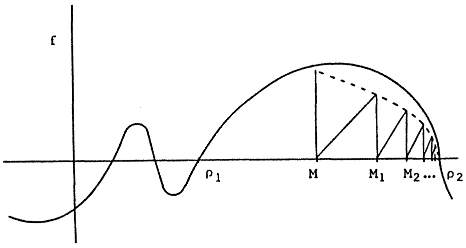


fig.5

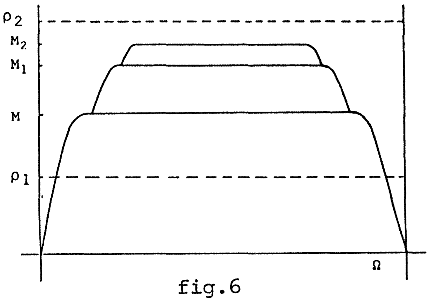


fig.6

Since $u > w(\lambda)$ one finds that

$$u(x) > M \text{ for } x \in \Omega(\theta.(\lambda/\mu)^{-\frac{1}{2}}).$$

The lemma then yields

$$u(x) > M_1 \text{ for } x \in \Omega((\theta.\mu^{\frac{1}{2}} + (\nu/\sigma_1)^{\frac{1}{2}}).\lambda^{-\frac{1}{2}}).$$

And after applying the lemma n times

$$u(x) > M_n \text{ for } x \in \Omega((\theta.\mu^{\frac{1}{2}} + \nu^{\frac{1}{2}}.\sum_{k=1}^{n} (\sigma_k)^{-\frac{1}{2}}).\lambda^{-\frac{1}{2}}).$$

By the definition of σ_k one finds, if $\alpha \neq 1$, that

$$\sum_{k=1}^{n}(\sigma_k)^{-\frac{1}{2}} = c^{-\frac{1}{2}}.(\tfrac{1}{2}(\rho_2-M))^q.\sum_{k=1}^{n}(\tfrac{1}{2})^{qk} = c^{-\frac{1}{2}}.(\tfrac{1}{4}(\rho_2-M))^q.\ \frac{1-(\frac{1}{2})^{qn}}{1-(\frac{1}{2})^q} \quad \text{, with}$$

$$q = \frac{1}{2}(1-\alpha).$$

If $\alpha = 1$, then $\sum_{k=1}^{n} (\sigma_k)^{-\frac{1}{2}} = n.c^{-\frac{1}{2}}$.

CASE 1 : $0 < \alpha < 1$.

For every $n \in \mathbb{N}$, $u > M_n$ in $\Omega(C_1.\lambda^{-\frac{1}{2}})$, with $C_1 = \theta.\mu^{\frac{1}{2}} + \nu^{\frac{1}{2}}.c^{-\frac{1}{2}}(\frac{1}{4}(\rho_2-M))^q.$ $.\ (1-(\frac{1}{2})^q)^{-1}$. Hence $u = \rho_2$ in $\Omega(C_1.\lambda^{-\frac{1}{2}})$, which proves together with $u > w(\lambda)$ the first statement.

CASE 2 : $\alpha = 1$.

For every $n \in \mathbb{N}$, $u > M_n$ in $\Omega((\theta.\mu^{\frac{1}{2}}+n.\nu^{\frac{1}{2}}.c^{-\frac{1}{2}}).\lambda^{-\frac{1}{2}})$. The inequality $u > M_n$ is equivalent with

$$\rho_2 - u < (\rho_2-M).\exp(-n \ln 2).$$

By setting $n = [(c\lambda/\nu)^{\frac{1}{2}}.d(x,\partial\Omega) - \theta.(c\mu/\nu)^{\frac{1}{2}}]$, where $[.]$ denotes the integer function, one finds

$$\rho_2-u < (\rho_2-M) \exp(-\ln 2.((c\lambda/\nu)^{\frac{1}{2}}.d(x,\partial\Omega) - \theta(c\mu/\nu)^{\frac{1}{2}}-1)).$$

Together with $u > w(\lambda)$ this proves the second statement.

CASE 3 : $\alpha > 1$ (hence $q = \frac{1}{2}(1-\alpha) < 0$).

For every $n \in \mathbb{N}$, $u > M_n$ in $\Omega((\theta.\mu^{\frac{1}{2}}+C_1.2^{-qn})\lambda^{-\frac{1}{2}})$ with $C_1 = \nu^{\frac{1}{2}}.c^{-\frac{1}{2}}.$ $.\ (\frac{1}{4}(\rho_2-M))^q.(2^{-q}-1)^{-1}$. Then $u(x) > \rho_2-2(\rho_2-M).(C_1^{-1}.\lambda^{\frac{1}{2}}.d(x,\partial\Omega)-C_1^{-1}.\theta.\mu^{\frac{1}{2}})^{-p}$ with $p = (-q)^{-1} = 2.(\alpha-1)^{-1}$. Together with $u > w(\lambda)$ this proves the third statement. □

REFERENCES

BREZIS, H. and STRAUSS, W.A., 1973, *Semi-linear second-order elliptic equations in L^1,* J.Math.Soc.Japan, 25,4, 565-590.

CLEMENT, Ph. and SWEERS, G., 1986, *Existence and multiplicity results for a semilinear elliptic eigenvalue problem,* to appear in Annali della Scuola Normale Superiore di Pisa.

FIFE, P., 1973, *Semilinear boundary value problems with small parameters,* Arch. Rat. Mech. Anal. 52, 205-232.

GIDAS, B., NI, W.M., NIRENBERG,L., 1979, *Symmetry and related properties via the maximum principle,* Comm. Math. Phys.68, 209-243.

HESS, P., 1981, *On multiple solutions of nonlinear elliptic eigenvalue problems,* Comm. Part. Diff. Eq. 6, 951 - 961.

SERRIN, J., 1971, *Nonlinear equations of second order,* A.M.S.symposium in Partial Diff. Eq., Berkeley, August 1971.

SWEERS,G., 1986, *On the maximum of solutions for a semilinear elliptic problem,* submitted.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Recent Problems in Uniform Asymptotic Expansions of Integrals

Nico M. Temme
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Laplace type integrals are considered for large values of the Laplace variable. Additional parameters may have influence on the classical expansion based on Watson's lemma. In that case modifications of the lemma are needed. We construct several uniform expansions in which the extra parameters do not disturb the validity and the nature of the expansions. Applications and examples are discussed for several special functions.

1980 Mathematics Subject Classification: 41A60
Keywords and Phrases: asymptotic expansions, uniform asymptotics, Laplace integrals

Note: This paper will be presented at the First International Conference on Industrial and Applied Mathematics, Paris, June 29 - July 3, 1987.

1. Introduction

A well-known lemma in asymptotics is the following

LEMMA (Watson). *Consider the Laplace integral*

$$F(z) = \int_0^\infty e^{-zt} f(t) dt. \tag{1.1}$$

Assume that

(i) *f is locally integrable on* $[0, \infty)$*;*

(ii) $f(t) \sim \sum_{s=0}^{\infty} a_s t^{s+\lambda-1}$ *as* $t \to 0^+, \lambda$ *fixed,* $\mathrm{Re}\lambda > 0$*;*

(iii) *the abscissa of convergence of (1.1) is not* $+\infty$.

Then

$$F(z) \sim \sum_{s=0}^{\infty} \Gamma(s+\lambda) a_s z^{-s-\lambda} \tag{1.2}$$

as $z \to \infty$ *in the sector* $|\arg z| \leq \frac{1}{2}\pi - \delta (< \frac{1}{2}\pi)$, *where* z^λ *has its principal value.*

PROOF. See OLVER (1974, p.113). □

Observe that (1.2) is obtained by substituting (ii) into (1.1) and interchanging the order of summation and integration. In (iii) we assume that (1.1) converges if Rez is sufficiently large (and positive). In (ii) λ is fixed. When $\lambda=\Theta(z)$ (or larger) the expansion (1.2) has no meaning. In that case the ratio of consecutive terms is

$$\Gamma(s+\lambda+1)a_{s+1}z^{-s-1-\lambda}/\Gamma(s+\lambda)a_s z^{-s-\lambda} = \Theta(\lambda/z), \tag{1.3}$$

if $a_s, a_{s+1} \neq 0$. It follows that the expansion (1.2) looses this asymptotic nature when $\lambda = O(z)$.

In its paper we consider several cases in which Watsons's lemma is not applicable owing to large or small extra parameters in the Laplace integral. These parameters, of which λ in the above expansion is a special case, may disturb the given expansion, say (1.2), and their influence can be described in terms of the notion of uniformity. The above expansion is not uniformly valid for λ in an unbounded subdomain of Re$\lambda>0$.

We consider the following integrals

$$\int_0^\infty t^{\lambda-1}e^{-zt}f(t)dt \tag{1.4}$$

$$\int_\alpha^\infty t^{\lambda-1}e^{-zt}f(t)dt, \quad \alpha\geqslant 0, \tag{1.5}$$

$$\int_0^\infty t^{\lambda-1}e^{-zt-\alpha/t}f(t)dt, \quad \alpha\geqslant 0, \tag{1.6}$$

$$\int_0^\infty t^{\lambda-1}e^{-\frac{1}{2}zt^2+\alpha t}f(t)dt, \quad \alpha\in\mathbb{R}, \tag{1.7}$$

and we construct asymptotic expansions with z as large parameter. The parameters α and λ play the part of uniformity parameters. For a proper description of the asymptotic estimation of the above integrals we need several special functions as basic approximants, which are obtained by replacing f with a constant, say unity. Then the integrals reduce to:

- gamma function,
- incomplete gamma function,
- Bessel function,
- parabolic cylinder function,

respectively.

2. Gamma function as approximant

In fact, the non-uniform expansion (1.2) with λ in compact subsets of Re$\lambda>0$ also makes use of gamma functions. The uniform expansion does not need another function. We construct an expansion in which λ is allowed to range through the interval $[0,\infty)$. The integral (1.1) is not defined for $\lambda=0$, but instead we use the normalized version

$$F_\lambda(z) = \frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}f(t)dt. \tag{2.1}$$

Now we have, when f is regular at $t=0$,

$$F_0(z) = f(0), \quad \mathrm{Re}z>0.$$

In the following we assume that f is analytic is a domain Ω of the complex plane that contains a circle (with positive radius R) around the origin and a sector $S_{\alpha,\beta}$ defined by

$$S_{\alpha,\beta} = \{t|-\alpha<\arg t<\beta\} \tag{2.2}$$

where α,β are positive numbers. Furthermore, we assume that there is a real number p such that

$$f(t) = \mathcal{O}(t^p),$$

as $t\rightarrow\infty$ in $S_{\alpha,\beta}$. The Taylor coefficients of f at $t=0$ are denoted by a_s, that is,

$$f(t) = \sum_{s=0}^{\infty} a_s t^s, |t|<R. \tag{2.3}$$

The asymptotic expansion of (2.1) is obtained by "expanding" f around the point $t=\mu$, where $\mu=\lambda/z$. We write

$$F_\lambda(z) = z^{-\lambda}f(\mu) + \frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}[f(t)-f(\mu)]dt.$$

Integrating by parts, writing

$$t^\lambda e^{-zt}dt = -\frac{t}{z}\frac{d[e^{-zt}t^\lambda]}{t-\mu},$$

we obtain

$$F_\lambda(z) = z^{-\lambda}f(\mu) + \frac{1}{z\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}f_1(t)dt,$$

where

$$f_1(t) = t\frac{d}{dt}\frac{f(t)-f(\mu)}{t-\mu}.$$

Continuing this procedure, we obtain

$$F_\lambda(z) = z^{-\lambda}[\sum_{s=0}^{n-1} f_s(\mu)z^{-s} + z^{-n}E_n(z,\lambda)], \tag{2.4}$$

where

$$f_{s+1}(t) = t\frac{d}{dt}\frac{f_s(t)-f_s(\mu)}{t-\mu}, \ s = 0,1,..., \tag{2.5}$$

$f_0=f$ and the remainder E_n is given by

$$E_n(z,\lambda) = \frac{z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}f_n(t)dt. \tag{2.6}$$

From authors papers (1983, 1985) it follows that (2.4) is an asymptotic representation for $z \to \infty$ that holds uniformly with respect to $\mu = \lambda/z$ in a closed sector properly interior to $S_{\alpha,\beta}$ defined in (2.2), and in a disc around the origin. For real values of λ, z we can say that

$$E_n(z,\lambda) = O(1) \text{ as } z \to \infty,$$

uniformly with respect to λ or μ in $[0,\infty)$.

This is the generalized version of Watson's lemma that gives expansion (1.2), with values of f (and derivatives) at $t=0$. The new expansion concentrates on values of f at $t=\mu$, the point where $t^\lambda e^{-zt}$ attains a maximal value, and the expansion remains valid when $\mu \to 0$ (Watson's lemma) or when $\mu \to \infty$ independent of z.

Example (Exponential integral) We take $f(t) = 1/(t+1)$. It is easily seen that the function $F_\lambda(z)$ of (1.1) can be written as

$$F_\lambda(z) = z^{1-\lambda} e^z E_\lambda(z);$$

$E_\lambda(z)$ is the well-known exponential integral

$$E_\lambda(z) = \int_1^\infty t^{-\lambda} e^{-zt} dt.$$

The first few terms of (2.4) are easily computed and we obtain

$$E_\lambda(z) \doteq \frac{e^{-z}}{z+\lambda}[1 + \frac{\lambda}{(z+\lambda)^2} + \frac{\lambda(\lambda-2z)}{(z+\lambda)^4} + \frac{(z+\lambda)}{z^4} E_3(z,\lambda)]$$

with E_3 defined in (2.6) with the functions f_s given in (2.5) with $f_0(t) = 1/(t+1)$. From the first coefficients in this expansion it can be seen that large values of λ do not disturb the asymptotic properties of these first coefficients.

When the functions f_n in (2.6) are bounded on $[0,\infty)$, a bound of $|E_n(z,\lambda)|$ can be easily constructed. This gives an error bound for the asymptotic expansion: let positive numbers M_n exists such that, for $n = 0,1,2,....$,

$$|f_n(t)| \leq M_n, \quad t \geq 0.$$

Then for the remainder in (2.4) we obtain

$$|E_n(z,\lambda)| \leq M_n, \quad n = 0,1,2,\ldots.$$

This gives, in a way, an idea of the asymptotic nature of the expansion when μ is fixed. If the numbers M_n do not depend on μ (the functions f_n do!) then the expansion holds uniformly with respect to μ. However, it is more realistic to assume that f_n is not bounded on $[0,\infty)$ and/or that M_n depend on μ. This asks for a more detailed approach for constructing error bounds. See author's papers (1985, 1986).

3. Incomplete gamma function as approximant

We write (1.5) in the form

$$F_\lambda(z,\alpha) = \frac{1}{\Gamma(\lambda)} \int_\alpha^\infty t^{\lambda-1} e^{-zt} f(t) dt \tag{3.1}$$

and we consider z as the large parameter and α and λ as uniformity parameters in $[0,\infty)$; $\alpha=0$ gives the previous case. The saddle point of $t^\lambda e^{-zt}$ at $t=\mu=\lambda/z$ may be inside the domain of integration ($\alpha<\mu$) or outside the domain ($\alpha>\mu$). The transition occurs when α passes the value μ and this arises interesting asymptotic phenomena, for instance for several types of cumulative distribution functions. For certain combinations of the parameters α and μ the function $F_\lambda(z,\alpha)$ can be estimated in terms of the normal (i.e. Gaussian) distribution function or error function. For all possible situations in the parameter domain $(\alpha,\mu)\in[0,\infty)x[0,\infty)$ an incomplete gamma function is needed for a uniform expansion.

When f equals unity the function $F_\lambda(z,\alpha)$ reduces to

$$\frac{1}{\Gamma(\lambda)} \int_\alpha^\infty t^{\lambda-1} e^{-zt} dt = z^{-\lambda} Q(\lambda,\alpha z), \tag{3.2}$$

where Q is the incomplete gamma function

$$Q(a,x) = \frac{1}{\Gamma(a)} \int_x^\infty t^{a-1} e^{-t} dt. \tag{3.3}$$

The integration by parts procedure of the previous section now gives an integrated non-vanishing term at $t=\alpha$. So we obtain the formal expansion

$$F_\lambda(z,\alpha) \sim z^{-\lambda} Q(\lambda,\alpha z) \sum_{s=0}^\infty f_s(\mu) z^{-s} + \frac{\alpha^\lambda e^{-\alpha z}}{z\Gamma(\lambda)} \sum_{s=0}^\infty B_s(\alpha) z^{-s}, \tag{3.4}$$

where

$$B_s(\alpha) = \frac{f_s(\alpha)-f_s(\mu)}{\alpha-\mu}, \; s = 0,1,..., \tag{3.5}$$

and the functions f_s are the same as in (2.5). Observe that the first series in (3.4) also occurs in (2.4). In fact we can write

$$F_\lambda(\alpha,z) = Q(\lambda,\alpha z) F_\lambda(z) + \frac{\alpha^\lambda e^{-\alpha z}}{z\Gamma(\lambda)} B_\lambda(z,\alpha) \tag{3.6}$$

with $F_\lambda(z)$ defined in (2.1) as the complete integral and

$$B_\lambda(z,\alpha) \sim \sum_{s=0}^\infty \frac{B_s(\alpha)}{z^s}, \tag{3.7}$$

the second series in (3.4). It follows that the present case (3.1) makes use of (2.1) and (2.4) and that, hence, in this section only the function $B_\lambda(z,\alpha)$ matters.

In our (1986) paper we have constructed error bounds for the remainders associated with expansion (3.7). Furthermore, the expansions are applied to the incomplete beta function, which can be transformed into (3.1) by means of a rather complicated transformation.

4. Bessel function as approximant

The integral (1.6) reduces to a modified Bessel function in the case that f is a constant. Explicity we have

$$2(\alpha/z)^{\lambda/2}K_\lambda(2\sqrt{\alpha z}) = \int_0^\infty t^{\lambda-1}e^{-zt-\alpha/t}dt. \tag{4.1}$$

In this section we consider

$$F_\lambda(z) = \int_0^\infty t^{\lambda-1}e^{-zt-\alpha/t}f(t)dt, \tag{4.2}$$

which reduces to the above modified Bessel function in the case that f is a constant.

We construct an asymptotic expansion of the above integral for $z\to\infty$. Observe that for $\alpha=0$ the integral reduces to (2.1); for $\alpha>0$ application of Watson's lemma is not possible due to the essential singularity of $\exp(-\alpha/t)$ at $t=0$.

As a first attempt we may expand f as in Watson's lemma at $t=0$. If we substitute (2.3) into (4.2) we obtain

$$F_\lambda(z) \sim \sum_{s=0}^\infty a_s\Phi_s \tag{4.3}$$

with

$$\Phi_s = 2(\alpha/z)^{(\lambda+s)/2}K_{\lambda+s}(2\sqrt{\alpha z}).$$

Suppose α is a fixed positive number. Then

$$\Phi_{s+1}/\Phi_s = \mathcal{O}(\sqrt{\alpha/z}), \text{ as } z\to\infty. \tag{4.5}$$

On the other hand, if $\alpha z\to 0$, we have

$$\Phi_{s+1}/\Phi_s = \mathcal{O}(z^{-1}). \tag{4.6}$$

This gives an indication that (4.3) may yield an asymptotic expansion of $z\to\infty$, with α restricted to a domain $[0,\alpha_0]$, with $\alpha_0=o(z)$ as $z\to\infty$.

In (4.5), (4.6) we have used the well-known asymptotic estimates

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2z}}\,e^{-z}, \text{ as } z\to\infty,$$

$$K_\nu(z) \sim \frac{1}{2}\Gamma(\nu)(2/z)^\nu, \text{ as } z\to 0.$$

The procedure below gives an expansion that is, under suitable conditions on f, uniform with respect to $\alpha\in[0,\infty)$.

We consider (4.2) and we write $\mu^2=\alpha/z$. Saddle points of $\exp(-zt-\alpha/t)$

occur at $t=\pm\mu$, μ is supposed to be positive. The first step is the representation

$$f(t) = a_0 + b_0 t + (t-\mu^2/t)g(t) \tag{4.7}$$

where a_0,b_0 follow from substitution of $t=\pm\mu$. We have

$$a_0 = \frac{1}{2}[f(\mu) + f(-\mu)],\ b_0 = \frac{1}{2\mu}[f(\mu)-f(-\mu)].$$

So we obtain upon inserting (4.7) into (4.2)

$$F(z) = a_0\Phi_0 + b_0\Phi_1 + F_1(z) \tag{4.8}$$

where Φ_j is given in (4.4). An integration by parts gives

$$\begin{aligned} F_1(z) &= \int_0^\infty t^{\lambda-1}e^{-z(t+\mu^2/t)}(t-\mu^2/t)g(t)dt \\ &= -\frac{1}{z}\int_0^\infty t^\lambda g(t)de^{-z(t+\mu^2/t)} \\ &= \frac{1}{z}\int_0^\infty t^{\lambda-1}e^{-z(t+\mu^2/t)}f_1(t)dt, \end{aligned}$$

with $f_1(t)=t^{1-\lambda}\frac{d}{dt}[t^\lambda g(t)]=\lambda g(t)+tg'(t)$. We see that $zF_1(z)$ is of the same form as $F(z)$. The above procedure can now be applied to $zF_1(z)$ and we obtain for (4.2) the formal expansion

$$F(z)\sim\Phi_0\sum_{s=0}^\infty\frac{a_s}{z^s} + \Phi_1\sum_{s=0}^\infty\frac{b_s}{z^s},\ \text{as } z\to\infty, \tag{4.9}$$

where we define inductively $f_0(t)=f(t)$, $g_0(t)=g(t)$ and for $s=1,2,...,$

$$f_s(t) = t^{1-\lambda}\frac{d}{dt}[t^\lambda g_{s-1}(t)] = a_s + b_s t + (t-\mu^2/t)g_s(t).$$

$$a_s = \frac{1}{2}[f_s(\mu) + f_s(-\mu)],\ b_s = \frac{1}{2\mu}[f_s(\mu)-f_s(-\mu)].$$

By using the recursion relation

$$K_{\nu+1}(z) = K_{\nu-1}(z) + \frac{2\nu}{z}K_\nu(z)$$

the series in (4.3) can be rearranged in the form (4.9); however, then the coefficients are essentially different. In (4.3) a_s comes from f and derivatives of f at $t=0$; in (4.9) a_s and b_s come from function values of f and derivatives of f at $t=\mu$ and $t=-\mu$.

We still need to prove that (4.9) is uniformly valid for $\alpha\in[0,\infty)$. At the moment a proof is not available and the proper conditions on f have to be formulated.

An interesting application of the expansions (4.9) can be given for confluent hypergeometric functions. Let us consider

$$U(a,b,,x) = \frac{1}{\Gamma(a)}\int_0^\infty t^{a-1}(1+t)^{b-a-1}e^{-xt}dt \tag{4.10}$$

for $a \rightarrow +\infty$, with $x > 0$, $b \in \mathbb{R}$. A transformation to the standard form (4.2) is needed, but first we give a simple transformation. The function $[t/(1+t)]^a$ takes its maximal value (on $[0,\infty)$) at $t=+\infty$. This function plays the role of an exponential function. Therefore we take as a new variable of integration τ defined by $t/(1+t)=\exp(-\tau)$. Then (6.9) becomes

$$U(a,b,x) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-a\tau - x/(e^\tau-1)}\tau^{-b}f(\tau)d\tau \tag{4.11}$$

where $f(\tau)=[\tau/(1-e^{-\tau})]^b$. The easiest way to arrive at the standard form (4.2) is to write

$$U(a,b,x) = \frac{e^{\frac{1}{2}x}}{\Gamma(a)}\int_0^\infty \tau^{-b}e^{-a\tau - x/\tau}\tilde{f}(\tau)d\tau \tag{4.12}$$

where

$$\tilde{f}(\tau) = f(\tau)\exp\{x[1/\tau - 1/(e^\tau-1) - \tfrac{1}{2}]\}.$$

Now we can use the procedure leading to (4.9), with $\lambda = 1-b$. The result is an expansion for $a \rightarrow \infty$, which holds uniformly with respect to $x \in [0,x_0]$, where x_0 is a fixed positive number. It is not possible here to replace x_0 by ∞; the main reason is that $\tilde{f}(\tau)$ depends on x, in such a way that coefficients a_s and b_s in (4.9) grow too fast when $x \rightarrow \infty$.

A more powerful expansion is obtained (with respect to the uniformity domain of x) when we transform (4.11) into (4.2) by using the mapping $u : \mathbb{R} \rightarrow \mathbb{R}$ that is defined by

$$\tau + \frac{\nu}{e^\tau - 1} = u + \frac{\alpha}{u} + A \tag{4.13}$$

where $\nu = x/a$;α and A are to be determined. We compute them by the following condition on the mapping u: the critical points at the left-hand side of (4.13) ($\tau = \pm\gamma$, where γ is the positive number satisfying $\cosh\,\gamma = 1+\frac{1}{2}\nu$) should correspond with those at the right ($u=\pm\mu$, where $\mu^2=\alpha$). If follows that

$$A = -\tfrac{1}{2}\nu,\ \mu = \tfrac{1}{2}(\gamma + \sinh\gamma). \tag{4.14}$$

These choices make the mapping u regular with $u(0)=0$, $u(\pm\infty)=\pm\infty$. We now obtain from (4.13) and (4.11)

$$U(a,b,x) = \frac{e^{-\frac{1}{2}x}}{\Gamma(a)}\int_0^\infty u^{-b}e^{-a(u+\alpha/u)}f^*(u)du, \tag{4.15}$$

where

$$f^*(u) = (u/\tau)^b f(\tau)\frac{d\tau}{du},$$

with f as in (4.11). We expect that expanding (4.15) as in (4.9) will give $[0,\infty)$ as uniformity domain for x. Proofs are needed. The first thing to do is to prove the regularity of f^* in a fixed domain containing $\mathbb{R}$ in its interior. Observe that f^* depends on the uniformity parameter α.

After these preparations the asymptotic expansion of the integral in (4.14) can be constructed by computing the coefficients a_s, b_s that appear in (cf. (4.9))

$$e^{\frac{1}{2}x}\Gamma(a)U(a,b,\nu a) \sim \Phi_0 \sum_{s=0}^{\infty}\frac{a_s}{a^s} + \Phi_1 \sum_{s=0}^{\infty}\frac{b_s}{a^s}, \text{ as } a\to\infty, \tag{4.17}$$

where Φ_0,Φ_1 are given in (4.4) with $\lambda = 1-b$, $z = a$ and $\alpha = \mu^2$, with μ defined in (4.14). The first coefficients in (4.16) are

$$a_0 = \frac{1}{2}[f^*(\mu) + f^*(-\mu)],\ b_0 = \frac{1}{2\mu}[f^*(\mu) - f^*(-\mu)].$$

A few calculations based on (4.13) and l' Hôpital's rule give

$$\frac{d\tau}{du}\Big|_{u\,=\,\pm\mu} = [\frac{2(1-e^{-\gamma},}{\mu(1+e^{-\gamma},}]^{\frac{1}{2}},\mu = \frac{1}{2}[\gamma + \sinh\gamma].$$

To express the first coefficients a_0, b_0 in terms of γ, let $\zeta(\gamma)$ denote the above value of $d\tau/du$ at $u = \pm\mu$, and write

$$\eta(\gamma) = [\frac{\mu}{1-e^{-\gamma}}]^b.$$

Then we have

$$a_0 = \frac{1}{2}\zeta(\gamma)[\eta(\gamma) + \eta(-\gamma)],\ b_0 = \frac{1}{2\mu}\zeta(\gamma)[\eta(\gamma) - \eta(-\gamma)].$$

Observe that the coefficients contain function values of f^* at the negative axis, although f^* in (4.15) is only used for non-negative u-values.

5. Parabolic cylinder function as approximant

In the previous section an essential singularity at $t = 0$ is incorporated in the Laplace integral. By replacing the exponential function in (2.1) with $\exp(-zt + \alpha\sqrt{t})$ a simpler singularity occurs and in fact this type of singularity can be accepted in Watson's lemma. Then the function $\exp(\alpha\sqrt{t})f(t)$ has to be expanded in a power series. It is more interesting to couple the parameter α with the large parameter z and to consider the effect when α crosses the origin. A slight change of variables gives a quadratic polynomial in the exponential function. In fact we consider

$$I_\lambda(z,\alpha) = \int_0^\infty t^{\lambda-1} e^{-\frac{1}{2}zt^2 + \alpha t} f(t)dt \tag{5.1}$$

for a large values of z; λ is a fixed positive parameter and α a uniformity parameter in $\mathbb{R}$. The saddle point occurs at $t=\alpha/z$. When α is positive it lies inside the interval of integration, when α is negative it is outside the interval. The transition at $\alpha=0$ can be described by using parabolic cylinder functions as approximants, i.e. the above integral with $f(t)\equiv$ constant.

In [1] BLEISTEIN introduced an integration by parts procedure that produced what is now called a canonical expansion. In a way, the procedures of the previous sections are all based on this approach. We repeat the steps in Bleistein's procedure and we also consider a new method for obtaining a similar expansion.

Let $\beta=\alpha/z$ and write

$$f(t) = a_0 + b_0 t + t(t-\beta)g(t), \tag{5.2}$$

with

$$a_0 = f(0),\ b_0 = \frac{f(\beta)-f(0)}{\beta}.$$

Then we have

$$I_\lambda(z,\alpha) = a_0 W_{\lambda-1} + b_0 W_\lambda + J_\lambda(z,\alpha), \tag{5.3}$$

with

$$W_\lambda = \int_0^\infty t^\lambda e^{-\frac{1}{2}zt^2+\alpha t}dt, \tag{5.4}$$

a parabolic cylinder function, and

$$J_\lambda(z,\alpha) = \int_0^\infty t^\lambda (t-\beta) e^{-z(\frac{1}{2}t^2-\beta t)} g(t)dt. \tag{5.5}$$

Integrating by parts gives

$$J_\lambda(z,\alpha) = -\frac{1}{z}\int_0^\infty t^\lambda g(t) de^{-z(\frac{1}{2}t^2-\beta t)}$$

$$= \frac{1}{z}\int_0^\infty t^{\lambda-1} e^{-z(\frac{1}{2}t^2-\beta t)} f_1(t)dt$$

with

$$f_1(t) = t^{1-\lambda}\frac{d}{dt}[t^\lambda g(t)].$$

Repeating this process we obtain the above mentioned canonical expansion

$$I_\lambda(z,\alpha) = W_{\lambda-1}\sum_{s=0}^{n-1}\frac{a_s}{z^s} + W_\lambda \sum_{s=0}^{n-1}\frac{b_s}{z^s} + z^{-n}E_n \tag{5.6}$$

with

$$a_s = f_s(0),\ b_s = \frac{f_s(\beta)-f_s(0)}{\beta}$$

$$f_{s+1}(t) = t^{1-\lambda}\frac{d}{dt}[t^{\lambda}\frac{g_s(t)-a_s-b_st}{t(t-\beta)}],\ s=0,1,...,$$

$f_0=f$, and E_n is the remainder given by

$$E_n = \int_0^{\infty} t^{\lambda-1}e^{-z(\frac{1}{2}t^2-\beta t)}f_n(t)dt.$$

Bounds for $|E_n|$ and proofs for the validity of the expansion can be based on bounds for $|f_n(t)|$ on $[0,\infty)$. A complication is that f_n depends also on the uniformity parameter β.

In a forthcoming paper of Soni & Sleeman the above procedure is replaced with an approach that resembles the procedure in Watson's lemma. Recall that in Watson's lemma (1.2) is obtained by substituting the expansion in (ii). Soni and Sleeman introduce a set of polynomials $\{P_k\}$ satisfying

$$\begin{cases} P_0(t) = 1, \quad P_1(t) = t/(\gamma+1) \\ [t^{\lambda}P_n(t)]' = t^{\lambda}(t-\beta)P_{n-2}(t), \quad n=2,3,.... \end{cases} \tag{5.7}$$

Then they assume that g in (5.2) can be expanded in terms of $\{P_k\}$, writing

$$g(t) = \sum_{k=0}^{\infty} c_kP_k(t), \tag{5.8}$$

where c_k are independent of t and have to be determined. Substituting this expansion in (5.5), we obtain the formal expansion

$$J_{\lambda}(z,\alpha) = \sum_{k=0}^{\infty} c_k\phi_k,$$

$$\phi_k = \int_0^{\infty} t^{\lambda}(t-\beta)e^{-z(\frac{1}{2}t^2-\beta t)}P_k(t)dt.$$

By using the properties of $\{P_k\}$ given in (5.7) it follows that $\phi_k=z^{-1}\phi_{k-2}$ and, hence, that

$$J_{\lambda}(z,\alpha) = \phi_0\sum_{k=0}^{\infty} c_{2k}z^{-k} + \phi_1\sum_{k=0}^{\infty} c_{2k+1}z^{-k},$$

which is of the same form as the expansion in (5.6). There is a simple relation between c_k and a_k,b_k. The computation of c_k in (5.8) is not a simpler problem than the computation of a_s,b_s in (5.6). It is expected, however, that this new approach will give new methods for constructing bounds of the remainders in the asymptotic expansion. Soni and Sleeman's method can also be used for other types of integrals.

References

[1] Bleistein, N. 1966, Uniform asymptotic expansions of integrals with stationary point near algebraic singularity, Comm. Pure Appl. Math. **19**, 353-370.

[2] Olver, F.W.J. 1979, Asymptotics and Special functions, Academic Press.

[3] Soni, K. and B. Sleeman, On uniform asymptotic expansions and associated polynomials, to appear in J. Math. Anal. and Appl.

[4] Temme, N.M. 1983, Uniform asymptotic expansions of Laplace integrals, Analysis **3**, 221-249.

[5] Temme, N.M. 1985, Laplace integrals: transformation to standard form and uniform asymptotic expansions, Quart. of Appl. Math. 103-123.

[6] Temme, N.M. Incomplete Laplace integrals: uniform asymptotic expansions with application to the incomplete beta integral; to appear in SIAM J. Math. An.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Symmetry and Integrability

F. Verhulst
Mathematisch Instituut
Rijksuniversiteit Utrecht
P.O. Box 80010, 3508 TA Utrecht
The Netherlands

SUMMARY

The regular behaviour of many orbits of particles or fluid elements in models of continuum mechanics can be explained with the concept of approximate integrability of Hamiltonian systems. Using the technique of averaging and normal forms, a high degree of approximate integrability of these models follows from assumptions like axial and mirror symmetry.

1. INTRODUCTION

Hamiltonian sytems play an important part in fluid mechanics, celestial mechanics and astrophysics. Generically these systems are non-integrable i.e. for a Hamiltonian system with n degrees of freedom there exist in general less than n functionally independent integrals which are in involution; the cases with n or more integrals (for example the gravitational two-body problem) are exceptional. For a survey of these integrable cases see Lynden-Bell (1962).
The classical example of a non-integrable system was given by Hénon and Heiles (1964). In this example there are regular and irregular orbits. The regular orbits are periodic solutions and solutions moving on invariant tori around the stable periodic solutions. The irregular orbits follow no

regular geometric pattern and are sometimes called wild or stochastic. Near the equilibrium solution in phase space, the regular orbits dominate, further away from equilibrium the irregular orbits become more and more important. This is in agreement with the KAM-theorem.

In studying equilibrium models of galaxies, it is of fundamental importance to assess the influence of the irregular orbits on the over-all dynamics of the model, see Binney (1982) and Binney and Tremaine (1987). It turns out that in actual models of galaxies the behaviour of most numerically computed orbits is surprisingly regular, the irregular orbits often being restricted to certain zones in phase-space. We shall show that this observed regularity is to be expected for two degrees of freedom systems like axi-symmetric or planar galaxies and that this regularity is tied in with natural symmetry assumptions for three degrees of freedom systems like elliptical galaxies.

The Hamiltonian systems which we shall discuss are, apart from the symmetry assumptions, completely general. This means that the application to galactic modelling which we mention here, is just one of the many examples of possible application. For application to nonlinear wave equations see Stroucken and Verhulst (1987) and Van der Aa and Krol (1987).

2. NORMAL FORMS AND APPROXIMATE INTEGRALS.

In this section we summarize the technique to analyse nonlinear dynamical systems; for details and further references see Arnold(1983) or Sanders and Verhulst (1985). An introduction to normalisation is given in Verhulst (1987). Suppose we are considering a Hamiltonian system with n degrees of freedom, characterized by a Hamiltonian function H(q,p) which in a neighbourhood of an equilibrium point can be expanded as

$$H = H_2 + H_3 + \ldots + H_m + \ldots \qquad (1)$$

in which H_k, $k = 2,3,\ldots$ is homogeneous in q and p of degree k. We shall restrict ourselves to the case in which H_2 is a positive definite quadratic form so that the equilibrium point (0,0) is stable. To make quantitatively explicit that we expand in a neighbourhood of the equilibrium point we shall use the small parameter ϵ for the scaling $q = \epsilon\bar{q}$, $p = \epsilon\bar{p}$; so ϵ^2 is a measure for the energy. Introducing this scaling in (1), dividing by ϵ^2 and dropping the bars produces

$$H = H_2 + \epsilon H_3 + \ldots + \epsilon^{m-2} H_m + \ldots \qquad (2)$$

We simplify the Hamiltonian (2) and the corresponding equations of motion by an averaging transformation or, equivalently, Birkhoff-Gustavson normalisation. This is a canonical near-identity transformation which leaves H_2 invariant and which removes a large number of terms of H_3, $H_4,\ldots$ to higher order. Suppose we have normalised to degree m, the Hamiltonian in normal form to this degree is

$$\bar{H} = H_2 + \epsilon\bar{H}_3 + \ldots + \epsilon^{m-2}\bar{H}_m + \ldots \qquad (3)$$

Until now the description of the system is exact. Solving the equations of motion corresponding with (3) and inverting the normalising transformation produces the solutions corresponding with (2).

The next step involves an approximation: truncation of (3) at the level of terms of degree m

$$\bar{H}_t = H_2 + \epsilon\bar{H}_3 + \ldots + \epsilon^{m-2}\bar{H}_m \qquad (4)$$

It turns out that if m is taken large enough, periodic solutions of the equations of motion corresponding with (4) represent an approximation of periodic solutions which exist in the system corresponding with (3) and (2). Moreover, as H represents an integral of the phase-flow induced by (2),

$\bar{H}$ is an integral of the flow induced by (3), $\bar{H}_t$ is an integral of the flow induced by (4). Furthermore H_2 is a second independent integral of the equations of motion corresponding with (4). It is easy to show that this quadratic integral is conserved for the system corresponding with (3) or (2) with error $O(\epsilon)$ for all time.

If the equations of motion induced by (4) have more than two independent integrals we have for these additional integrals a slightly weaker estimate. Suppose $I(q,p)$ is such an integral, I_0 its constant value for given initial conditions. Then for the equations of motion induced by (3) (and after inverting the transformation, for the equations of motion induced by (2)) we have

$$I(q,p) - I_0 = O(\epsilon^{m-1}t) \qquad (5)$$

This estimate follows simply by calculating the orbital derivative of I for sytem (3); note that we can split the Poisson bracket by using $\bar{H} = \bar{H}_t + \epsilon^{m-1}\bar{H}_{m+1} + \ldots$ and the fact that the orbits are bounded. We find

$$\frac{dI}{dt} = [I,\bar{H}] = [I,\bar{H}_t] + O(\epsilon^{m-1}) = O(\epsilon^{m-1}).$$

Integration produces the estimate (5).

More subtle extimates can be obtained concerning the individual orbits, see Sanders and Verhulst (1979, 1985).

It follows from the estimate (5) that in the worst case, $m = 3$, the expression $I(q,p)$ is conserved for the flow induced by (3) (or (2)) with error $O(\epsilon)$ on the long time-scale $1/\epsilon$. If $m > 3$, the estimate improves; one can use this to obtain approximations on a longer time-scale or to have more precision on the time-scale $1/\epsilon$.

In this sense the (exact) integral $I(q,p)$ of (4) is an approximate integral of the original system corresponding with (3) or (2). Note that I is not a formal integral but an approximation in the rigorous mathematical sense of the word.

An important consequence is the following. If the phase-flow induced by (4) has n independent integrals, the phase-flow corresponding with the original Hamiltonian (3) or (2) is approximately integrable in the sense described above. This means that the occurrence of irregular orbits in such a system is limited by the given error estimates and must be a small-scale phenomenon on a long time-scale.

3. RESONANCE

The quadratic part of the Hamiltonian can be written as

$$H_2 = \frac{1}{2}\omega_1(q_1^2 + p_1^2) + \ldots + \frac{1}{2}\omega_n(q_n^2 + p_n^2)$$

The numbers $\omega_1,\ldots,\omega_n$ are positive and they are the frequencies of the linearized flow around equilibrium; $(\omega_1,\ldots,\omega_n)$ is called the frequency vector. We shall take the ratios of the frequencies to be rational; irrational ratios ρ can always be approximated arbitrarily close by a rational number m/n where the detuning ρ-m/n is taken as a perturbation factor. The actual choice of m and n is determined by ρ and by the energy level we are considering; for smaller values of the energy we have to take a more accurate approximation of ρ. Put in a different way; increasing the energy around equilibrium, more important resonances may be encountered as m + n can be smaller.

The presence of resonances is indicated by the annihilation vector $(k_1,\ldots,k_n)$ where $k_1,\ldots,k_n$ are integer numbers such that the frequency vector and the annihilation vector are orthogonal. Counting the independent annihilation vectors for Hamiltonian (4) with $k = |k_1| + |k_2| + \ldots + |k_n| \le m$ we are assessing the part played by resonance. Symmetry assumptions diminish in general the number of annihilation vectors and reduce the part played by resonance.

Considering two degrees of freedom systems, n = 2, it is easy to see that at H_3 only the 1 : 2 resonance is effective (annihilation vector (2,-1) with k = 3); this is called a first order resonance. At H_4 level there are the resonances 1 : 3 and 1 : 1, we call these second order resonances. The implication is that if $\omega_1 : \omega_2$ is 1 : 3 or 1 : 1 we have to calculate the normal form to H_4, m = 4, to have nonlinear interaction. If however, $\omega_1 : \omega_2 = 1 : 2$ with discrete or mirror symmetry in the second degree of freedom (replacing q_2, p_2 by $-q_2$, $-p_2$ leaves the system invariant) the first annihilation vector we encounter is (4,-2) and we have to calculate the nornal form to H_6 to study the nonlinear interaction caused by the resonance.

Repeating the analysis for three degrees of freedom we collect the first order resonances in table 1; in these cases two independent annihilation vectors can be found at H_3 level. Second order resonances are the cases where we have to go to H_4 to find at least two independent annihilation vectors; they are collected in table 2. This classification holds for the general Hamiltonian without assumptions of symmetry.

Number	Resonance
1	1:2:2
2	1:2:4
3	1:2:1
4	1:2:3

Table 1.
First order resonances for systems with three degrees of freedom.

Number	Resonance
5	1:1:1
6	1:1:3
7	1:2:6
8	1:3:4
9	1:2:5
10	1:3:7
11	1:3:6
12	2:3:4
13	1:3:3
14	1:3:5
15	1:3:9
16	2:3:6

Table 2
Second order resonances for systems wtih three degrees of freedom.

4. TWO DEGREES OF FREEDOM

In stellar dynamics, the assumption of axial symmetry or restricting the motion of stars to a plane, presents us with a two degrees of freedom Hamiltonian system. The study of Hénon and Heiles (1964) was inspired by this.

It follows from our results in section 2, that the phase flow of such a system is always approximately integrable. This holds for values of the energy not too high. In Verhulst (1979) the phase flow of an axial symmetric galaxy with mirror symmetry with respect to the equatorial plane is analysed by normalisation to H_4. The origin of phase space has been chosen in reference frames which are comoving with the various circular orbits in the equatorial plane. It turns out that because of the mirror symmetry only the 1 : 1 resonance plays a part; this resonance arises near the centre and near the edge of the galaxy. In between, the flow looks like a separable one because of the mirror symmetry.

Normalisation to higher order terms than H_4 plays a part in the paper by Sanders and Verhulst (1979). They carried out computations for a cubic potential proposed by Contopoulos at the same time explaining the meaning of the so-called formal integral.

5. THREE DEGREES OF FREEDOM.

In section 2 we have pointed out that the Hamiltonian in normal form (4) always has at least two independent integrals. If we can find a third independent integral, Hamiltonian (4) is integrable and the original Hamiltonian (3) or (2) is approximately integrable. For a survey of the results without symmetry assumptions see Verhulst (1983) and Sanders and Verhulst (1985).

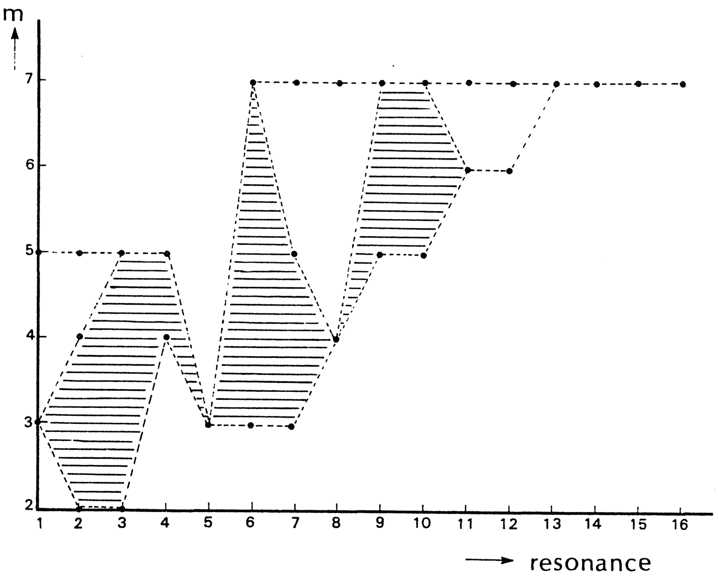


Figure 1

Horizontally the resonances have been given according to tables 1 and 2. Vertically the highest values of m have been given for which the truncated Hamiltonian in normal form (4) is known to be integrable. The lowest dashed line is based on the assumption of discrete symmetry in the first degree of freedom. The shaded area indicates the improvement derived from the assumption that one has also discrete symmetry in the second degree of freedom. The upper dashed line assumes discrete symmetry in three degrees of freedom.

Here we shall present results for the 16 basic resonances with discrete (mirror) symmetries. For other resonances the results are stronger. It should be noted that the integrability is given in so far as it is known at present; it is possible that in some cases our results can be improved upon. In most cases the integrability is obtained by counting the number of independent annihilation vectors. In figure 1 the lowest dashed line indicates for each resonance the highest value of m for which the truncated

Hamiltonian in normal form (4) is integrable if one assumes discrete symmetry in the first degree of freedom. The upper: dashed line indicates the integrability if one assumes discrete (mirror) symmetry in all three degrees of freedom; this is the usual assumption in models for elliptical galaxies. In the last case the original Hamiltonian (2) is for first order resonances approximately integrable with error $O(\epsilon^4 t)$; in the case of second order resonance (with the exception of the 1 : 1 : 1 resonance) we have approximate integrability with error $O(\epsilon^6 t)$.

It is interesting to see the improvement of approximate integrability when symmetry assumptions are added. Also it turns out that in these models the 1 : 1 : 1 resonance presents the most difficult case. This causes the papers of de Zeeuw (1985 ab) and de Zeeuw and Lynden-Bell (1985) to be of particular interest.

In constructing figure 1 we have started with the first degree of freedom, then adding the second, which is rather arbitrary. Therefore, in table 3 we are listing the integrability properties for the three groups of symmetry cases. The group of discrete (mirror) symmetry in one degree of freedom has three cases for each resonance (equal frequencies have not been identified) so there are 48 cases for the basic resonances; the same holds for discrete symmetry in two degrees of freedom. The numbers in the columns indicate how many cases are integrable when normalized to $H_3 \ldots H_8$. The numbers of the last column are trivial; they have been included to remind us where approximate integrability stops for the 16 basic resonances.

	Normal form integrable to					
Discrete symmetry in	H_3	H_4	H_5	H_6	H_7	H_8
one degree of freedom (48 cases)	45	31	25	15	10	0
two degrees of freedom (48 cases)	47	42	36	29	26	0
three degrees of freedom (16 cases)	16	15	15	11	11	0

Table 3.

Number of approximate integrability cases for the 16 basic resonances of three degrees of freedom systems as known at present

6. AN EXAMPLE: THE 1:3:7-RESONANCE

To obtain the results of section 5, the 16 basic resonances of three degrees of freedom systems have to be analysed in detail. To demonstrate the analysis we shall discuss the 1:3:7-resonance.

In this case we have

$$H_2 = \frac{1}{2}(q_1^2+p_1^2)+\frac{3}{2}(q_2^2+p_2^2)+\frac{7}{2}(q_3^2+p_3^2).$$

It is convenient to use action-angle variables r, ϕ which are given by the canonical transformation

$$q_i = \sqrt{2r_i}\,\sin\phi_i$$

$$p_i = \sqrt{2r_i}\,\cos\phi_i, \qquad i = 1,2,3.$$

In action-angle variables we can write

$$H_2 = r_1 + 3r_2 + 7r_3.$$

We calculate the normal form (4) to m = 8; the functions $h_p(r_1,r_2,r_3)$

are homogeneous of degree p with terms, containing factors only of the form r_i^n and $r_i^{n+\frac{1}{2}}$, $n \in \mathbb{N}$. We find

$$\bar{H}_3 = 0$$

$$\bar{H}_4 = h_2(r_1,r_2,r_3) + a_1\, r_1^{3/2}\, r_2^{1/2}\cos(3\phi_1-\phi_2+a_2) +$$
$$a_3\, r_1^{1/2}\, r_2\, r_3^{1/2}\cos(\phi_1+2\phi_2-\phi_3+a_4).$$

$$\bar{H}_5 = h_{5/2}(r_1,r_2,r_3).$$

$$\bar{H}_6 = h_3(r_1,r_2,r_3) + b_1\, r_1\, r_2^{3/2} r_3^{1/2}\, \cos(2\phi_1-3\phi_2+\phi_3+b_2) +$$
$$b_3 r_1^2 r_2^{1/2} r_3^{1/2}\cos(4\phi_1+\phi_2-\phi_3+b_4).$$

$$\bar{H}_7 = h_{7/2}(r_1,r_2,r_3).$$

$$\bar{H}_8 = h_4(r_1,r_2,r_3) + c_1 r_1^3 r_2\, \cos(6\phi_1-2\phi_2+c_2) +$$
$$c_3 r_1 r_2^2 r_3\, \cos(2\phi_1+4\phi_2-2\phi_3+c_4)+c_5 r_1^{7/2} r_3^{1/2}\, \cos(7\phi_1-\phi_3+c_6)+$$
$$c_7 r_1^{1/2} r_2^{5/2} r_3\, \cos(\phi_1-5\phi_2+2\phi_3+c_8).$$

Note that we have used the annihilation vectors (3,-1,0) and (1,2,-1) for H_4, (2,-3,1) and (4,1,-1) for H_6 etc.

The equations of motion are

$$\dot{r}_i = -\frac{\partial H}{\partial \phi_i},\ \dot{\Phi}_i = \frac{\partial H}{\partial r_i},\quad i = 1,2,3.$$

If one calculates the normal form to level m, as in (4), and if there is no combination angle present, the actions r_1,r_2,r_3 are integrals of the system. If one combination angle is present in the normalized Hamiltonian, the normalized system is also integrable. The integrals are $\bar{H}_t$, H_2 and the integral generated by the fact that only one combination angle arises in the equations. If at least two combination angles are present in the normal form, the system may be either integrable or not, but the analysis is not easy.

We discuss some cases used for the construction of table 3 and figure 1.

a. Mirror symmetry in the first-degree of freedom.

We find $a_1 = a_3 = 0$, $b_1, b_3 \neq 0$. The Hamiltonian normalized till H_5 is still integrable; integrals are the actions r_1, r_2, r_3; they are approximate integrals of the original Hamiltonian with error $O(\epsilon^4 t)$.

b. Mirror symmetry in the second degree of freedom.

We find $a_1 = 0$, $a_3 \neq 0$, $b_1 = b_3 = 0$, c_1, $c_3 \neq 0$, $c_5 = c_7 = 0$.

The Hamiltonian normalized till H_7 is still integrable; the integrals are approximate integrals of the original Hamiltonian with error $O(\epsilon^6 t)$.

c. Mirror symmetry in the first and second degree of freedom.

In this case $a_1 = a_3 = b_1 = b_3 = c_5 = c_7 = 0$.

Again we have integrability till $\bar{H}_7$, the actions are approximate integrals of the original Hamiltonian with error $O(\epsilon^6 t)$.

d. The case with mirror symmetry in three degrees of freedom coincides for the 1:3:7-resonance with case c.

More straightforward is the analysis of the periodic solutions of the equations of motion. It should be noted that if one has mirror symmetry in three degrees of freedom, the three normal modes are (exact) periodic solutions of the normalized and the original Hamiltonian system. Also, that in this case there exist three fourdimensional invariant sets, corresponding with the respective two degrees of freedom systems imbedded in the three degrees of freedom system. According to section 4, the normalized flow in these invariant subsets is integrable for all time.

It should be noted that the analysis of these invariant subsets requires normalization to a very high order. Taking for instance the case of mirror symmetry in the first degree of freedom, the second and third degree of freedom form an invariant subset involving the 3:7-resonance. To describe this system in some detail we have to normalize to H_{10}; the annihilation vector is (7,-3). If the second and/or the third degree of freedom also has mirror symmetry, we have to normalize to H_{20}; annihilation vector (14,-6).

ACKNOWLEDGEMENT.

Writing this paper was stimulated by the discussion when giving a seminar at the Institute for Advanced Study, Princeton, November 1986. R. Cushman, J.J. Duistermaat and H.C. van de Hulst commented on the manuscript.

REFERENCES

Arnold, V.I., Geometrical methods in the theory of ordinary differential equations, Springer-Verlag, New York etc. 1983.

Binney, J.J., Ann. Rev. Astron. Astrophys. 20, 399-429, 1982

Binney, J.J. and Tremaine, S.D., Galactic Dynamics, 1987.

de Zeeuw, T., M.N. 215, 731-760 (1985a).

de Zeeuw, T., M.N. 216, 273-334 (1985b).

de Zeeuw, T. and Lynden-Bell, D., M.N. 215, 713-730 (1985).

Hénon, M. and Heiles, C.,Astron.J. 69, 73-79, 1964.

Lynden-Bell, D., M.N. 124, 9-123, 1962.

Sanders, J.A. and Verhulst, F., Lecture Notes Math. 711 (F. Verhulst ed.), Springer-Verlag, 1979.

Sanders, J.A. and Verhulst, F., Averaging methods in nonlinear dynamical systems, Applied Math. Sciences 59, Springer-Verlag, 1985.

Stroucken, A. and Verhulst, F., Math. Methods Applied Sciences, 1987.

Van der Aa, E., and Krol, M., Preprint Rijksuniversiteit Utrecht,1987.

Verhulst, F., Phil. Trans. roy. Soc. London 290, 435-465, 1979.

Verhulst, F., Lecture Notes Math. 985 (F. Verhulst ed.), 137-183, Springer-Verlag, 1983.

Verhulst, F., Nonlinear differential equations and dynamical systems, Universitext, Springer-Verlag, Heidelberg etc. 1987.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Numerical Improvement of the Gauss-Jordan Algorithm

T.J. Dekker and W. Hoffmann
Department of Mathematics
University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

Abstract. In this paper a Gauss-Jordan matrix inversion algorithm with column interchanges is presented and analysed. This analysis gives theoretical evidence that the solutions are as good as those obtained by Gaussian elimination and the residuals mostly are equally small. Moreover, the algorithm presented has good vectorisation properties. The results of numerical experiments and timing experiments on a Cyber 205 are fully satisfactory.

1. INTRODUCTION

The Gauss-Jordan algorithm for the solution of linear systems has received renewed interest because of its supposedly good properties with respect to parallelization. For inverting matrices, the Gauss-Jordan algorithm requires the same number of operations as Gaussian elimination plus backsubstitution, but is simpler and better vectorisable.

As has been shown by Peters and Wilkinson [6], the Gauss-Jordan algorithm with the usual column pivoting strategy and row interchanges, may yield large residuals corresponding to the calculated solutions when the matrix is ill conditioned. We showed in [1], however, that the algorithm with row pivoting and correspondingly interchanging of columns, is much more satisfactory, and yields residuals which in most cases are not larger than those corresponding to the Gaussian elimination solutions. In this paper we use the improved Gauss-Jordan algorithm for the development of a fast and reliable routine for matrix inversion.

In the next three sections we present our formulation of the basic Gauss-Jordan algorithm, the concept of the row pivoting strategy, and implementation aspects of our inversion routine. In the remaining two sections we give an error analysis and an overview of numerical experiments on a Cyber 205 vector computer.

2. BASIC FORM OF THE GAUSS-JORDAN ALGORITHM

Let A be a given matrix of order n. The Gauss-Jordan algorithm consists of n consecutive transformation steps reducing the matrix to a diagonal matrix. We first describe a basic form of the algorithm (without pivoting) for which the resulting diagonal matrix is the identity matrix. Putting $A^{(1)} = A$, the k-th transformation step, $k = 1, \dots, n$, is given by:

$$\begin{array}{lll} \delta_k & := & A^{(k)}{}_{kk} \\ D_k & := & I + (\delta_k - 1)\, e_k e_k^T \qquad \{ = \text{diag}(1,\dots,1,\delta_k,1,\dots,1) \} \\ g_k & := & \delta_k e_k - A^{(k)} e_k \\ G_k & := & g_k e_k^T \\ A^{(k+1)} & := & (I + G_k) D_k^{-1} A^{(k)} \end{array} \tag{2.1}$$

The effect of this k-th step is that the k-th column of the matrix is transformed into e_k which then remains invariant during the subsequent transformation steps, because for all $j \neq k$ we have $(I + G_k)D_k^{-1}e_j = e_j$. So after n transformation steps we obtain:

$$(I + G_n)D_n^{-1} \cdots (I + G_1)D_1^{-1} A = I, \tag{2.2}$$

so that:

$$A^{-1} = (I + G_n)D_n^{-1} \cdots (I + G_1)D_1^{-1}. \tag{2.3}$$

Consequently, the inverse matrix, A^{-1}, is obtained by applying the same transformation steps, starting from $Z^{(0)} = I$, as follows:

$$Z^{(k+1)} := (I + G_k)D_k^{-1} Z^{(k)}, \quad k = 1, \dots, n, \tag{2.4}$$

which yields $Z^{(n+1)} = A^{-1}$.

3. ROW PIVOTING AND COLUMN INTERCHANGES

For numerical stability it is necessary that the pivot element, δ_k, of the k-th transformation step is not too small in magnitude. This is usually achieved by selecting an element of largest magnitude in the sub-triangular part of the k-th <u>column</u> of $A^{(k)}$ and correspondingly interchanging the pivotal row and the k-th row of the matrix. This is called (partial) <u>column pivoting</u>. Similarly one can perform <u>row pivoting</u>, i.e. select an element in the upper-triangular part of the k-th <u>row</u> of $A^{(k)}$ and correspondingly interchange the pivotal column and the k-th column of the matrix.

For Gaussian elimination these pivot strategies both yield essentially the same numerical stability. For the Gauss-Jordan algorithm, however, row pivoting is much more satisfactory then column pivoting, as we have shown in [1].

The application of row pivoting requires an adaption of the transformation rules preceding formulae 2.1. such that for the calculation of G_k in the k-th step, the selected pivotal column must replace the k-th column. Consider P_k as the permutation matrix that describes the proper column interchanges in the k-th step, then formula 2.1 should be changed into:

$$A^{(k+1)} := (I + G_k)D_k^{-1} A^{(k)} P_k \qquad (3.1)$$

and 2.3 should be changed accordingly into:

$$A^{-1} = P\,(I + G_n)D_n^{-1} \cdots (I + G_1)D_1^{-1} \quad , \qquad (3.2)$$

where P stands for the product $P_1 \cdots P_n$.

4. IMPLEMENTATION ASPECTS

The factorization 3.2 can serve as a starting-point for the explicit calculation of A^{-1}. For the k-th column of $(AP)^{-1}$ we observe

$$(AP)^{-1}e_k = (P^{-1}A^{-1})\, e_k = (I + G_n)D_n^{-1} \cdots (I + G_k)D_k^{-1}e_k \,. \qquad (4.1)$$

As a consequence, the inverse can be calculated in situ; i.e. the columns of A^{-1} are calculated and delivered in the same location where the original matrix A was stored. In the k-th step the k-th column of matrix $A^{(k)}P_k$ is replaced by e_k and the transformation step is carried out on all columns of the matrix. In this way $Z = P^{-1}A^{-1}$ will overwrite A. Finally, to produce A^{-1}, matrix Z is successively premultiplied by P_k, for $k = n, n-1, \dots ,1$.

If the routine is to be implemented on a vector machine, as we did, it is advantageous to avoid the alteration between row operations and column operations. Premultiplication with D_k^{-1}, as in formula (3.1), gives rise to row operations. However, the algorithm admits this premultiplication to be postponed till after the last transformation step; this is a consequence of the next observation:

$$(I + G_k)D_k^{-1} = (D_k^{-1} + \delta_k^{-1}G_k) = D_k^{-1}(I + \delta_k^{-1}G_k) \,. \qquad (4.2)$$

Our implementation will be such that next to the row pivoting operations and column interchanges in the k-th step, the premultiplication with $(I + \delta_k^{-1}G_k)$ is performed only. The accumulated row scaling operations, i.e. the premultiplications by $D_n^{-1} \cdots D_1^{-1}$, can be applied in the form of column operations at the end of the algorithm. Summarising, the implementation of our Gauss-Jordan inversion routine is described in the following piece of informal programming code; the notation is introduced in [8] and clarifies the use of storage.

```
perm  ←  (1, ... , n) ;
For  k = 1, ... ,n
        Determine p = $p_k$ with $k \le p \le n$ and $|A_{kp}| = \max_{k \le j \le n} |A_{kj}|$
        $A_{\bullet p}$    ↔    $A_{\bullet k}$              {i.e. interchange columns nr k and p}
        $perm_p$  ↔    $perm_k$
        $d_k$     ←    $1 / A_{kk}$
        $A_{\bullet k}$    ←    $g_k \; = \; e_k - d_k A_{\bullet k}$
        For    j = 1, ... ,k-1, k+1, ... ,n
               $A_{\bullet j} \leftarrow A_{\bullet j} + A_{kj}\, g_k$
        $A_{kk}$    ←    1
{form P D $A_{\bullet k}$ as follows}
For  k = 1, ... ,n
        $A_{\bullet k}$    ←    diag(d) $A_{\bullet k}$
        permute elements of $A_{\bullet k}$ according to perm
```

5. ERROR ANALYSIS

For the error analysis of the Gauss-Jordan algorithm we use the factorization of the inverse matrix as given in formula (3.2). With respect to its effect on the elements of the lower-triangular part of the matrix, the Gauss-Jordan algorithm is identical with Gaussian elimination. Therefore we know that in the sub-triangular part of the vectors g_k the columns of a unit lower triangular matrix L are generated for which $U := L^{-1}AP$ has unit upper triangular form. From inspecting the Gauss-Jordan algorithm we observe that in the upper-triangular part of the vectors g_k the columns of matrix U^{-1} are generated. The calculation of the k-th column of matrix $(AP)^{-1}$ appears to be numerically equivalent with solving $Ly_k = e_k$ for y_k, followed by the explicit matrix-vector multiplication $U^{-1}y_k$ to yield $(AP)^{-1}e_k$. The solution stage for solving y_k and the multiplication stage for calculating $U^{-1}y_k$ are intertwined in the Gauss-Jordan algorithm.

For a complete error analysis we refer the reader to [1]; the main results are repeated here.

Matrices L and U and vector y_k, as introduced above, exactly satisfy:

$$LU = AP + E_1,$$

and $(L + E^{(k)}_2)\, y_k = e_k$,

with $||E_1|| \leq \phi_1(n)\, g\, ||A||\, \mu$ and $||E^{(k)}_2|| \leq \phi_2(n)\, ||L||\, \mu$,

where g is the growth factor and ϕ_1 and ϕ_2 are low degree polynomials, (See e.g. [3],[8]).

The row pivoting strategy ensures that the elements of the (implicitly calculated) unit upper-triangular matrix U are bounded by one. As a consequence, it can be shown (see [1] for details) that matrix V, the calculated version of U^{-1}, satisfies

$$VU + E_3 = I,$$

with $||E_3|| \leq \phi_3(n)\, ||V||\, \mu$, for a low degree polynomial ϕ_3.

The columns z_k of Z, the calculated version of $(AP)^{-1}$, satisfy

$$(V + E^{(k)}_4)\, y_k = z_k,$$

with $||E^{(k)}_4|| \leq \phi_4(n)\, ||V||\, \mu$, for a low degree polynomial ϕ_4.

With the use of w_k for $(I - E_3 + E^{(k)}_4 U)^{-1} z_k$ we have $y_k = U w_k$ so that the following relation holds:

$$e_k = (AP + E_1 + E^{(k)}_2 U)\, w_k\ ;\ k = 1, \ldots, n. \qquad (5.1)$$

If we use $E^{(k)}_5$ for $(E_3 - E^{(k)}_4 U)$ then the difference between w_k and z_k satisfies

$$||z_k - w_k|| \,/\, ||z_k|| \leq ||E^{(k)}_5|| \,/\, (1 - ||E^{(k)}_5||), \qquad (5.2)$$

provided that $|| E^{(k)}_5|| < 1$.

For $||E^{(k)}_5||$ we can derive the following bound

$$|| E^{(k)}_5|| \leq \phi_5(n)\, ||U^{-1}||\, \mu \,/\, \{1 - \phi_3(n)\, ||U^{-1}||\, \mu\}, \qquad (5.3)$$

for a low-degree polynomial ϕ_5, provided that the denomenator is positive.

Summarizing, the k-th column of the calculated inverse Z is close to the solution w_k of a linear system with right-hand side vector e_k and a coëfficient matrix that is close to A as specified in (5.1).

For the k-th column of the residual matrix $R = I - APZ$ we have according to these formulae

$$r_k = e_k - AP\, z_k$$

which yields

$$r_k = (AP + E_1 + E^{(k)}_2 U)\, w_k - AP(I - E^{(k)}_5) w_k. \qquad (5.4)$$

This gives the following bound for the columns of the residual matrix:

$$||r_k|| \leq (||E_1|| + ||E^{(k)}_2 U|| + ||A||\, ||E^{(k)}_5||)\, ||z_k|| \,/\, (1 - ||E^{(k)}_5||). \qquad (5.5)$$

In this bound the contribution $||A||\ ||E^{(k)}{}_5|| / (1 - ||E^{(k)}{}_5||)$ creates the essential difference with the formula for the residual bound for Gaussian elimination.
As long as $||E^{(k)}{}_5|| << 1$, this term has order of magnitude $||A||\ ||U^{-1}||$. As a consequence of our pivoting strategy, U will mostly be well-conditioned, even when A itself is ill-conditioned, so that the contribution of this term is nearly always harmless.

6. NUMERICAL EXPERIMENTS

Experiments on accuracy and timing were carried out on the Cyber 205 computer of the Academic computer centre SARA in Amsterdam. This machine has one vector pipe, resulting in a peak performance of 50 Mflops for general vector operations and 100 Mflops for so called linked triads, vector constructions of the form $x := y + \alpha z$. The arithmetic precision of this machine is about 10^{-14}.
The programs were written in FORTRAN200, which is the Cyber205 version of FORTRAN 77 with extensions for explicite use of vectorization features.
We compared two matrix inversion algorithms and calculated for a large number of matrices the norm of the residual matrix I - AX, where X stands for the calculated inverse in either of the two methods. The first method is with the use of INVGJ, our FORTRAN200 implementation of inversion by the Gauss-Jordan algorithm with row pivoting, to be published in [7]. The second method is with the use of LINPACK routines SGEFA and SGEDI [2], where Gaussian elimination with forward and backward substitution is performed. The size of the residual matrix is the only measurement we took. In all cases tested, a norm of this residual matrix when using INVGJ was of the same size or smaller then this norm when using LINPACK's routines. We used the following types of matrices.
Case a. The matrices are constructed from a given diagonal matrix (the singular values chosen) which is pre- and post- multiplied by random orthogonal matrices. These left and right orthogonal factors are the product of $\sqrt{n}$ random Householder reflections. The singular values are chosen in various ways; the largest between 1 and 10^{+5}, the smallest between 10^{-5} and 1 and the remaining ones either distributed equally, or clustered on one end of the spectrum, or on the other end. We used hundreds of matrices of order $n = 25$ or $n = 50$ with condition number varying between 1 and 10^{+10}.

Case b. The matrices have upper triangular form. We tested several random matrices of this form of various order; in particular an upper triangular matrix of order 50 where the diagonal elements have the value +1 exept A_{33} and A_{44} which have the value 10^{-5} and the elements in the strictly upper triangular part have random values between -1 and +1. This type of matrices is used by Peters and Wilkinson [6] to show that the Gauss-Jordan algorithm with column pivoting in stead of row pivoting as we do, can produce larger residual vectors than Gaussian elimination.

A special upper triangular matrix in our test is an upper triangular matrix with ones on the diagonal and having all elements in the strictly upper triangular part equal to -1. These matrices have increasing bad condition for growing values of n (the condition number being of the order 2^n). We used order n=25 and n=50.

Case c. Matrices W and W^T for which maximal growth is obtained during Gaussian elimination with partial pivoting. Matrix W is given by

$$W_{ij} = -1 \text{ for } j > i \text{ ; } W_{jj} = W_{nj} = 1 \text{ for all } j \text{ and } W_{ij} = 0 \text{ elsewhere.}$$

$$W = \begin{pmatrix} 1 & -1 & \cdots & \cdots & -1 \\ 0 & 1 & & & -1 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & -1 \\ 1 & \cdots & \cdots & 1 & 1 \end{pmatrix}$$

We used matrices W and W^T for n = 50; the conditionnumber of W roughly equals 1700 and the element growth is 2^{49}.

As was pointed out above in a different formulation: no matrix was found for which the calculated inverse via the Gauss-Jordan algorithm with row pivoting has a larger residual than the inverse matrix calculated via Gaussian elimination.

With respect to the CP-time used, we carried out a comparison between four routines.

1) Routine INVGJ; our Cyber205 implementation of the Gauss-Jordan algorithm using row pivoting.
2) A Cyber205 implementation of the Gauss-Jordan algorithm with column pivoting written by Johnson [4].
3) The LINPACK routines SGEFA and SGEDI [2], for the LU decomposition of the matrix and for the forward and backward substitution, respectively.

4) The NAG routine F01AAF [5], which implements LU decomposition followed by forward and backward substitution. This routine uses an extra two-dimensional array; it does not calculate the inverse in situ.

CPU-TIME AND MFLOPS ON THE CYBER205

n =	25		50		100		200	
	cpu	**mflops**	cpu	**mflops**	cpu	**mflops**	cpu	**mflops**
INVGJ (NUMVEC)	0.0023	**13.6**	0.0098	**25.5**	0.0486	**41.1**	0.2722	**58.8**
JOHNSON	0.0028	**11.2**	0.0122	**20.5**	0.0586	**34.1**	0.3126	**51.2**
LINPACK	0.0061	**5.1**	0.0253	**9.9**	0.1193	**16.8**	0.5049	**31.7**
NAG	0.0082	**3.8**	0.0321	**7.8**	0.1371	**14.6**	0.6467	**24.7**

REFERENCES

[1] T.J. DEKKER and W. HOFFMANN, Rehabilitation of the Gauss-Jordan algorithm; Report 86-28, Dept. of Mathematics, University of Amsterdam, 1986.

[2] J.J. DONGARRA, C.B. MOLER, J.R. BUNCH and G.W. STEWART, LINPACK User's guide; SIAM, Philadelphia 1979.

[3] G. H. GOLUB and C.F. VAN LOAN, Matrix Computations; North Oxford Academic, Oxford 1983.

[4] Ch. H. J. JOHNSON, Matrix Arithmetic on the Cyber 205; Supercomputer 8/9, July, September 1985, pp. 28-48.

[5] NAG Library Manual; Numerical Algorithms Group, Oxford 1982.

[6] G. PETERS and J.H. WILKINSON, On the Stability of Gauss-Jordan Elimination with Pivoting; Communications of the ACM 18, 1975, pp. 20-24 .

[7] RIELE, H.J.J. te (ed.) , NUMVEC, a library of numerical software for Vector and parallel computers in FORTRAN; Centre for Mathematics and Computer Science, Amsterdam.

[8] G.W. STEWART, Introduction to matrix computations; Academic Press 1973.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

An Introduction to the Stabilized Galerkin Method

P.P.N. de Groen
Dept. of Mathematics & Comp. Sc., Vrije Universiteit
Pleinlaan 2, Brussels, Belgium
and
M. van Veldhuizen
Dept. of Mathematics & Comp. Sc., Vrije Universiteit
P.O. Box 7161, 1007 MC Amsterdam, The Netherlands

1. Preliminaries

It is well known that the standard finite element method for convection-diffusion problems with boundary layers or internal layers gives rise to spurious oscillations and bad results. This holds true also for central finite difference schemes.

A remedy against these oscillations is an upwind type discretization of the convective term. The disadvantage of this technique is twofold. First, the order of accuracy drops to one. Second, the artificial diffusion introduced by the upwinding smoothes the solution in all directions. In particular, this artificial diffusion gives too much smoothing in the direction perpendicular to the streamlines.

A natural improvement is the restriction of the upwinding to directions parallel to the streamlines. This so-called streamline upwinding gives better results, but the order of accuracy is still low. The accuracy in a streamline upwind method can be enhanced by changing it to a streamline upwind/Petrov-Galerkin method. This is a non-conforming finite element method, which gives a streamline upwinding in the discretization of the operator, plus a consistent treatment of the inhomogeneous term. Such a method is described in Brooks and Hughes[1]. This paper also contains an excellent introductory survey of the streamline upwind/Petrov-Galerkin method.

The aim of this paper is the description of another discretization method for convection-diffusion problems. In Section 2 we describe the application of the method in a special problem. We shall indicate why the method works well. A more extensive treatment of the method is given in De Groen and Van Veldhuizen [4]. In the new method the finite element test space is extended by a suitable set of test functions. This results in an overdetermined system, which is solved in a least squares sense. For piecewise linear elements on a rectangular domain the normal equations have a special form. These equations consist of a term corresponding to the standard finite element discretization plus a term generated by the extension. The latter term is reminiscent to streamline upwind methods. In addition, due to the least squares technique, the new method treats an inhomogeneous right-hand side in a consistent manner. For a discussion of the appropriate norms used in the least squares technique we refer to [4].

In Section 3 we give some numerical examples. The results are similar to the results obtained by the streamline upwind/Petrov-Galerkin method described by Brooks and Hughes [1].

2. Bilinear Elements on Rectangles

On the unit square $\Omega := [0,1] \times [0,1]$ we consider the convection-diffusion equation ($\epsilon > 0$)

$$-\epsilon \Delta u + p\, u_x + q\, u_y = f \qquad (2.1)$$

The boundary Γ of Ω consists of two parts Γ_D and Γ_N, with homogeneous boundary conditions of Dirichlet and Neumann type respectively,

$$u(x,y) = 0, \qquad (x,y) \in \Gamma_D = \{x = 0\} \cup \{y = 0\} \qquad (2.2a)$$

$$\frac{\partial}{\partial n} u(x,y) = 0, \qquad (x,y) \in \Gamma_N = \{x = 1\} \cup \{y = 1\} \qquad (2.2b)$$

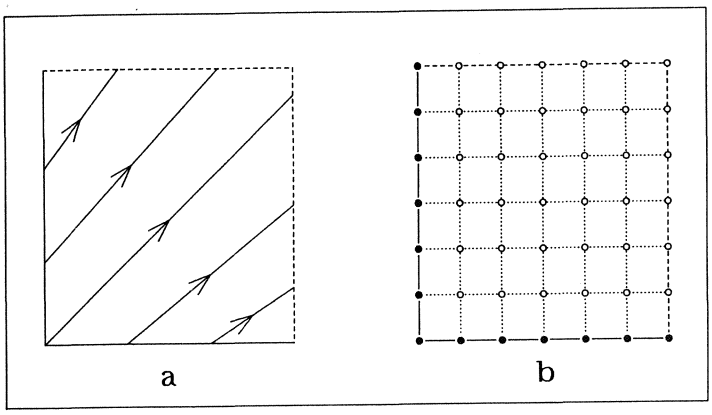


Fig.1. The domain Ω with the field velocity v (fig.1a), and the regular mesh with $n = m = 6$ (fig.1b). The black dots correspond to nodal values on the Dirichlet boundary, the open dots correspond to unknown nodal values.

The velocity vector $v = (p, q)$ may depend on the position, but is non-zero everywhere on the closure of Ω. Moreover, its direction is such that Γ_N is an outflow boundary, i.e. $p \geqq 0$ and $q \geqq 0$. The bilinear form associated to this problem is

$$a_\epsilon(u,v) := \epsilon[(u_x, v_x) + (u_y, v_y)] + (pu_x, v) + (qu_y, v) \tag{2.3}$$

where $u \in H^1(\Omega)$ must satisfy the essential boundary condition (2.2a). The test functions v belong to the subspace $H_E(\Omega)$ of functions in $H^1(\Omega)$ which vanish on Γ_D. The weak formulation of the boundary value problem is given by

$$a_\epsilon(u,v) = (f,v), \quad \forall\, v \in H_E(\Omega). \tag{2.4}$$

For the finite element discretization, cf. Ciarlet[2], we choose a rectangular mesh as shown in Fig.1b. It is obtained by taking the product of the subdivisions

$$\{x_0 := 0, x_1, \ldots, x_n := 1\}\ , \quad \{y_0 := 0, y_1, \ldots, y_m := 1\}$$

with mesh sizes $h_j := x_j - x_{j-1}$ and $k_j := y_j - y_{j-1}$ in $x-$ and $y-$direction respectively. The finite element space S_E^h is chosen as the space of bilinear functions on this mesh, which vanish on the essential boundary Γ_D. In S_E^h we choose the basis of nodal finite elements $\phi_{i,j}$, where $\phi_{i,j}(x_k, y_l) = \delta_{i,k}\delta_{j,l}$. The standard Finite Element discretization of (2.4) solves $u^h \in S_E^h$ from the set of equations

$$a_\epsilon(u^h, \phi_{i,j}) = (f, \phi_{i,j}), \qquad i = 1 \ldots n\,, j = 1 \ldots m \tag{2.5}$$

In matrix-vector notation these equations may be written as

$$A_1 \vec{u} = \vec{b}_1 \tag{2.6}$$

where $\vec{u}$ is the coordinate vector of u^h with respect to the basis $\{\phi_{i,j}\}$. The coordinate of $\vec{u}$ corresponding to $\phi_{i,j}$ will be denoted by $u_{i,j}$.

The stability of this type of discretization is a problem for $\epsilon < h\,\|\mathbf{v}\|/2$. In numerical experiments one observes fast oscillations which spoil the solution. It is the aim of this paper to describe a new remedy for this problem.

In our approach we improve the stability by augmenting the test space. In (2.5) the test space is identical to the space S_E^h in which the approximation is sought. Now we choose a test space T_E^h of greater dimension than S_E^h by augmenting the basis of S_E^h with the locally biparabolic functions given on the rectangle $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$ by

$$\psi_{i,j}(x,y) = \xi(1-\xi)\eta(1-\eta), \qquad \xi = \frac{x - x_{i-1}}{h_i}\,, \quad \eta = \frac{y - y_{j-1}}{k_j} \tag{2.7}$$

Outside this rectangle we put $\psi_{i,j}(x,y) = 0$. Clearly, the $\psi_{i,j}$ belong to $H_E(\Omega)$. By choosing the testfunctions in (2.5) in the larger space T_E^h we now obtain the overdetermined set of equations, $i = 1, .. , n\,, j = 1, .. , m$

$$a_\epsilon(u^h, \phi_{i,j}) = (f, \phi_{i,j}), \tag{2.8a}$$

$$a_\epsilon(u^h, \psi_{i,j}) = (f, \psi_{i,j}), \tag{2.8b}$$

In matrix vector notation these equations may be written as

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \vec{u} = \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \end{bmatrix} \tag{2.9}$$

where $\vec{u}, \vec{b}_1, \vec{b}_2 \in \mathbb{R}^{nm}$. The stabilizing equations corresponding to the extensions are given by (2.8b). The overdetermined equations (2.9) are solved in a least squares sense. A detailed discusion of possible inner product norms is given in [4].

In order to get an idea why this provides a stabilization we consider as an example the case $p = q = 1$ and $h_i = k_j = h$ for all relevant indices i, j. The left-hand side of one of the equations (2.8b) is easily computed resulting in

$$a_\epsilon(u^h, \psi_{i,j}) = \frac{h}{36}(u_{i,j} - u_{i-1,j-1}) \tag{2.10}$$

Clearly, this is a discretized differentiation in the direction of the vector field **v**. The diffusion term disappears in this formula due to the choice of the functions $\psi_{i,j}$ in relation to the nodal elements $\phi_{i,j}$. The right-hand side of an equation (2.8b) is given by

$$[\vec{b}_2]_{i,j} = (f, \psi_{i,j}) = \frac{h^2}{36} f_{i,j},$$

where $f_{i,j}$ is a mean value of f on $supp\ \psi_{i,j}$. Hence, the equations (2.8b) are discretizations of the reduced equation $pu_x + qu_y = f$. See Eckhaus[3] for the concept and the role of the reduced equation in this type of problems. Since $u^h \in S_E^h$ is prescribed at the inflow boundary only, the equations (2.8b) are solved exactly by the vector $\vec{w}$,

$$w_{i,j} = h \sum_{k=0}^{min(i,j)-1} f_{i-k,j-k}$$

Apart from the $O(\epsilon/h)$ diffusion terms, A_1 represents a discretization of the reduced equation $pu_x + qu_y = f$. Hence, apart from the diffusion term, w^h is also an (approximate) solution of (2.8a). Indeed, arguing as in the computation of a local discretization error we find for coordinates of $A_1\vec{w} - \vec{b}_1$ corresponding to nodes outside the boundary layers a small value of order $O(\epsilon h + h^3)$. In a boundary layer the value of a coordinate is much larger, of the order $O(\frac{\epsilon}{h} + h)$ or so. However, the boundary layer regions are very small. Hence, the euclidean norm $\| A_1\vec{w} - \vec{b}_1 \|$ is small for $h \rightarrow 0$ and small ϵ. As a consequence, the true least squares solution $\vec{u}$

of

$$\min_{\vec{x}} \| A_1\vec{x} - \vec{b}_1 \|^2 + \| A_2\vec{x} - \vec{b}_2 \|^2$$

must have a residual $\vec{r}$ of norm less than $\| \vec{r} \| \leqq \| A_1\vec{w} - \vec{b}_1 \|$, i.e. $\| \vec{r} \|$ is small. Hence the least squares solution $\vec{u}$ satisfies exactly

$$u_{i,j} - u_{i-1,j-1} = [\vec{b}_2]_{i,j} - [\vec{r}_2]_{i,j}$$

where each component of the residual $\vec{r}_2$ is small, since the euclidean norm of $\vec{r}_2$ is at most equal to the norm of the residual vector, and this norm is at most equal to the norm of $\vec{r}$, which is small. This implies that strong spurious oscillations in the direction of **v** are not possible with this type of discretization. Indeeed, large spurious oscillations in the direction of **v** correspond to a large residual vector $\vec{r}_2$, and we have shown that this vector is small.

The overdetermined system (2.9) has to be solved in a least squares sense. In the usual Euclidean vector norm the equations are not invariant under scaling. Rescaling of a basis function changes the weight of the corresponding node and thus the discretization method. In addition, the Euclidean vector norm of a coordinate vector is not compatible with the L^2-norm of the corresponding function in S_E^h, because the basis in S_E^h is not orthogonal. A more natural norm is obtained from the interpretation of (2.8ab) as a projection method in a suitable Hilbert space. This interpretation is given in [4]. Here we only mention the result. It turns out that a natural norm is the $L^2(\Omega) \times L^2(\Omega)$-norm. This norm induces a norm in the discrete spaces we are working with. For the overdetermined set of equations (2.9) this results in the normal equations

$$(A_1^T \; A_2^T) \begin{bmatrix} M_1^{-1} & 0 \\ 0 & M_2^{-1} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \vec{x} = (A_1^T \; A_2^T) \begin{bmatrix} M_1^{-1} & 0 \\ 0 & M_2^{-1} \end{bmatrix} \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \end{bmatrix} .$$

The matrix

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$$

is the matrix corresponding to the finite element discretisation of the Laplace operator $-\Delta$ in the space T_E^h. By the orthogonality this matrix splits up in the diagonal blocks M_1 and M_2 corresponding to the space S_E^h and the extensions (2.7) respectively. Since the supports of the basis functions in the extension $T_E^h - S_E^h$ do not overlap, M_2 is a diagonal matrix. However, the inverse M_1^{-1} of the discretized Laplacean is a full matrix. As a consequence, the normal equations form a large full system. This makes the method quite expensive.

A more practical variant is obtained if we use only the diagonal of M_1. In this way we replace the matrix M by a diagonal matrix. It can be shown that this choice of the inner product does not affect the accuracy of the method, and that it keeps the method scaling invariant. More details are given in [4].

The resulting normal equations can be solved by a preconditioned conjugate gradient algorithm. The matrix of the normal equations is not an M-matrix. I.e. incomplete decompositions do not necessarily exist. Nevertheless, we have been able to solve equations with as many as 22500 unknowns in a reasonable time.

3. An Example

We consider the problem (2.1) with $\epsilon = 10^{-6}$, and $\mathbf{v} = (p,q) = (cos\,\phi, \sin\phi)$. At the outflow boundary we take the homogeneous Neumann boundary conditions (2.2a), but at Γ_D we take the inhomogeneous boundary condition

$$u(x,y) = \begin{cases} 0 & (x,y) \in \Gamma_D \;\; y < \tfrac{1}{4} \\ 1 & (x,y) \in \Gamma_D \;\; y > \tfrac{1}{4} \end{cases}$$

The exact solution is not known explicitly. We compare the results of our Stabilized Galerkin method (SG-method) and the streamline upwind/Petrov-Galerkin method of Brooks and Hughes[1] (BH-method) with the asymptotic solution (indicated by A)

$$u_{asymp}(x,y) = H(tg(\phi)x - y - \tfrac{1}{4}),$$

where H is Heaviside's function

$$H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The results for the angles $\phi = \tfrac{1}{8}\pi,\ \tfrac{3}{4}\pi,\ \tfrac{3}{8}\pi$ are displayed graphically in Fig.2. We see that both methods show a small overshooting at the border of the jump discontinuity.

References

[1] Brooks, A.N., Hughes, Th.J.R.: Streamline Upwind/Petrov-Galerkin Formulations for Convection Dominated Flows with Particular Emphasis on the Incompressible Navier-Stokes Equations. Comp. Meth. Appl. Mech. Engrg. 32, 199-259 (1982).

[2] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems.* North-Holland (1979).

[3] Eckhaus, W.: *Matched Asymptotic Expansions and Singular Perturbations.* Mathematics Studies 6, North-Holland 1973.

[4] Groen, P.P.N. de, Veldhuizen, M.van: A Stabilized Galerkin Method for Convection-Diffusion Problems. Tech.Report VU 320, 1986.

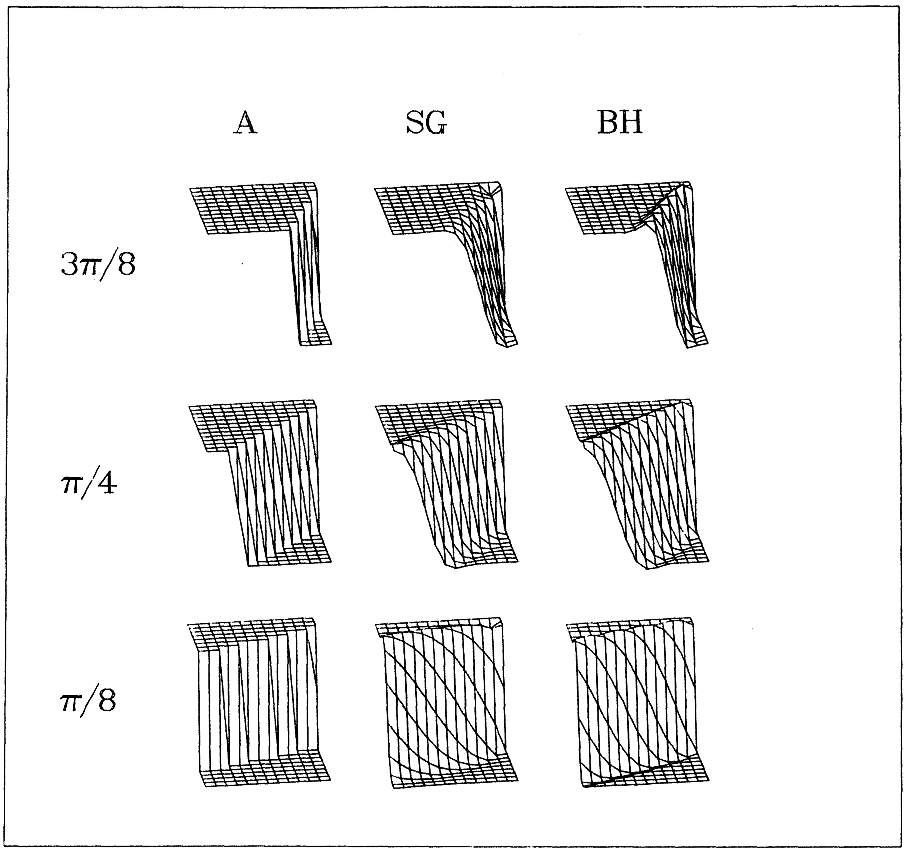


Fig.2. Results for the problem (2.1) as specified in section3. The mesh is regular with $n = m = 12$. The viewpoint is at infinity in the direction of the vector (3,20,8) in (x,y,z)-space.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Piecewise C^1-Approximation, with Application to Water/Steam Thermodynamic Functions

C.R. Traas and R.H.J. Gmelig Meyling
Faculty of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

ABSTRACT

The piecewise approximation technique is very suitable for application to approximation problems in which irregular stuctures occur.
The domain of a considered problem is triangulated. On each triangle a cubic polynomial is defined such, that polynomials on neighbouring triangles match in C^1 sense. An efficient representation for these so called bivariate cubic C^1 splines is presented. Arbitrary curved lines in the domain, over which discontinuities exist, can be taken into account. Given a set of data, a solution to the spline coefficients is found such that the spline is a suitable approximation to the data.
As an application the approximation of the entropy of a water/steam mix as a function of temperature and pressure is presented. The saturation line is a line of discontinuity.

1. INTRODUCTION

The flexibility of the piecewise approximation technique renders it very suitable for application to approximation problems in which irregular structures occur.
Let a set of data be given over a domain in $\mathbb{R}^2$. After having adopted an

efficient representation for the bivariate cubic C^1 spline over a triangulation of this domain, the spline coefficients are solved such that a suitable approximation to the data is obtained. The numerical linear algebra in this process is arranged such that the sparseness of the matrices under consideration is fully exploited. As an application the approximation of the entropy of a water/steam mix as a function of temperature and pressure is presented. In this function a line of discontinuity occurs: the saturation line. This application was taken from the practice of an industry involved with engines and tools driven by steam. The usual method up to now for storing the properties of water/steam was by partitioning the pressure-temperature region into (effectively) four subregions, and to construct for each subregion a very complicated analytical expression. Small discontinuities over the boundaries of these regions could then not be avoided. Our splines method is not only much more uniform and elegant, but is also expected to save considerable computer time. In addition, the thermodynamic functions represented by these C^1-splines can be differentiated, giving unique first derivatives. This is of importance for deriving other functions, e.g. specific heat, from the approximants.

2. EFFICIENT REPRESENTATION FOR BIVARIATE CUBIC C^1-SPLINES.

Let a two-dimensional domain Ω in x,y-space be triangulated. We assume that any two triangles from this triangulation share a common edge, or a common vertex or are disjunct, and the union of all triangles covers the domain (i.e. the domain is "properly" triangulated).

We introduce barycentric coordinates in each of the triangles. Let a point P(x,y) be given in a triangle and let the vertices of this triangle be labeled 0, 1 and 2, respectively. Then the barycentric coordinates λ_0, λ_1 and λ_2 of the point P are defined by the relations

$$x = \lambda_0 x_0 + \lambda_1 x_1 + \lambda_2 x_2$$

$$y = \lambda_0 y_0 + \lambda_1 y_1 + \lambda_2 y_2$$

$$1 = \lambda_0 + \lambda_1 + \lambda_2$$

where x_i, y_i are the coordinates of the vertex i, i = 0,1,2.
We define a cubic polynomial in λ_0,λ_1 and λ_2, and thus in x and y, over this triangle:

$$\begin{aligned} p(x,y) = \lambda_0^2[p_0\cdot(3-2\lambda_0) &+ \lambda_1\{g_0\cdot(x_1 - x_0) + h_0\cdot(y_1 - y_0)\} + \\ &\lambda_2\{g_0\cdot(x_2 - x_0) + h_0\cdot(y_2 - y_0)\}] + \\ \lambda_1^2[p_1\cdot(3-2\lambda_1) &+ \lambda_2\{g_1\cdot(x_2 - x_1) + h_1\cdot(y_2 - y_1)\} + \\ &\lambda_0\{g_1\cdot(x_0 - x_1) + h_1\cdot(y_0 - y_1)\}] + \\ \lambda_2^2[p_2\cdot(3-2\lambda_2) &+ \lambda_0\cdot\{g_2\cdot(x_0 - x_2) + h_2\cdot(y_0 - y_2)\} + \\ &\lambda_1\cdot\{g_2\cdot(x_1 - x_2) + h_2\cdot(y_1 - y_2)\}] + \\ 6b_{012}\,\lambda_0\lambda_1\lambda_2 \, , & \end{aligned} \tag{1}$$

with the ten parameters p_i, g_i, $h_i(i = 0,1,2)$ and b_{012} (Gmelig Meyling, 1986). The parameters p_i, g_i and h_i represent

$$p_i = p(x_i,y_i),$$

$$g_i = \left.\frac{\partial p}{\partial x}\right|_{x_i,y_i} \quad , \quad h_i = \left.\frac{\partial p}{\partial y}\right|_{x_i,y_i} .$$

The parameter b_{012} is the so-called Bézier ordinate, associated with the center of gravity of the triangle with the vertices 0, 1 and 2.

Considering an adjacent triangle 1,2,3, sharing a commom edge $\overline{1,2}$ with the above triangle, we require that the cubic polynomials over both triangles match in C^1-sense. For this reason the following condition is imposed to hold over the edge $\overline{1,2}$:

$$\begin{aligned} &\tau_1\cdot[3p_1 + g_1\cdot(x_2 - x_1) + h_1\cdot(y_2 - y_1)] \\ &- \tau_2\cdot[3p_2 + g_2\cdot(x_1 - x_2) + h_2(y_1 - y_2)] \\ &+ 3(\tau_3 b_{123} - \tau_0 b_{012}) = 0 \end{aligned} \tag{2}$$

with

$$\tau_0 = \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} , \quad \tau_1 = \begin{vmatrix} 1 & 1 & 1 \\ x_2 & x_3 & x_0 \\ y_2 & y_3 & y_0 \end{vmatrix} ,$$

$$\tau_2 = \begin{vmatrix} 1 & 1 & 1 \\ x_3 & x_0 & x_1 \\ y_3 & y_0 & y_1 \end{vmatrix} , \quad \tau_3 = \begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \end{vmatrix} .$$

For an arbitrary (proper) triangulation of the domain the above representation is optimal in the sense that the total number n_c of smoothness constraints and the total number n_p of parameters are minimal, except for very regular triangulations in rectangular regions where even better representations are possible. Another feature of the representation (1) is the fact that it already incorporates first order differentiability at the vertices, due to the common use of g_i and h_i for all triangles sharing the vertex i. Only a single condition (2) over each internal edge then suffices to actually obtain overall smoothness of class C^1.

The total number of smoothness constraints is E_0, the number of interior edges in the triangulation. Possibly, part of this set of constraints is redundant; this depends on the precise geometry of the triangulation. The total number of parameters is 3V + T, where V is the total number of vertices, and T the number of triangles in the triangulation. Comparing with a few other representations we find (Gmelig Meyling, 1986):

1. Euclidean coordinates: $n_c = 7E_0 - 3V_0$, $n_p = 10T$
2. Bézier-Bernstein : $n_c = 3E_0 - 2V_0$, $n_p = V + 2E + T$,

where V_0 is the number of internal vertices, and E is the total number of edges.

When dealing with triangulations which are not very small, i.e. consisting of several dozens of triangles at least, the present representation thus leads to considerable savings.

In the case of a (possibly curved) line passing through the domain Ω, across which a discontinuity exists, the condition (2) is locally abandoned and in each vertex on this line two sets of parameters p, g and h are introduced, each set valid at one side of the line.

3. APPROXIMATION OF A SET OF DATA

Let be given a set of data f_i, $i = 1,2,\cdots,n_d$, scattered over the domain Ω. The aim is to approximate these data in a C^1-continuous way (apart from a line of discontinuity passing through Ω) using the bivariate cubic splines described in section 2.

Given an arbitrary data point, the first step is to determine the triangle containing this point, and next to use (1) in order to fill one of the rows of the system matrix. After having treated all data points in this way, a large and sparse linear algebraic system

$$Av = f \tag{3}$$

has been obtained, where v is the vector of the n_p parameters, and f the vector of the n_d data, $n_d > n_p$. Solving (3) in the sense of least squares gives an approximation to the data of class C^0 (with unique first derivatives in the vertices). Let this solution to (3) be written as v_0.

Let the sytem with smoothness constraints (based upon (2)) be written as

$$Bv = 0 \tag{4}$$

where B is a $(E_0 \times n_p)$ matrix (or, in fact, with less rows due to the discontinuity line in Ω), $E_0 < n_p$.

In general it will be true that $Bv_0 = r \neq 0$, where r is a vector of residuals.

We will construct a correction e such, that $B\cdot(v_0 + e) = 0$, i.e. we solve e from

$$Be = - r. \tag{5}$$

In particular, we solve that vector e which is of minimal Euclidean length. This gives the smallest possible correction (in the sense of the Euclidean norm) to be added to v_0 in order to obtain an approximation of class C^1 to the data.

The minimum norm vector e can be found in the range of B^T, the transpose of B. Hence, writing $e = B^Ty$, where y is some vector in the space $\mathbb{R}^{E_0}$, y must be solved from

$$BB^Ty = - r \qquad (6)$$

in which BB^T is symmetric and positive semi-definite. It may occur that BB^T is singular. In that case an arbitrary solution to (6) will do. Finally e is computed: $e = B^Ty$ (which is unique, even in the case that BB^T is singular), and the corrected vector is computed: $v = v_0 + e$. Strictly speaking, the obtained solution v is not, in general, the least squares solution to (3) constrained with (4). The latter solution could be obtained by computing the singular value decomposition of the matrix B, and eliminating a part of the elements of v from (3). This method, however, does not take advantage of the sparsity of A and B, and is therefore prohibitive with respect to computer memory in the case of extended triangulations.

To exploit the sparsity of the matrices a method similar to the conjugate gradient method has been applied to solve (3) and (6). No matrix-matrix multiplications occur in these algorithms, so the sparsity is fully preserved (Paige, Saunders, 1982).

4. APPLICATION TO WATER/STEAM THERMODYNAMIC FUNCTIONS

The entropy of a water/steam mix is considered, as a function of temperature and pressure. In the current practice in industry the relevant temperature-pressure region is partitioned into four subregions (fig. 1), (Schmidt,Grigull, 1982).

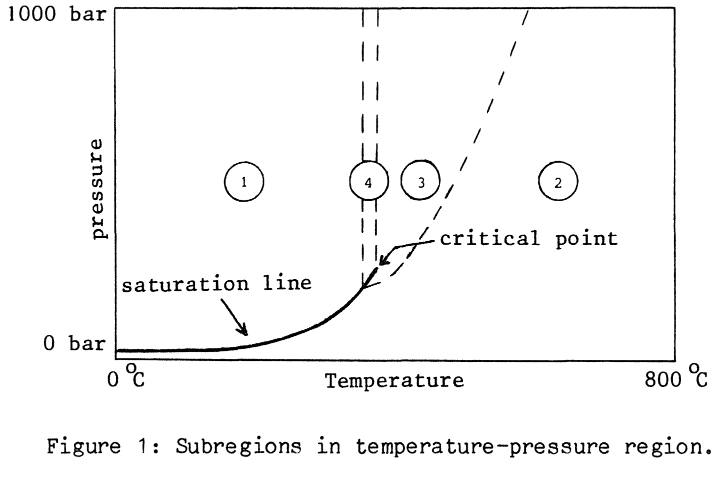


Figure 1: Subregions in temperature-pressure region.

In each subregion an analytical expression is given for approximating the canonical thermodynamic functions, from which other functions are derived (e.g. entropy as the temperature derivative of the Gibbs function).
These analytical expressions are very complicated, as can be expected since they approximate the relevant function over a large subregion, and they need to match the expression in the adjacent subregion. From these analytical expressions extensive tables with numerical values are derived (Schmidt, Grigull, 1982).
In discussions with engineers from industry the need for an alternative approach to the storage of thermodynamic data became relevant. This inspired us to adopt the splines technique, as described in the sections 2 and 3, for this purpose. As a first test case, which should serve as a "proof of principle" for the industry, we have approximated the reduced entropy in a region around the critical point (fig. 2)

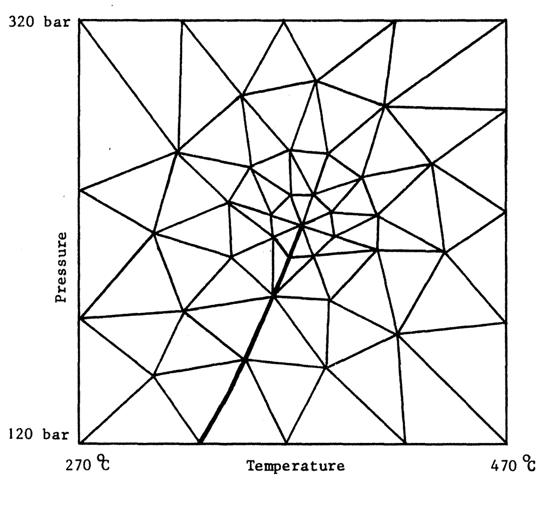


Figure 2: Triangulation of a region around the critical point.

The number of triangles in the considered region is 75, the number of vertices is 46 and the number of internal edges is 105 (101 of which produce a smoothness constraint). The geometry of the saturation line is approximated as a univariate cubic spline of class C^1.
The data are taken from some of the tables in (Schmidt, Grigull, 1982). The total number of data used in the present test case is $n_d = 1220$, and

the total number of parameters in the spline respresentation is $n_p = 225$. The solution of the system (3), subject to (4), along the lines described in section 3 gives a C^1-class approximation depicted in figure 3. The discontinuity over the saturation line is clearly visible (although the plotting process smoothes the discontinuity a little). Around the critical point a tiny wavy structure is visible. We interpret this structure as a consequence of a shortage of data points in the immediate vicinity of the critical point. In this region very steep gradients occur and the triangulation therefore has been taken locally relatively fine. We have not evaluated the analytical formulas from (Schmidt, Grigull, 1982) to obtain data but, instead, we have used the tabulated values, which are given at fixed intervals. The density of these data is probably insufficient, locally. We expect that with additional data and, possibly, a locally still finer triangulation, this defect can be eliminated.

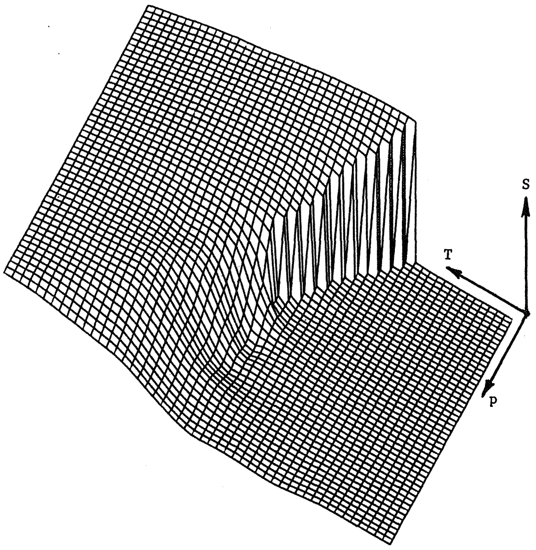


Figure 3: Approximation of the entropy S as a function of temperature and pressure.

5. ERRORS IN THE APPROXIMATION

The largest errors occur near the critical point. We have determined the maximum absolute difference between the data points and the corresponding points of the approximation, in a region around the critical point. The result is:

$$\underset{i}{\text{MAX}}\left|f_i - s_{0,i}\right| = 0.056$$

where f_i is the given value of the entropy in the i-th data point, and $s_{0,i}$ is the value of the entropy in this point, as reproduced by the unconstrained approximation to the data (i.e. corresponding with the least squares solution v_0 to (3)).

For the constrained approximation (i.e. corresponding with a solution v to (3), subject to (4), obtained by minimum norm correction to v_0) the result is:

$$\underset{i}{\text{MAX}}\left|f_i - s_i\right| = 0.245.$$

Both maximum errors occur in the point T = 370 °C, p = 220 bar, thus very near to the critical point (T_c = 374.15 °C, p_c = 221.20 bar). Since $\underset{i}{\text{MAX}}\left|f_i\right|$ = 6.383, in the considered region, the relative differences are 1% and 4%, respectively. Neglecting the above data point the maximum absolute differences are 0.031 and 0.115, respectively, in two different points, a little farther away from the critical point. These results illustrate the problematic nature of the critical point (and its vicinity), with respect to approximation.

REFERENCES

1. Gmelig Meyling, R.H.J., A respresentation for piecewise cubic C^1-splines on arbitrary triangulations.
   Memorandum 574, July 1986, University of Twente. To appear in Numerische Mathematik.

2. Paige, C.C., Saunders, M.A., LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Transactions on Mathematical Software, 8, 1 (1982), pp. 43-71.

3. Schmidt, E., Grigull, U., Properties of water and steam in SI-units. Springer-Verlag Berlin Heidelberg New York and R. Oldenbourg München, 1982.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

On Polygonal Approximations of an Invariant Curve

M. van Veldhuizen
Dept. of Mathematics & Computer Science
Vrije Universiteit
P.O. Box 7161, 1007 MC Amsterdam, The Netherlands

1. The Approximation of the Invariant Curve

In this paper we consider a map Φ from $\mathbb{R}^2$ to $\mathbb{R}^2$, and we want to compute an invariant curve of this map. I.e. we want to approximate by a numerical technique a curve $\gamma \in \mathbb{R}^2$ such that $\Phi\gamma \subset \gamma$. Such problems arise in the study of oscillatory motion of an ordinary differential equation like

$$\frac{dx}{dt} = g(x) + f(t) \tag{1.1}$$

where $x(t) \in \mathbb{R}^2$ and where f is a smooth periodic map with period p. For such systems of ordinary differential equations one often investigates the Poincaré map $\mathbf{P}$.

Under reasonable conditions the differential equation (1.1) with initial vector x_0 at $t = 0$ has a unique solution $x(t;x_0)$ on $[0,p]$. Hence, by assigning to x_0 at $t = 0$ the vector $x(p;x_0)$ at $t = p$ we define a map from (part of) $\mathbb{R}^2$ to $\mathbb{R}^2$, the Poincaré map $\mathbf{P}$ related to the differential equation (1.1) and the time-step p. The Poincaré map describes how $\mathbb{R}^2$ is transformed by the differential equation (1.1) in one period p. Often the map Φ is actually a Poincaré map. See [5] for an example.

There exist some algorithms for the numerical approximation of an invariant curve, cf. Thoulouze-Pratt and Jean[4], Chan[2], Kevrekidis et al.[3] and Van Veldhuizen[5]. In this paper we investigate an algorithm essentially due to Kevrekidis et al.[3].

Let the Jordan curve γ be an attracting invariant curve of the map Φ. The curve γ will be approximated by a polygon in $\mathbb{R}^2$. A polygon is completely determined by its vertices. The N vertices, vectors in $\mathbb{R}^2$, are denoted by $x_1, x_2, \cdots, x_N$, and the polygon $\mathbf{p}(\{x_i\}_{i=1}^N)$ is obtained as the set of line segments $[x_1,x_2]$, $[x_2,x_3]$, ..., $[x_{N-1},x_N]$ and $[x_N,x_1]$. By considering images of vertices, a polygon $\mathbf{p}(\{x_i\}_{i=1}^N)$ approximating the curve γ is mapped to another approximating polygon $\mathbf{p}(\{\Phi x_i\}_{i=1}^N)$. For an attracting curve γ we expect that $\mathbf{p}(\{\Phi x_i\}_{i=1}^N)$ is a better approximation than $\mathbf{p}(\{x_i\}_{i=1}^N)$. By means of piecewise linear interpolation we project $\mathbf{p}(\{x_i\}_{i=1}^N)$ onto $\mathbf{p}(\{\Phi x_i\}_{i=1}^N)$, thus completing one iteration step. To that end we assume that the invariant curve γ to be approximated can be parametrized as a circle

$$\gamma : \theta \in [0,2\pi) \rightarrow x_C + r(\theta)\,(\cos(\theta),\sin(\theta))^T \tag{1.2}$$

where $r(\theta) > 0$. Observe x_C must belong to the interior of γ and $\mathbf{p}(\{x_i\}_{i=1}^N)$. We need the following assumption.

Assumption (radial coordinates). In an annular neighborhood of the curve γ the nonlinear coordinate transformation

$$(\varrho,\theta) \rightarrow x_C + (r(\theta) + \varrho)\,(\cos(\theta),\sin(\theta))^T \tag{1.3}$$

is a smooth one-to-one map.

The composition of Φ and the interpolation will be denoted by K. The map K is defined as follows. Let x_i be a vertex in $\mathbf{p}(\{x_i\}_{i=1}^N)$. Determine the angle θ_i such that

$$x_i = x_C + \|x_i - x_C\|\,(\cos(\theta_i),\sin(\theta_i))^T$$

Assume that the point x_C is in the interior of the polygon $\mathbf{p}(\{\Phi x_i\}_{i=1}^N)$. The vertices of the polygon $\mathbf{p}(\{\Phi x_i\}_{i=1}^N)$ may then be written in this coordinate system as

$$\Phi x_j = x_C + \hat{r}_j\,(\cos(\hat{\theta}_j),\sin(\hat{\theta}_j))^T \tag{1.4}$$

Assume $\theta_i \in [\hat{\theta}_j,\hat{\theta}_{j+1}]$ taking into account the identification of 2π and 0. Then define K as the intersection of the half-line in the direction θ_i and the line segment $[\Phi x_j, \Phi x_{j+1}]$,

$$K x_i = \{x \mid x = x_C + r\,(\cos(\theta_i),\sin(\theta_i))^T, r > 0\} \cap [\Phi x_j, \Phi x_{j+1}] \tag{1.5}$$

It is easily seen that the set of equations $\mathbf{p}(\{x_i\}_{i=1}^N) = \mathbf{p}(\{Kx_i\}_{i=1}^N)$ is almost the discretization as described in Kevrekidis et al.[3]. Clearly, the map K is not necessarily differentiable. Therefore we restrict ourselves to stable invariant curves, and we solve the equation $\mathbf{p}(\{x_i\}_{i=1}^N) = \mathbf{p}(\{Kx_i\}_{i=1}^N)$ by iteration.

Let $d(x;\gamma)$ denote the euclidean distance of the point x to the smooth curve γ. By the attractivity of the invariant curve γ we mean the existence of a constant $0 \leqq \varkappa < 1$ such that $d(\Phi x;\gamma) \leqq \varkappa\, d(x;\gamma)$ for all x in a tubular neighborhood γ. One can now prove the following result, cf. [6].

Theorem 1. Let $\mathbf{p}(\{x_i\}_{i=1}^N)$ belong to a sufficiently small annular neighborhood of γ. Then, for $\varkappa$ sufficiently small, the sequence $\mathbf{p}(\{x_i\}_{i=1}^N)$, $\mathbf{p}(\{Kx_i\}_{i=1}^N)$, $\mathbf{p}(\{K^2x_i\}_{i=1}^N)$,... converges to a unique polygon $\mathbf{p}(\{\overline{x}_i\}_{i=1}^N)$. The discretization error satisfies the estimate

$$\max_{i=1,..N} d(\overline{x}_i;\gamma) = O(\max_{j=1,..,N} \|\Phi\overline{x}_j - \Phi\overline{x}_{j+1}\|^2)$$

In practice the constant $\varkappa$ can be made arbitrarily small by using the map Φ^p, for some integer $p > 0$, instead of Φ.

2. Inclusion of the Invariant Curve

In addition to the smoothness and attractivity of γ, let γ also be convex. We also assume that Φ maps the interior of γ in the interior, and the exterior in the exterior. This is a reasonable assumption for maps Φ which are the Poincaré map corresponding to an ordinary differential equation like (1.1). The assumption then follows from the unique solvability of an initial value problem. Now one may prove the following result, cf.[6].

Theorem 2. Let $\mathbf{p}(\{x_i\}_{i=1}^N)$ be a polygon approximating the convex invariant Jordan curve γ. Let every vertex of $\mathbf{p}(\{x_i\}_{i=1}^N)$ belong to a line segment $[\Phi x_j, \Phi x_k]$. Then $\mathbf{p}(\{x_i\}_{i=1}^N)$ is in the interior of γ or on γ.

The Theorem gives a one-sided error estimate for the polygon $\mathbf{p}(\{x_i\}_{i=1}^N)$. Observe that both the polygonal Kevrekidis algorithm of Section 1 and the method described in [5] satisfy the requirement about x_i belonging to a line segment between images of vertices.

Once we have a one-sided error estimate, we might obtain an inclusion. Suppose we can find a transformation (an invertible map) $\Psi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that the interior of γ is mapped by Ψ in the exterior of $\Psi\gamma$, and such that the exterior of γ is mapped in the interior of $\Psi\gamma$. Now approximate the invariant curve $\Psi\gamma$ of the map $\Phi \circ \Psi$. If Theorem 2 is applicable, the result is an approximation in the interior of $\Psi\gamma$. But then we also have an approximation in the exterior of γ, obtained as the image of the approximating polygon under the map Ψ^{-1}. A simple example of a map Ψ is given in [6].

3. Numerical Illustration

We consider the delayed logistic map ϕ defined by

$$\phi \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ \lambda y (1-x) \end{pmatrix} \tag{3.1}$$

This map has been studied in great detail in Aronson et al.[1]. Here we consider the simple case $\lambda = 2.02$. For this value of the parameter λ there is a convex invariant curve, and a map Ψ as required in Section 2 leads also to a convex invariant curve. For $\lambda > 2$ there is a source in the interior of the invariant curve given by (x_S, y_S) where

$$x_S = y_S \qquad x_S = \frac{\lambda - 1}{\lambda}$$

For the map $\Phi = \phi^p, p = 48$ we compute the invariant curve γ by means of the algorithm described in Section 1. In this case the center x_C is chosen as the source (x_S, y_S). We also compute an of the invariant curve of the map $\Phi \circ \Psi$, where the inverse of Ψ is given by

$$\Psi^{-1}(x,y) = \frac{(x - x_S, y - y_S)}{(x - x_S)^2 + (y - y_S)^2}$$

By means of the approximation in the interior and the exterior of γ we are able to obtain an upperbound for the error in each of the approximations. We measure the distance between the two approximations along the radial vectors centered in (x_S, y_S). The maximum distance $E_{incl}(N)$ is mentioned in the table below. The integer N denotes the number of vertices.

N	$E_{incl}(N)$	$N^2 E_{incl}(N)$
33	1.3E-02	14.2
65	5.0E-03	21.1
131	1.8E-03	30.9
244	4.4E-04	26.2
464	1.2E-04	25.8

For nicely distributed vertices the theory predicts a global error of the order N^{-2}. The error estimates support this behavior of the error.

References

[1] Aronson, D.G., Chory, M.A., Hall, G.R., McGeehee,R.P.: Bifurcations from an invariant circle for two-parameter families of maps of the plane. *Commun. Math. Phys.* **83** (1982), 303-354.

[2] Chan, Tze Ngon : Numerical Bifurcation Analysis of Simple Dynamical Systems. Thesis (unpublished), Dept. of Computer Science, Concordia University, Montreal.

[3] Kevrekidis, I.G., Aris, R., Schmidt, L.D., Pelikan, S.: Numerical Computations of Invariant Circles of Maps. *Physica 16 D* (1985), 243-251.

[4] Thoulouze-Pratt, E., Jean,M.: Analyse Numérique du Comportement d'une Solution Presque Périodique. *Int. J. Non-Linear Mechanics* **17** (1982), 319-326.

[5] Veldhuizen, M. van: A New Algorithm for the Numerical Approximation of an Invariant Curve. (1985) To appear in SIAM J. Sci. Stat. Comp.

[6] Veldhuizen, M. van: On Polygonal Approximations of an Invariant Curve, Report WS314, Dept. of Math. & Comp. Sc., Vrije Universiteit, Amsterdam (1986).

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

A Multigrid Method for Elliptic Equations with a Discontinuous Coefficient

P. Wesseling
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands

ABSTRACT

A multigrid method for elliptic equations with a discontinuous coefficient is proposed. This method is simpler than methods in current use, and is easier to justify theoretically.

1. INTRODUCTION

Consider

$$- \frac{\partial}{\partial x_1}\left(a \frac{\partial \phi}{\partial x_1}\right) - \frac{\partial \phi}{\partial x_2}\left(a \frac{\partial \phi}{\partial x_2}\right) = f, \quad (x_1,x_2) \in \Omega = (0,1) \times (0,1),$$

$$\phi|_{\partial\Omega} = 0. \tag{1.1}$$

The coefficient $a(x_1,x_2) > 0$ is not continuous everywhere. This precludes application of standard multigrid methods. Alcouffe et al. (1981), Kettler and Meijerink (1981) (see also Kettler (1982)) have developed special multigrid methods that work well for the problem considered here. In these methods the prolongation and restriction operators depend on the discrete approximation to (1.1). Until now, theoretical justification is lacking and seems hard to come by. In the following, a multigrid method is proposed for (1.1) that also works in practice, and that can be justified theoretically. The difference with the methods just mentioned is, that prolongation and restriction are not problem-dependent, and that grid coarsening is done finite volume fashion rather than finite difference fashion. What this means will be made clear in the sequel.

Within the confines of this short paper only a brief synopsis of theoretical and practical aspects can be given. A more complete account will be published elsewhere.

2. FINITE VOLUME DISCRETIZATION

The domain Ω is subdivided in finite volumes Ω^{ℓ}_{ij}, which are squares of size h^{ℓ} with centers at the points

$$(x_1,x_2): x_1 = (i - 1/2)h, \quad x_2 = (j - 1/2)h, \quad i,j = 1,2,\dots,2^{\ell}, \quad h^{\ell} = 2^{-\ell}. \tag{2.1}$$

This grid is called Ω^{ℓ}. Coarser grids Ω^{k}, $k = 0,1,2,\dots,\ell-1$ are defined by (2.1) with ℓ replaced by k.

The finite volume approximation of (1.1) on Ω^{ℓ} is obtained in the usual way. $\Phi^{k} = \{\Omega^{k} \to \mathbb{R}\}$ is the set of grid functions on Ω^{k}, with domain the centers of the finite volumes in Ω^{k} and range $= \mathbb{R}$. The backward and forward divided differences in the x_i-direction on Ω^{k} are denoted by ∇^{k}_{i} and Δ^{k}_{i}. For example,

$$\nabla^{k}_{1}\phi^{k}_{ij} = \left(\phi^{k}_{ij} - \phi^{k}_{i-1,j}\right)/h^{k}, \tag{2.2}$$

where $\phi^{k} \in \Phi^{k}$, and ϕ^{k}_{ij} is the value at the center of Ω^{k}_{ij}. The finite volume discretization of (1.1) on Ω^{ℓ} is given by

$$A^{\ell}\phi^{\ell} \equiv -\left(\nabla^{\ell}_{1}w^{\ell}_{1}\Delta^{\ell}_{1} + \nabla^{\ell}_{2}w^{\ell}_{2}\Delta^{\ell}_{2}\right)\phi^{\ell} = f^{\ell}, \tag{2.3}$$

$$w^{\ell}_{1,ij} = 2a^{\ell}_{ij}a^{\ell}_{i+1,j}/\left(a^{\ell}_{ij} + a^{\ell}_{i+1,j}\right), \quad w^{\ell}_{2,ij} = 2a^{\ell}_{ij}a^{\ell}_{i,j+1}/\left(a^{\ell}_{ij} + a^{\ell}_{i,j+1}\right), \tag{2.4}$$

where a^{ℓ}_{ij} is the average of $a(x_1,x_2)$ over Ω_{ij}. Note that

$$\min\left(a^{\ell}_{ij},a^{\ell}_{i+1,j}\right) \leqslant w^{\ell}_{1,ij} \leqslant 2\min\left(a^{\ell}_{ij},a^{\ell}_{i+1,j}\right) \tag{2.5}$$

and similarly for w^{ℓ}_{2}. Hence $w^{\ell}_{i} > 0$. This, together with Dirichlet boundary conditions (assumed here) makes A^{ℓ} symmetric positive definite. For details on the derivation of (2.3), see for example Wesseling (1987).

3. PROLONGATION, RESTRICTION AND COARSE GRID APPROXIMATION

Prolongation operators $P^k: \Phi^{k-1} \rightarrow \Phi^k$ are defined by

$$
\begin{aligned}
(P^k\phi^{k-1})_{2i,2j} &= (P^k\phi^{k-1})_{2i-1,2j} = (P^k\phi^{k-1})_{2i,2j-1} = \\
&= (P^k\phi^{k-1})_{2i-1,2j-1} = \phi^{k-1}_{ij}. \qquad (3.1)
\end{aligned}
$$

A prolongation operator that interpolates polynomials of degree m exactly is said to be of order m+1. Therefore P^k is of order 1.

Restriction operators $R^{k-1}: \Phi^k \rightarrow \Phi^{k-1}$ are defined by

$$
\begin{aligned}
(R^{k-1}\phi^k)_{ij} = \Big(&\phi^k_{2i,2j-2} + \phi^k_{2i+1,2j-2} + 2\phi^k_{2i-1,2j-1} + \\
&+ 3\phi^k_{2i,2j-1} + \phi^k_{2i+1,2j-1} + \phi^k_{2i-2,2j} + 3\phi^k_{2i-1,2j} + \\
&+ 2\phi^k_{2i,2j} + \phi^k_{2i-2,2j+1} + \phi^k_{2i-1,2j+1}\Big) /16. \qquad (3.2)
\end{aligned}
$$

The stencil of this operator is given by

$$
\frac{1}{16}\begin{bmatrix} 1 & 1 & & \\ 1 & 3 & 2 & 1 \\ & 2 & 3 & 1 \\ & & 1 & 1 \end{bmatrix}.
$$

Apart from a scaling factor, this is the adjoint of linear interpolation in triangles, cf. the following figure. This adjoint interpolates first degree polynomials exactly. Therefore R^{k-1} is said to be of order 2. The sum of the orders of P^k and R^{k-1} should exceed the order of the differen-

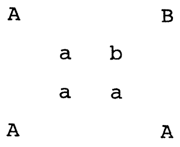


Figure 3.1. Linear interpolation in triangles. Points A, B, a, b are centers of coarse respectively fine grid finite volumes. Values at a are found by linear interpolation in triangle AAA, at b in AAB.

tial equation (Brandt (1977), Hackbusch (1985)). This condition is satisfied here.

Coarse grid approximations A^k to A^ℓ, $k < \ell$ are defined by

$$A^k = R^k A^{k+1} P^{k+1}, \quad k = \ell-1, \ell-2, \ldots, 0. \tag{3.3}$$

It turns out that A^k is given by

$$A^k \phi^k = -\left(\nabla_1^k w_1^k \Delta_1^k + \nabla_2^k w_2^k \Delta_2^k\right) \phi^k \tag{3.4}$$

$$w_{1,ij}^k = \frac{1}{2}\left(w_{1,2i,2j}^{k+1} + w_{1,2j,2j-1}^{k+1}\right), \quad w_{2,ij}^k = \frac{1}{2}\left(w_{2,2i,2j}^{k+1} + w_{2,2i-1,2j}^{k+1}\right). \tag{3.5}$$

The availability of this simple explicit representation of A^k makes the development of a rate of convergence theory feasible. With a more accurate P^k the stencil of A^k becomes larger than the stencil of A^{k+1}, which is why we do not consider more accurate prolongations. Equivalent results are obtained by taking P^k as the adjoint of the present R^{k-1}, and R^k as the adjoint of the present P^{k+1}. Note that although in (3.3) R^k is not the adjoint of P^{k+1}, the resulting A^k still turns out to be self-adjoint.

4. MULTIGRID METHOD AND PRACTICAL RESULTS

For smoothing incomplete LU-decomposition (ILU) is used. For the use of ILU in multigrid methods, see for example Hemker (1982), Kettler (1982), Wesseling (1982a,b), Hemker et al. (1983), Hackbusch (1985), Sonneveld and Wesseling (1985).

The W-cycle is used with post-smoothing.

Computations have been performed for the following problem. Inside Ω we have a square with side $L < 1$ and center coinciding with the center of Ω, where $a = a_1$ = constant. In the rest of Ω, $a = a_2$ = constant. The boundary condition is

$$\phi|_{\partial\Omega} = x_1^2 + x_2^2, \tag{4.1}$$

and the right hand side is

$$f = x_1 x_2. \tag{4.2}$$

The starting iterand is zero.

Two cases have been studied; case 1: $a_1 = .333 * 10^5$, $a_2 = 2$ and case 2: $a_1 = 2$, $a_2 = .333 * 10^5$. The average reduction factor ρ of the residue was measured over the iterations required to make the residue so small that rounding errors become apparent, with a maximum of 12 iterations. Computations have been made for $\ell = 3,4,5,6$, with $L = n2^{-\ell}$, $n = 0,2,4,\ldots,2^{\ell}$. The definition of ρ is $\rho = (|\text{final residue}|_0/|\text{initial residue}|_0)^{1/it}$, with it the number of iterations carried out, and $|\cdot|_0$ the ℓ_2-norm. The following table gives ρ for n = 0 (i.e. the Laplace equation) and for that value of n for which ρ obtains its largest value for a given ℓ.

Case	ℓ	3	4	5	6
1	n,ρ	0, .058	0, .099	0, .100	0, .110
1	n,ρ	6, .390	10, .640	18, .547	34, .532
2	n,ρ	0, .049	0, .060	0, .066	0, .069
2	n,ρ	2, .071	2, .087	20, .092	44, .097

Table 4.1. Average reduction factor ρ.

Table 4.1 indicates that multigrid is functioning properly, because ρ seems to be independent of h. That convergence for case 1 is slower than for case 2 is to be ascribed to the fact that case 1 is appreciably worse conditioned than case 2. This is because in case 1 we are solving almost a pure Neumann problem for the interior region with $a = a_1$, due to the "weak coupling" with the exterior. An extensive comparison of rate of convergence with the methods proposed by Alcouffe et al. (1981) and Kettler and Meijerink (1981), Kettler (1982) has not yet been made. These methods seem less sensitive to the value of n than the present method, and may have smaller ρ. However, an iteration with the present method is cheaper, and the cost of obtaining A^k, $k < \ell$, which is appreciable for the former methods, is negligible for the present method, due to the availability of the explicit expression (3.4). Furthermore, the present method can be justified theoretically, as will be shown next.

5. THEORETICAL CONSIDERATIONS

In this section an (incomplete) outline will be given of a convergence theory for the problem and the method under consideration. This section also highlights some aspects of multigrid convergence theory in general for finite difference and finite volume discretizations.

The framework for multigrid convergence analysis for finite difference and element methods as it has developed over the past 20 years is very well described by Hackbusch (1985). We use this framework also in the present case. It suffices to study two-grid convergence. Multigrid convergence then follows in the standard way described in Hackbusch (1985).

Consider two grids Ω^k and Ω^{k-1}. On Ω^{k-1} we solve exactly. Smoothing consists of ν iterations with a smoothing method with iteration matrix S^k on Ω^k, following coarse grid correction. Coarse grid correction is defined by:

$$\phi^k := \phi^k + P^k(A^{k-1})^{-1}R^{k-1}(f^k - A^k\phi^k). \tag{5.1}$$

To improve readability, the superscript k is dropped, and the superscript k-1 is replaced by an overbar. The two-grid iteration matrix is given by

$$M = S^\nu(I - P\bar{A}^{-1}RA). \tag{5.2}$$

Following Hackbusch (1985), instead of M we study $\hat{M} = AMA^{-1}$, and we introduce the following splitting:

$$\hat{M} = (AS^\nu)(A^{-1} - P\bar{A}^{-1}R). \tag{5.3}$$

Norms $|\cdot|_s$ (on Φ or $\bar{\Phi}$, as the case may be) are defined by

$$|\phi|_s = |L^{s/2}\phi|_0, \tag{5.4}$$

where L is A with $w_i \equiv 1$ (i.e. the familiar discrete 5-point Laplace operator with homogeneous Dirichlet boundary condition), and where $|\cdot|_0$ is defined by

$$|\phi|_0^2 = (\phi,\phi), \tag{5.5}$$

where

$$(\phi,\phi) = h^2 \sum_{i,j=1}^{1/h} \phi_{ij}^2. \tag{5.6}$$

Let Q: $\Phi \rightarrow \Phi$ be some operator. As usual we define

$$|Q|_{t\leftarrow s} = \sup_{\phi\neq 0} |Q\phi|_t/|\phi|_s. \tag{5.7}$$

We have

$$|\hat{M}|_{s\leftarrow s} \leqslant |AS^{\nu}|_{s\leftarrow s+\sigma} \; |A^{-1} - P\bar{A}^{-1}R|_{s+\sigma\leftarrow s} \tag{5.8}$$

where s and σ remain to be chosen. The purpose of two-grid convergence theory is to show that

$$|\hat{M}|_{s\leftarrow s} \leqslant C_1 < 1 \tag{5.9}$$

for suitable s and ν, with C_1 independent of h. Equation (5.9) follows if we can show that the following two properties hold:

Smoothing property: $|AS^{\nu}|_{s\leftarrow s+\sigma} \leqslant \eta(\nu)h^{-\alpha}$ for all $1 \leqslant \nu \leqslant \bar{\nu}(h)$,

with $\eta(\nu) \rightarrow 0$ as $\nu \rightarrow \infty$ and $\bar{\nu}(h) = \infty$ or $\bar{\nu}(h) \rightarrow \infty$ as $h \rightarrow 0$.

Approximation property: $|A^{-1} - P\bar{A}^{-1}R|_{s+\sigma\leftarrow s} \leqslant C_A h^{\alpha}$, with C_A a constant independent of h.

The exponent α is the same in both properties. The smoothing property has been introduced by Hackbusch (1985). Uniform validity of both properties on all grids is sufficient for multigrid convergence.

We proceed with the approximation property. Since

$$A^{-1} - P\bar{A}^{-1}R = A^{-1}(I - P'R)(I - AP\bar{A}^{-1}R) \tag{5.10}$$

with P': $\bar{\Phi} \rightarrow \Phi$ arbitrary we can write

$$|A^{-1} - P\bar{A}^{-1}R|_{s+\sigma\leftarrow s} \leqslant |A^{-1}|_{s+\sigma\leftarrow\tilde{s}} \; |I - P'R|_{\tilde{s}\leftarrow s} \; |I - AP\bar{A}^{-1}R|_{s\leftarrow s}$$

where $|\cdot|_{\tilde{s}}$ will be specified shortly.

Because P is of order 1 we can handle the last term only for $s \leqslant 1$. We choose $s = 1$. Due to the irregularity of a we cannot estimate $|A^{-1}|_{s+\sigma\leftarrow s+\sigma-2}$, unless $s+\sigma = 1$. With $s+\sigma = 1$ we have to estimate $|I - P'R|_{-1\leftarrow -1}$. However, this does not allow us to gain a power of h, which is needed to have $\alpha > 0$, which is required by the smoothing property, as it turns out. This is the reason for the introduction of the norm $|\cdot|_{\tilde{2}}$, defined by

$$|\phi|_{\tilde{2}}^2 = |\nabla w_1 \Delta_1 \phi|_0^2 + |\nabla_2 w_2 \Delta_2 \phi|_0^2, \tag{5.11}$$

with appropriate modifications of the operators $\nabla_i w_i \Delta_i$ near the boundaries to take into account the Dirichlet boundary condition. We also define

$$|\phi|_{-\tilde{2}} = \sup_{|\theta|_{\tilde{2}} \leqslant 1} (\phi,\theta). \tag{5.12}$$

We choose $\sigma = 1$. It is not difficult to show that

$$|R|_{-1\leftarrow -1} \leqslant C_1, \tag{5.13}$$

$$|\bar{A}^{-1}|_{1\leftarrow -1} \leqslant \underline{a}^{-1}, \qquad |A|_{-1\leftarrow 1} \leqslant \bar{a}, \tag{5.14}$$

with $\underline{a}$ = inf (a), $\bar{a}$ = sup (a)

$$|P|_{1\leftarrow 1} \leqslant C_2, \tag{5.15}$$

$$|A^{-1}|_{0\leftarrow -\tilde{2}} \leqslant \underline{a}. \tag{5.16}$$

More difficult is to show (choose P' = P):

$$|I - PR|_{-\tilde{2}\leftarrow -1} \leqslant C_3 \bar{a} h. \tag{5.17}$$

Details will be given elsewhere. With these results, the approximation property easily follows. C_A depends on $\bar{a}/\underline{a}$, which is not disturbing, since this is also the case when a is smooth.

Following Hackbusch (1985) chapter 6, the smoothing property is also easily verified.

ACKNOWLEDGEMENT

The author is indebted to C. Cuvelier, J. van Kan and P. Sonneveld for useful discussions.

REFERENCES

ALCOUFFE, R.E., BRANDT, A., DENDY Jr., J.E., PAINTER, J.W., 1981, *The multigrid method for the diffusion equation with strongly discontinuous coefficients*, SIAM J. Sci. Stat. Comp. 2, 430-454.

BRANDT, A., 1977, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp. 31, 333-390.

HACKBUSCH, W., TROTTENBERG, U., (eds.), 1982, *Multigrid Methods*. Proceedings, Köln-Porz, Lecture Notes in Mathematics 960, Springer-Verlag, Berlin.

HACKBUSCH, W., 1985, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin.

HEMKER, P.W., 1982, *On the comparison of Line-Gauss-Seidel and ILU relaxation in multigrid algorithms*. In: Miller (1982), 269-277.

HEMKER, P.W., KETTLER, R., WESSELING, P., DE ZEEUW, P.M., 1983, *Multigrid methods: development of fast solvers*, Appl. Math. Comp. 13, 311-326.

KETTLER, R., MEIJERINK, J.A., 1981, *A multigrid method and a combined multigrid-conjugate gradient method for elliptic problems with strongly discontinuous coefficients in general domains*, SHELL Publ. 604, KSEPL, Rijswijk, The Netherlands.

KETTLER, R., 1982, *Analysis and comparison of relaxation schemes in robust multigrid and preconditioned conjugate gradient methods*. In: Hackbusch and Trottenberg, 1982, 502-534.

McCORMICK, S., (ed.), 1987, *Multigrid Methods*, Frontiers in Applied Mathematics, 5, SIAM, Philadelphia.

MILLER, J.J.H., (ed.), 1982, *Computational and Asymptotic Methods for Boundary and Interior Layers*, Proc. BAIL II Conference, Boole Press Conference Series 4, Boole Press, Dublin.

PADDON, D.J., HOLSTEIN, H., (eds.), 1985, *Multigrid Methods for Integral and Differential Equations*, The Institute of Mathematics and its Applications Conference Series, New Series Number 3, Clarendon Press, Oxford.

SONNEVELD, P., WESSELING, P., 1985, *Multigrid and conjugate gradient methods as convergence acceleration techniques*. In: Paddon and Holstein (1985), 117-168.

WESSELING, P., 1982a, *A robust and efficient multigrid method.* In: Hackbusch and Trottenberg (1982), 164-184.

WESSELING, P., 1982b, *Theoretical and practical aspects of a multigrid method*, SIAM J. Sci. Stat. Comp. 3, 387-407.

WESSELING, P., 1987, *Linear Multigrid Methods*. In: McCormick (1987).

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Applications and Problems of Error Correction Coding with respect to Storage Channels

C.P.M.J. Baggen
Philips Research Laboratories
P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands

Abstract

The Compact Disc system can be seen as a particular implementation of a digital storage channel. It will be shown that the use of error-correction coding is inevitable in almost all possible applications of modern digital mass storage systems.

Traditionally much mathematical effort has been invested into the design of good codes. In the current industrial environment, more emphasis is put on the decoding algorithms and performance evaluations of codes.

The use of Reed-Solomon codes belonging to the class of Maximum Distance Separable Codes will be elucidated.

Finally it will be shown that product codes offer interesting possibilities although both the optimal decoding strategy and the performance evaluation are still open problems.

Introduction

Recording of information has been practised for almost hundred years now since the introduction of Edisons Phonograph. Traditionally analog signals like audio and video have been recorded in an analog way (e.g. the grammophone record, the audio tape and the VCR tape). Computer data, being digital of nature, always have been stored digitally since the introduction of paper tape during the second world war.

Developments in physics allow an ever increasing information storage capacity for example on magnetic or optical media. It appears that at some point in the development one switches to digital recording for any source, even those which are analog of origine. We have been able to observe this event at the introduction of Compact Disc, and it will happen in the near future with video. There are mainly two reasons for doing this.

The first reason is connected with quality. When storage capacity is increasing, one is willing to trade 'excess playing duration' for quality. A really high quality (to be maintained through all phases of editing and duplication) can only be achieved if the signals are recorded digitally including error-correction.

The second reason is inherent to the increasing recording density itself. As less surface; volume or atoms are used to store a given amount of information, one becomes more liable to disturbances or noise generated in real life systems. The only known way to cope with these disturbances effectively forces us to record digitally, by which we may apply information theoretic concepts like error-correcting codes.

It turns out that the error rates of almost all existing digital storage systems fall short many orders of magnitude to the desired reliabilities. Coding offers the only known reasonable solution.

Current solid state technology offers the possibility to implement advanced coding schemes although we still want to minimize the complexity for a given performance.

In the remainder we will clarify the use of particular coding schemes. We will start with some general remarks on block codes, followed by an analysis of Reed-Solomon codes. Finally we will treat product codes, where we will address some mathematical problems connected with their applications and evaluations.

Block Codes in General

The storage of digital information is usually done by successive recording of elements of a finite alfabet called symbols. It turns out that for high density recording, this alfabet must have a reasonable cardinality q (e.g. $q = 256$) which implies that errors occur in the form of q-ary symbol errors. It is advantageous to give the alfabet the structure of a finite field, which allows application of algebraic codes defined over GF(q).

In general an error-correcting code uses redundant information to detect and/or correct errors (see Peek's contribution in this issue). In our applications we assign redundancy to information according to a fixed algorithm which preferably maximizes the error-correction capacity for a given fraction of redundancy.

Important parameters of a block code $C(n,k,d)$ are the number of symbols per codeword called the length n, the number of symbols containing useful information called the dimension k and the minimum number of symbols in which any two codewords differ at least called the distance d. The rate R of a code is defined as: $R = k/n$.

Generally decoding consists of finding the codeword which is closest (in Hamming distance sense) to the received word. The minimum (Hamming) distance of a code [1,3] is related to the error-correction capacity t by

$$2t \leq d-1 \text{ symbols / codeword.} \tag{1}$$

Sometimes we might know that certain symbols on given positions are unreliable. These symbols are called erasures. In that case a code can correct t errors and e erasures simultaneously if

$$2t+e \leq d-1. \tag{2}$$

If we restrict the decoding to error patterns not exceeding the above mentioned parameters (which is called bounded distance decoding), it appears that efficient decoding algorithms exist for certain classes of codes e.g. BCH codes [4].

In a recording environment (e.g: CD) the physical dimensions of the defects often are such that many consecutive symbols are destroyed (burst errors). In order to cope effectively with burst errors a number of codewords are interleaved [2,3], i.e. consecutive recorded symbols are assigned to different codewords. A particular implementation might be envisioned by regarding the codewords as the rows of a two dimensional array, which is written and read from the disc columnwise.

Reed-Solomon Codes

It can be proven [1] that for any code:

$$d \leq n-k+1. \tag{3}$$

This so-called Singleton Bound imposes an upperbound on correction capabilities of a code given the number of its redundant symbols. Reed-Solomon (RS) codes are a class of codes which achieve this bound with equality, they are said to be Maximum Distance Separable

(MDS). Therefore RS codes are optimal candidates for correcting symbol errors and burst errors.

Because RS codes also belong to the class of BCH codes, relative efficient encoding and decoding algorithms are known. Roughly, the decoding complexity is quadratic in d for RS codes. However for very small d (say $d \leq 7$) decoding algorithms may degenerate, which makes them extremely simple and fast in that case.

It appears that consensus has been reached on the use of RS codes or codes constructed from them for error-correction applied to current digital storage channels.

Performance Evaluation of RS Codes

The reliability performance of a code (assuming that (1) holds with equality) usually is expressed in terms of uncorrectable or undetectable error probabilities after decoding. These probabilities depend strongly both on the chosen code and on the error characteristics of the channel. We will evaluate the performance of some interleaved RS codes assuming the channel producing random symbol errors.

In a real life situation, storage of computer data usually is done in sectors of e.g. 512 bytes, on top of which the redundancy and some protocol information must be written. We will fix the rate of the considered codes, hence all codes will use the same amount of storage capacity for a sector. A sector will consist of I codewords each of length n where:

$$I \times k \approx 512 \quad \text{(user information per sector)}$$
$$\frac{k}{n} \approx .85 \quad \text{(code rate or efficiency)}$$

If the error-correction capacity is varied (with a corresponding change in the number and length of codewords), we may study the influence of error-correction capacity on the reliability performance. Note that in this comparison increased error-correction capacity is traded only against increased complexity.

Given a random symbol error rate $p_{s,}$ we can easily calculate the probability P_e that a sector is uncorrectable:

$$P_e = 1 - (1 - P_u)^I \tag{4}$$

where:

$$P_u = \sum_{i=t+1}^{n} \binom{n}{i} p_s^i (1-p_s)^{n-i} \ ; \tag{5}$$

$$t = \frac{(n-k)}{2} \ ; \tag{6}$$

P_u being the probability that a single codeword of a sector is uncorrectable. For $p_s << 1$ we may use first order approximations:

$$P_e \approx I \binom{n}{t+1} p_s^{t+1} \qquad (p_s << 1) \ . \tag{7}$$

In fig. 1 we plotted the performance of some RS codes as a function of the symbol error rate with the code's distance as parameter. Note that the performance improves with increasing distance.

Product Codes

A two dimensional product code $PC(n_p, k_p, d_p)$ can be envisioned as a two dimensional array, where each row is a codeword from a first code $C1(n_1, k_1, d_1)$ and each column is a codeword from a possibly different second code $C2(n_2, k_2, d_2)$. Mostly C1 and C2 are simple (low-distance) codes. It can be shown [1] that the following identities hold for the parameters of the product code:

$$\begin{aligned} n_p &= n_1 \times n_2 \ ; \\ k_p &= k_1 \times k_2 \ ; \\ d_p &= d_1 \times d_2 \ . \end{aligned} \tag{8}$$

An advantage of product codes consists of reduced decoding complexity, because these codes may be decoded by successive row and column decodings of relatively simple codes. Product codes are also known to have good burst error correcting capabilities (note the correspondence with the above mentioned interleave).

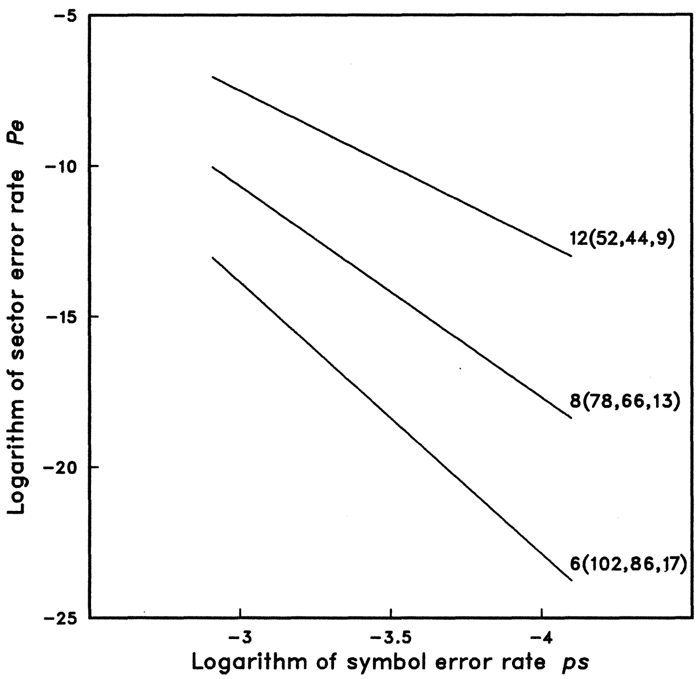


FIG. 1. Performance of Interleaved Reed Solomon Codes Each code is characterized by its parameters *I(n,k,d)*

However, the optimal decoding strategy as well as the performance evaluation in terms of remaining error rate are still open problems.

In general the performance of any (product) code, given a decoding algorithm, can be expressed as a power series in p_s if we assume random symbol errors:

$$P_e = \sum_{i=t+1}^{n_p} A_i p_s^i (1 - p_s)^{n_p - i}. \tag{9}$$

For $p_s << 1$ the first few terms dominate.
The exponent of the first term $(t + 1)$ is determined by the minimum weight of the uncorrectable error patterns. The corresponding coefficient A_{t+1} takes into account all

possible configurations of t + 1 errors, each weighted with its probability of being uncorrectable.

Although a product code has a low minimum distance, a good strategy leads to a very small coefficient A_{t+1}, which may result in desirable performance characteristics. Traditionally only the minimum distance of a code is considered, because it determines which code performs best in the limit $p_s \rightarrow 0$. However in practice we are interested in the performance of a code in a fixed range of nonzero values of p_s.

We will elucidate this by a simple example which appears to be sufficiently tractable. In general the determination of the coefficients of the above mentioned power series is yet unsolved for product codes.

A Simple Product Code

We may construct a product code by taking RS codes as elementary row and column codes. Let $PC(625,529,9)$ over $GF(2^8)$ be a $RS(25,23,3) \times RS(25,23,3)$ over $GF(2^8)$. Note that the important user parameters like dimension and rate are comparable to the previously considered sectors encoded with interleaved RS codes.

Although the distance is rather small it will turn out that this code has a surprisingly good performance because it can correct many error patterns exceeding the guaranteed error-correction capacity.

A Decoding Strategy for a Simple Product Code

We will first fix a decoding strategy for the above mentioned product code. Subsequently the reliability performance will be estimated. It is not claimed that the strategy is the optimal one. However it can easily be implemented and it shows already the strength of a product code. The choice of this strategy is based on the following thoughts:

- during an elementary decoding operation one must use reliability information obtained by previous decoding operations;
- actual correction of symbols must occur in order of increasing risk, i.e. symbols which can be corrected with high certainty are corrected first.

The decoding algorithm consists of the following steps:

1 Calculate all row and column syndromes;

2 Assign indicators to rows and columns containing nonzero syndromes;

3 Repeated row/column single error correction using indicators as reliability information.

4 If necessary and possible try 2-erasure correction using indicators as pointers.

Note the simplicity of the elementary decoding operations in the above strategy compared to 4, 6 or 8 error-correction of previous examples.

Performance Evaluation of a Simple Product Code

It can be shown that in the above case all error patterns of weight at most 4 will always be corrected. The error patterns of weight 5 must be studied carefully. Because the code's distance is 9, we know already that some of these error patterns must be uncorrectable. Note that for an error pattern of weight 5 to be uncorrectable, at most 3 columns and at most 3 rows should be affected by the errors.

Because the constituent row and column codes are MDS codes, we can estimate fairly accurate the probabilities of correct decoding, miscorrection and error detection [5]. Detailed analysis (Appendix A) of all error configurations of weight 5 in a 3 by 3 array reveals that the contribution of weight 5 error patterns to the error probability can be estimated by:

$$P_e \approx 2 \times 10^4 p_s^5 \qquad (p_s << 1) \tag{10}$$

It turns out that in this particular situation the second term of the series expansion of P_e cannot be neglected for $p_s \approx 10^{-3}$. Taking into account the next most important term (Appendix A), we may approximate the performance by:

$$P_e \approx 2 \times 10^4 p_s^5 + 3 \times 10^7 p_s^6 \qquad (p_s \leq 10^{-3}) \tag{11}$$

In fig. 2 we have plotted the performance of the product code as a function of the symbol error rate (curve indicated by 'PC(BASIC)'). For a comparison we also indicated the results of fig. 1.

We might even extend the given decoding algorithm by attempting eventually maximum likelihood decoding in a 3 by 3 array indicated by nonzero syndromes, which is quite efficient as the code's distance is 9. Although we do not describe this procedure in detail, it can be shown that the performance increases to the curve indicated by 'PC(ML)'.

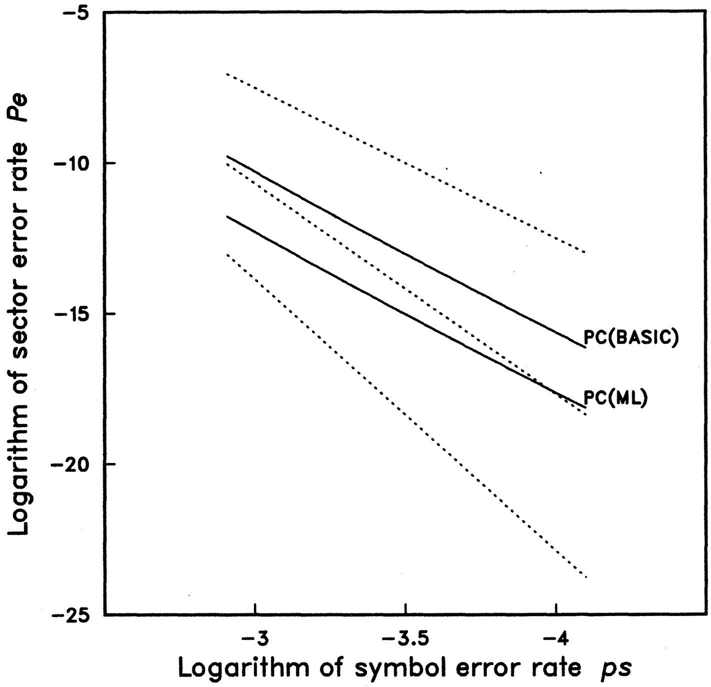


FIG. 2. Performance of the Product Code. The dotted curves correspond to FIG. 1.

Note that the slope of the curve corresponding to the product code indeed equals the slope of a 4 error-correcting code. However, decoding of the product code mainly consists of repeated single error-correction, where single error-correction consists of only one division given the syndromes. Nevertheless the performance is orders of magnitude better than the performance of comparable d = 9 interleaved RS codes.

Note also that the performance curve of the PC intersects the performance curves of interleaved RS codes with a high distance. This implies that for reasonable good channels a high distance code performs better than a simple product code. On the other hand for bad channels (when performance is most critical), a product code might outperform a high distance code.

What really matters in practice is the absolute performance under most adverse channel conditions (say $p_s \approx 10^{-3}$). It is less interesting how much overkill is present under good channel conditions. If we look at the performance that way, a low distance may have the advantage of making a code less sensitive to the exact channel characteristics.

Conclusion

Reed-Solomon codes or codes constructed with RS codes are optimally suited for error-correction applied to storage channels because of the MDS property of RS codes combined with the occurrence of q-ary symbol errors which even may be bursty. Product codes of RS codes offer interesting possibilities because they are easy to decode while retaining a very good performance as we have shown for random symbol errors. This may be attributed to the product code's capability to correct many error patterns exceeding its designed error-correction capacity.

Appendix A

Because the product code is a linear code, we assume the allzero word to be transmitted. As usual we consider an error pattern to be additive to the transmitted codeword.

Because each row codeword belongs to a distance 3 MDS code over GF(q), there exist exactly q-1 row codewords which are nonzero in three specified positions and zero in all other positions [1]. These codewords are multiples of each other. The same holds for the column codewords.

Because the weight 9 codewords of the product code have their nonzero symbols at the intersections of three rows and three columns, it can be seen that in 9 such specified positions there are exactly q-1 nonzero codewords of weight 9 which are multiples of each other.

Miscorrection of row (column) decoder

An error pattern of weight 2 in a row codeword might be detected or it might be miscorrected by single error-correction, thereby changing the error pattern into a weight 3 codeword.

More specifically if a row (or column) codeword contains two random errors and a third specified (flagged) position which the decoder may alter, this situation will be miscorrected with probability:

$$\Pr(\text{miscorrection} \mid 2 \text{ errors}, 1 \text{ extra flag}) = (q-1)^{-1}$$

while detection will happen with probability:

$$\Pr(\text{detection} \mid 2 \text{ errors}, 1 \text{ extra flag}) = (q-2)(q-1)^{-1}$$

This can be verified easily by counting arguments using the above mentioned properties of weight 3 codewords.

Misdetection of row (column) decoder

Another probability we need is the probability that an error pattern of weight 3 in three specified positions will lead to an allzero syndrome. This event is called an undetected error or a miscorrection. It can only happen if the error pattern is equal to a codeword. The probability that one of the q-1 codewords, nonzero in exactly those positions, agrees to the error pattern is:

$$\Pr(\text{misdetection} \mid 3 \text{ errors}) = (q-1)^{-2}$$

Product code analysis

With the above knowledge and some combinatorics, we are able to estimate the probability of uncorrectable errors.

Because the code is a product code of MDS codes, it can be verified by counting arguments that the average decoding result of a configuration of 5 or 6 error positions over all possible error values is not dependent on any row or column permutations. This fact may be used to classify error patterns into equivalence classes. Two error patterns are called equivalent if there exists a combination of row and column permutations, which maps the error positions of the first pattern into the error positions of the second pattern. Two error patterns which are not equivalent are called different.

Weight 5 error patterns

We will analyse all different error patterns of weight 5 which lead to uncorrectable errors. For each pattern the contribution to the probability of an uncorrectable error is derived.

The first components of the formulae describe the number of ways in which erroneous rows resp. columns can be chosen from the codeword (including the interchange of rows and columns). The following part describes how many configurations of the error pattern exist given the rows and columns that were chosen. The product of the first and second part corresponds to the size of the equivalence class. The third part finally approximates the probability that decoding will fail due to the occurrence of the error pattern which is representative for the equivalence class.

We distinguish between the presence of errorwords containing initially undetected errors (indicated by S = 0) and the case that all initial flags are set correctly.

Presence of errorwords containing initial undetected errors (S = 0):

x x x → S=0
. x x

$$P \approx 2\binom{25}{3}\binom{25}{2}\binom{6}{5}\frac{1}{(q-1)^2}\,p_s^5$$

```
x   x x  → S=0
x   . .
x   . .
↓
S=0
```

$$P \approx \binom{25}{3}^2 3^2 \frac{1}{(q-1)^4} p_s^5$$

```
x x x  → S=0
x . .
. x .
```

$$P \approx \binom{25}{3}^2 2 \binom{3}{1}\binom{3}{1}\binom{2}{1} \frac{1}{(q-1)^2} \frac{2}{(q-1)} p_s^5$$

The following error patterns correspond to the case where all codewords containing errors are initially detected. In order to be uncorrectable at least two errorwords should be miscorrected:

```
x x x
x . .
. x .
```

$$P \approx \binom{25}{3}^2 \binom{3}{1}\binom{3}{1}\binom{2}{1} \frac{5}{(q-1)^2} p_s^5$$

```
x x .
x x .
. . x
```

$$P \approx \binom{25}{3}^2 \binom{3}{2}^2 \frac{6}{(q-1)^2} p_s^5$$

```
x x .
x . x
. x .
```

$$P \approx \binom{25}{3}^2 \binom{3}{2}\binom{3}{2} \bullet 2 \bullet 2 \bullet \frac{5}{(q-1)^2} p_s^5$$

Summation of the above contributions leads to an estimate of the error probability due to 5 random symbol errors:

$$P \approx 2 \times 10^4 p_s^5$$

Weight 6 error patterns

There is one weight 6 error configuration which is always uncorrectable. Because all other error patterns of weight 6 need at least 1 miscorrection or misdetection of a row or column decoder in order to be uncorrectable, we only need to take into account the following configuration:

```
x x .
x . x
. x x
```

$$P \approx \binom{25}{3}^2 3! \; p_s^6$$

References

[1] F.J. MacWiliams, N.J.A. Sloane;
The Theory of Error Correcting Codes;
North Holland Publishing Company, Amsterdam.

[2] G.C. Clarc, Jr., J. Bibb Cain;
Error-Correction Coding for Digital Communications;
Plenum Press, New York.

[3] J.B.H. Peek;
Communications Aspects of the Compact Disc Digital Audio System;
IEEE Comm. Mag., Vol.23, No.2, Febr. 1985.

[4] Elwyn R. Berlekamp;
Algebraic Coding Theory;
McGraw-Hill Book Company, 1968.

[5] Robert J. McEliece and Laif Swanson;
On the Decoder Error Probability for Reed-Solomon Codes;
IEEE Trans. Inform. Theory, IT-32, No. 5, p. 701, Sept. 1986.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Introduction to Nilpotent Approximation Filtering

Michiel Hazewinkel
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The socalled reference probability of unnormalized probability method for nonlinear filtering problems leads to a (robust) infinite dimensional filter of bilinear type. If the associated Lie algebra is topologically solvable or nilpotent an infinite dimensional version of Wei-Norman theory applies. If not then ideas of nilpotent approximation lead to (potential) approximation filters. This note is not so much a definite report on results as on outline of a research program.

1980 Mathematics subject classification: 93E11, 93B30, 93E10, 93D25, 60H15, 93B15, 17B65, 17B99, 57R25, 35H05.
Keywords and phrases: nonlinear filtering, estimation Lie algebra, topologically nilpotent Lie algebra, Duncan-Mortenson-Zakai equation, unnormalized conditional probability density, robustness, nilpotent approximation, Wei-Norman theory, stochastic differential equations.

1. Statement of the problem

In full generality *filtering* is concerned with obtaining estimates concerning a stochastic process $\{x_t\}$, the *signal process,* on the basis of another related process $\{y_t\}$, the *observation process.* In this paper we have the following realization of this situation in terms of stochastic differential equations.

$$dx_t = f(x_t)dt + G(x_t)dw_t \ , \ \ x_t \in \mathbb{R}^n, \ w_t \in \mathbb{R}^m \tag{1.1}$$

$$dy_t = h(x_t)dt + dv_t, \ \ y_t \in \mathbb{R}^p, \ v_t \in \mathbb{R}^p \tag{1.2}$$

where f, G, h are vector and matrix valued functions of the right dimensions and w_t and v_t are independent Wiener noise processes also independant of the initial state x_0. The problem is the following. For a given (interesting) function $\phi(x)$ of the state x, give a calculation procedure for the best estimation $\widehat{\phi(x_t)}$ at time t given the observations y_s, $0 \leq s \leq t$. More generally one also considers finding $\widehat{\phi(x_t)}$ given y_s, $0 \leq s \leq t_1$, $t_1 < t$ *(prediction)* and finding $\widehat{\phi(x_t)}$ given y_s, $0 \leq s \leq t_2$, $t < t_2$ *(smoothing).* Of particular importance is finding $\hat{x}_t$ (*state estimation*).

Ideally one would like the calculation procedure to be *finite dimensional, exact, recursive,* and *robust.* The first three adjectives here mean (more or less by definition) that the calculation procedure, the *filter,* should be of the form

$$dm_t = \tilde{\alpha}(m_t)dt + \sum_{j=1}^{r} \tilde{\beta}_j(m_t)d\tilde{\zeta}_j(y_t) \tag{1.3}$$

$$\widehat{\phi(x_t)} = \gamma(m_t, y_{1t}, \ldots, y_{pt}) \tag{1.4}$$

Here $\tilde{\alpha}, \tilde{\beta}_j, \tilde{\zeta}_j, \gamma$ are known functions and vectorfields and m_t evolves over a finite dimensional manifold (finite dimensionality); recursiveness is embodied by the fact that (1.3) is directly driven by the observations and that $\widehat{\phi(x_t)}$ only depends on the filter state m_t; and the current observations; (1.4) of course also reflects exactness. For robustness one requires that the filter equations be driven by y_t itself instead of also involving the dy_t. I.e. one requires (1.3) to be replaced by an equation

$$\frac{dm_t}{dt} = \alpha(m_t) + \sum_{j=1}^{r} \beta_j(m_t)\zeta_j(y_{1t}, \ldots, y_{pt}). \tag{1.5}$$

Thus while (1.3) is a stochastic differential equation its robust version (if it exists) (1.3) can be treated pathwise and makes sense as a family of differential equations, one for each possible observation path $\{y_t\}$.

The problem now is: given a system (1.1), (1.2) and a function ϕ how to find a filter (1.4), (1.5); i.e. how to determine the functions γ and ζ_j and vectorfields α and β_j occurring in (1.4), (1.5).

2. The DMZ filter

Under mild regularity assumptions on f,G,h and reachability and observability conditions on the system (1.1), (1.2) the conditional state $\hat{x}_t = E[x_t \,|\, y_s, 0 \leqslant s \leqslant t]$ has a density $\pi(x,t)$.

Theorem 2.1. (Duncan [2], Mortensen [6], Zakai [9]). Under appropriate regularity conditions there exists an unnormalized version $\rho(x,t)$ of $\pi(x,t)$ (i.e. $\rho(x,t) = \sigma(t)\pi(x,t)$ for some unknown function $\sigma(t)$) which satisfies the stochastic partial differential equation

$$d\rho = \mathcal{L}\rho dt + \sum_{i=1}^{p} h_i(x) dy_{it}. \tag{2.2}$$

Here $\mathcal{L}$ is the second order partial differential operator defined by

$$\mathcal{L}\psi = \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}((GG^T)_{ij}\psi) - \sum_{i=1}^{n}\frac{\partial}{\partial x_i}(f_i\psi) - \frac{1}{2}\sum_{j=1}^{p} h_j^2\psi. \tag{2.3}$$

Here G^T is the transpose of the matrix valued function G and $(GG^T)_{ij}$ is the (ij)-th entry of the matrix GG^T, f_i is the i-th component of the function f and h_j the j-th component of the function h.

The stochastic PDE (2.2) is to be regarded as a Fisk-Stratonovic stochastic PDE. To obtain the equivalent Ito version remove the term $-\frac{1}{2}\sum h_j^2\psi$ in (2.3).

Consider the time dependant gauge transformation

$$\tilde{\rho}(x,t) = \exp(-h_1(x)y_{1t} - \ldots - h_p(x)y_{pt})\rho(x,t). \tag{2.4}$$

Substituting this into (2.2) yields an equation

$$\frac{\partial\tilde{\rho}(x,t)}{\partial t} = \mathcal{L}\tilde{\rho} - \sum_{i=1}^{p} y_i(t)\mathcal{L}_i\tilde{\rho} - \sum_{i,j=1}^{p} y_i(t)y_j(t)\mathcal{L}_{ij}\tilde{\rho} \tag{2.5}$$

where

$$\mathcal{L}_i = [h_i,\mathcal{L}] := h_i\mathcal{L} - \mathcal{L}h_i, \quad \mathcal{L}_{ij} = \mathcal{L}_{ji} = \frac{1}{2}[h_i,[h_j,\mathcal{L}]]. \tag{2.6}$$

Given $\phi(x)$ and $\tilde{\rho}(x,t)$ the best estimate $\widehat{\phi(x_t)}$ can be calculated by

$$\rho(x,t) = \exp(h_1(x)y_{1t} + \ldots + h_p(x)y_{pt}) \tag{2.7}$$

$$\widehat{\phi(x_t)} = (\textstyle\int \rho(x,t)dx)^{-1}\int\phi(x)\rho(x,t)dx. \tag{2.8}$$

Note that (2.5) together with the output map (2.7), (2.8) is a recursive, exact and robust filter. The only trouble with it (from the calculation point of view) is that it is infinite dimensional.

3. Wei-Norman theory [8].

For the moment let us consider control systems of the form

$$\dot{x} = u_1A_1x + \ldots + u_kA_kx, \quad x \in \mathbb{R}^n \tag{3.1}$$

where the A_i are $n \times n$ matrices and the u_i are inputs (known functions of time). Adding a few more terms (with $u_j = 0$, $j > k$) we may as well assume that $A_1, \ldots, A_k$ are a basis of a Lie algebra of $n \times n$ matrices (under the commutator difference product $[A,B] = AB - BA$). Let us look for solutions of the form

$$x(t) = e^{g_1 A_1} \dots e^{g_k A_k} x(0) \tag{3.2}$$

where the $g_i(t)$ are still to be determined functions of time. By differentiating (3.2), inserting $\exp(-g_1 A_1) \cdots \exp(-g_i) \exp(g_i A) \cdots \exp(g_1 A)$ just after $\dot{g}_{i+1} A_{i+1}$ in the result, using the Baker-Cambell-Hausdorff formula, using (3.1) and collecting terms, one finds a set of equations

$$\dot{g}_i + \sum_{j=1}^{k} \dot{g}_j h_{ji}(g_1, \dots, g_k) = u_i, \quad i = 1, \dots, k \tag{3.3}$$

with $h_{ij}(0, \dots, 0) = 0$ and the following properties of the $h_{ij}(g_1, \dots, g_k)$:

$$h_{ij} \text{ only involves } g_1, \dots, g_{i-1} \tag{3.4}$$

and if $A_{l+1}, \dots, A_k$ are a basis of an ideal or $\mathfrak{a} \subset \mathfrak{g}$ (so that $[A_i, \mathfrak{a}] \subset \mathfrak{g}$ for all i) then

$$h_{ji} = 0 \text{ for } i = i, \dots, l;\ j = l+1, \dots, k \tag{3.5}$$

so that the equations for $g_1, \dots, g_l$ do not involve $g_{l+1}, \dots, g_k$ at all. It is also important to note that the h_{ij} are universal functions depending only on the Lie algebra $\mathfrak{g}$ and the chosen basis and totally independent of the particular matrix realization (representation) we may be dealing with. In particular if $\mathfrak{a}$ is an ideal of $\mathfrak{g}$ and $A_1, \dots, A_k$ is a basis as above then

$$\text{equations for } g_1, \dots, g_l \text{ only depend on } \mathfrak{g} / \mathfrak{a}. \tag{3.6}$$

In case that $\mathfrak{g}$ is nilpotent (or more generally solvable) equations (3.3) therefore take a particularly pleasant triangular form which can be solved just using quadratures. Indeed if L is nilpotent, so that

$$L \underset{\neq}{\supset} [L, L] = L_2 \underset{\neq}{\supset} [L, L_2] = L_3 \underset{\neq}{\supset} \cdots \underset{\neq}{\supset} [L, L_r] = L_{r+1} = 0$$

and if we choose a basis

$$A_1, \dots, A_{k_1}, A_{k_1+1}, \dots, A_{k_2}, \dots, A_{k_{r-1}+1}, \dots, A_{k_r}, \quad k_r = k$$

such that

$$A_{k_{i-1}+1}, \dots, A_{k_i}, \quad k_0 = 0$$

is a basis for L_i, $i = 1, \dots, r$, then the equations take the form

$$\begin{aligned}
\dot{g}_1 &= u \\
&\dots \\
\dot{g}_{k_1} &= uk_1 \\
\dot{g}_{k_1+1} &= u_{k_1+1} + \alpha_{k_1+1}(u_1, \dots, u_{k_1}; g_1, \dots, g_{k_1} \\
&\dots \\
\dot{g}_{k_2} &= u_{k_2} + \alpha_{k_2}(u_1, .., u_{k_1}; g_1, \dots, g_{k_1}) \\
\dot{g}_{k_2+1} &= u_{k_2+1} + \alpha_{k_2+1}(u_1, \dots, u_{k_2}; g_1, \dots, g_{k_2}) \\
&\dots \\
\dot{g}_{k_3} &= u_{k_3} + \alpha_{k_3}(u_1, \dots, u_{k_2}; g_1, \dots, g_{k_2}). \\
&\dots
\end{aligned} \tag{3.6}$$

Now note that the robust DMZ filter equation (2.5) is of the form (3.1) except that it takes place in a function space. So in particular if the Lie algebra generated by the operators $\mathfrak{L}, \mathfrak{L}_i, \mathfrak{L}_{ij}$ in (2.5) is nilpotent (solvable) and finite dimensional with basis $A_1, \dots, A_k$ and we have given an initial density $\rho_0(x)$ and function ϕ then equations (3.6) together with the output equation

$$(g_1,\ldots,g_k)\mapsto\tilde{\rho}(x,t) = \exp(g_1A_1)\cdots\exp(g_kA_k)\rho_0(x)$$
$$\tilde{\rho}(x,t)\mapsto\rho(x,t) = \exp(h_1(x)u_1)\cdots\exp(h_p(x)y_p)\tilde{\rho}(x,t)$$
$$\widehat{\phi(x)} = (\int\rho(x,t)dx)^{-1}\int\phi(x)\rho(x,t)dx$$

constitute a recursive exact robust filter for $\widehat{\phi(x_t)}$. It is not really finite dimensional because the A_i here are operators and calculating $\exp(g_iA_i)$ (for known $g_i(t)$) amounts to solving $\frac{d}{dt}B_i = \dot{g}_iA_iB_i$, $B_0 = id$ which is again a partial differential equation.

4. The identification case

The problem of identifying a linear system

$$dx_t = Ax_tdt+Bdw_t,\ \ dy_t = Cx_t+dv_t \tag{4.1}$$

i.e. the problem of determining the unknown matrices A,B,C on the basis of the observations, can be viewed as a nonlinear filtering problem for the system with state vector (x,A,B,C) obtained by adding the equations $dA=0$, $dB=0$, $dC=0$ to (4.1). It can be proved that the Lie algebra generated by the $\mathfrak{L},\mathfrak{L}_i,\mathfrak{L}_{ij}$ in this case is topologically solvable. I.e. there is a sequence of ideals $\mathfrak{a}_i$ such that $\mathfrak{g}/\mathfrak{a}_i$ is finite dimensional solvable for all i and $\bigcap_i\mathfrak{a}_i = \{0\}$. Because of (3.6) this yields a sequence of approximate filters via

$$e^{g_1A_1}\cdots e^{g_1A_{k_1}}\rho_0,\ \ e^{g_1A_1}\cdots e^{g_{k_2}A_{k_2}}\rho_0,\ \cdots$$

where $A_1,\ldots,A_{k_1},A_{k_1+1},\ldots,A_{k_2},\cdots$ are such that the equivalence classes of $A_1,\ldots,A_{k_r}$ mod $\mathfrak{a}_r$ are a basis for $\mathfrak{g}/\mathfrak{a}_r$. Cf [5] for more details.

5. Nilpotent and solvable approximations

However, in many cases, the Lie algebra generated by $\mathfrak{L},\mathfrak{L}_i,\mathfrak{L}_{ij}$ will not be topologically solvable. For instance in the case of perturbed linear systems

$$dx = (Ax+\epsilon P_A(x)dt+(B+\epsilon P_B(x)dw_t,\ \ dy = (C+\epsilon P_C(x)dt+dv_t \tag{5.1}$$

where the $P_A(x)$, $P_B(x)$, $P_C(x)$ are polynomial higher order disturbances. In this case the Lie algebra tends to be $W_n = \mathbb{R}<x_1,\ldots,x_n;\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n}>$, the Lie algebra of all differential operators (any order) with polynomial coefficients. In this case the higher order operations come with higher powers of ϵ in the sense that

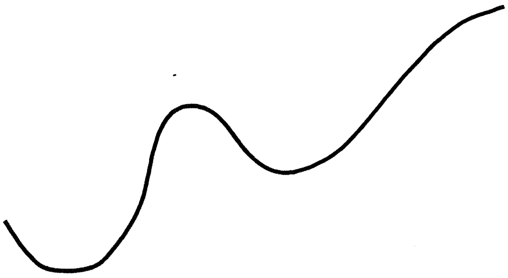
$$\mathrm{Lie}(\mathfrak{L},\mathfrak{L}_i,\mathfrak{L}_{ij})\bmod\epsilon^n\text{is finite dimensional for all } n \tag{5.2}$$

(and these algebras are solvable). Again there result approximate filters and they seem to perform well [3,4]. Still more generally there is no small parameter at all, but there still is a natural gradation structure on the Lie algebra. To see why this might be the case and why this will give us possibilities for constructing approximate filters observe that the operators $\mathfrak{L},\mathfrak{L}_i,\mathfrak{L}_{ij}$ are of the general forms

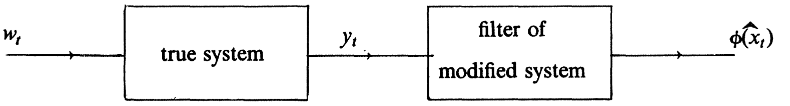
$$\mathfrak{L} = \sum a_{ij}\frac{\partial^2}{\partial x_i\partial x_j}+\sum b_j\frac{\partial}{\partial x_j}+c$$
$$\mathfrak{L}_i = \sum d_{ij}\frac{\partial}{\partial x_j}+e_i$$
$$\mathfrak{L}_{ij} = f_{ij}$$

where the $a_{ij},b_{ij},f_{ij},e_i,c$ are explicit functions of the G_{ij},f_i,h_j and their derivatives. Commuting various $\mathfrak{L}$'s brings at least one derivative of the G_{ij},f_i,h_j in each term, third order brackets bring second order derivatives or products of first order derivatives, ...

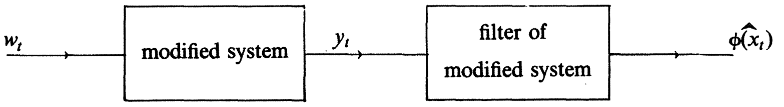
Now if the system described by the f_i, h_j, G_{ij} is supposed to model some real world phenomenon then we can not assume that we know these functions perfectly. In general one would expect that the values of the functions would be known very well, their derivatives less so, their second derivatives still less, etc., and by the time r-th derivatives come into play their values are almost totally unknown.



For $r=2$ the kind of approximation involved is somewhat like illustrated on the above, i.e. something like a piecewise linear approximation with rounded corners. One expects a system close to real one in this sense of diminishing importance of higher derivatives (globally) to behave much like the true one. The comulative effect of small inaccuracies in first derivatives, larger ones in second derivatives, ..., very large ones in r-th derivatives will be such that r order brackets are almost totally unknown. And thus a system approximation which just happened to have all these zero would perform much as the original one but that one would have a filter as in section 3 above and this filter should also give reasonable results for the true system by considering the stability properties of the composed system



which is close to the system with exact filter



Now such a modified system which just happens to have all terms in r-th order brackets of the $\mathcal{L}, \mathcal{L}_i, \mathcal{L}_{ij}$ equal to zero will probably not as a rule exist. But the corresponding filters can certainly be constructed. It suffices to introduce a counting mechanism and to consider the Lie algebra generated by the operator $z\mathcal{L}, z\mathcal{L}_i, z\mathcal{L}_{ij}$. This one is topologically nilpotent and so Wei-Norman theory can be applied to Lie $(z\mathcal{L}, z\mathcal{L}_i, z\mathcal{L}_{ij})$ mod z^n for all n (after which one sets $z=1$.) Here z is an extra parameter.

The argument above indicates that such a procedure could work well. Another not unrelated argument can be based on Volterra series expansions. These ideas have of course a good deal to do with nilpotent and solvable approximation ideas [1], [7].

REFERENCES

1. P.E. CROUCH, *Solvable approximation to control systems,* SIAM J. Control and Opt. 32 (1984), 40-54.
2. T.E. DUNCAN, *Probability densities for diffusion processes with applications to nonlinear filtering,* Ph.D. thesis, Standord, 1967.
3. M. HAZEWINKEL, *On deformations, approximations and nonlinear filtering,* Systems and Control Lett. 1 (1982), 32-36.
4. M. HAZEWINKEL, *Lie algebraic methods in filtering and identification,* Report PM-R86-6, November 1986, CWI, Amsterdam; to appear in Proc. 8-th Int. Symp. IAMP (Luminy, 1986), World Scientific and in Proc. 1-st World Congress Bernouilli Society (Taskent, 1986), VNU Science Press.
5. P.S. KRISHNAPRASAD, S.I. MARCUS, M. HAZEWINKEL, *Current algebras and the identification problem,* Stochastics 11 (1983), 65-101.
6. R.E. MORTENSEN, *Optimal control of continuous time stochastic differential equations,* Ph.D. thesis, Berkeley, 1966.
7. CH. ROCKLAND, *Intrinsic nilpotent approximation,* preprint MIT, LIDS-R-1482, 1985; to appear Acta Appl. Math., 1987.
8. J. WEI, E. NORMAN, *On the global representation of the solutions of linear differential equations as products of exponentials,* Proc. AMS 15 (1964), 327-334.
9. M. ZAKAI, *On the optimal filtering of diffusion processes,* Z. Wahrsch. und verw. Gebiete 11 (1969), 230-243.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

The Kalman Filter in Dendroclimatology

J. Molenaar
Mathematics Consulting Department
Catholic University
Toernooiveld, 6525 ED Nijmegen, The Netherlands
and
H. Visser
Research and Development Division
N.V. KEMA
P.O. Box 9035, 6800 ET Arnhem, The Netherlands

Abstract

The extended Kalman filter is used to study the response of trees to the weather conditions as reflected in tree-ring series.In contrast to the traditional multiple regression models the present approach allows for the detection of time-dependent variations in tree response. These changes may be of a natural origin (e.g. ageing) or due to anthropogenic influences (e.g. environmental pollution). An essential feature of the method is the simultaneous estimation of both trends and weather contributions . As an example a ring width series of an European Silver Fir is analyzed, leading to the conclusion that this tree died because of competition rather than of pollution effects.

§1. Introduction

Only the last decade it is quite commonly realized that the poor condition of considerable parts of the forests in Europe might be disastrous for the quality of future life. Nowadays, one takes for granted that environmental pollution, and in particular the phenomenon of "acid rain", is one of the main causes of the menacing catastrophe. In spite of impressive research exertion, the involved mechanisms are not yet fully understood. This is, of course, partly due to the fact that most laboratory experiments are not representative for the open field.

Already long before "acid rain" became a topic in environmental sciences, dendroclimatologists studied the relations between weather conditions and tree growth. Statistical techniques are frequently used in this discipline after the exploring and innovative work by Fritts (1976). Most research in this field concentrates on the regression of tree ring data on weather data. Ring width data reflect in an easily measurable way the growth and thus health of trees in history. They can be considered as realizations of a stochastic process, superposed on an age related trend, an environment related trend, and a weather signal. It is to be expected that environmental pollution manifests itself both in the age related trend and the weather signal, because it may lead to an overall reduction of ring width and a changing sensitivity of the tree to the weather . Separation of these effects is of particular importance in all regression methods, but on this point many questions are still unanswered.

Current techniques to remove trends from tree ring series are fitting by polynomials, splines or a negative exponential curve, application of high- and low-pass filters and ARIMA modelling (see e.g. [2],[3],[5]). Competition effects are usually diminished by averaging several series from one stand. A common shortcoming of these methods is the presence of subjective elements depending on the experience and insights of the researcher, although improvements have been made [15].

Current techniques to analyze the climatic part of ring width series are based on linear regression models with constant coefficients. The restriction of time independence seriously obstructs the study of air pollution effects, because changes in tree response likely contain essential information, see e.g. [12] and [16].

In this paper we present and investigate a method to estimate trends and weather response on the same footing, at the same time allowing for time dependent coefficients. Key techniques are Kalman filtering and maximum likelihood estimation, which have proven to be successful in numerous other applications, e.g. econometrics ([6],[7]). The mathematical model to be used and the relevant filtering formulae are given in §2. In §3 we point out how tree growth can be modelled within this context. An example of a rather complicated ring width series is given in §4. In the last section we discuss some general implications of the present approach as far as selection of variables ,modelling of the trend, and use of principal components is concerned.

§2 The Univariate Kalman Filter

Here we recapitulate the main features of the Kalman filter technique. In view of the applications we are aiming at, we do not present the filter to its full extent. For example, only the discrete and univariate case, i.e. one observed quantity, is dealt with. For derivations of the formulae and an overview of the relevant ideas we refer to [1], [6], [7], [9], [10], [11], [13], and [14].

2a. State Space Formulation.

The system to be studied is assumed to be characterized by a stochastic vector $\mathbf{a}_t$ of dimension M, say. This state vector is not directly observable and to be estimated from successive observations of a measurable quantity y_t. The relationship between $\mathbf{a}_t$ and y_t is taken to be linear and given by the measurement equation

$$y_t = \mathbf{z}_t^T \mathbf{a}_t + u_t + v_{y,t}, \tag{1}$$

with $\mathbf{z}_t$ the vector of explanatory variables, $v_{y,t}$ the observation error and u_t a known input signal representing an external influence. From (1) the vector of response variables,-or state vector $\mathbf{a}_t$ is seen to act as the vector of regression coefficients in a conventional linear regression model. The essential difference is given by the time dependence of $\mathbf{a}_t$. This dynamic feature is assumed to be governed by the transition equation

$$\mathbf{a}_t = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{v}_{a,t} \tag{2}$$

with $\mathbf{T}$ the known $M \times M$ transition matrix.

The disturbances $v_{y,t}$ and $\mathbf{v}_{a,t}$ are taken to be serially uncorrelated. They have zero mean and respective variances R_t and $\mathbf{Q}_t$. For all t, they are uncorrelated with each other and with the initial state $\mathbf{a}_0$. The filter consists of an iterative scheme to calculate the minimum mean square estimate $\mathbf{a}_{t/t'}$ of $\mathbf{a}_t$, based on the information contained in the y_t, $t=1,\ldots,t'$. The covariance matrix of $\mathbf{a}_{t/t'}-\mathbf{a}_t$ is denoted by $\mathbf{P}_{t/t'}$. The cases $t>t'$, $t=t'$ and $t<t'$ correspond with prediction, filtering and smoothing respectively.

2b. Prediction and Filtering

The prediction formulae assume the filtered quantities $\mathbf{a}_{t-1/t-1}$ and $\mathbf{P}_{t-1/t-1}$ to be known:

$$\begin{aligned} \mathbf{a}_{t/t-1} &= \mathbf{T}_t \mathbf{a}_{t-1/t-1}, \\ \mathbf{P}_{t/t-1} &= \mathbf{T}_t \mathbf{P}_{t-1/t-1} \mathbf{T}_t^T + \mathbf{Q}_t, \end{aligned} \tag{3}$$

The one-step-ahead prediction errors or innovations ν_t, defined by

$$\nu_t = y_t - \mathbf{z}_t^T \mathbf{a}_{t/t-1} - u_t, \tag{4}$$

play a central role in the theory. They are serially uncorrelated, have zero mean, and variance f_t given by

$$f_t = \mathbf{z}_t^T \mathbf{P}_{t/t-1} \mathbf{z}_t + R_t \; . \tag{5}$$

The filtered estimates follow from the predicted ones via the equations

$$\mathbf{a}_{t/t} = \mathbf{a}_{t/t-1} + \mathbf{P}_{t/t-1}\mathbf{z}_t \mathrm{v}_t / f_t ,$$
$$\mathbf{P}_{t/t} = \mathbf{P}_{t/t-1} - \mathbf{P}_{t/t-1}\mathbf{z}_t \mathbf{z}_t^T \mathbf{P}_{t/t-1} / f_t \; . \tag{6}$$

To start the scheme the initial values $\mathbf{a}_{0/0}$ and $\mathbf{P}_{0/0}$ have to be specified. If exact information is available the estimates $\mathbf{a}_{t/t}$ are unbiased. In general this is not the case. However, as shown by Jazwinski (1970), the prior data are eventually forgotten and a bias stemming from initial uncertainties damps out after sufficient observations having been processed. So, in practice, it suffices to choose $\mathbf{a}_{0/0}$ arbitrarily and $\mathbf{P}_{0/0}$ large. The first, say N_s, iteration steps then serve as a transient period, in which the filter itself constructs appropriate starting values for the rest of the process.

2c. **Smoothing**

Once $\mathbf{a}_{t/t}$ has been estimated for $t=1,\ldots,N$, these estimates can be smoothed using all information instead of only the foregoing y_t values. The smoothing procedure works backwards :

$$\mathbf{a}_{t/N} = \mathbf{a}_{t/t} + \mathbf{P}_t^*(\mathbf{a}_{t+1/N} - \mathbf{T}_{t+1}\mathbf{a}_{t/t}),$$
$$\mathbf{P}_{t/N} = \mathbf{P}_{t/t} + \mathbf{P}_t^*(\mathbf{P}_{t+1/N} - \mathbf{P}_{t+1/t})\mathbf{P}_t^{*T},$$
$$\mathbf{P}_t^* = \mathbf{P}_{t/t}\mathbf{T}_{t+1}^T \mathbf{P}_{t+1/t}^{-1} \; . \tag{7}$$

Contrary to the filtered estimates, the smoothed quantities are not sensitive to transient phenomena and are thus reliable also for $t<N_s$. It can be shown (e.g. Otter (1978)) that when $\mathbf{Q}_t = 0$, the Kalman estimate $\mathbf{a}_{t/N}$ and $\mathbf{P}_{t/N}$ are independent of t and equal to the ordinary least squares (OLS) estimates. So the OLS fitting procedure is a special case of the Kalman filter .

2d. **Optimality and Maximum Likelihood Estimation**

The estimate $\mathbf{a}_{t/t'}$ is the best linear estimator in the sense that the error covariance matrix of any other linear estimator exceeds the Kalman $\mathbf{P}_{t/t'}$ by a positive definite matrix. If, in addition, $v_{y,t}$,$\mathbf{v}_{a,t}$ and $\mathbf{a}_0$ are normally distributed, then $\mathbf{a}_{t/t'}$ is the best, in the sense of minimum variance, of all possible estimates.

At the start of the filtering process the transition matrix $\mathbf{T}_t$ and the disturbance variances R_t and $\mathbf{Q}_t$ are, generally, unknown. In the present univariate case there is no need to estimate R_t and $\mathbf{Q}_t$ separately. The Kalman filter formulation only depends on the quotient $\mathbf{Q}_t/R_t$ and it suffices to choose $R_t \equiv 1$. If $v_{y,t}$,$\mathbf{v}_{a,t}$, and $\mathbf{a}_0$ are normally distributed, a convenient way to estimate the unknown parameters is to follow the maximum likelihood approach. The likelihood function L of the observations y_t for $t>N_s$ is given by the so-called prediction error decomposition [6] :

$$-2 \log L = (N-N_s)\log 2\pi + \sum_{t=N_s+1}^{N} [\log f_t + \mathrm{v}_t^2/f_t]. \tag{8}$$

As discussed above, the determination of N_s is in most cases trivially obtained from inspection (see e.g. [17]). In this approach the unknown parameters are found by maximizing log L as a function of $\mathbf{T}_t$, R_t and $\mathbf{Q}_t$.

The transition matrix $\mathbf{T}_t$ can also alternatively be estimated by considering its elements as an intrinsic part of the state vector $\mathbf{a}_t$. This method merely requires a convenient extension of the original state vector and a straightforward redefinition of the transition matrix as used in the formulae (3)-(7).

While the maximum likelihood approach may require a considerable number of filter evaluations, the latter method may imply manipulation of large matrices. Therefore, to gain computational speed and simplicity it is always desirable to pose, in advance, restrictions on the dimensions and parameters of the model. The effects of inaccurate modelling can be studied following an analysis by Jazwinski (1970), but will be omitted here in view of the restricted purpose of this paper.

§3. Modelling Tree Growth

In this section we show how application of the Kalman filter may enrich the conventionally used analysis of tree growth. In dendroclimatology ring width series d_t are usually modelled as the product of a trend g_t and a stochastic signal w_t. The series g_t and w_t are usually referred to as the growth curve and the tree-ring index respectively. It is common practice to estimate g_t first , after which the quotient d_t/g_t is analyzed. In view of the difficulties in separating these effects we prefer to consider both contributions as being stochastic in nature and to be estimated on an equal footing. So we adopt the following model :

$$d_t = g_t(1+\bar{\mathbf{z}}_t^T\bar{\mathbf{a}}_t+\bar{v}_{y.t}) \equiv g_t w_t. \tag{9}$$

The elements of $\bar{\mathbf{z}}_t$ represent weather data. They are standardized to have zero mean and unit variance. To complete the model we have to specify the dynamic behaviour of g_t and $\bar{\mathbf{a}}_t$. Hardly any biological information is available at this point. Therefore, we describe the stochastic nature of $\bar{\mathbf{a}}_t$ as a random walk process, i.e.

$$\bar{\mathbf{a}}_{t+1} = \bar{\mathbf{a}}_t + \mathbf{v}_{\bar{a}.t} \ . \tag{10}$$

The trend is modelled as a local trend in which the level g_t and slope s_t vary slowly in time, both driven by random walk processes, i.e. (cf. Harvey (1984))

$$\begin{aligned} g_{t+1} &= g_t + s_t + v_{g.t}, \\ s_{t+1} &= \qquad s_t + v_{s.t} \ . \end{aligned} \tag{11}$$

In equations (10) and (11) the disturbances are assumed to be normally and independently distributed with zero mean. To estimate g_t and $\bar{\mathbf{a}}_t$ via Kalman filtering equation (9) has to be linearized using the local approximation

$$d_{t+1} = g_{t+1}w_t + g_t w_{t+1} - g_t w_t \ . \tag{12}$$

Now the right-hand-side of (9) is in the form of equation (1) if we identify

$$\begin{aligned}
\mathbf{a}_{t+1}^T &= ((\bar{\mathbf{a}}_{t+1})^T,\ g_{t+1},\ s_{t+1}),\\
\mathbf{z}_{t+1}^T &= ((g_t\bar{\mathbf{z}}_{t+1})^T,\ w_t,\ 0),\\
u_{t+1} &= -g_t(\bar{\mathbf{z}}_t^T\bar{\mathbf{a}}_t+\bar{v}_{y.t}),\\
v_{y.t+1} &= g_t\bar{v}_{y.t+1}\ .
\end{aligned} \tag{13}$$

So the first N elements of $\mathbf{a}_t$ and $\mathbf{z}_t$ correspond with the weather parameters and the last two ones with the trend. The transition matrix $\mathbf{T}_t$ in equation (2) is given by the $(N+2)\times(N+2)$ unit matrix with an additional one at entry $(N+1,N+2)$. The disturbance $\mathbf{v}_{a.t}$ is in the same manner defined by

$$\mathbf{v}_{a.t}^T = ((\mathbf{v}_{\bar{a}.t})^T,\ v_{g.t},\ v_{s.t}) \tag{14}$$

Because monthly averaged weather data show little correlation, we used for $\mathbf{Q}_t$, the covariance matrix of $\mathbf{v}_{a.t}$, a diagonal matrix. To reduce the required computer time needed in the optimization procedure described in section 2.d, $\mathbf{Q}$ is taken to be constant over time.

From equations (13) it is seen that at time $t+1$ the quantities $\mathbf{z}_{t+1}, u_{t+1}$, and $\mathbf{v}_{y.t+1}$ are constructed from known input data at time $t+1$ and parameters estimated at time t. In this so-called extended Kalman filter approach no optimality and even no convergence of the filter is guaranteed [1]. How valuable the estimates still are clearly depends on the reliability of approximation (12). In tree ring analysis no serious problems on this point arise because strong variations in the trend or the weather response are not expected to occur within very short periods .

§4. **Example**

As an example we analyze data from a European silver fir (Abies alba Mill.) cut in Wörth (G.F.R.). This ring width series runs over the period 1900-1977 and is plotted in figure 1. As explanatory variables we use monthly averaged temperature and precipitation data, running from May prior to the year of growth through August in the year of growth. In order to reduce this considerable number (32) of variables a selection procedure is used which is discussed in §5. In this way the May, July and September precipitations prior to the year of growth and the January precipitation together with the February temperature in the year of growth were found to be significantly related with ring width . The corresponding variances or diagonal values of the $\mathbf{Q}$ matrix, obtained by maximum likelihood estimation described in §2.d, are 0.4300, 0.0022, 0.0000, 0.0016 and 0.0000 respectively. The estimated trend is given by the solid line in figure 1. The estimated response parameters are drawn in figure 2. The results in figure 2 show the following interesting features:

- The May precipitation parameter varies wildly. The explaining value of this variable is rather small, as can be concluded from its confidence limits. This strongly suggests that the established correlation with tree growth is rather fortuitous than of a biological origin.
- The other parameters hardly show any time dependence. In particular they do not reflect the decline in growth after about 1960 as being present in the trend in figure 1. This suggests competition rather than illness as cause of this fall.
- All parameters precede the growing season, usually running from March through September. The weather conditions during this period are apparently not influencing growth.
- The precipitation in the preceding summer is clearly of great importance, probably because then the tree lays in the necessary nutritious matter. An outstanding example is the impressive growth in 1959, thanks to the very wet summer of 1958.
- The negative signs of the January and February parameters indicate that a moist and relatively warm winter are highly unfavourable to tree growth. These conditions might initiate the growth too early. Extreme examples are in 1929 and 1956 (see figure 1) : these narrow rings coincide with the highest February temperatures of this century.

From figure 3 it can be seen how the data are fitted by the present model. The solid curve is calculated by substituting the trend and parameter estimates in the right-hand side of equation (9). The relatively small model explains the data remarkably well: the variance of the original ring series is reduced with 79%.

§5. **Concluding Remarks.**

We conclude with some remarks and suggestions on the present and future use of the Kalman filter in regression analysis .

5a. **Modelling the Trend**

In model (9) the trend is introduced in a multiplicative way, as is usual in standardizing raw tree ring data ([3] and [4]). This choice may be doubted, because in many cases the amplitudes of fluctuations are not exactly proportional to the trend level. Therefore, further analysis on this assumption may be crucial for future research. We intend to use an additive trend model, based on annual increments of basal area instead of ring width .

5b. Selection of Explanatory Variables

For regression models with constant parameters several selection procedures are available to reduce the number of explanatory variables see e.g. Thompson (1978) and Hughes et al (1982). In the present context an alternative procedure is needed such as e.g. described in [17] and [18]. This procedure is also used in the example of section 4. It contains three fitting parameters to be adjusted by computer simulation. Although the results are satisfactory, it still has the disadvantage of selecting isolated variables. For example, in section 4 the July precipitation is selected on its own, whereas the influence of June and August precipitations are fully ignored. From a biological point of view refinement of the procedure is desirable.

5c. Principal Components

In most applications of multiple regression analysis the explanatory variables are not independent. In those cases one usually resorts to the application of the principal components technique in order to avoid numerical inaccuracies or even collapse of the regression method. Application of this technique in the context of the present approach is still rather straightforward as pointed out in [17].

Acknowledgement

The authors want to thank D. Eckstein from Hamburg University for putting at their disposal the data analyzed in section 4. This work has been supported by the Ministry of Housing, Physical Planning and Environment, the Ministry of Agriculture and Fisheries, the Ministry of Economic Affairs, Oil refineries and the co-operating Electricity Boards.

References

1 Bagchi, A, *Stochastic filter and identification theory,* University Twente report nr. 57016, 1982.

2 Briffa, K.R., Wigley, T.M.L., and Jones, P.D., *Towards an objective approach to standardization,* Paper presented at the Task force meeting on tree-ring analysis, Krakow, Poland, 1986.

3 Fritts, H.C., *Tree rings and climate,* Academic Press, New York, 1976.

4 Graybill, D.A., *Chronology development and analysis,* in [8].

5 Guiot, J., *ARMA techniques for modelling tree-ring response to climate and for reconstructing variations of paleoclimates,* Ecol. Modelling Vol. 33, pp. 149-171, 1986.

6 Harvey, A.C., *Time series models,* Philip Allan Publishers Limited, Oxford, 1981.

7 Harvey, A.C., *A unified view of statistical forecasting procedures,* J. of Forecasting, Vol. 3, pp. 245-275, 1984.

8 Hughes, M.K., Kelly, P.M., Pilcher, J.R. and Lamanche, V.C., *Climate from tree rings,* Cambridge University Press, 1982.

9 Jazwinski, A.H., *Stochastic processes and filtering theory,* Academic Press, New York, 1970.

10 Kalman, R.E., *A new approach to linear filtering and prediction problems,* Trans. ASME, series D, Vol.82, pp.35-45, 1960.

11 Kwakernaak, H. and Sivan, R., *Linear optimal control systems,* Wiley, New York, 1972.

12 McClenahen, J.R., and Dochinger, L.S., *Tree-ring response of white oak to climate and air pollution near the Ohio river valley,* J. Environmental Quality Vol. 14, pp. 274-280, 1985

13 Otter, P.W., *The discrete Kalman filter applied to linear regression models: statistical considerations and an application,* Statistica Neerlandica, Vol. 32, pp. 41-56, 1978.

14 Sage, A.P. and Melsa, J.L., *Estimation theory with applications to communications and control,* McGraw Hill, New York, 1971.

15 Thompson, M.L., *Selection of variables in multiple regression,* Inst. Stat. Rev., Vol. 46, pp. 1-19 and 129-146, 1978

16 Visser, H., *Analysis of tree-ring data using the Kalman filter technique,* IAWA bulletin nr. 7(4), pp. 289-297, 1986.

17 Visser, H. and Molenaar, J., *Time dependent responses of trees to weather variations: an application of the Kalman filter,* KEMA report 50385-MOA 86-3041, Arnhem, 1986.

18 Visser, H. and Molenaar, J., *Kalman filter analysis in Dendroclimatology,* Biometrics, 1987 (submitted).

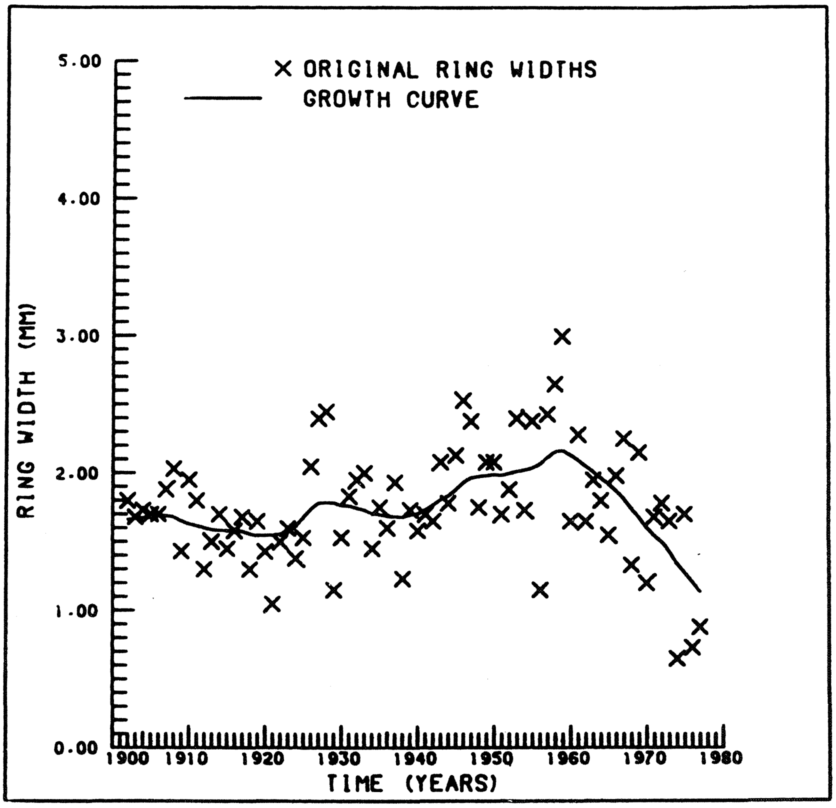
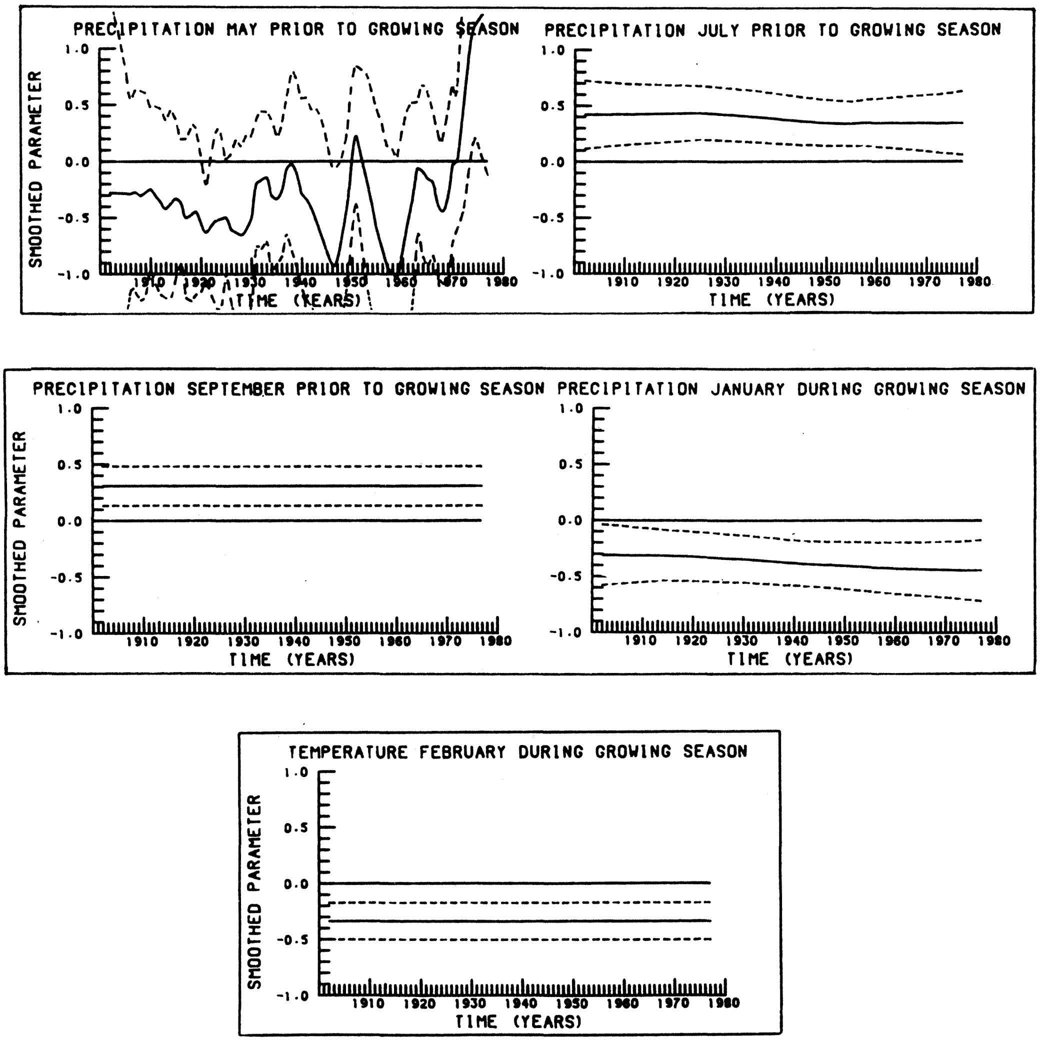


Figure 1. Ring width series in mm of a European silver fir (Abies alba Mill.), cut in Wörth (GFR) in 1978. The length of the series $N = 78$. The solid line represents the trend g_t or growth curve as estimated by the model and filtering methods explained in the text.



PRECIPITATION MAY PRIOR TO GROWING SEASON

PRECIPITATION JULY PRIOR TO GROWING SEASON

PRECIPITATION SEPTEMBER PRIOR TO GROWING SEASON

PRECIPITATION JANUARY DURING GROWING SEASON

TEMPERATURE FEBRUARY DURING GROWING SEASON

Figure 2. Response coefficients $a_{t/N}$ of the five weather parameters which appeared to be significantly correlated with the time series in figure 1. The variances of the corresponding disturbances in equation (2) (diagonal elements of Q) were 0.4300, 0.0022, 0.0000, 0.0016, and 0.0000 for the May, July, September, January and February parameters respectively. The dashed lines represent 95 percent confidence limits.

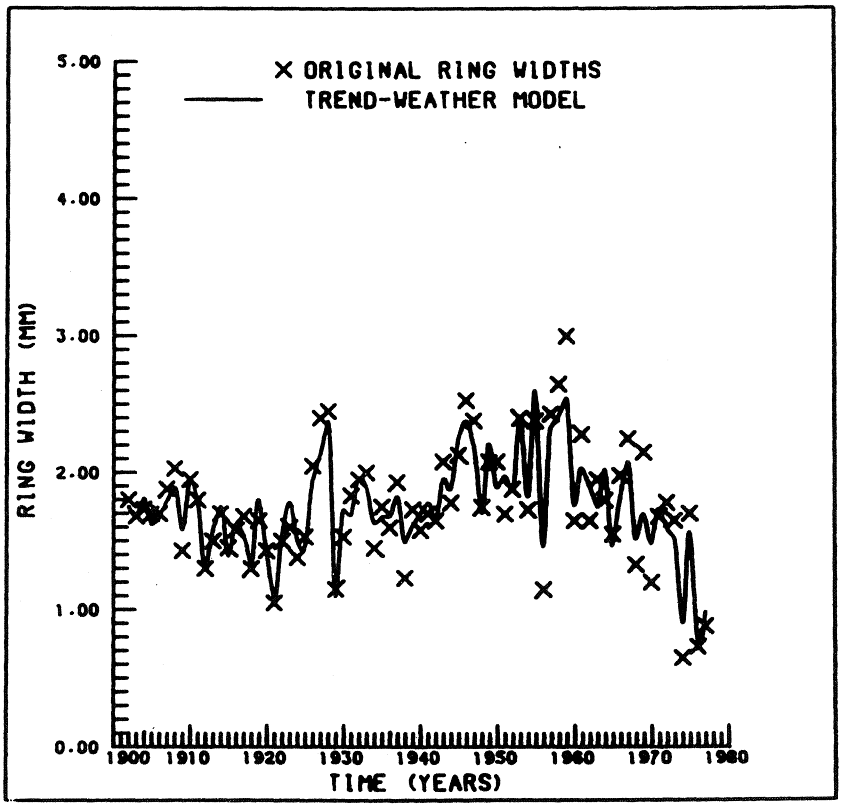


Figure 3. The crosses are as in figure 1. The solid curve represents the model fit. This fit is obtained by substitution of the estimated trend g_t and parameters $a_{t/N}$ into the right-hand-side of equation (9).

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Some Features of the Compact Disc Digital Audio System

J.B.H. Peek
Philips Research Laboratories
P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands

ABSTRACT

The following aspects of the Compact Disc digital audio system will be described: the laser optical scanning of the disc, the use of control and display data and the insensitivity to dust, scratches and fingerprints compared with the normal gramophone. One of the reasons of less sensitivity to imperfections is the use of two Reed-Solomon error correcting codes and the application of interleaving in order to deal with burst errors. A further important characteristic that will be described is that the audio signals are digitally recorded on the disc.

INTRODUCTION

The Compact Disc (CD) digital audio system has brought about a revolution in the audio world. The system is a unique combination of a number of advanced techniques, offering new capabilities to the listener and a deaper enjoyment of music. In further developments of this system use is made of mathematical methods and insights which are described in the two following articles. This introductory article outlines, in simple terms, some of the principal features of the Compact Disc digital audio system.

1. DIMENSIONS

The most striking feature is the compactness of the metallized CD disc (photo 1). The outer diameter of the disc is only 12 cm. With one hand the disc can therefore be removed from the cassette and inserted in the CD player. Because of the small dimensions of the disc the player can also be of compact construction.

2. LASER OPTICAL SCANNING

As opposed to the conventional gramophone, there is no mechanical contact between the disc and the pick-up. The laser beam used to read out the information does causes no wear on the disc. Figure 1 shows a schematic representation of the disc and the optical read-out system. The digital information is recorded on the Compact Disc in the form of a spiral track consisting of a succession of pits. The intervals between the pits are known as 'lands'. Each pit and each land represents a series of bits called channel bits.

Photo 2 shows a micrograph of part of the surface of the information layer in the Compact Disc. The information density on the Compact Disc is exceptionally high. The minimum length of a pit and of the land between two pits is 0.9 micron; the maximum length is 3.3 micron. The width of the beam as it scans the optical disc (i.e. the diameter of the scanning light-spot) is only 1 micron. The tangential bit density is about 0.9 µ/bit. The distance between two tracks (the pitch) is 1.6 micron.

The tracks are optically scanned by the laser beam from below the disc. The laser beam is generated by a solid-state laser, shown at the bottom of fig.1, the light from which is passed by a half-silvered mirror and then focussed by an objective lens on to the information layer in the Compact Disc. When the spot falls on an interval between two pits (a land) the light is almost totally reflected and is reflected by the half-silvered mirror and reaches the photodiode shown on the left in the figure. When the spot falls in a pit, the depth of which is about one quarter of the wavelength of the light in the plastic material, interference and extinction cause less light to be reflected and therefore less light reaches the photodiode. The optical system is mounted on a pivoting arm that can rotate around the indicated axis, so that the optical beam is able to scan the whole surface of the disc. A tracking servosystem in every CD player ensures that the laser beam accurately follows the spiral track. By means of this servosystem, and by means of control and display information (to be touched upon later) the beam can also be directed to any required part of the track. Without counter measures, defocussing would occur as a result of considerable variations in the distance between the objective lens and the disc, making it impossible to read out the information. The objective lens can therefore be displaced by means of an electric actuator so that, by means of a second servosystem, correct focussing can be maintained at all times.

3. PROGRAMMING

Another feature of the system is that the player can be programmed. This is made possible by adding to the audio information control and display information in the form of C & D bits. These C & D bits carry information which the listener may need, such as playing time and the number a piece of music has been given on the disc, thus enabling the player to be preprogrammed so that different sections of the music on the disc can be played in the order selected by the user.

4. INSENSITIVITY TO DUST, SCRATCHES AND OTHER IMPERFECTIONS

A further feature of the Compact Disc system is that it is less sensitive to scratches, dust and other imperfections than the conventional gramophone record. There are three main reasons for this:

The first reason can best be illustrated by referring to figure 2. This shows the laser beam as it is focussed by the objective lens on to the information layer in the Compact Disc. To reach the information layer the light first passes through a 1.2 mm-thick transparent protective layer. The various layers are shown here magnified for the sake of clarity. The information layer is coated with a thin metal reflector only 60 nm thick. When the light beam strikes the surface of the disc the beam has a diameter of 0.7 mm. This implies that tiny particles of dust and small scratches on the surface will not significantly affect the light beam, so that the information contained in the information layer will be correctly read out. This is the first reason why the system is less sensitive to scratches, fingermarks and dust.

The second reason why the system is less sensitive to errors is that errors can be corrected up to a certain extent. The occurrence of channel errors can be attributed to a variety of causes. First of all, in the manufacturing process air bubbles may be left behind in the plastic material or damage to the pits during the pressing of the discs can occur. Flaws of this kind can distort the information so that errors are read out. Errors can also result from fingermarks left on the plate during handling and from mayor surface scratches.

A typical feature of the errors is that they occur in groups, called error bursts. Without counter measures, these errors could ultimately result in incorrect samples, which in turn would give rise to audible disturbances of the audio signal. In the Compact Disc system, however, measures are taken to combat such errors, in the first place by means of error-correction codes. As we shall see, two Reed Solomon codes are applied. The error bursts are further countered by a technique known as interleaving, (as will be explained in more detail) whereby the errors that first occur in groups, or bursts, and can therefore affect a number of successive frames, are spread over a larger number of other frames so that the errors per frame are easier to correct. These measures, then, are the second reason why the Compact Disc system is better able to handle errors.

The third reason why the system is less sensitive to errors is that interpolation is used. If the magnitude of the erros assumes such proportions, however, that the error-correction codes are no longer able to cope with them, there is still a last means of tackling the errors and that is by making use of the error-detecting capabilities of the code. The error-detection circuit identifies in principle the samples that are unreliable, which are then replaced by interpolated samples.

5. DIGITAL AUDIO

Perhaps the most conspicuous feature of the system is that the audio signal is digitally recorded on the disc. Upon analog-digital conversion in the studio the left and right channels are scanned at a sampling frequency of 44.1 kHz. Consequently, both for the left and the right channel a bandwidth of 20 kHz is available.

The choice of a sampling frequency of 44.1 kHz has to do with the fact that recordings in studios and concert halls used to be made, and to some extent still are today, with a video cassette recorder which is able to record the digital audio signals by means of an interface unit.

Upon analog-digital conversion a signal sample is further quantized by the method of uniform (linear) quantization to produce 16 bits, which means that the $2^{16} = 65536$ levels are all equally spaced. This makes it possible to achieve a signal-to-noise ratio of more than 90 dB. The quantization error that remains can very well be described by treating the error signal as white noise which is added to the audio signal and which, because of the signal-to-noise ratio of 90 dB, is inaudible.

In order to maintain the high quality of the audio signals it is necessary that the audio samples are recovered in the player at a rate which has quartz crystal accuracy. The data flow control needed to achieve this goal is schematically indicated in figure 3. The rate at which the bit stream leaves the demodulator, which fills the buffer, is determined by the speed of revolution of the disc, which is controlled by a motor. The rate at which the bit stream leaves the buffer memory to enter the digital signal-processing block is controlled by a quartz crystal C. If the average bit rate from the demodulator, which fills the buffer, is higher than the fixed rate at which the bit stream enters the digital signal processing, the buffer will eventually overflow.

If, however, the average bit rate from the modulator to the buffer is lower than the bit rate (determinded by the quartz crystal C) that fills the digital signal-processing block, then the buffer will ultimately be emptied.

The data stream is kept flowing smoothly by controlling the speed of revolution of the disc in such a way that the buffer memory is on the average filled to 50% of its capacity.

6. CORRECTION AND DETECTION OF ERRORS AND INTERPOLATION

Figure 4 shows the Compact Disc digital audio system considered as a transmission system. On the left we see the studio equipment and on the right the player. Between them is the transmission channel. In telecommunication terms the equipment on the left may be seen as the sender and that on the right as the receiving end. In the studio the left and right stereo channels enter an analog-digital converter, resulting in digital signal samples of 16 bits each, which, as described, are recorded on a digital recorder. In the channel encoder, parity bits are then added and interleaving is applied. The function of the modulator will be touched upon later; for the present it is sufficient to note that the modulator maps the bits out of the encoder onto patterns of pits and lands.

In the channel we see first of all the recording of the digital signal on the master disc. This master disc serves ultimately for the making of stampers, which are used for manufacturing the Compact Discs. On the player side we see the rotating Compact Disc, which is scanned by a laser beam.

Any errors that occur are thus to be found in this channel. At the receiving end, i.e. on the player side, we find a number of operations that are the inverse of those on the sender side. First of all there is the demodulator, followed by the channel decoder. Errors are corrected as far as possible in the channel decoder, and those that cannot be corrected are in principle detected and passed on to the interpolator.

In the interpolator the unreliable samples are replaced by estimates of the audio signal. Some players, finally, provide additional signal processing to ease the digital to analog conversion before the digital signals are finally converted back to the analog audio signals from the two audio channels L and R.

By the mapping in the modulator the frequency spectrum can be manipulated. This is needed e.g. to minimize disturbances of the tracking control system and for recovering the clock frequency in the player. Because of the mapping the intersymbol interference can be minimized which leads to the high information density. In the method of modulation used in the Compact Disc sytem blocks of 8 bits, i.e. of 1 byte, are converted into blocks of 14 channel bits. This and other measures have the effect of making the bit rate in the channel equal to 4.32 megabits per second.

Before going deeper into the method of detecting and correcting errors in the Compact Disc system, it will be useful to look at some of the principles underlaying error correction and detection, with particular reference to block codes.

A binary block code is a code whereby a block (n-k) of parity bits is added to a block of k information bits (fig.5). The manner in which these (n-k) bits are obtained depends on the mathematical structure of the code.

The total block thus adds up to n bits. A block code is often specified by indicating the values (n,k).

A simple example of a binary block code, for single-error correction, is given in Table 1. When an 0 or 1 is transmitted, it is repeated three times, in other words two parity bits are added. The code is thus specified by the values (3,1). With this code we can correct at the most one error occurring in transmission. This can be understood from the third column, which lists the number of possible channel outputs for at most one error. We see that in each word received the number of zeros (in the transmission of a zero) is in the majority, as also is the number of ones in the transmission of a one. Thus, by taking a majority decision we can correct one error. This code cannot be used, however, for correcting two errors, though it can serve for detecting a double error, because if three zeros or three ones are not observed at the output of the channel, then either one error or two errors must have occurred.

We now come to an important point. With this code it is not possible at the same time to correct a single error and detect a double error. In general there is a trade-off between the error-correction and error-detection capabilities of a code. That is to say that, the more use is made of the error-correction capacity, the less is left over for error-detection which in turn results in a greater probability of an error going undetected. An undetected error, as has been mentioned earlier, gives rise in the Compact Disc system to an erroneous audio sample, which may result in an audible 'click'.

The next concept to be touched upon is that of erasures. It might be that some bits (by methods which will become clear presently) are known to be unreliable. A decoder may consider very unreliable bits as being erased i.e. the value of such a bit is completely uncertain. If, in our example, two given bit positions in a received word are erased, then correction is possible. For of course, the value of the bit that is not erased should correspond to the value of the data bit that has been transmitted. The code in our example can thus correct at most two erasures. In summary, the block code can correct one single error or detect maximally two errors, but it can also correct two erasures. These operations, however, cannot be performed simultaneously.

The last concept I wish to introduce is the "Hamming distance". The Hamming distance between two bit sequences of length n is equal to the number of positions at which the bit values of the sequences differ from each other. In coding theory these bit sequences are usually regarded as vectors in a n-dimensional space. If we have a collection of codewords each being a bit sequence we can make a list of the Hamming distances between all these codewords. The smallest distance between two codes is defined as d_{min} of the code. Now there is an important relation between this d_{min} and the maximum number of errors that always can be corrected, t. If at the most t errors occur in a transmitted codeword x, then all possible received words lie, as shown in fig.6, within or on a sphere of radius t. This applies therefore not only to the codeword x but equally to the codeword y and to all other codewords. If these spheres do not intersect each other, then it is always possible to correct t errors by searching for the centrepoint of the sphere within which the received word lies.

A necessary and sufficient condition for these spheres not to intersect each is:

$$d_{min} \geqslant (2t+1).$$

The audio samples that are recorded on the Compact Disc are first grouped into what are called frames (fig.7). A frame consists of twelve audio samples, six samples of the left channel and six samples of the right channel. Since each audio sample consists of 16 bits, that is to say 2 bytes, we can also say that a frame consists of 24 bytes.

The concepts of error detection and correction as explained in the foregoing can be extended from the bit level to the byte level.

In the Compact Disc system two Reed-Solomon codes are used. In the first encoder 4 parity bytes are added to the 24 audio

bytes, producing a C_2 word of 28 bytes. In the second encoder another 4 parity bytes are added to the 28 bytes coming from the C_2 encoder, resulting in a C_1 word of 32 bytes.
For a Reed-Solomon code the minimum distance d_{min} equals the number of parity bytes plus one, hence in our case

$$d_{min} = 4 + 1 = 5$$

Since it further holds that

$$d_{min} \geqslant (2t+1)$$

it follows that we can correct a maximum of 2 (byte) errors. Because however there is also a trade-off between error correction and erasure error correction, it can be shown that each code can correct each number of errors t and erasures e simultaneously provided we satisfy the condition $2t + e \lesssim 4$ (e,t in bytes).

Figure 8 gives a schematic diagram of a CD decoder. The operations here are the inverse of those in the encoder. For convenience we shall forget for the moment the delay lines, which are marked with a capital D and are situated before the input of the C_1 decoder. At the input of C_1 a succession of words arrives, each consisting of 32 bytes. It is the task of the C_1 decoder to correct as many errors as possible and indicating by flag signals the reliablility of the bytes leaving the C_1 decoder.
Between the C_1 and C_2 decoders the de-interleaving takes place, which conceptually consists of a set of delay lines. The effect of interleaving is illustrated in figure 9. On the left side we see a number of successive words, each consisting of 28 bytes, as they leave the C_1 decoder. In the last C_1 word (see fig.9) all bytes indicated by small circles are unreliable. On the right we see the constitution of words after they have passed the de-interleaving set of delay lines and as they appear at the input of the C_2 decoder. We now see that the 28 bytes of a C_1 word, each of them flagged, are spread over 28 successive C_2 words in such a way that each C_2 word contains only one unreliable byte.

Since the Hamming distance of the C_2 code, like the C_1 code, equals 5 bytes, one erasure per frame can easily be corrected by the C_2 decoder. In principle, indeed, the C_2 decoder can even correct 4 erasures. Basically the task of the C_2 decoder is to correct all errors which were uncorrectable by the C_1 decoder (in particular the burst errors). Hereby the C_2 decoder makes use of the reliability information generated by the C_1 decoder. The error correction capabilities, used by the C_2 decoder, are limited by the fact that the probability of an undetected error must be kept small enough. Errors exceding the afore mentioned capabilities are in principle again being flagged as unreliable.

In the block designated by Δ in figure 8 the bytes from the C_2 decoder are reshuffled. The purpose of this operation is to scatter unreliable samples in such a way that each unreliable sample is surrounded as much as possible by reliable samples.

In figure 10 an example of the influence of the operation in the block designated by Δ is shown. Only the third and fifth samples of the left audio channel are unreliable. A reasonably reliable estimate of these two samples can be obtained by means of a first-order linear interpolation (fig.10).

Although linear interpolation is a good and satisfactory solution in the CD system, the question arises as to whether there might be other interpolation methods with which a larger number of missing samples could be estimated. This question is the starting consideration of the last article in this series, by Veldhuis and Janssen.

Acknowledgment

I would like to thank C.P.M.J. Baggen for reading the manuscript and making helpful suggestions.

For Further Reading

1 C.P.M.J. Baggen, "Applications and problems of error correction coding with respect to storage channels", This issue.

2 R.N.J. Veldhuis, A.J.E.M. Janssen, "Adaptive restoration of unknown samples in certain discrete-time signals; mathematical aspects", This issue.

3 J.B.H. Peek, "Communications aspect of the compact disc digital audio system", IEEE Communications Magazine, pp. 7-15, February 1985, Vol. 23, No. 2.

4 M.G. Carasso, J.B.H. Peek and J.P. Sinjou, " The compact disc digital audio system", Philips Tech. Rev., (special issue), vol. 40, no. 6, pp. 151-156, 1982.

5 S. Miyaoka, "Digital audio is compact and rugged," IEEE Spectrum, pp. 35-39, March 1984.

6 "Draft Compact Disc Digital Audio System". now under discussion in the International Electrotechnical Commission as document 60A (Central Office) 82. For text, reference is made to doc. 60A (Secretariat 96).

7 B.A. Blesser, "Digitization of audio", J. Audio Eng. Soc., vol. 26 no. 10, pp. 739-771, Oct. 1978.

8 T. Doi, Y. Tsuchiga, and A. Iga, "On several standards for converting PCM signals into video signals, "J. Audio Eng. Soc. , vol. 26, no. 9, pp. 641-649, Sept. 1978.

9 G.C. Clark and J.B. Cain, Error-Correction Coding for Digital Communications, Plenum Press, 1981.

10 K.A. Immink, "Modulation systems for digital audio discs with optical readout, "IEEE Int. Conf. on Acoust., Speech and Sig. Process, pp. 587-589, Atlanta, GA, March 30-April 1, 1981.

11 J.P.J. Heemskerk and K.A. Schouhamer Immink, "Compact disc: system aspects and modulation, "Philips Tech.Rev. (Special Issue). vol. 40, no. 6, pp. 157-164, 1982.

12 W. Verkaik, "Compact Disc (CD) mastering-an industrial process, "The AES Premier Conference- The New World of Digital Audio, New York, June 3-6, 1982.

13 A.B. Carlson, Communication Systems, McGraw-Hill, 1975.

14 L.B. Vries and K. Odaka, "CIRC- the error correcting code for the Compact Disc, "The AES Premier Conference, New World of Digital Audio, New York, June 3-6, 1982.

15 L.M.H.E. Driessen and L.B. Vries, "Performance calculations of the compact disc error correcting code on a memoryless channel," Int. Conf. Video and Data Recording, University of Southampton, April 20-23, 1982.

16 H. Hoeve, J. Timmermans and L.B. Vries, "Error correction and concealment in the compact disc system," Philips Tech. Rev. (Special Issue), vol. 40, no. 6, pp. 166-172, 1982.

17 T. Doi, "Error correction for digital audio recordings," The AES Premier Conference - The New World of Digital Audio, New York, June 3-6 1982.

18 A.J.E.M. Janssen, R.N.J Veldhuis and L.B. Vries, "Adaptive interpolation of time-discrete signals that can be modelled as autoregressive Processes," IEEE Trans. Acoust., Speech, and Sig. Process, ASSP-34, no.2 April 1986.

19 D. Goedhart, R.J. van de Plassche, and E.F. Stikvoort, "Digital-to-analog conversion in playing a compact disc," Philips Tech. Rev., (Special Issue), vol. 40, no. 6 pp. 174-179, 1982.

20 T.A.C.M. Claasen, W.F.G. Mecklenbräuker, J.B.H. Peek, and N. van Hurck, "Signal processing method for improving the dynamic range of A/D and D/A converters," IEEE Trans. Acoust., Speech, and Sig. Process., ASSP-28, no. 5 Oct. 1980.

Photo 1. A compact disc.

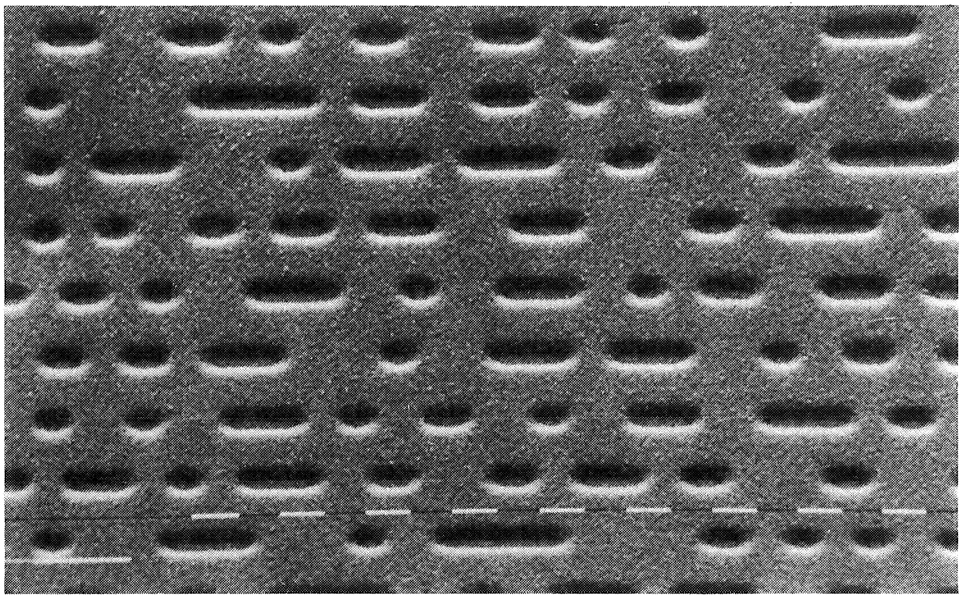


Photo 2. Microphoto of part of the surface of the information layer in the Compact Disc.

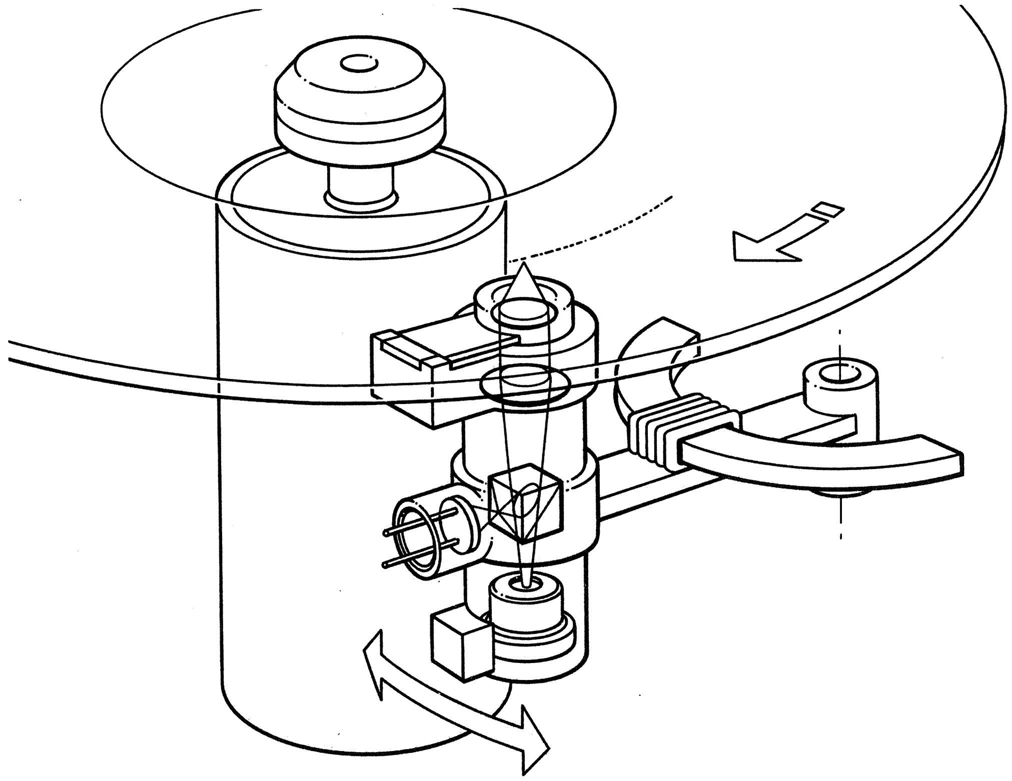


Fig. 1. Schematic representation of the disc and the optical read-out system.

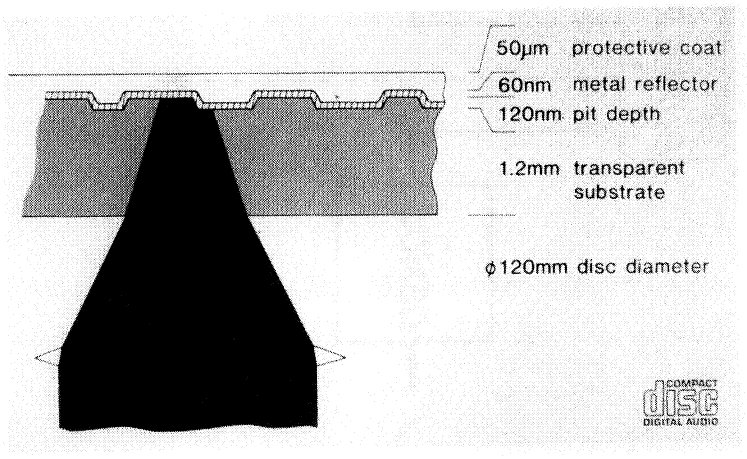


Fig. 2. Path of laser beam.

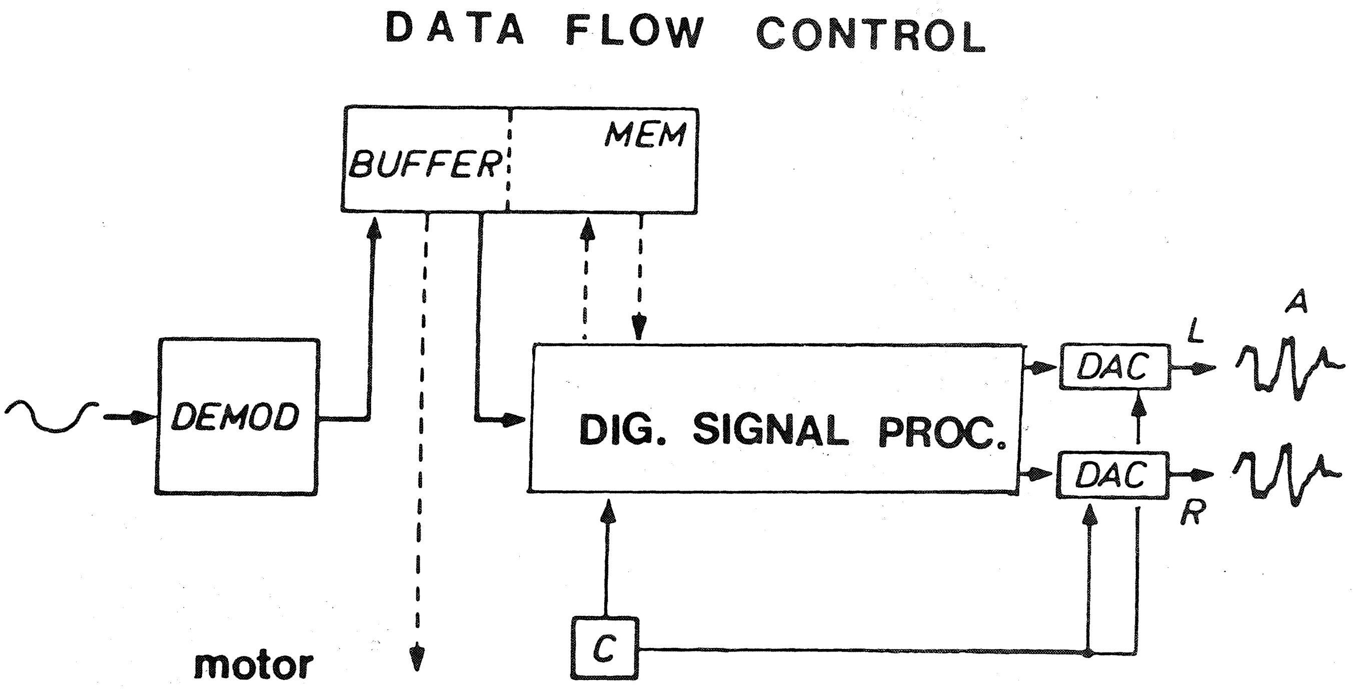


Fig. 3. Schematic diagram of data flow control.

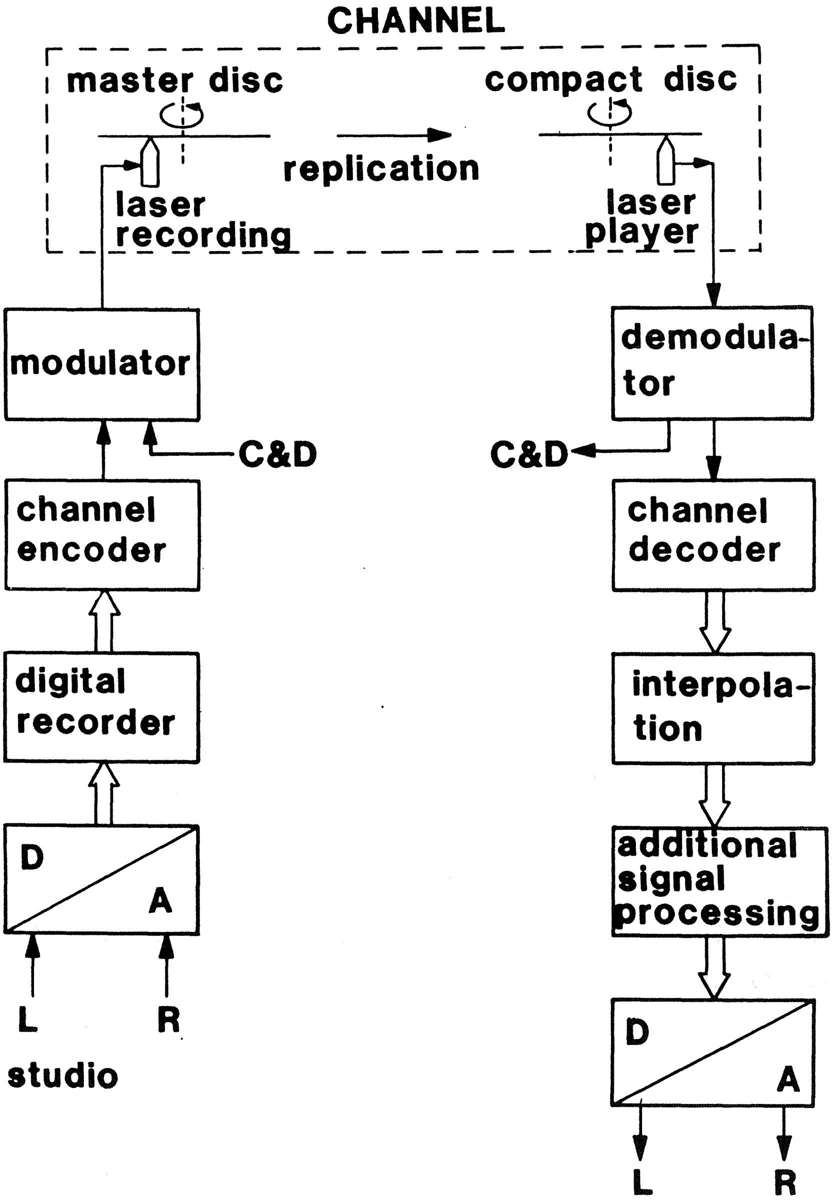


Fig. 4. The compact disc digital audio system, considered as a transmission system

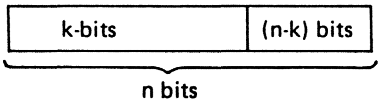


Fig. 5. A block code

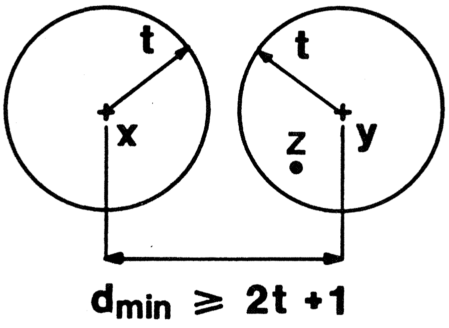


Fig. 6.

Relation between minimum distance d_{min} and the number of maximal correctable errors t

Fig. 7.

L_{6n} ,R_{6n} ,L_{6n+1} , R_{6n+1} , ...,L_{6n+5} ,R_{6n+5}

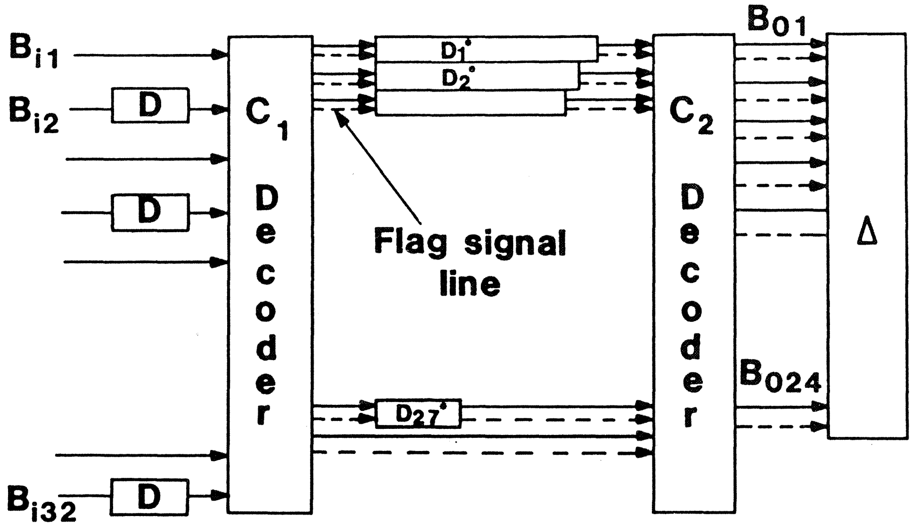


Fig. 8. **Scheme of the CD-decoder**

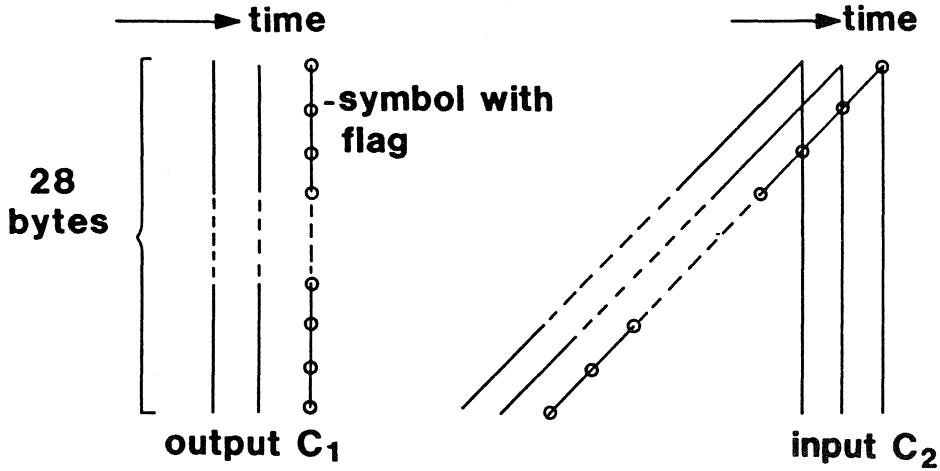


Fig. 9. **Effect of deinterleaving**

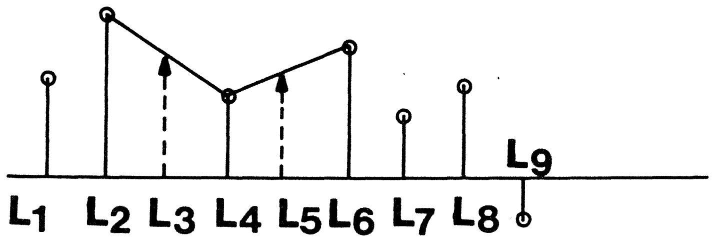


Fig. 10. **First-order linear interpolation**

data bit	single error correcting code	channel outputs (max. one error)

0	0 0 0	0 0 0
		0 0 1
		0 1 0
		1 0 0

1	1 1 1	1 1 1
		1 1 0
		1 0 1
		0 1 1

Table 1. Example of single error correcting code

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Adaptive Restoration of Unknown Samples in Certain Discrete-Time Signals; Mathematical Aspects

R.N.J. Veldhuis and A.J.E.M. Janssen
Philips Research Laboratories
P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands

Abstract

This paper presents an adaptive algorithm for the restoration of lost sample values in discrete-time signals that can locally be modelled as autoregressive processes. The only restrictions are that the positions of the unknown samples should be known and that they should be embedded in a sufficiently large neighbourhood of known samples. The estimates of the unknown samples are obtained by iteratively minimizing the sum of squares of the residual errors that involve estimates of the autoregressive parameters. A statistical analysis shows that, for a burst of lost samples, the expected quadratic restoration error per sample converges to the signal variance when the burst length tends to infinity. The numerical robustness of the method is investigated. The method has been developed to be used on digitized music and speech signals.

I. Introduction

This paper treats the problem of restoring (or interpolating) unknown or lost sample values in a discrete-time signal. An algorithm is presented that is capable of restoring satisfactorily unknown samples with known positions occurring in bursts and more general patterns. Examples of both cases are shown in Fig. I.1. To restore the unknown samples the algorithm uses the information contained in the known neighbouring samples.

Until rather recently the problem of estimating unknown sample values in digitized music signals in real-time could only be solved by relatively simple, non-adaptive methods, such as Lagrange type curve fitting. These methods are not well-suited for audio signals, since these signals primarily contain harmonic components. Severe audible errors can be expected, when the number of samples in the periods of the harmonic components is less than the number of unknown samples. For instance, linear interpolation gives already audible restoration errors for bursts in a digital audio signal of length 5. Because of the progress made in the field of chip design, one can now contemplate more complicated real-time restoration methods, that may also involve some signal model. Examples of these methods, can be found in [1,2,3,21]. In [22] a general technique to derive from a signal model a restoration method for lost samples is presented. An extensive description of the method described here can be found in [20].

In this paper the same point of view as in [2], Section II is taken for the restoration of more general patterns of unknown samples than single ones or bursts. That is, it is assumed that the signals to be interpolated can be modelled as autoregressive (AR) processes of finite order.

The method is adaptive in the sense that, from a finite segment of data, it estimates the AR-parameters as well as the unknown samples. The restoration is done in such a way that the restored signal fits the estimated model as well as possible.

The choice of the autoregressive process as a model for the signal can be motivated by the fact that many signals that are encountered in practice can be modelled in this way. Therefore, it is expected that the restoration method presented here can be applied successfully in many practical situations.

The organization of this paper is as follows. In Section II the restoration method is presented and a statistical analysis is given. The restoration error is analyzed under the assumption that the AR-parameters are known. This analysis is detailed for the case that Gaussian probability density functions are assumed. In Section III computational aspects of the method are considered. Also, the numerical properties of certain parts of the algorithm are discussed. In Section IV some results are presented. Section V presents some conclusions.

II. Presentation and analysis of the restoration method

In this section it is assumed that the sequence s_k, $k=-\infty,\dots,\infty$, is a realization of a stationary autoregressive process $\tilde{s}_k$, $k=-\infty,\dots,\infty$, (the tilda $\tilde{}$ indicates that a variable is a stochastic variable). This means that there exist a finite positive integer p, the prediction order, numbers $a_0, a_1, \dots, a_p$, $a_0=1$, the prediction coefficients, and a zero mean white noise process $\tilde{e}_k$, $k=-\infty,\dots,\infty$, the excitation noise, with variance σ_e^2, such that

$$\text{(II.1)} \qquad a_0\tilde{s}_k + a_1\tilde{s}_{k-1} + \ldots + a_p\tilde{s}_{k-p} = \tilde{e}_k,\ k=-\infty,\ldots,\infty .$$

For notational convenience, it shall be agreed that $a_k=0$ for $k<0$ or $k>p$. The AR-spectrum $S(\theta)$ of $\tilde{s}_k$, $k=-\infty,\ldots,\infty$, is given by

$$\text{(II.2)} \qquad S(\theta) = \frac{\sigma_e^2}{\left|\sum_{l=0}^{p} a_l \exp(-j\theta l)\right|^2} = \frac{\sigma_e^2}{\sum_{l=-p}^{p} b_l \exp(-j\theta l)},$$

where

$$\text{(II.3)} \qquad b_l = \sum_{k=0}^{p} a_k a_{k+l} .$$

In part A of this section the algorithm for estimating the AR-parameters and the unknown samples from a finite sequence of samples is presented. A statistical analysis of the restoration error is given in part B of this section.

A. Presentation of the restoration method

The available data consists of a segment s_k, $k=0,\ldots,N-1$, of a realization of an AR-process $\tilde{s}_k$, $k=-\infty,\ldots,\infty$. It is assumed throughout that the unknown samples occur at the known time instants $t(1),\ldots,t(m)$, where $0<p\leq t(1)<\ldots<t(m)\leq N-p-1$. The problem is to estimate the values of the unknown samples $s_{t(1)},\ldots,s_{t(m)}$ and the AR-parameters $p,a_1,\ldots,a_p$ and σ_e^2 from the available data in such a way that the restored segment fits the assumed model as well as possible in a quadratic sense. That is, the restoration is such that the sum of the squares of the residual error $e_p,\ldots,e_{N-1}$ is minimal.

Although methods to estimate the order of an autoregressive process have been reported [4], it has been decided, if p is unknown, to choose p as a function of the number m of unknown samples. The rather arbitrary relation $p\cong 3m$ has proved to give good restoration results. For notational convenience the vector notation $\underline{a}=[a_1,\ldots,a_p]^T$, $\underline{x}=[s_{t(1)},\ldots,s_{t(m)}]^T$ (the superscipt ^{T} denotes vector or matrix transposition) shall be adopted. The estimation of $\underline{a}$ and $\underline{x}$ is expressed as a minimization problem, where the estimates $\hat{\underline{a}}$ for $\underline{a}$ and $\hat{\underline{x}}$ for $\underline{x}$ are chosen such that

$$\text{(II.4)} \qquad Q(\underline{a},\underline{x}) = \sum_{k=p}^{N-1} \left| \sum_{l=0}^{p} a_k s_{k-l} \right|^2 = \sum_{k=p}^{N-1} |e_k|^2$$

is minimal as a function of $\underline{a}$ and $\underline{x}$. Once $\hat{\underline{a}}$ and $\hat{\underline{x}}$ have been determined, σ_e^2 is estimated by

$$\text{(II.5)} \qquad \hat{\sigma}_e^2 = \frac{1}{N-p-m}\, Q(\hat{\underline{a}},\hat{\underline{x}}).$$

Since $Q(\underline{a},\underline{x})$ involves 4^{th} order terms, such as $a_1^2 s_{t(m)}^2$, the minimization with respect to $\underline{a}$ and $\underline{x}$ is a non-trivial problem. The following iterative approach can then be applied succesfully. One chooses an initial estimate $\hat{\underline{x}}^{(0)}$, for instance $\hat{\underline{x}}^{(0)}=\underline{0}$, for the vector $\underline{x}$ of the unknown samples. Next, one minimizes $Q(\underline{a},\hat{\underline{x}}^{(0)})$ as a function of $\underline{a}$ to obtain an estimate $\hat{\underline{a}}^{(1)}$. Secondly, one minimizes $Q(\hat{\underline{a}}^{(1)},\underline{x})$ as a function of $\underline{x}$ to obtain an estimate $\hat{\underline{x}}^{(1)}$ for the unknown samples.

Both minimizations are feasible, since $Q(\underline{a},\underline{x})$ is a quadratic form in both $\underline{a} \in \mathbb{R}^p$ and $\underline{x} \in \mathbb{R}^m$. In fact, it can be shown that

$$\text{(II.6)} \qquad Q(\underline{a},\underline{x}) = \underline{a}^T C(\underline{x})\underline{a} + 2\underline{a}^T \underline{c}(\underline{x}) + c_{00}(\underline{x}).$$

Here

$$\text{(II.7)} \qquad C(\underline{x}) = (c_{ij}(\underline{x}))_{i,j=1,\ldots,p},$$

$$\underline{c}(\underline{x}) = [c_{01}(\underline{x}),\ldots,c_{0p}(\underline{x})]^T,$$

where

$$\text{(II.8)} \qquad c_{ij}(\underline{x}) = \sum_{k=p}^{N-1} s_{k-i}s_{k-j}, \quad i,j=0,1,\ldots,p.$$

Hence, $C(\underline{x})$ is the p x p-autocovariance matrix, estimated from s_k, $k=0,\ldots,N-1$. At the same time it can be shown that

$$\text{(II.9)} \qquad Q(\underline{a},\underline{x}) = \underline{x}^T B(\underline{a})\underline{x} + 2\underline{x}^T \underline{z}(\underline{a}) + D(\underline{a}).$$

Here

$$\text{(II.10)} \qquad B(\underline{a}) = (b_{t(i)-t(j)})_{i,j=1,\ldots,m},$$

$$\underline{z}(\underline{a}) = [z_1(\underline{a}),\ldots,z_m(\underline{a})]^T,$$

b_l, $l=-p,\ldots,p$, has been defined in (II.3), and

$$\text{(II.11)} \quad z_i(\underline{a}) = \sum_{k=-p}^{p} b_k s_{t(i)-k}, \quad i=1,\dots,m,$$

and $D(\underline{a}) \in \mathbb{R}$ depends on $\underline{a}$ and the known samples only. Hence $\hat{\underline{a}}^{(1)}$ and $\hat{\underline{x}}^{(1)}$ are given by

$$\text{(II.12)} \quad C(\hat{\underline{x}}^{(0)})\hat{\underline{a}}^{(1)} = -\underline{c}(\hat{\underline{x}}^{(0)}),$$

and

$$\text{(II.13)} \quad B(\hat{\underline{a}}^{(1)})\hat{\underline{x}}^{(1)} = -\underline{z}(\hat{\underline{a}}^{(1)})$$

respectively. The above method for calculating prediction coefficients from a sequence of samples is known as the autocovariance method [5]. On substitution of (II.12) into (II.5) it easily follows that

$$\text{(II.14)} \quad \hat{\sigma}_e^2 = \frac{1}{N-p-m}\,(c_{00}(\hat{\underline{x}}) + \hat{\underline{a}}^T\underline{c}(\hat{\underline{x}})).$$

The procedure described above is the first step of an iterative procedure, in which in every step new prediction coefficients $\hat{\underline{a}}^{(i)}$ are estimated as in (II.12) by using $\hat{\underline{x}}^{(i-1)}$ instead of $\hat{\underline{x}}^{(0)}$. These prediction coefficients can be substituted into (II.13) to obtain new estimates $\hat{\underline{x}}^{(i)}$ for the unknown samples. It is clear that in this way $Q(\underline{a},\underline{x})$ decreases to some non-negative number. One may hope that the sequence thus obtained converges to a point where $Q(\underline{a},\underline{x})$ attains its global minimum. Unfortunately, it seems very hard to prove any definite result in this direction. In [20] it is shown that this iterative minimization procedure closely resembles a maximum likelihood parameter estimation algorithm, well-known in statistics: the EM algorithm [6,7,8].

B. Statistical analysis of the restoration error

In this subsection some statistical properties of the restoration error are discussed. It is assumed that p, $\underline{a}$ and σ_e^2 are known. Since, in practice, these parameters are estimated from the data, this assumption may be a simplification from reality. However, it has the advantage that the results take a pleasant form.

The restoration error is defined as the stochastic vector $\underline{\tilde{d}}$, (II.15) $\underline{\tilde{d}} = \underline{\tilde{x}} - \underline{\hat{x}} = \underline{\tilde{x}} + (B(\underline{a}))^{-1}\underline{\tilde{z}}(\underline{a})$.

Note that the realization $\underline{z}(\underline{a})$ of (II.13) is replaced by a stochastic vector $\underline{\tilde{z}}(\underline{a})$. It follows easily from (II.15) and from the fact that $E[\tilde{s}_k]=0$ that $E[\underline{\tilde{d}}]=\underline{0}$ and that the estimator $\underline{\hat{x}}$ is unbiased. The (stochastic) relative quadratic restoration error per sample, $\tilde{e}$, is defined by

(II.16) $$\tilde{e} = \frac{\underline{\tilde{d}}^T\underline{\tilde{d}}}{m\ E[\tilde{s}_k^2]} .$$

To evaluate the expectation $E[\tilde{e}]$ of $\tilde{e}$, it is noted that

(II.17) $\underline{\tilde{d}} = (B(\underline{a}))^{-1}(\underline{\tilde{z}}(\underline{a}) + B(\underline{a})\underline{\tilde{x}}) =: (B(\underline{a}))^{-1}\underline{\tilde{w}}$,

and that, for i=1,...,m,

(II.18) $$\tilde{w}_i = \sum_{k=-p}^{p} b_k\tilde{s}_{t(i)-k} = \sum_{l=0}^{p} a_l\tilde{e}_{t(i)+l},$$

as follows straightforwardly from the definitions in (II.3), (II.10), (II.11). Thus,

(II.19) $E[\underline{\tilde{d}\tilde{d}}^T] = (B(\underline{a}))^{-1}\ E[\underline{\tilde{w}\tilde{w}}^T]\ (B((\underline{a}))^{-1}$.

Since $(E[\underline{\tilde{w}\tilde{w}}^T])_{ij}=E[\tilde{w}_i\tilde{w}_j]= \sigma_e^2 b_{t(i)-t(j)}$, one has that $E[\underline{\tilde{w}\tilde{w}}^T]= \sigma_e^2 B(\underline{a})$ and that

(II.20) $E[\underline{\tilde{d}\tilde{d}}^T] = \sigma_e^2(B(\underline{a}))^{-1}$.

Finally, $E[\tilde{e}]$ is given by

(II.21) $$E[\tilde{e}] = \frac{\sigma_e^2}{m\ E[\tilde{s}_k^2]} \text{ trace } ((B(\underline{a}))^{-1}).$$

For the expected relative quadratic restoration error of the i^{th} unknown sample one has

(II.22) $E[\tilde{d}_i^2] = \sigma_e^2\ ((B(\underline{a}))^{-1})_{ii}$, i=1,...,m.

The case of a burst of m consecutive unknown samples deserves somewhat more attention than the general case. Then the matrix $B(\underline{a})$ is Toeplitz and therefore has some properties that facilitate a further analysis of the restoration error. Toeplitz matrices are persymmetric: an n x n-matrix M is persymmetric if $M_{ij} = M_{n+1-j,n+1-i}$, i,j=1,...,n. It is a property of persymmetric matrices that their inverses are also persymmetric. If $B(\underline{a})$ is Toeplitz then $(B(\underline{a}))^{-1}$ is persymmetric, and

$$\text{(II.23)} \quad E[\tilde{a}_i^2] = E[\tilde{a}_{m+1-i}^2], \quad i=1,\dots,m.$$

Extensive observations for the case of a burst of m unknown samples have revealed that the $((B(\underline{a}))^{-1})_{ii}$, $i=1,\dots,m$, seem to depend quadratically on i for m not too large, and that the $((B(\underline{a}))^{-1})_{ii}$ tend to have their maximum for $i \cong m/2$, i.e. in the middle of the burst. Hence, much of the error energy is usually concentrated in the middle of the burst.

In case of a burst the asymptotic behaviour of $E[\tilde{e}]$ as goes to infinity can be determined by applying the Szegö limit theorem [10]. From (II.21) one has

$$\text{(II.24)} \quad E[\tilde{e}] = \frac{\sigma_e^2}{m\,E[\tilde{s}_k^2]} \sum_{i=1}^{m} \lambda_i^{-1},$$

where λ_i is the i^{th} eigenvalue of $B(\underline{a})$. According to the Szegö limit theorem one has for any function F, continuous on the set $\{\, \Sigma_{k=-p}^{p}\, b_k \exp(-j\theta k) \mid |\theta| < \pi \,\}$,

$$\text{(II.25)} \quad \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} F(\lambda_i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\Big(\sum_{k=-p}^{p} b_k \exp(-j\theta k)\Big) d\theta.$$

Taking $F(a)=a^{-1}$, one finds by using (II.2),

$$\text{(II.26)} \quad \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \lambda_i^{-1} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{\sum_{k=-p}^{p} b_k \exp(-j\theta k)}\, d\theta$$

$$= \frac{1}{\sigma_e^2} \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\theta)\, d\theta$$

$$= \frac{E[\tilde{s}_k^2]}{\sigma_e^2}.$$

Hence,

$$\text{(II.27)} \quad \lim_{m \to \infty} E[\tilde{e}] = 1.$$

This shows that for long bursts of consecutive samples the quadratic restoration error per sample approaches the signal energy per sample.

The result (II.27), derived for the burst case, is also useful for finding a bound on the restoration error in the general case. Indeed, the matrix $B(\underline{a})=(b_{t(i)-t(j)})_{i,j=1,\ldots,m}$ is a principal submatrix of the $(t(m)-t(1)+1) \times (t(m)-t(1)+1)$ Toeplitz matrix $B'(\underline{a})=(b_{k-l})_{k,l=1,\ldots,t(m)-t(1)+1}$. Denoting the first m eigenvalues of $B(\underline{a})$ and $B'(\underline{a})$ in increasing order by $\lambda_1,\ldots,\lambda_m$ and $\lambda'_1,\ldots,\lambda'_m$, one has by [11], Section 3.5, Theorem 5.6 that $0< \lambda'_i< \lambda_i$, $i=1,\ldots,m$. Hence,

$$\text{(II.28)}\quad \operatorname{trace}((B(\underline{a})^{-1}) = \sum_{i=1}^{m} \lambda_i^{-1} < \sum_{i=1}^{m} \lambda'^{-1}_i < \operatorname{trace}((B'(\underline{a})^{-1}),$$

and it follows that $E[\tilde{e}]$ is asymptotically bounded by $\limsup m^{-1}(t(m)-t(1)+1)$. Although this bound is not as good as for the burst case, the restoration error in the case of m randomly positioned unknown samples usually turns out to be smaller than in the case of a burst of length m.

The restoration error can be analyzed in some more detail if $\tilde{e}_k$ has a Gaussian probability density function. It then follows that $\underline{\tilde{d}}$ has a probability density function

$$\text{(II.29)}\quad p_{\underline{\tilde{d}}}(\underline{d}) = \frac{|B(\underline{a})|^{1/2}}{(2\pi)^{m/2}\sigma_e^m} \exp\left(-\frac{\underline{d}^T B(\underline{a})\underline{d}}{2\sigma_e^2}\right).$$

It is a rather tedious but straightforward exercise to calculate the variance, $\operatorname{var}(\tilde{e})=E[(\underline{\tilde{d}}^T\underline{\tilde{d}} - E[\underline{\tilde{d}}^T\underline{\tilde{d}}])^2]$, of $\tilde{e}$:

$$\text{(II.30)}\quad \operatorname{var}(\tilde{e}) = \frac{\sigma_e^4}{m\, E[\tilde{s}_k^2]^2} \sum_{i=1}^{m} \lambda_i^{-2}.$$

In the case of a burst of unknown samples of length m, one can use the Szegö limit theorem (II.25) with $F(a)=a^{-2}$. For large m one finds

$$\text{(II.31)}\quad \operatorname{var}(\tilde{e}) = 2\,\frac{\dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} |S(\theta)|^2\, d\theta}{\left|\dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} S(\theta)\, d\theta\right|^2}.$$

It can be observed that $\operatorname{var}(\tilde{e})$ is larger if the signal spectrum $S(\theta)$ is more peaky.

III. Computational aspects of the restoration algorithm

In this section the computational aspects of the calculation of $\underline{\hat{a}}$ in (II.12) and $\underline{\hat{x}}$ in (II.13) are considered. It should be noted that a linear system needs to be solved for the calculation of both $\underline{\hat{a}}$ and $\underline{\hat{x}}$. If p is chosen 3m, which is done when p is unknown, the need for efficiency is more urgent for the calculation of $\underline{\hat{a}}$ than for the calculation of $\underline{\hat{x}}$. The calculation of $\underline{\hat{a}}$ in (II.12) is in fact a well-known problem. It is often referred to as the autocovariance method and is discussed in great detail for instance in [5]. In this reference also an efficient algorithm is given for solving $\underline{\hat{a}}$ from (II.12) in $O(p^2)$ operations as well as a summary of the various methods to estimate $\underline{a}$ from a sequence of samples. The numerical stability of some of these methods is discussed in [13].

For the calculation of $\underline{\hat{x}}$ in (II.13) it makes sense to analyze the matrix $B(\underline{\hat{a}})$ defined in (II.10) and (II.3) in some detail. It can be seen from (II.10) that $B(\underline{\hat{a}})$ has constant values b_0 on its main diagonal. Furthermore, the matrix $B(\underline{\hat{a}})$ is positive definite, as can be seen from the expression

$$\text{(III.1)}\quad \sum_{i=1}^{m}\sum_{j=1}^{m}(B(\underline{\hat{a}}))_{ij}v_iv_j = \sum_{k}\left|\sum_{i=1}^{m}\hat{a}_{k+t(i)}v_i\right|^2 ,$$

which follows on inserting (II.10) and (II.3) into the left-hand side of (III.1). Indeed, when i' is the largest index with $v_{i'}\neq 0$, the term in the right-hand sum of (III.1) with $k=-t(i')$ equals $v_{i'}^2$, as $\hat{a}_l=0$ for $l<0$, $a_0=1$ and $v_i=0$ fo $i>i'$. Hence, if $\underline{v}$ has non-zero elements, the right-hand sum of (III.1) consists of non-negative terms of which at least one is positive. This shows that $B(\underline{\hat{a}})$ is positive definite.

The fact that $B(\underline{\hat{a}})$ is positive definite allows one to use Cholesky decomposition [14] of $B(\underline{\hat{a}})$ for solving $\underline{\hat{x}}$ from (II.13) in $O(m^3)$ operations. In case of a burst of unknown samples, $B(\underline{\hat{a}})$ is Toeplitz and (II.13) can be solved in $O(m^2)$ operations by the Levinson algorithm [15]. Even in the case of a more general pattern of unknown samples $B(\underline{\hat{a}})$ is related to a Toeplitz matrix, so that the system in (II.13) can be solved more efficiently by using generalized Levinson algortihms [16]. However, this requires rather involved mathematics and does not lead to a less complicated hardware implementation, since the generalized Levinson algorithm to be used strongly depends on the pattern of unkown samples. For these reasons in this paper only the solution of $\underline{\hat{x}}$ from (II.13) by using Cholesky decomposition is considered.

In a Cholesky decomposition the matrix $B(\underline{\hat{a}})$ is decomposed as a product

$$\text{(III.2)} \quad B(\underline{\hat{a}}) = LL^T,$$

or as a product

$$\text{(III.3)} \quad B(\underline{\hat{a}}) = \mathcal{L}D\mathcal{L}^T.$$

In (III.2) L is a lower triangular m x m-matrix, in (III.3) $\mathcal{L}$ is a lower triangular m x m-matrix with constant values 1 on its main diagonal, D is a diagonal m x m-matrix with $D_{ii}=L_{ii}^2$, $i=1,\ldots,m$. The systems $B(\underline{\hat{a}})\underline{\hat{x}}=LL^T\underline{\hat{x}}=-\underline{z}(\underline{\hat{a}})$ and $B(\underline{\hat{a}})\underline{\hat{x}}=\mathcal{L}D\mathcal{L}^T=-\underline{z}(\underline{\hat{a}})$ are now solved by subsequently solving by back substitution $\underline{y}$ and $\underline{\tilde{y}}$ from $L\underline{y}=-\underline{z}(\underline{\hat{a}})$ and from $\mathcal{L}\underline{\tilde{y}}=-\underline{z}(\underline{\hat{a}})$ respectively, and $\underline{\hat{x}}$ from $L^T\underline{\hat{x}}=\underline{y}$ and $\mathcal{L}^T\underline{\hat{x}}=D^{-1}\underline{\tilde{y}}$ respectively. Both forms of Cholesky decomposition take $O(m^3)$ operations. A drawback of the decomposition in (III.2) is that it requires the calculation of square roots. On the other hand, as is shown further on, the elements of L in (III.2) satisfy bounds that are more convenient if one has a fixed point implementation in mind.

For the elements of the matrices L and D one has the following results:

$$\text{(III.4)} \quad 1 \leq L_{jj} = D_{jj}^{1/2} \leq b_0^{1/2}, \quad j=1,\ldots,m,$$

$$\text{(III.5)} \quad \sum_{i=1}^{m} L_{ij}^2 = b_0, \quad j=1,\ldots,m,$$

so that,

$$\text{(III.6)} \quad |L_{ij}| \leq (b_0 - 1)^{1/2}, \quad i=1,\ldots,j-1, \; j=1,\ldots,m.$$

On substitution of $L_{ij}=\mathcal{L}_{ij}D_{jj}^{1/2}$ into (III.6) and by using (III.4) one obtains

$$\text{(III.7)} \quad |\mathcal{L}_{ij}| \leq (b_0 - 1)^{1/2}, \quad i,1,\ldots,j-1, \; j=1,\ldots,m.$$

The bounds in (III.5) and (III.6) and the right-hand bound of (III.4) can be derived by using results of [17], Section 7 and by the fact that $(B(\underline{\hat{a}}))_{jj}=b_0$, $j=1,\ldots,m$. The left-hand bound in (III.4) was not known to the authors. It can be derived as follows. First remark that

$$\text{(III.8)} \quad B(\underline{a}) = A^T A,$$

where $A=[\underline{a}_1,\ldots,\underline{a}_m]$ is a $(t(m)-t(1)+p+1)$ x m-matrix, defined by

(III.9) $A_{ij} = (\underline{a}_j)_i = a_{t(j)-i-t(1)+p+1}$,

the a_i being the prediction coefficients of (II.1). Note that $a_i=0$ for $i<0$ or $i>p$. Since A has full rank, A can be decomposed as a product $A=QR$, where Q is a $(t(m)-t(1)+p+1)$ x m-matrix, consisting of m orthogonal columns and R is an upper triangular m x m-matrix. On substituting $A=QR$ into (III.8), one obtains

(III.10) $B(\underline{a}) = R^TQ^TQR = \mathcal{L}D\mathcal{L}^T$,

where $\mathcal{L}$ and D are as in (III.3). Clearly $D_{jj}=|\underline{q}_j|^2$. The QR decomposition of A can be done iteratively. In every iteration step $\underline{q}_j$ is found by subtracting from $\underline{a}_j$ the projection of $\underline{a}_j$ on to the space spanned by $\underline{q}_1,\ldots,\underline{q}_{j-1}$:

(III.11) $$\underline{q}_j = \underline{a}_j - \sum_{k=1}^{j-1} \frac{\underline{a}_j^T\underline{q}_k}{|\underline{q}_k|^2}.$$

The space $sp\{\underline{q}_1,\ldots,\underline{q}_{j-1}\}$ spanned by $\underline{q}_1,\ldots,\underline{q}_{j-1}$ is the same as the space $sp\{\underline{a}_1,\ldots,\underline{a}_{j-1}\}$ spanned by $\underline{a}_1,\ldots,\underline{a}_{j-1}$. Therefore,

(III.12) $$\begin{aligned} |\underline{q}_j|^2 &= \min_{\underline{v}\in sp\{q_1,\ldots,q_{j-1}\}} |\underline{a}_j-\underline{v}|^2 \\ &= \min_{\underline{v}\in sp\{a_1,\ldots,a_{j-1}\}} |\underline{a}_j-\underline{v}|^2 \\ &= \min_{\underline{w}\in\mathbb{R}^{j-1}} |\underline{a}_j + \sum_{k=1}^{j-1} w_k\underline{a}_k|^2. \end{aligned}$$

Since $(\underline{a}_j)_{t(j)-t(1)+p+1}=a_0=1$ and $(\underline{a}_k)_{t(j)-t(1)+p+1}=0$ for $k=1,\ldots,j$, by (III.9) it follows easily that $|\underline{q}_j|^2\geq 1$. This proves the left-hand inequality of (III.4)

In a fixed point implementation it is more convenient to solve the system $B'(\underline{\hat{a}})\underline{\hat{x}}=-\underline{z}'(\underline{\hat{a}})$, where $B'(\underline{\hat{a}})=B(\underline{\hat{a}})/b_0$ and $\underline{z}'(\underline{\hat{a}})=\underline{z}(\underline{\hat{a}})/b_0$, than the system in (II.13), because the absolute values of the elements $B'(\underline{\hat{a}})$ are all bounded by 1. Then $B'(\underline{\hat{a}})=L'L'^T=\mathcal{L}D'\mathcal{L}^T$, where $L'=L/b_0$ and $D'=D/b_0$. On substituting this into (III.4), (III.5) and (III.6) one obtains

(III.13) $1/b_0^{1/2} \leq L'_{jj} = D'^{1/2}_{jj} \leq 1$, $j=1,\ldots,m$, (III.14) $\sum_{i=1}^{m} L'^{2}_{ij} = b_0, \quad j=1,\ldots,m,$

or,

(III.15) $|L'_{ij}| \leq 1, \; i,j=1,\ldots,m.$

Now the $L'L'^T$ decomposition of $B'(\underline{\hat{a}})$ has the advantage over the $CD'C^T$ decomposition that the absolute values of all elements of L' are bounded by 1 and that all fixed point multiplications can be performed without prescaling.

The lower bound in (III.13) is important because the elements L'_{jj}, j=1,...,m, are used as divisors in the process of back substitution and accuracy will be lost if they are too small. It is the experience of the authors that, for digitized music, b_0 usually has rather modest values, say $b_0<4$, so that the L'_{jj} do not become too small.

IV. Results

In this section the performance of the adaptive restoration method discussed in this paper is considered for the following test signals:

1) Artificially generated realizations of an autoregressive process of 10th order with a peaky and with a smooth spectrum. Fig. IV.1 shows the AR-spectrum. Ten statistically independent sequences of 512 samples each have been used. The excitation noise sequences are uncorrelated pseudo-random sequences with a Gaussian probability density function with zero mean and unit variance. The patterns of the unknown samples were bursts of lengths m=16,50.

2) Multiple sinusoids. A sequence of 512 samples, given by

(IV.1) $s_n = 100 \sin(0.23\pi n + 0.3\pi) + 60 \sin(0.4\pi n + 0.3\pi)$

has been used. The patterns of the unknown samples were bursts of lengths m=16.

3) Digital audio signals. Bursts of 16 unknown samples, occurring at a rate of 10 s^{-1} in a fragment of 36s taken from a Compact Disc ^{R} recording of Beethoven's Violin Concert have been interpolated. The sample frequency of the signal is 44100 Hz, so that a burst of 16 samples has a duration of 0.36ms.

For all test signals the performances of the adaptive restoration methods are judged by means of the relative quadratic restoration error e

$$\text{(IV.2)}\qquad e = \frac{\frac{1}{m}\sum_{i=1}^{m}(\hat{s}_{t(i)}-s_{t(i)})^2}{\frac{1}{N}\sum_{i=1}^{N-1} s_i^2}.$$

This is the realization of the stochastic relative quadratic restoration error $\tilde{e}$, defined in (II.16). Diagrams of some typical restoration results are presented in Figures IV.2-8, together with the original signals, in which the correct values of the unknown samples have been substituted. In the diagrams the original signal is marked by a (1), the restoration result is marked by a (2), the positions of the unknown samples are indicated on the time axis. Besides the diagrams, the performances of the adaptive restoration method on the music signals are also evaluated by listening tests.

The figures give rise to the following remarks. For large N and small m ususally one iteration is sufficient. However, if the segment length N is smaller, continuing the iterations gives an improvement. In general, the restoration errors for autoregressive processes with a peaky spectrum are substantially smaller than for processes with a smooth spectrum.

For sinusoids $\sigma_e^2=0$, so that, theoretically, the restoration error is also zero. Indeed, Fig IV.7 shows very small restoration errors for methods c_1 and c_3. The order of prediction, p, must not be chosen too high. For after more than one iteration the autocovariance matrix will become nearly singular and the prediction coefficients can no longer be calculated straightforwardly by solving the system (II.12).

For the music signal it was found that the relative quadratic restoration errors for the adaptive restoration methods were of the same orders of magnitude as those for the autoregressive processes with a peaky spectrum.

Listening tests have revealed that the restoration errors in these test signals and in many other signals are practically inaudible. After increasing the burst length from 16 to 50 the restoration results are still quite good for most music signals, although some restoration errors become audible.

V. Conclusions

In this paper an adaptive method has been presented for the restoration of general patterns of unknown samples occurring in discrete-time signals that can be modelled as autoregressive processes. It has been demonstrated that this method gives satisfactory results for digital audio signals. Roughly speaking, the method amounts to trying to minimize, as a function of the unknown samples and the unknown prediction coefficients, a sum of squares of residual errors involving the unknown samples, the prediction coefficients and the known samples from a sufficiently large neighbourhood.

For a small amount of lost samples in a large segment of data, a single iteration is sufficient. More iterations give an improvement in restoration quality if a relatively small segment of data is available.

It has been shown that the various minimizations can be carried out by efficiently solving in a stable manner, certain systems of linear equations. This indicates that the restoration method is suitable for a fixed point implementation in an integrated circuit. However, in that case the number of unknown samples should not be too high (up to 16, say).

References

[1] R. Steele, F. Benjamin, "Sample Reduction and Subsequent Adaptive Interpolation of Speech Signals", Bell Syst. Tech. J. 62 (6), pp. 1365-1398, 1983.

[2] S.M. Kay, "Some Results in Linear Interpolation Theory", IEEE Trans. ASSP, 31 (3), pp. 746-749, 1983.

[3] A.J.E.M Janssen, L.B. Vries, "Interpolation of Band-Limited Discrete-Time Signals by Minimizing Out-of-Band Energy", Proc. ICASSP '84, San Diego, 1984.

[4] H. Akaike, "A New Look at the Statistical Model Identification", IEEE Trans. AC, 19 (6), pp 716-728, 1974.

[5] S.M. Kay, S.L. Marple Jr., "Spectrum Analysis- A Modern Perspective", Proc. IEEE, 69 (11), pp. 1380-1419, 1981.

[6] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Roy. Stat. Soc., Series B,

39, pp. 1-38, 1977.

[7] C.F.J. Wu, "On the Convergence Properties of the EM Algorithm", The Annals of Statistics, 11 (1), pp. 95-103, 1983.

[8] R.A. Bayles, "On the Convergence of the EM Algorithm", J. Roy. Stat. Soc., Series B, 45 (1), pp. 47-50, 1983.

[9] F. Itakura, S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", Electron. Commun. Japan, 53-A, pp.36-43, 1970.

[10] I.I. Hirschmann Jr., "Recent Developments in the Theory of Finite Toeplitz Operators", Advances in Probability and Related Topics, Vol. 1, Dekker, New York, 1971.

[11] M. Marcus, H. Minc, Introduction to Linear Algebra, MacMillan, New York, 1965.

[12] J. Durbin, "The Fitting of Time-Series Models", Rev. Inst. Int. Stat., 28, pp. 233-243, 1960.

[13] G. Cybenko, "The Numerical Stability of the Levinson-Durbin Algorithm for Toeplitz Systems of Equations", SIAM J. on Sci. Stat. Comp., Vol. 1, pp. 303-320, 1980.

[14] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Oxford Clarendon Press, 1965.

[15] N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction", J. Math Phys., 25, pp. 261-278, 1947.

[16] P. Delsarte, Y. Genin, Y. Kamp, "A Polynomial Approach to the Generalized Levinson Algorithm, Based on the Toeplitz Distance", IEEE Trans. IT 29, pp. 268-278, 1983.

[17] J.H. Wilkinson, Error Analysis of Direct Methods of Matrix Inversion", J. Assoc. Comp. Mach. 8, pp. 281-330, 1961.

[18] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.

[19] S.M. Kay, "Recursive Maximum Likelihood Estimation of Autoregressive Processes", IEEE Trans. ASSP, 31 (1), pp. 56-65, 1983.

[20] A.J.E.M. Janssen, R.N.J. Veldhuis, L.B. Vries, "Adaptive Interpolation of Discrete-Time Signals That Can Be Modeled as Autoregressive Processes", IEEE Trans. ASSP, 34 (2), pp. 317-330, 1986.

[21] R.N.J. Veldhuis, "A Method for the Restoration of Burst Errors in Speech Signals", SIGNAL PROCESSING III: Theories and Applications, I.T. Young e.a. (ed.), pp. 403-406, (North-Holland, Amsterdam, 1986).

[22] R.N.J. Veldhuis, A.J.E.M. Janssen, "A Unified Approach to the Restoration of Lost Samples in Discrete-Time Signals", to appear in the Proceedings of the 1986 Asilomar Conference.

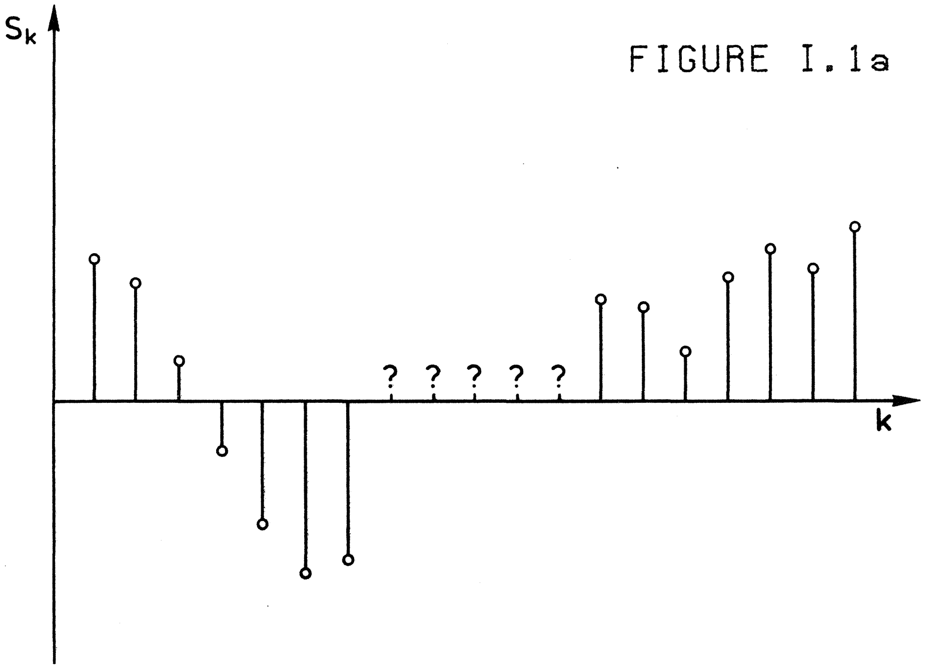


Fig. I.1a Sequence containing a burst of unknown samples.

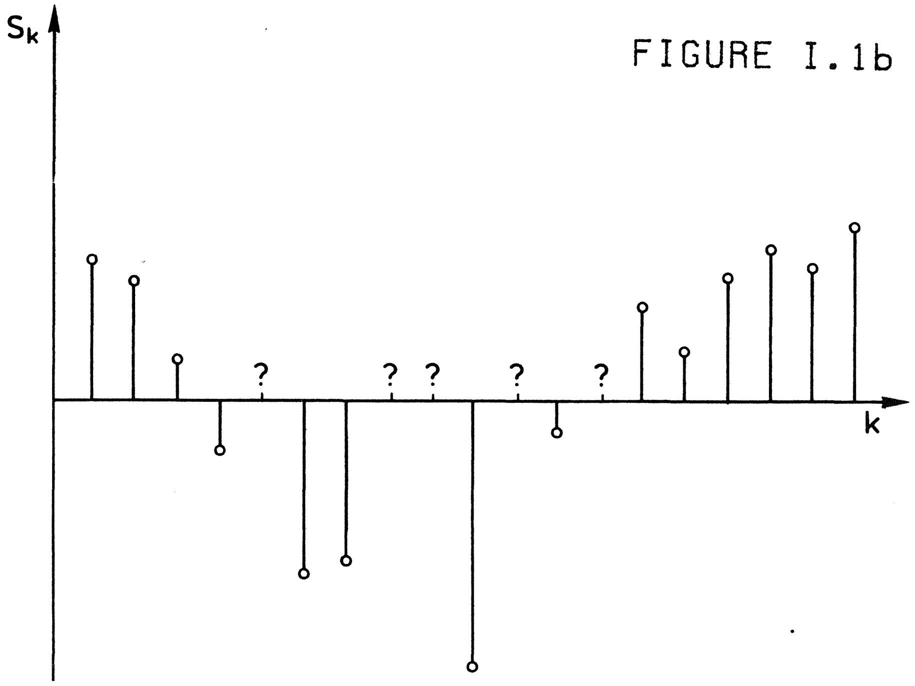


Fig. I.1b Sequence containing a random pattern of unknown samples.

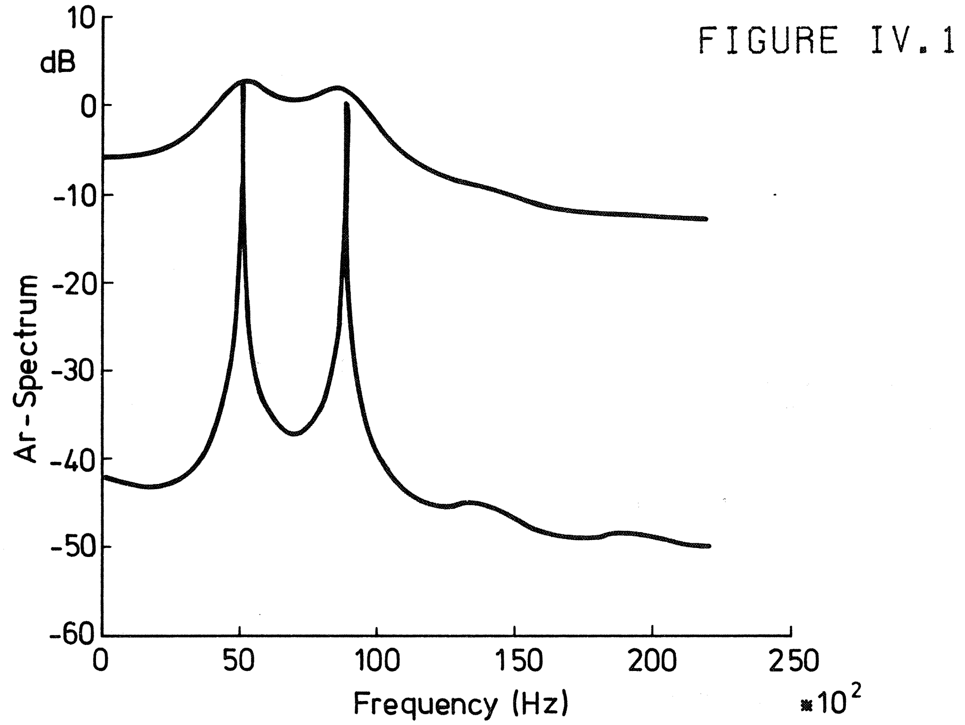


Fig. IV.1 AR-spectra of test signals 1 and 2.

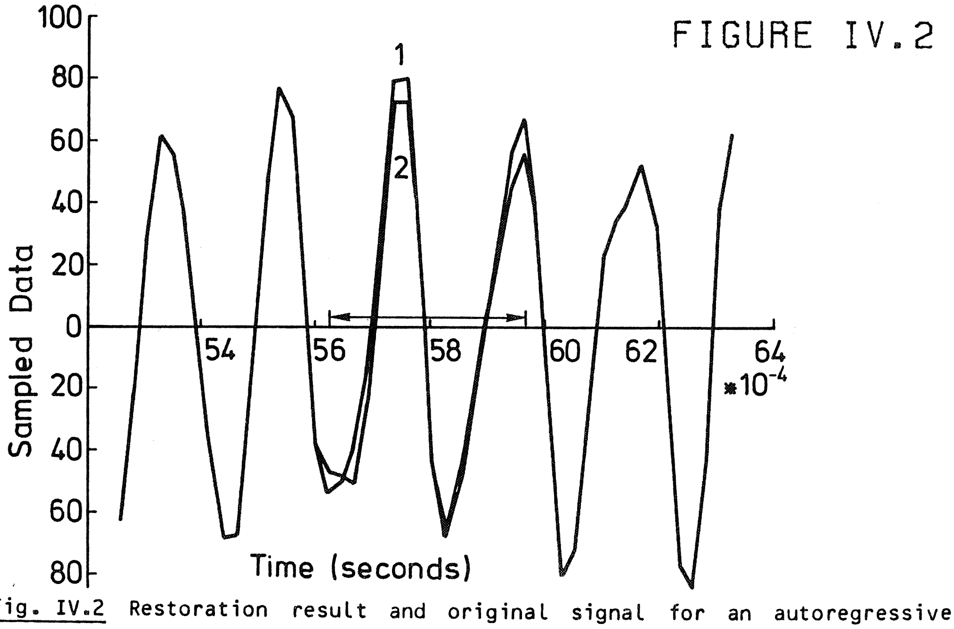


Fig. IV.2 Restoration result and original signal for an autoregressive process with a peaky spectrum, m=16, p=10, N=512, after 1 iteration. Restoration error e=0.23E-01.

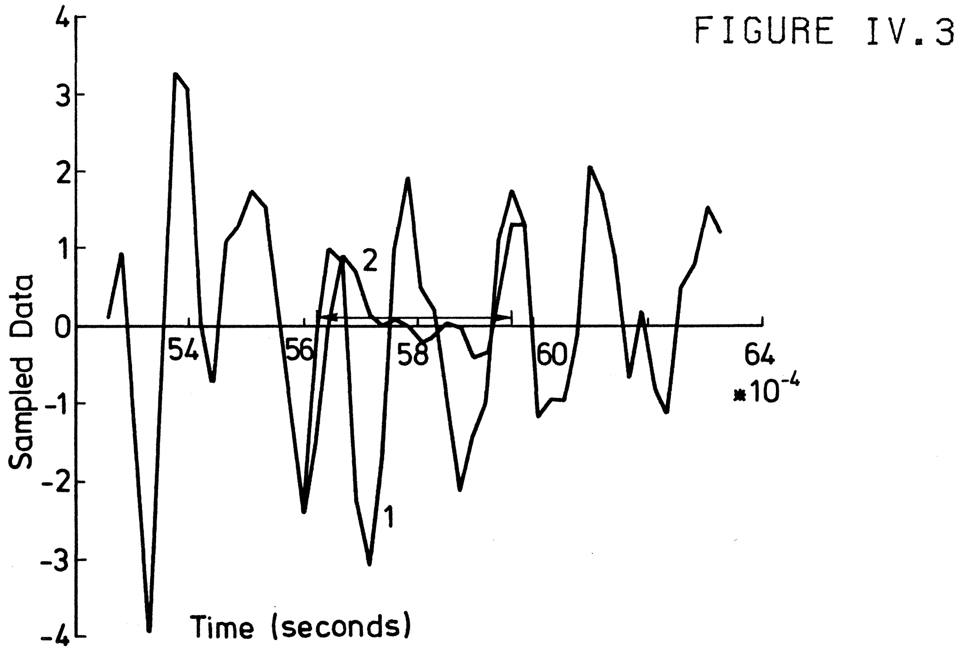


Fig. IV.3 Restoration result and original signal for an autoregressive process with a smooth spectrum, m=16, p=10, N=512, after 1 iteration. Restoration error e=0.10E+01.

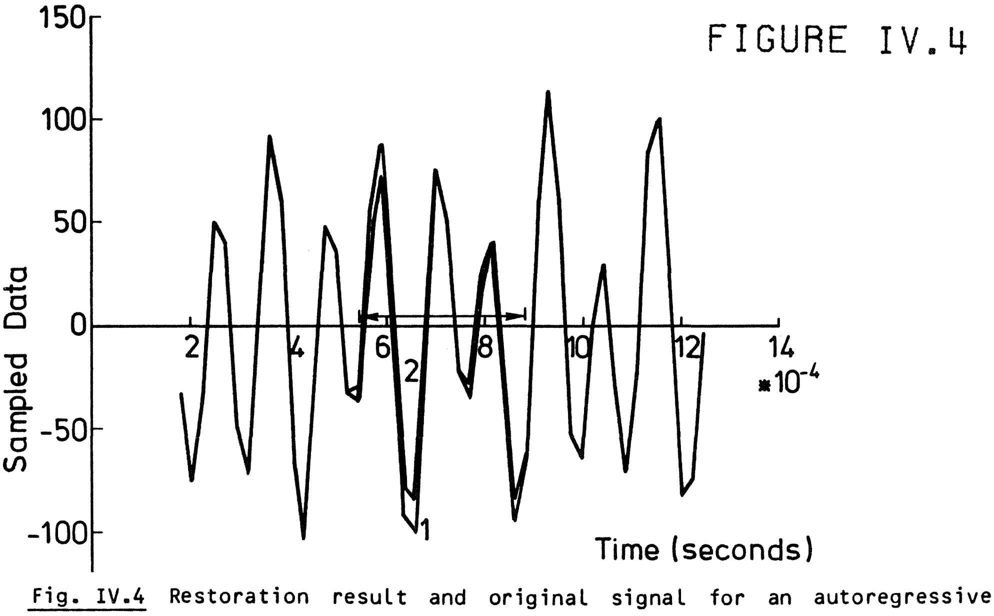


Fig. IV.4 Restoration result and original signal for an autoregressive process with a peaky spectrum, m=16, p=10, N=64, after 1 iteration. Restoration error e=0.38E-01.

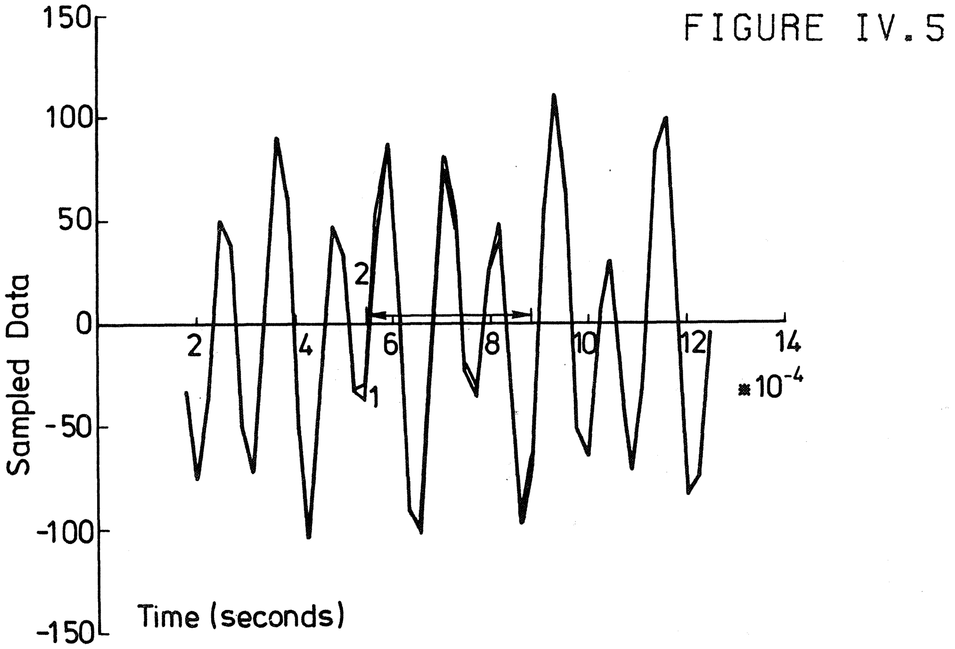


Fig. IV.5 Restoration result and original signal for an autoregressive process with a peaky spectrum after 3 iterations, m=16, p=10, N=64, after 3 iterations. Restoration error e=0.12E-01.

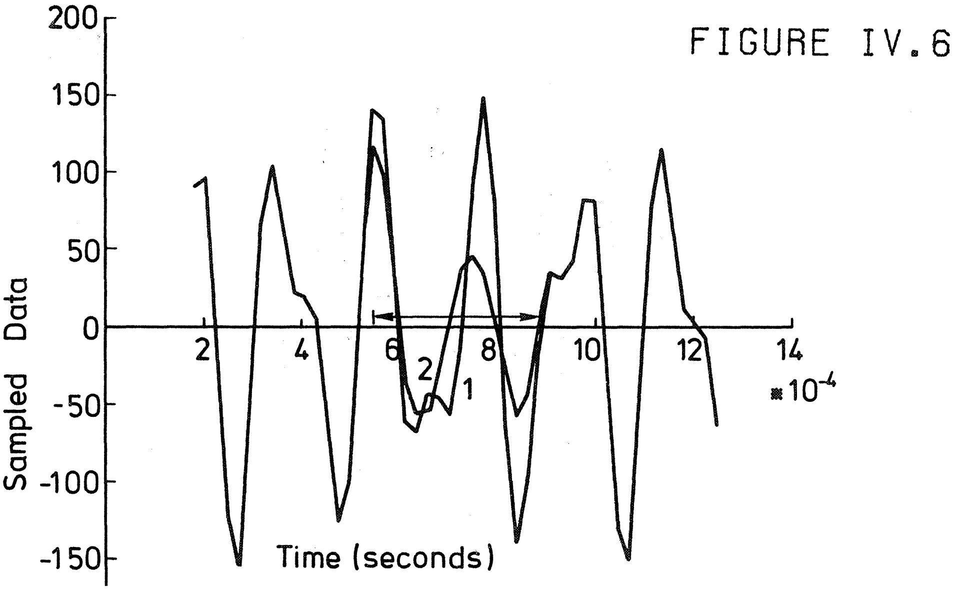


Fig. IV.6 Restoration result and original signal for a sum of 2 sine waves, m=16, p=10, N=64, after 1 iteration. Restoration error e=0.52E+00.

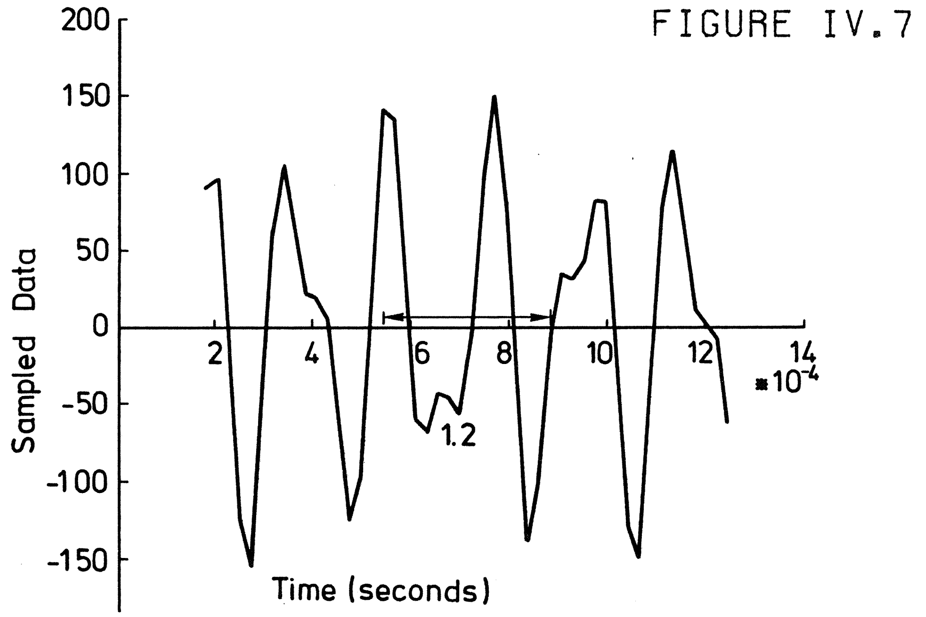


Fig. IV.7 Restoration result and original signal for a sum of 2 sinusoids after 3 iterations, m=16, p=10, N=64, after 3 iterations. Restoration error e=0.58E-02.

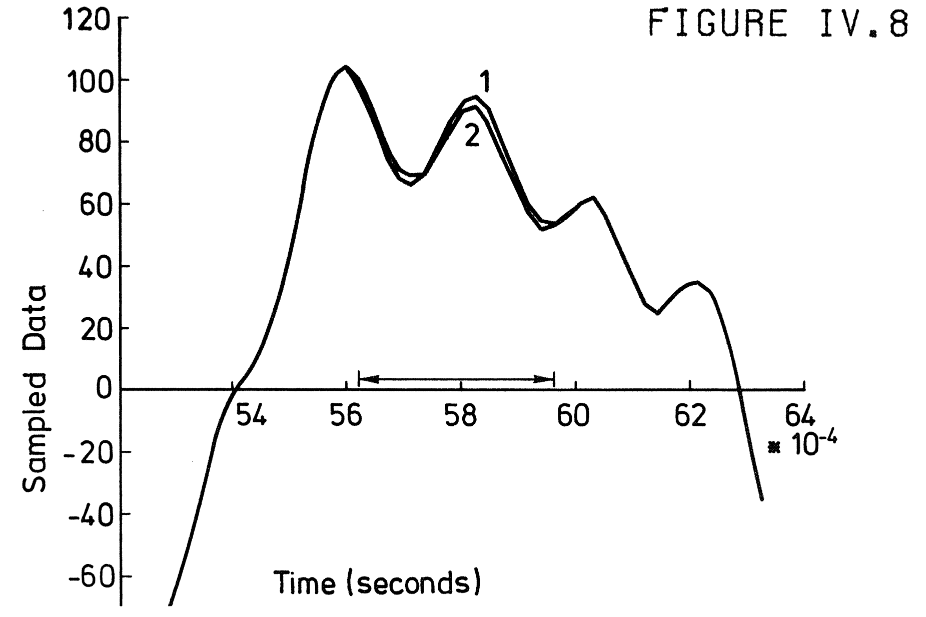


Fig. IV.8 Restoration result and original signal for a music signal, m=16, p=50, N=512, after 1 iteration. Restoration error e=0.22E-01.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Steps Into a Geometer's Workbench

Varol Akman
Department of Interactive Systems
Center for Mathematics and Computer Science
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

ABSTRACT

As computational geometry matures, it becomes crucial to use its techniques in the professional environment where graphics and robotics are natural candidates. This however is a nontrivial task since (i) computational geometry concerns itself with asymptotic analysis, and (ii) in search of elegance it ignores the special cases which are the bugbear of practical applications. I see experimentation as a way to resolve these difficulties and propose a software system to act as a "Geometer's Workbench." This entails the integration of geometric knowledge with algorithm animation and object-oriented graphics. The workbench should allow improvisation with geometric objects and is expected to broaden the way geometry is used in the style Macsyma[†] *accomplished for algebra.*

INTRODUCTION

Like other key areas of mathematics of the old times (most noticeably algebra and number theory) geometry is being revitalized after a long period of dormancy. The counterpart of "classical" geometry in the age of computers is "computational" geometry. In the latter, we are interested in designing efficient algorithms for tasks of geometric nature. For example, we may want to know the inherent complexity of identifying the region in a subdivision of plane (respectively space) by algebraic curves (respectively surfaces), enclosing a given point — a problem popularized as *point-location.* Here classical geometry must be augmented to deal with a new notion, namely, the complexity of computation. In the lack of computers old geometers did not concern themselves with such efficiency problems.

I am persuaded that if computational geometry is maturing (as many people say) then it is crucial to use its techniques in the professional environment where computer graphics and robotics are natural candidates. This however is difficult since computational geometers are traditionally most concerned with asymptotic analysis, and in search

[†] Macsyma is a product of Symbolics, Inc.

of elegance, underplay the special cases which are the bugbear of real problems[§].

My proposal, admittedly the first one that comes to mind, is then a software system, a so-called "Geometer's Workbench," which incorporates geometric knowhow and interactive graphics to assist in experimenting with geometric algorithms. It is a remarkable fact that computational geometry has advanced so much since its inception a decade ago. There are now literally hundreds of references (including several books) on computational geometry and a substantial number of these works deal with graphics and robotics problems. The reader is referred to Edelsbrunner et al [6] and van Leeuwen [17] who present fundamental ideas on the relationship between computational geometry and computer graphics. As for the ties of robotics and computational geometry, my dissertation [1] may be a good place to start.

1. MACSYMA AS AN ALGEBRAIST'S WORKBENCH

My preliminary thoughts about a geometer's workbench owe to Macsyma [18], a sophisticated computer algebra system built to assist researchers in solving mathematical problems. A user enters symbolic input to Macsyma which in return yields symbolic output. A great deal of knowledge has been stored into Macsyma's knowledge base. The user has access to mathematical techniques which he may not even fully understand but can easily employ to solve his problem. Conjectures can be tested easily and fast with Macsyma. The system is simple to use but not at the expense of being simplistic; several problems may require serious programming in the Macsyma command language and mastering the "insides" of the system. In short, Macsyma gives the user room to study problems from a more intellectual viewpoint, i.e. leaving the low-level, uninteresting computational details to the computer. It offers an extensible and exploratory programming environment.

Doing geometry, like algebra and many other research endeavours, is an iterative process. We define problems, draw figures, pose conjectures, redraw things, revise our ideas, etc. This exploratory process must be equipped with effective aids to graph data, to draw 2- and 3-dimensional figures to convey as many relationships as possible, to discover properties, and to store all this information in a meaningful and easily retrievable format. These tools must not require a large amount of initial training but have to be powerful. Generally, they should only make their functionality visible to a user, but when required the internals of the system should support reprogrammability and editability. These requirements are satisfied in one way or another by Macsyma. The Lisp programming environment that has evolved during the past two decades of artificial intelligence research also delivers these. Lisp machines, for instance, combine the Lisp programming environment with powerful graphics. Since Macsyma's base language is Lisp, these machines naturally support Macsyma. At the top level of a Lisp system is a read-

[§] cf. Forrest [8] for an excellent account of special cases. Don Knuth's detailed analyses of algorithms in his classical books run counter to the big-oh trends of today.

evaluate-print loop that reads expressions from the input stream, evaluates them, and prints the outcome on the output stream. Flexible structure editors, debuggers, and execution tracers provide a rich environment for *rapid prototyping,* an emerging pragmatical philosophy in software development. Windows enhance the interaction and, menus and use of a pointing device such as a mouse frees the user from being keyboard-bound. All of the above features must be present in the envisaged geometer's workbench.

2. HANDLING SPECIAL CASES

What kind of geometric knowhow would one expect from a workbench built upon such a workstation? First and foremost, it must be possible to perform conceptually trivial operations such as Voronoi partitions (cf. Figure 1), convex hulls, polyhedral boolean operations, and so on without undue emotional trauma. It is known that implementation of even the simplest geometric algorithms is difficult because of numerical problems and the number of special cases that warrant special care [8]. Franklin et al [14] mention the case of intersecting two polygons, a seemingly trivial operation which can result in about 1,000 lines of Fortran code once all the cases, such as polygons with multiple components that may or may not intersect the other polygon and whose edges may coincide with other edges or vertices, are taken into account.

Algorithm and data structure animation techniques [4] are found to be of crucial assistance in this respect. Since a personal workstation has a much friendlier user interface, the user may gain an insight by real-time observation of the outputs from algorithms. A good example for the need for the latter is derived from the weakness of a bare asymptotic analysis of an algorithm. It is not clear how efficient some of the asymptotically optimal say, Voronoi and point-location algorithms are when applied to scenes with moderate complexity. For instance, a theoretically ingenious algorithm of Richard Lipton and Robert Tarjan for planar point-location had a notice for the reader stating that the authors didn't think of it as suitable for implementation.

3. IMPROVISING WITH GEOMETRIC OBJECTS

Another key requirement for a geometer's workbench is the availability of good update facilities for the underlying geometric model. This refers to the users ability to add new geometric objects or delete the existing ones. It also embodies the concept of modifying (operating on) existing objects to obtain new ones. For instance, one should be able to take a cube, slice it in the corners and drill a hole in its middle to obtain a new object, and give it a name. One must be able to take the convex hull of say three polyhedra and create a new polyhedron. I call this kind of liberal approach in dealing with geometric objects "improvisation." In a very elegant early work Baumgart [2] and recently, Fogg and Eades [7] made some efforts in this respect. Pentland's [19,20] Supersketch^{TM} system depicts probably the state-of-art in supporting improvisation.

The preceding operations require that the system has a good understanding of what an object is. For example, if a cube has a hole it means that one can have a sufficiently small object pass through that hole; this would be trivial knowledge had the system possess a pair of eyes but in the lack of that it has to be stored in some way along with the cube. Similarly, it is normally an illicit operation to take the convex hull of two polyhedra, since the convex hull operation is defined for a set of points. However, the operation makes sense and must be allowed once it is understood that one is in fact dealing with the vertices of the polyhedra under consideration.

4. ADAPTIVE GRID AS AN IMPLEMENTATION TOOL

Adaptive grid is a data structure invented by Franklin [9] and can be thought of as a sort of hashing for geometric objects (instead of character strings). It is used to alleviate the problem of comparing everything with everything in order to detect the intersections among them. Several implementations using adaptive grid exist; Franklin and Akman [12] deal with hidden line removal via haloed lines, and Franklin and Akman [9, 13] give a hidden surface program for flat-faced polyhedra.

Let G be an integer ≥ 1 and assume that all the polyhedra are projected into the xy-plane. (That is, we fixed our viewpoint and carried out the preliminary transformations.) Without loss of generality, assume that the initial screen is a square of side 1, coordinates limited to $0 \leq x, y < 1$ and real. Essentially adaptive grid is a uniform $G \times G$ grid overlaid on the scene. The fineness $(\frac{1}{G})$ of the grid is some heuristically determined function of the statistics of the scene, e.g. average edge length, average face area, number of edges, number of faces, etc. The idea is to isolate the geometric objects (line segments or faces) into different cells so that they won't be compared to each other, as much as this is possible. Ways of roughly determining G are imaginable and we won't concern ourselves with it anymore. Besides, it is an experimental fact that changing the fineness within a factor of two makes little difference [9]. Now faces in the projection plane are entered into cells of the grid. Thus if a face has a common part with a grid cell, it is added to the list of faces in that cell. Note that this is *not* done by comparing the face under consideration against all the cells; the bounding box of the face will suffice. This way a few extra cells will be included but the algorithm will still perform correctly albeit a bit slower. For integrity reasons, each grid cell is held responsible for its interior, and additionally, its bottom (south) and left (west) sides. Since we excluded the coordinates with x or y equal to 1, the above provision partitions the scene into G^2 squares which are pairwise disjoint.

Obviously, within a cell visibility computation is carried out only with the faces that are in the face list of that cell. Hence we filtered out all those faces in the scene which are far away from this cell — thus the envisioned localization.

It turns out that the adaptive grid is especially perfect in determining which few pairs of a large number of short edges intersect. In this case the average execution time is linear in the expected number of intersections *plus* the number of edges, thus optimal within a multiplicative constant. Furthermore, since practical scenes tend to be resolution-limited[†] and frequently homogeneous, adaptive grid is also powerful even when the above assumption about short edges is relaxed. While one can think of using a hierarchical grid to accommodate regions of the scene where the edges are clustered more and while this would save time in scenes with orders of magnitude variation in edge density, as soon as cells become hierarchical formerly easy tasks such as determining the cells spanned by an edge become more complicated.

Figure 2 shows a set of cuoes spelling CS and is from a haloed line program. Haloed lines are used by drafting people in complicated drawings. Briefly we assume that each line has a "halo" that runs along it on both sides. If a more distant line intersects this first line, then part of the farther line that passes through the first line's halo is blotted out. We divided the haloed line computation into two disjoint steps. The first uses the adaptive grid to find all edge crossings fast and writes a set containing all the locations where each edge is intersected in front by another. The second step sorts the intersections along each edge and computes where the visible and hidden transitions take place. Dividing the computation into two steps means that redrawing a plot with a different halo width is quick since only the latter step need be rerun. Figure 3 shows a hidden line picture of a set of random blocks. Figure 4 was computed by the same program but painted on a raster scene. Figure 9 is from another hidden surface algorithm working with octrees. This algorithm uses another fast technique first to build the octree from a set of parallelepipeds and then to compute the visible voxels in back-to-front order. Since octrees support set operations efficiently by their nature, they are useful in interference detection problems.

5. SUPPORT FOR ROBOTICS

How does the Voronoi diagram on the boundary of a convex polyhedron change when the source point moves? Theoretically, this would amount to parametrizing the diagram's edge set with respect to the source coordinates so that how they change while the source moves on the boundary can be guessed. Note however that the change in the diagram will by no means be continuous, i.e. there will be certain "jump" points at which the diagram on a given face of the polyhedron will gain a new topology. Accounting for this effect seems messy. Randolph Franklin in a private communication (1985) suggested that one can make movies showing the effect of different locations of the source, to study this problem experimentally.

Given a boundary description for a polyhedron, one may be required to determine where the holes are. This problem has been completely solved with the well-known classification of 2-manifolds; however I am not aware of any practical program doing this for a given polyhedral description. Also note that, as long as the source and/or the goal is not inside it, a *bounded cavity* cannot contribute to the minimal path computation and thus can be "filled."[#] Curved objects make minimal path computations extremely

[†] People don't create scenes with enormous variations; they either detail blank expanses or simplify crowded regions with clarifying annotations.

[#] The implicit assumption here is that the entrance of the cavity is planar. When this is violated difficult, e.g. there may be an innumerable number of minimal paths. This is a domain where utilization of the variational calculus techniques may prove useful. Minimal paths on fancy objects such as Möbius bands and Klein bottles are also confusing.

When subdividing the space to compute minimal paths to any goal around polyhedra [10] we are particularly interested in finding the intersection curve of two arbitrary surfaces efficiently and reliably. The latter requirement necessarily dictates a symbolic approach to the problem since there may be all kinds of degeneracies. Another relevant problem is to enumerate the regions of space separated by several surfaces which may intersect each other in all conceivable ways. Although there are many relevant results on the intersections of algebraic varieties in the area of algebraic geometry, their introduction to the realm of computational geometry has been started only recently by George Collins and his students.

6. INTERACTION AND THE MVC TRIAD

The meaning of interaction is just too wide to be employed without some explanation. We accordingly offer a description of what we mean by this term and then offer a more formal viewpoint based on Smalltalk's Model-View-Controller paradigm.

Imagine yourself looking at a graphics screen. You normally see a hidden surface picture of say a machine part or a building. There are several regions on the 2-dimensional screen which have different colors, shadows, transparencies, etc. The important thing is that they are all disjoint since the hidden surface remover already handled the overlapping parts suitably. A particularly interesting interaction is then as follows. You point with a mouse to a region on the screen and pick it. There are several alternatives to what happens next. The following lists them in increasing order of sophistication in terms of user-friendliness:

- Picked region is highlighted.
- Boundary of the face which gave rise to this region is highlighted.
- Besides this region all other visible regions which are parts of the face which gave rise to this region are also highlighted.

All alternatives assume that the visible regions are kept not as a set of pixel values (as in ray tracing algorithms) but as geometric data*, e.g. polygons. Performing the second feedback operation is then easy since one keeps an identifier with each visible region. However the last operation may be inefficient; one must go through all the visible regions just to keep the necessary ones. It is my understanding that an ideal system should give the user the last feedback [16].

As another exercise in friendly visual interface, consider the following problem.

the filling must be done with care and only partially.

* This in turn dictates that one is using an object-space hidden surface algorithm in contrast to an image-space algorithm. See Sutherland et al [21] for details.

Construct the Voronoi tesselation of the 3-dimensional space by a given set of points. The question arises. How can one present the output in the most meaningful manner? Color would help, transparency would help, and finally the ability to selectively review regions would help.

As a formal model of interaction, Cunnigham's work [5] on the construction of Smalltalk [15] applications is relevant to the graphical interface that a geometer's workbench should provide. For other insightful views on graphical interfaces and their "power" the reader is referred to Williams [23] and Bier and Stone [3]. (Williams' paper is also very instructive in that it describes a workbench for economists.)

According to Cunnigham the right approach to building an application is three-fold:

- *Model* This consists of problem data and operations to be performed on it.
- *View* This presents information from the Model to the user via the display.
- *Controller* This interprets inputs from the user and modifies the Model or View accordingly.

In fact, it is quite correct to say that the Model represents the application while the View and Controller represent its user interface. An application may have several of the latter. Windows often provide several Views of a single Model, each different and each with a different Controller to deal with the inputs to that window.

Due to the object-oriented philosophy, any kind of object could represent a Model, View, or Controller as long as it obeys the demanded protocols. A View is not really concerned about the nature of a Model; all it cares is that the Model offers it some information to fill the screen. Similarly, a Model is only slightly aware of being viewed. It just provides answers to questions by its View(s). A Controller has the responsibility for receiving user input in the context of its corresponding View. Input may come from mouse or keyboard. The Controller detects the input and makes something happen. Mouse buttons and key strokes take on different meanings in different windows because different Controllers are listening to them. A Model should redraw when its model changes. There is no magic associated with this. Views are dependents of their Models. As a dependent, a View is sent a message "redraw" whenever its model is altered. Either a Model generates this message itself (as part of a modification protocol) or the change is dictated by a Controller following an editing operation.

7. SUMMARY OF REQUIREMENTS

I have only touched upon some key functionalities that a future geometer's workbench will have to provide. Only experience in developing prototypes will demonstrate the validity and completeness of my views. However, I believe that the main philosophy will stay more or less the same: a window- and menu-oriented user interface, a set of geometric functions similar in scope and generality to the algebraic functions of Macsyma, ability to pursue several computational activities in parallel using MVC-like paradigms, and algorithm and data structure animation.

Appendix: SP — A PROGRAM TO COMPUTE MINIMAL PATHS

SP consists of a family of programs written in Franz Lisp and Macsyma command language to experiment with minimal paths in the presence of polyhedral obstacles in 3-dimensional space. The following description is only cursory and the reader is referred to [1] for details of the system.

The program was designed with the following philosophy in mind. Let a workspace including a set of polyhedra be given. SP, using the geometric descriptions of the specified polyhedra, computes minimal paths in this workspace. It has some interactive graphics facilities and can supply the user with the views of the workspace so that he can have an intuitive feeling about the correctness of a particular computation. I believe that in geometric computations visual debugging is very effective.

In this sense, SP resembles to Verrilli's [22, 11] Voronoi-based system; it provides the user with facilities to carry out the needed computations, once in a while asking for his intervention here and there. To see the effectiveness of Verrilli's system, consider a minimal path following robot (idealized as a point) that must avoid a set of given walls in the plane. The robot starts from a fixed source point each time but goes to a different goal point. Using the *locus method* of computational geometry one can partition the plane into a set of regions (which turn out to be delimited by a collection of edges and hyperbolic portions) such that for every goal in a given region, the sequence of wall corners that must be followed to obtain the minimal path is the same. Thus in Figure 7 taken from Verrilli's thesis, if the goal is inside the shaded region then one knows that the minimal path is via corners 24 followed by 7 followed by 4 followed by 3. The problem is essential in manufacturing where there is a pile of parts in a location and a robot is supposed to carry the parts to many different locations (or in a fast food joint where you have to deliver hamburgers from a fixed location to many windows).

Following the prototyping approach I either simply excluded from SP those computations which I do not currently know how to perform effectively, or reformulated them to be controlled by user advice at certain points. Due to its loosely coupled structure, it is easy to upgrade SP with new algorithms when they become available.

Currently, one can work with a single convex polyhedron using Franz part of SP. There are facilities to solve Boundary Findpath, Exterior Findpath, and Boundary Findpath (locus). It is also possible to implement an approximate Findpath algorithm for a workspace with several convex polyhedra. Using Macsyma part of SP it is feasible to compute minimal paths in a general workspace although this is not fully automated in the light of the combinatorial explosion that known Findpath algorithms have. Nevertheless, if the user specifies the list of edges that the minimal path must touch, then the problem is solvable using a Newton-Raphson like method. There are also facilities based on Macsyma functions to deal with general Findpath (locus) but this is not automated yet.

Since it was built as a research tool, the prospective user is expected to know the internals of SP. Fortunately, the interactive nature of Lisp comes into play whenever one wants to debug or inspect the current computation and data structures. Working with SP is incremental in the sense that one computes things, stops and studies them (by plotting if necessary), and continues. To make a rough analogy, it is useful to visualize SP as a sophisticated calculator tailored for minimal path computations.

SP has facilities to read and check the consistency of polyhedral objects. It can also give extensive statistical information about an object. Once a polyhedron is read, SP builds the edge, vertex, and face data structures to access it easily. SP has a facility to unfold (develop) a given face sequence onto the xy-plane. In such a development all polygons must have z-coordinates either 0 or within the ε-neighborhood of 0. SP checks whether this constraint holds true. Figures 5 and 6 depict respectively an example path computed from a development and another computed similarly and then mapped back to the surface of the object.

For Exterior Findpath, facilities exist to compute visibility relationships and to construct the silhouettes. Then a new object is created and the minimal path computation proceeds routinely. For approximate path planning SP has a function to find the intersections of the given polyhedra with the source-to-goal line segment and to return a list of point pairs for each polyhedron intersecting the segment. Once these tuples are available a Boundary Findpath is performed for each pair and its associated polyhedron. Further path optimization can also be incorporated. For Boundary Findpath (locus), SP uses a naive Voronoi program. Figure 8 was generated by this program. Since the system is graphical, I needn't implement a point-location routine. For this figure the analogy is as follows. Assume that you have a set of construction sites on a mountain and a fixed location where you keep your tools. The idea is to efficiently compute the route of a truck carrying the tools to different sites. In this case the regions of the boundary of the polyhedron under consideration are delimited only by edges. Once the sequence of faces that a minimal path must visit is known, obtaining the path itself is trivial by unfolding the involved faces to the plane. The main cost of computation is then incurred in constructing the diagram itself since querying with different goals is just point-location which is asymptotically much cheaper.

I regard SP as a first-order, very modest approximation to the large and more general workbench I have proposed above.

Acknowledgments

I thank Randolph Franklin, Jan van Leeuwen, and Paul ten Hagen for providing the environments in which the ideas in this paper took shape. They haven't seen the manuscript and the responsibility for any shortcomings or errors is solely mine. Except Figure 7, all the pictures shown herein were computed by software developed by myself in collaboration with Franklin. Mention of commercial products in this paper does not imply endorsement.

References

1. V. Akman, ''Shortest paths avoiding polyhedral obstacles in 3-dimensional Euclidean space,'' Rep. IPL-TR-075, Image Processing Lab., Rensselaer Poly. Inst., Troy, New York (1985). (also: PhD thesis, Electrical, Computer, and Systems Eng. Dept.; Springer-Verlag, LNCS Series, to appear.)

2. B.G. Baumgart, ''Geomed: Geometric editor,'' Rep. STAN-CS-74-414, Computer Science Dept., Stanford Univ., Stanford, Calif. (1974).

3. E.A. Bier and M.C. Stone, ''Snap-dragging,'' *ACM SIGGRAPH'86 Proc.*, Dallas, Texas, pp. 233-240 (1986).

4. M.H. Brown and R. Sedgewick, ''Techniques for algorithm animation,'' *IEEE Software* **2**(1), pp. 28-39 (1985).

5. W. Cunningham, ''The construction of Smalltalk-80 applications,'' Manuscript, Computer Research Lab, Tektronix, Inc., Beaverton, Oregon (1985).

6. H. Edelsbrunner, M.H. Overmars, and R. Seidel, ''Some methods of computational geometry applied to computer graphics,'' *Computer Vision, Graph., and Image Processing* **28**, pp. 92-108 (1984).

7. I. Fogg and P. Eades, ''Ged — A graphics editor for computational geometers,'' Rep. No. 68, Computer Science Dept., Univ. of Queensland, Australia (1986).

8. A.R. Forrest, ''Computational geometry in practice,'' pp. 707-724 in *Fundamental Algorithms for Computer Graphics,* ed. R.A. Earnshaw, Springer-Verlag, Berlin (1985).

9. W.R. Franklin and V. Akman, ''Adaptive grid for polyhedral visibility in object space: An implementation,'' *Computer J.* **14**(3), pp. 117-123, to appear (1987).

10. W.R. Franklin and V. Akman, ''Euclidean shortest paths in 3-space, Voronoi diagrams with barriers, and related complexity and algebraic issues,'' pp. 895-917 in *Fundamental Algorithms for Computer Graphics,* ed. R.A. Earnshaw, Springer-Verlag, Berlin (1985).

11. W.R. Franklin, V. Akman, and C. Verrilli, ''Voronoi diagrams with barriers and on polyhedra for minimal path planning,'' *Visual Computer* **1**, pp. 133-150 (1985).

12. W.R. Franklin and V. Akman, ''A simple and efficient haloed line algorithm for hidden line elimination,'' *Computer Graph. Forum,* to appear (1986).

13. W.R. Franklin and V. Akman, ''Reconstructing visible regions from visible segments,'' *BIT* **26**, pp. 430-441 (1986).

14. W.R. Franklin, P.Y.F. Wu, S. Samaddar, and M. Nichols, ''Prolog and geometry projects,'' *IEEE Computer Graph. and Applic.* **6**(11), pp. 46-55 (1986).

15. A. Goldberg, *Smalltalk-80, The Interactive Programming Environment,* Addison-Wesley, Reading, Mass. (1984).

16. P.J.W. ten Hagen, A.A.M. Kuijk, and C.G. Trienekens, ''Display architecture for

VLSI-based graphics workstations,'' Rep. CS-R8637 , Center for Math. and Computer Science, Amsterdam (1986).

17. J. van Leeuwen, ''Graphics and computational geometry,'' Rep. RUU-CS-81-18, Computer Science Dept., Univ. of Utrecht, Netherlands (1981).

18. R. Pavelle, ''Macsyma: Capabilities and applications to problems in engineering and sciences,'' pp. 1-61 in *Applications of Computer Algebra*, ed. R. Pavelle, Kluwer Academic Publ., Boston (1985).

19. A.P. Pentland, ''Perceptual organization and the representation of natural form,'' Rep. TN-357, AI Center, SRI Int'l, Menlo Park, Calif. (1985).

20. A.P. Pentland, ''Supersketch^{TM},'' User manual, AI Center, SRI Int'l, Menlo Park, Calif. (1985).

21. I.E. Sutherland, R.F. Sproull, and R. Schumaker, ''A characterization of ten hidden surface algorithms,'' *ACM Computing Surveys* **6**(1), pp. 1-55 (1974).

22. C. Verrilli, ''One-source Voronoi diagrams with barriers — A computer implementation,'' Rep. IPL-TR-060, Image Processing Lab., Rensselaer Poly. Inst., Troy, New York (1984). (also: MSc thesis, Electrical, Computer, and Systems Eng. Dept.)

23. T.L. Williams, ''A graphical interface to an Economist's Workstation,'' *IEEE Computer Graph. and Applic.* **4**(8), pp. 42-47 (1984).

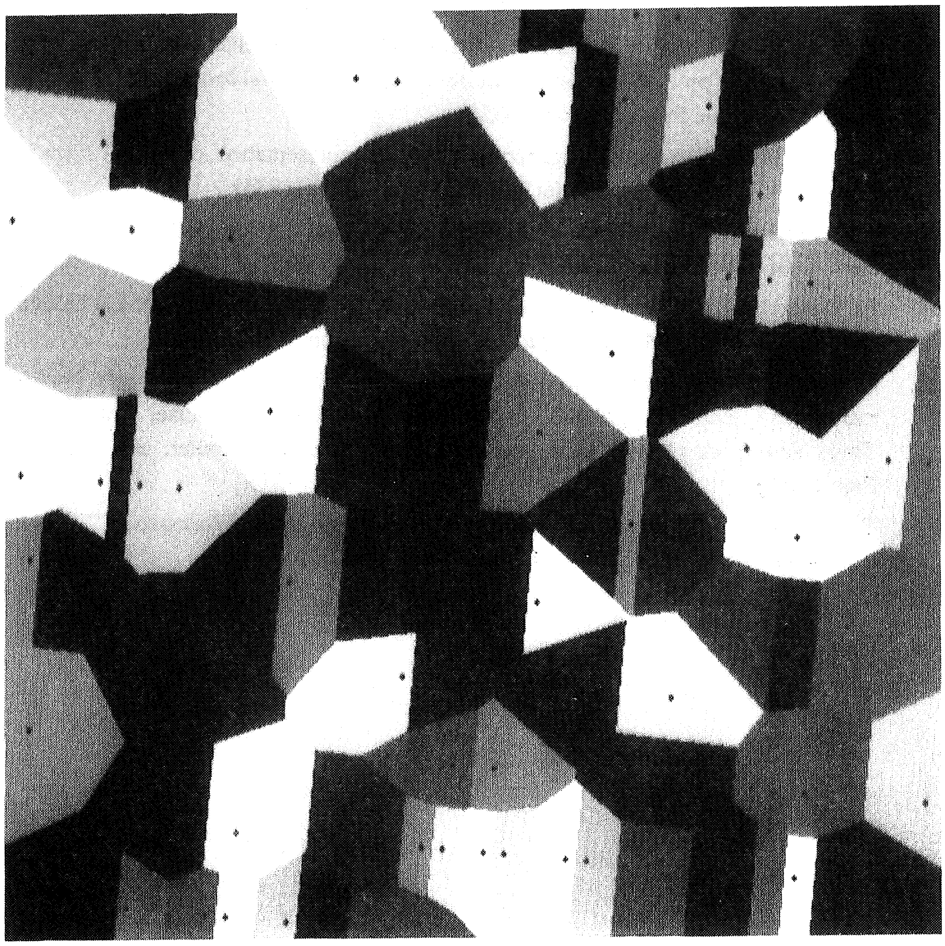


Fig. 1 Voronoi diagram of a set of random points

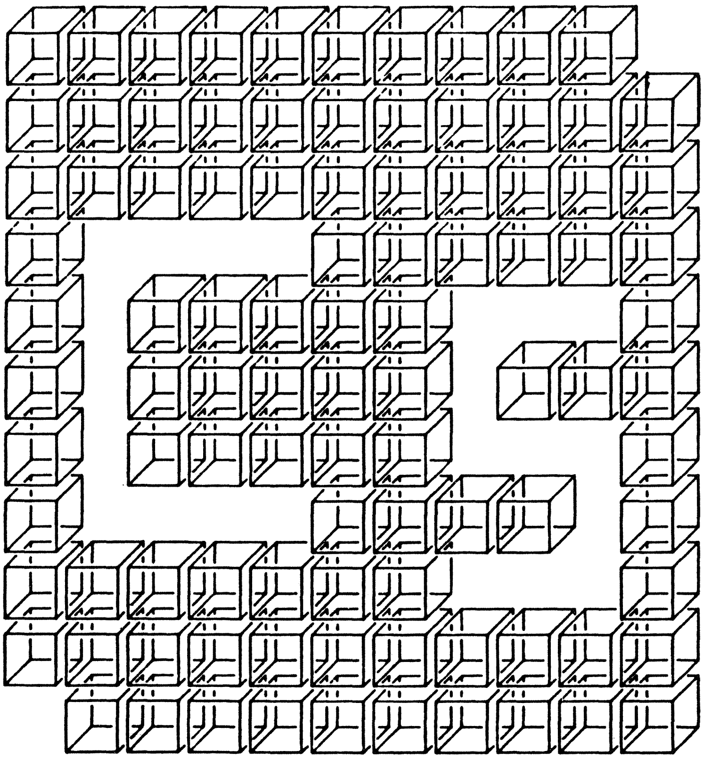


Fig. 2 An arrangement of cubes after haloed line computation

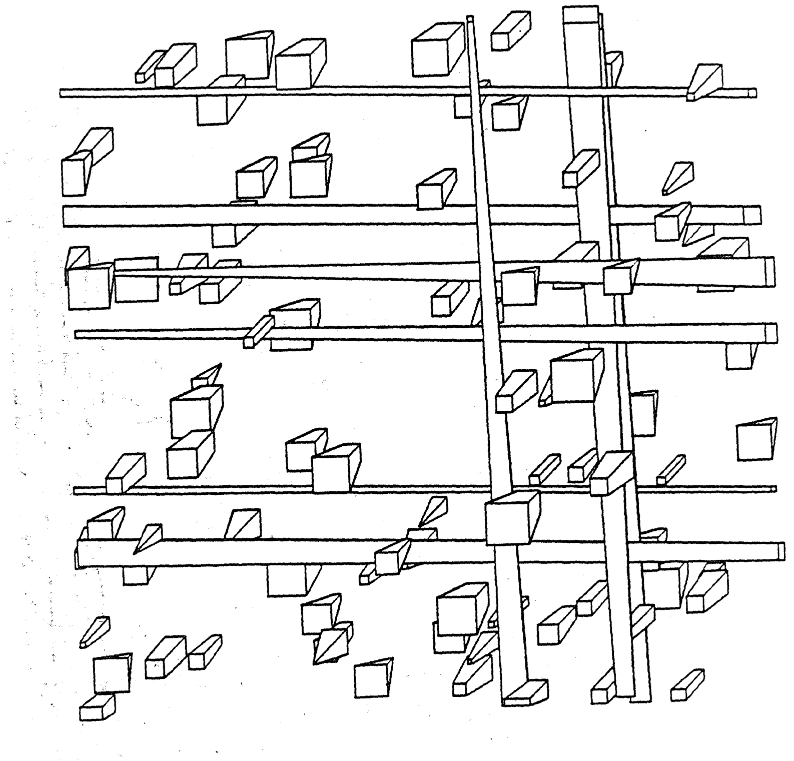


Fig. 3 A random family of blocks after hidden line computation

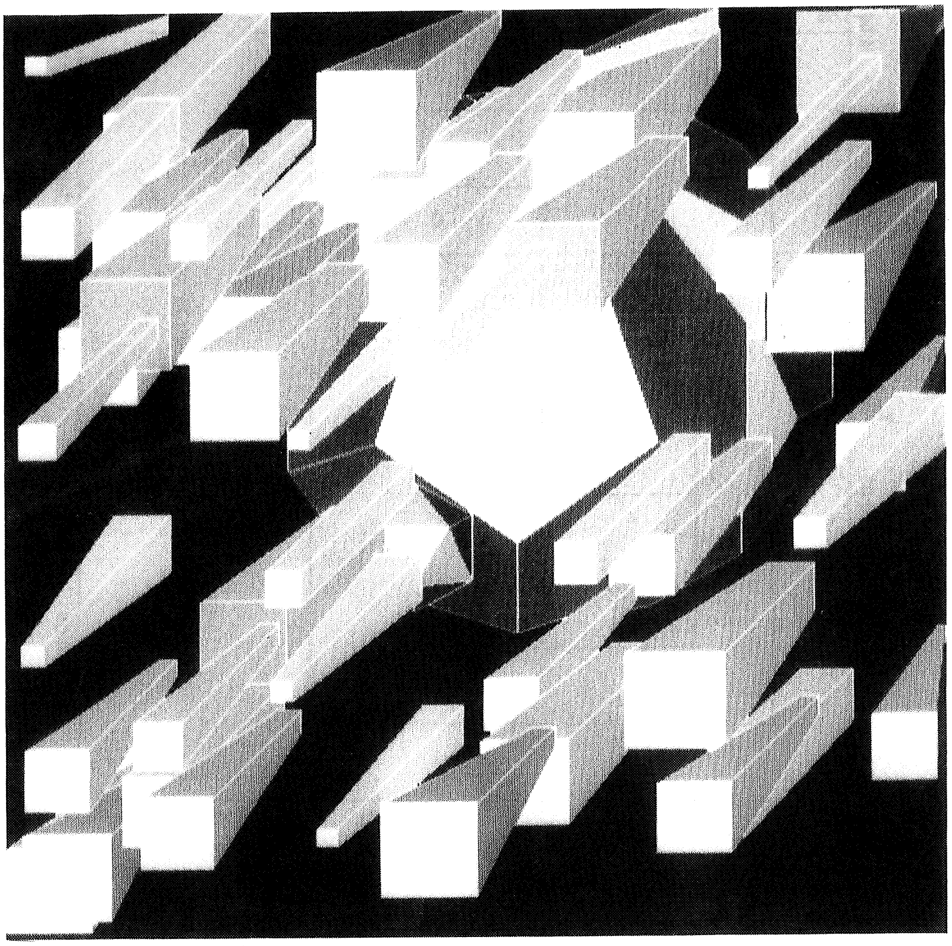
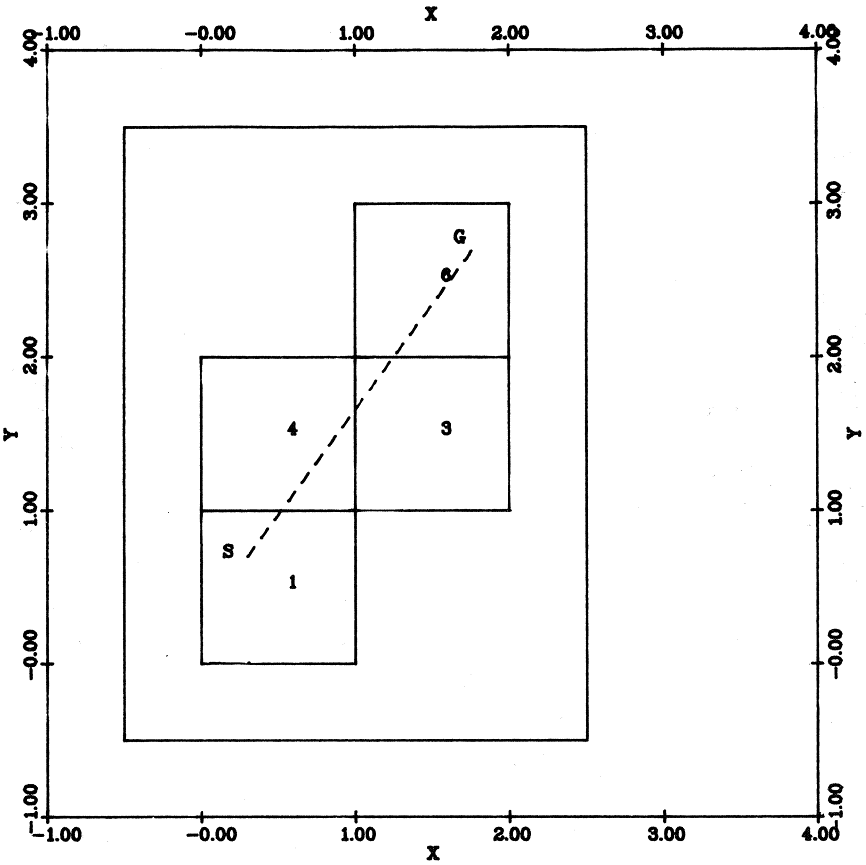


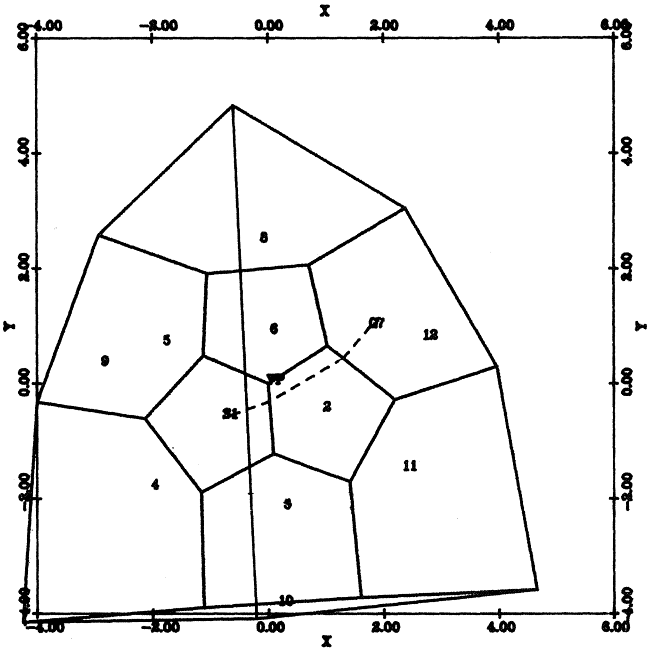
Fig. 4 A random family of blocks after hidden surface computation



```
Face development sequence:       1  4  3  6
Source face no.:                       1
Goal face no.:                         6
Source point coords.:                 0.000     0.700     0.300
Goal point coords.:                   1.000     0.200     0.250
Source point coords. (plane):         0.300     0.700
Goal point coords. (plane):           1.800     2.750
Source-to-goal distance:              2.540
Bounding box extents:                -0.500     2.500    -0.500     3.500
```

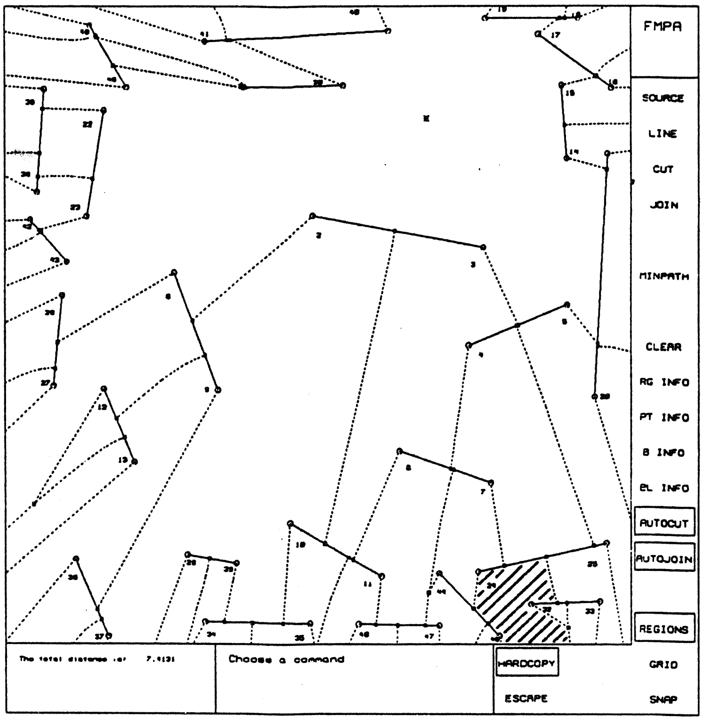
Fig. 5 A planar development for a face sequence of a cube

```
Visible face nos.:
 9
12
Invisible face nos.:
 1
 2
 3
 4
 5
 6
 7
 8
10
11
```



```
Viewpoint:                          1.500    0.800    2.500
Source face no.:                    1
Goal face no.:                      7
Source point:                        0.688    0.500    0.000
Goal point:                         -0.118   -0.085    1.611
Face development sequence:     1  2  7
Shortest path length:                2.618
Shortest path bend points:           0.688    0.500    0.000
                                     0.000    0.276    0.000
                                    -0.496   -0.085    0.996
                                    -0.118   -0.086    1.611
Bounding box extents:               -4.000    4.000   -4.000    4.000
```

Fig. 6 A minimal path mapped to boundary of an object in perspective

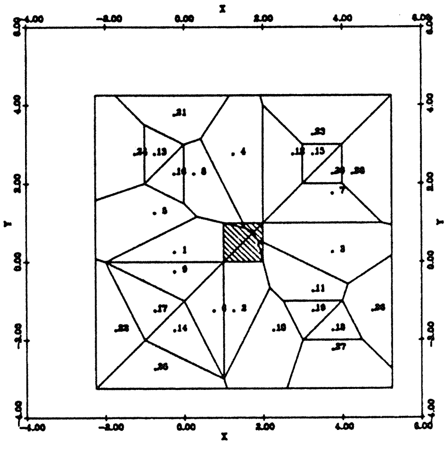


```
The total distance is:    7.4131
```

```
Choose a command
```

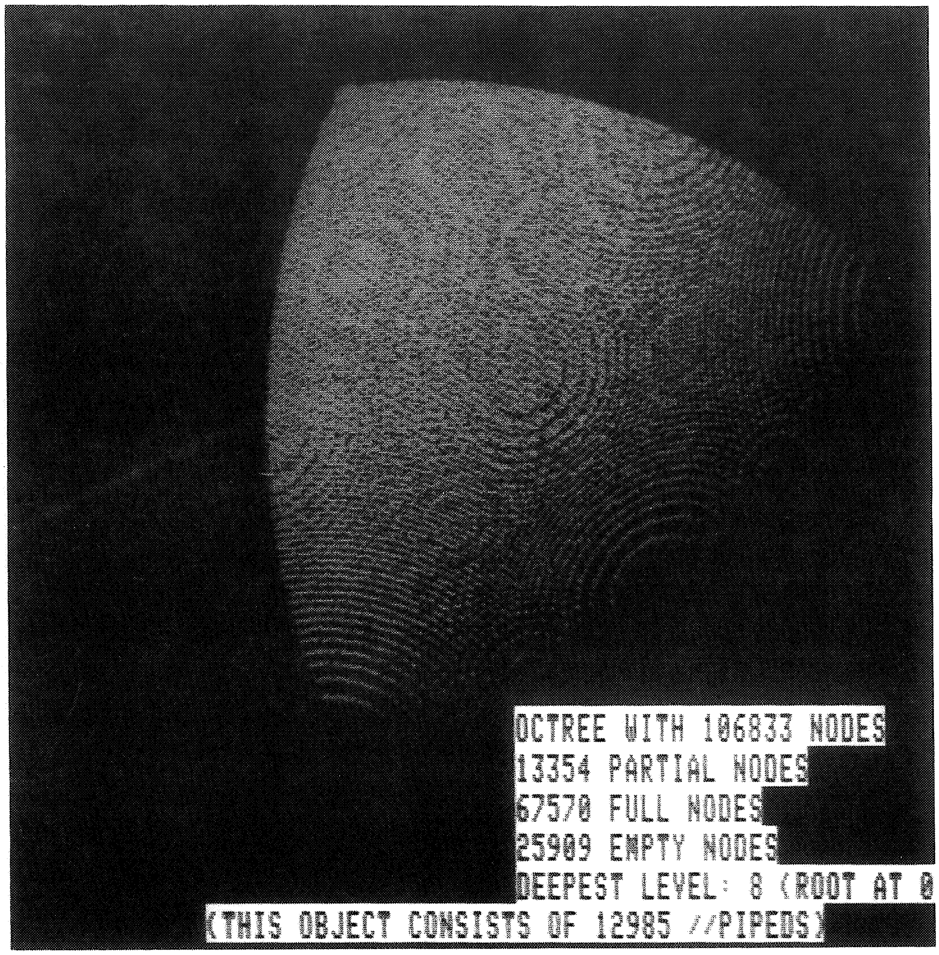
Fig. 7 Verrilli's single-source many-goals Voronoi system

```
Development sequences:
 1:   6   5   1
 2:   6   2   1
 3:   6   3   1
 4:   6   4   1
 5:   6   5   4   1
 6:   6   2   5   1
 7:   6   3   4   1
 8:   6   4   5   1
 9:   6   5   2   1
10:   6   2   3   1
11:   6   3   2   1
12:   6   4   3   1
13:   6   5   4   3   1
14:   6   2   5   4   1
15:   6   3   4   5   1
16:   6   4   5   2   1
17:   6   5   2   3   1
18:   6   2   3   4   1
19:   6   3   2   5   1
20:   6   4   3   2   1
21:   6   5   4   3   2   1
22:   6   2   5   4   3   1
23:   6   3   4   5   2   1
24:   6   4   5   2   3   1
25:   6   5   2   3   4   1
26:   6   2   3   4   5   1
27:   6   3   2   5   4   1
28:   6   4   3   2   5   1
```



```
Source face no.:                    1
Goal face no.:                      6
Source point coords.:                0.000     0.250     0.250
No. of developments:                28
Bounding box extents:               -2.250     5.250    -3.250     4.250
```

Fig. 8 Partitioning the face of a cube for several goals



```
OCTREE WITH 106833 NODES
13354 PARTIAL NODES
67570 FULL NODES
25909 EMPTY NODES
DEEPEST LEVEL: 8 (ROOT AT 0
(THIS OBJECT CONSISTS OF 12985 //PIPEDS)
```

Fig. 9 An octree hidden surface display of a spherical part

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

On the Use of Digit Distributions in Pattern Recognition

J. van der Pol
Postbank N.V.
P.O. Box 21009, 1000 EX Amsterdam, The Netherlands
and
H.P.M. Essink and R. de Jager
Dr. Neher Laboratory
P.O. Box 421, 2260 AK Leidschendam, The Netherlands

ABSTRACT

Currently the (Dutch) Postbank processes on average 3.5 million giro orders per working day. Roughly 2 million of these orders are received in the shape of booking forms filled in by hand, or typewritten. In order to process the documents automatically the booking data is encoded on the document in Ocr-b writing. An automatic check on the code is performed with an optical reading machine (CODAL). It matches in real-time handwriting and Ocr-b code. Using a pattern classification method, the current average recognition rate per character is 98.5%, with an average document acceptance rate of 75%. Further upgrading of the average recognition rate is possible by implementing information on digit distributions as functions of the position in the coded numbers. The first-digit-problem is relevant in this context.

Key words: Pattern classification, contextual information, first-digit-problem.

1. INTRODUCTION

In most of the currently available commercially manufactured optical reading systems for the recognition of hand-, or typewritten character sequences, the detection and matching of the sequence is based essentially on the recognition of each individual character separately. This implies that in order to be able to process the character classification automatically, some separation procedure is performed on the sequence previous to the actual recognition procedure. Dependent on the conditons of the application the recognition procedure can be a one-one type of matching procedure in a sequential processing computer environment, a pattern classifier in a parallel processing computer environment, ect. A lot of branching is possible here both in the choice of 'perception model', the technical implementation of the model and the processing of the documents read. But most systems share the methodology of recognition on the basis of feature shapes of the separated characters (see e.g. Essink, 1986-a). Though restrictive as perceptual 'model', this procedure can be shown to give commercially acceptable results. The CODAL optical reading machine, for instance, used by a Dutch bank to match automatically handwriting and Ocr-b coding of booking data on optically readable giro order forms (account number of the creditor, transaction amount), currently reads 300 documents per minute with an average acceptance rate of 98.5% per character. With an average length of six digits for the (handwritten) account number of the creditor and on average five digits for the transaction amount the average document acceptance rate is 75% (Essink, 1986-b).

From the common observation, however, that each individual has a style of handwriting and that type set letters are grouped in fonts, it is obvious that a statistical pattern classification model treating characters as independently distributed observations is restrictive. Therefore, a further upgrading of the average recognition rate per character may be expected from generalizations of the recognition model employing the covariance structure in shape patterns over the total character sequence. The testing of 'multiple character pattern classifiers' is one of the current topics of research at the Dutch Dr. Neher Laboratory with respect to further improvement of the optical reading machine CODAL.

In this paper, however, the implementation and testing of another type of contextual information is considered. Here we will investigate the possibility of improving the pattern classifier using distributional characteristics of digits in number sequences. The motivation of this research lies in the often established fact that in large 'natural' sets of 'randomly' generated numbers, the digits 0,1,..,9 are generally not uniformly distributed over the number fields. This effect, if established, is particularly prominent in the first significant position (disregarding decimal point, leading zeroes and non-numerical preceding separators as stripes, stars, ect.). In such sets the relative frequency of numbers having leading digit i decreases monotonously with i, i=1,2,..,9, and the 'probability' of a number having a 1 as first digit is more than six times that of a number starting with the digit 9. Here it will be shown that the existence of the first digit phenomenon can also be established in the number collections of handwritten account numbers and transaction amounts as generated in the (Dutch) giro traffic. As a consequence it is clear that the addition of positional parameters to the character feature vector in a conventional single character pattern classification model must lead to an improved average recognition rate and lower digit confusion values. Preliminary results obtained from a test version of such a combined model support this view.

The structure of the paper is as follows. Section 2 starts with a short review of the history and some statistical aspects of the first-digit-problem. Then some examples of the phenomenon in financial data are reviewed which are relevant to the discussion of the pattern classifier. In section 3 the linear pattern recognition model currently used in CODAL is presented and the modification to 'contextual classification' is indicated. In section 4, finally, preliminary results of the modified classification model are presented, using a test set of approximately 40,000 characters.

2. The first-digit-problem

If an extensive collection of 'naturally' generated decimal numbers is classified according to the first significant digit of the numbers, i.e. without regard to the position of the decimal point, the nine resulting classes are usually not of equal size. Instead of finding relative frequencies of approximately 1/9 for each class of numbers starting with digit i, i=1,..,9, we often find class sizes decreasing with i, starting with approximately a 30% frequency for digit 1, an 18% for digit 2, and so on, ending with a less than 5% frequency for digit 9. More surprisingly, the class percentages p_i can be shown to have a simple logarithmic relation, namely

$$p_i = {}^{10}\log((i+1)/i),\ i = 1,2,..,9. \qquad (1)$$

Or, in a suggestive statistical notation,

$$\text{prob}\,\{x \epsilon D_i\} = {}^{10}\log(i+1),\ i = 1,2,..,9, \qquad (1')$$

with x from some number collection and D_i the set of real positive numbers whose decimal expansion begins with an integer ≦ i. Examples of large number collections obeying (1) are numerous, and can be found in almost any field where 'natural' numerical data can be accumulated (lengths of phone-calls, credit balances of bank accounts, area of rivers and lakes, number of newspaper items, ect., see Table 1).

Once recognized the first digit phenomenon usually puzzles the observer, and since its (probably) first report in Newcomb (1881) it has given rise to an extended literature (see Raimi, 1976, for a bibliography on the topic). We will not go into much detail here, but restrict the discussion of the problem to some general remarks with respect to its history and possible origins.

First some history. Newcomb (1881) noted "That the ten digits do not occur with equal frequency must be evident to any one making use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones". He formulated and investigated the problem, but dit not give a quantitative relation.
The logarithmic 'law' is presented in Benford (1938), together with many examples of the phenomenon (number of footnotes in books, black body radiation, area of rivers, death rates, ect.). The almost 'universal' presence of Benford's Law, as (1) is now usually coined, inspired Benford, and later also Furlan (Furlan, 1946), to formulate a universal principle underlying (1). According to Benford the phenomenon does not follow from the structure of our decimal number system, but is 'evoked' by 'Nature' itself. Nature counts geometrically e^o, e^a, e^{2a}, e^{3a},...., instead of arithmetically as man does a,2a,3a,... And a geometric series can be shown to obey (1) under relatively mild conditions (pg. 560 ff.). To support this view, Benford presents a variety of real-life examples of the phenomenon from medicine (growth curves), psychology (Fechner's Law), astronomy (star brightness scale), ect.

Table 1. Distributions of first significant digits in empirical number collections, and a X^2-test against Benford's Law

		Case(*)							
First digit	Benford's Law	1	2	3	4	5	6	7	8
1	30.1	30.0	31.0	31.0	29.5	29.0	20.8	30.4	31.2
2	17.6	18.0	16.4	16.2	16.2	19.9	20.5	21.6	18.3
3	12.5	12.0	10.7	12.4	12.0	16.0	19.6	10.7	12.3
4	9.7	10.0	11.3	9.5	10.1	14.6	17.9	8.0	9.2
5	7.9	8.0	7.2	7.5	9.6	7.9	12.8	8.9	7.6
6	6.7	6.0	8.6	6.7	6.1	4.1	2.0	6.5	6.2
7	5.8	6.0	5.5	6.1	2.3	2.6	2.2	6.1	5.5
8	5.1	5.0	4.2	5.8	9.6	3.2	2.0	4.1	5.0
9	4.6	5.0	5.1	4.8	4.6	2.7	2.1	3.7	4.6
X^2_8	-	0.2	1.5	0.3	6.7	8.2	26.1	2.0	0.2

(*) Case 1: Number of newspaper items (from Benford, 1938, Table 1)

2: Area of rivers and lakes (ibid.)

3: Lengths of phone-calls (Netherlands) (from Lisman, 1986)

4: Processed giro account numbers of creditors (Postbank, Netherlands, 1985)

5: Processed giro account numbers of debtors (Postbank, Netherlands, 1985)

6: Collection of giro account numbers (Postbank, Netherlands, 1985)

7: Giro transaction amounts (Postbank, Netherlands, 1985)

8: Credit balances of bank accounts (Postbank, Netherlands, ultimo Januari 1986)

Furlan (1946) presents a similar view by stating that "the inherent spectrum of natural number collections is the harmonic spectrum", where the harmonic spectrum stands for Benford's Law (*).

In later 'explanations' the problem is tackled from a probabilistic, or a number-theoretic model. Out of many contributions we would like to mention here the one using density arguments in connection with the notion of equidistributed sequences (Raimi, 1976, pg. 524 ff.). Let $\{t_n\}$ be a sequence of real numbers in the interval $[0,1]$. Then $\{t_n\}$ is equidistributed if for each subinterval $[a,b) \subset [0,1)$ we have

$$\lim_{k} k^{-1} \sum_{n=1}^{k} \delta(n) = b - a, \qquad (2)$$

with $\delta(n) = 1$ if $t_n \in (a,b)$ and $\delta(n) = 0$ otherwise.

Now let $\{s_n\}$ be a sequence in $\mathrm{IR}^{+}\backslash\{0\}$ and define $t_n = {}^{10}\log s_n \pmod 1$.

Then s_n has first digit $\leq p$ iff. $0 \leq t_n < {}^{10}\log (p+1)$. Thus if $\{t_n\}$ is equidistributed on $[0,1)$ then (2) holds with $[a,b) = [0, {}^{10}\log(p+1))$, and $\{s_n\}$ obeys Benford's Law. $\{s_n\}$ is then sometimes called a strong Benford sequence. Examples of strong Benford sequences are geometric sequences $\{ar^n\}$, with a a constant and r a non-rational power of 10, and asymptotically geometric sequences as, for instance, the Fibonacci numbers. Note that the sequence of natural numbers, IN, is not a strong Benford sequence.

A third and final class of 'explanations' of the first digit problem we would like to mention here shortly, is the class of probabilistic models using the scale invariance principle. The model then says that if Benford's Law is universally true, it must be scale invariant, since Nature is not known to prefer any unit of measurement. For example, if Benford's Law holds for areas of rivers and lakes in the English system, it must also hold for the same data measured in the metric system (Case 2, Table 1). To express this statistically, let D_i denote the set of all numbers of IR^{+} whose standard decimal expansion begins with an integer $\leq i$, $i = 1,2,..,9$.

(*) "Die den natürlichen Zahlenkollektiven eigentümlichen form des Spektrums ist das harmonische Spektrum". (pg. 443)

The scale-invariance of a set $\{x_n\} \epsilon IR^+$ is then defined by

$$(\text{prob } (x \epsilon D_i) = \text{prob } (x \epsilon kD_i), \text{ for all } k>0.$$

Now suppose that $\{x_n\}$ is drawn randomly from a distribution F: $IR^+ \rightarrow [0,1]$, so that prob $(x \leqq a) = F(a)$. F is chosen continuous and, for convenience, differentiable. F is an approximation of the finite situation, and need only be consistent with the sample. Define further $G(x) = F(10^x)$, and $H(x) = \sum_{k=-\infty}^{+\infty} (G(k+x) - G(k))$ for all $x \epsilon [0,1]$. Thus, if X is a random variable and F its distribution function, then G is the cumulative distribution for the random variable $^{10}\log X$ and H for the variable $^{10}\log X \pmod 1$.

Then we have

$$\text{prob } (x \in D_i) = \sum_{k=-\infty}^{+\infty} [F((p+2)10^k) - F(10^k)] =$$

$$= \sum_{k=-\infty}^{+\infty} [G(k + {}^{10}\log(p+1)) - G(k)] ,$$

which can be rewritten as

$$\text{prob } (x \epsilon D_i) = H({}^{10}\log(p+1)), \ p = 1,2,..,9. \tag{3}$$

If $\{x_n\}$ follows Benford's Law then it must hold that

$$H(u) = u, \ u = {}^{10}\log 2, {}^{10}\log 3,.., {}^{10}\log 9. \tag{4}$$

The question now is to find distribution functions F whose corresponding H has property (4). Or, more specifically, to find functions F whose corresponding H is uniformly distributed on [0,1],

$$H(x) = x \ , \ x \epsilon [0,1]. \tag{5}$$

Obviously, (5) implies (4) and is therefore a sufficient condition for the exactness of Benford's Law. Now let X be a random variable having distribution function F and suppose $H = {}^{10}\log X \pmod 1$ obeys (5). Then the variables 1/X and cX, where c is any positive constant, have the same property (Adhikari & Sarkar, 1968). Thus, if $\{x_n\}$ obeys Benford's Law, so does $\{1/x_n\}$ and $\{cx_n\}$, for any positive constant c. In other words, the first digit phenomenon is preserved under scale transformation. Conversely it is also clear that once a collection violates the Law no improvement can be expected from rescaling, or taking reciprocals.

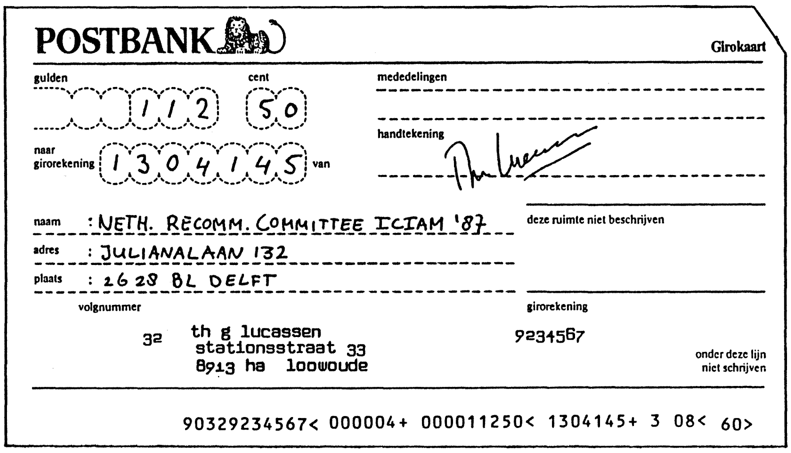
Surveying the three notions mentioned here in connection with the first-digit-phenomenon ('universality' of occurence, the relation with exponential and logarithmic sequences, scale invariance), and turning once again to empirical number collections, it follows that in large collections of financial data Benford's Law may also be expected to apply. Examples that lend themselves typically to investigation in that respect are the 'number flows' produced by retail institutes as insurance companies, retail banks, giro banks, ect. Selling their financial services and products mostly in terms of 'paperwork', daily large flows of numbers as a result of processing, evaluation and identification are generated (e.g. transaction amounts of cheques, credit balances of saving accounts, giro account numbers). In Table 1 it is shown that Benford's Law can hold very nicely in such number collections (Cases 4, 7, 8). Possibly there may also be some connection here with the well-known fact that financial data often have a log-normal distribution (see e.g. Johnson & Kotz, 1970).

On the other hand it is also obvious that in many cases Benford's Law does not hold. To convince oneself of that, it usually suffices to take, for instance, a telephone-book and survey the numbers. Also in financial data many counter-examples can be given (e.g. Case 6 in Table 1 representing the first-digit-distribution of a set of almost 5 million giro account numbers ranging from 1 to about 5,700,000). Summarizing, it shows that over a relatively long period of investigation on the first-digit-phenomenon and its origin(s) various mathematical and empirical arguments have been presented, but considered separately, or in combination, they do not provide a satisfying overall view of the problem. To illustrate this we conclude with just one example of what (in our opinion) remains curious and unexplained. Observe in Table 1 the Cases 4, 5, and 6. Each Case represents one and the same type of data (namely giro account numbers), but the numbers registered enter the collection in three different ways. Case 6 is the first-digit distribution of the entire collection of existing numbers, the Cases 4, and 5 are obtained by sampling account numbers from processed order forms. Now Benford's Law holds approximately for both the creditor - and debtor collection but it is definitely violated in Case 6. Apparently a (random?) dynamic component in the number generating process is somehow responsible for the appearance of the phenomenon, but the origin of this dynamical 'mechanism' remains unclear.

3. A linear classification model using contextual information

Having established significant departures from a uniform digit distribution in giro account numbers and transaction amounts, the question rises whether it is commercially meaningful to implement contextual information of this kind in a pattern classifier designed for the recognition of (handwritten) numbers. To investigate this an extension of the classification model must be developed and tested against the simpler alternative. Both steps will be discussed here, and illustrated in the setting of an existing pattern classification system. For a better understanding of the formal discussion to follow, first an introductory outline is presented of this system.

Currently the Dutch Postbank processes on average 3,5 million giro transactions per working day. Roughly 2 million of the orders are received in the shape of booking forms filled in by hand, or typewritten. The actual booking of these orders is computerized, so the handwritten information must be encoded in a computer readable form. For that purpose the OCR-b character set is used. The document, say a cheque, is processable once all the necessary booking information (transaction amount, the account number of creditor and debtor, ect.) is encoded on the document in the code-line (bottom line in Figure 1). This encoding is performed manually by (women) typists at a high speed (on average approximately 300 documents per hour). The account number of the debtor is preprinted, the account number of the creditor and the amount are added. As a result of the high encoding speed an average 2% of the documents contains one or more coding errors. Thus to prevent erroneous booking, the coding must be verified. A straightforward method to perform such a verification is to repeat the encoding and to match both results automatically. Though easy to implement as a practical procedure, such a manually performed verfication process is labour intensive, slow in verification speed and of a relatively low quality, because individual perception and/or typing errors tend to correlate.



POSTBANK

Girokaart

gulden 112 cent 50

mededelingen

naar girorekening 1304145 van

handtekening

naam : NETH. RECOMM. COMMITTEE ICIAM '87

adres : JULIANALAAN 132

plaats : 2628 BL DELFT

deze ruimte niet beschrijven

volgnummer

girorekening

32 th g lucassen
stationsstraat 33
8913 ha loowoude

9234567

onder deze lijn
niet schrijven

90329234567< 000004+ 000011250< 1304145+ 3 08< 60>

Fig.1 Postbank cheque. On the bottom line, in Ocr-b, the preprinted owners account number and the manually encoded transaction amount and creditor account number.

Developing methods towards a total automization of the verification process is therefore an obvious strategy in trying to improve the speed and quality of document processing. To this purpose the Postbank uses pattern recognition techniques. Technically this implies that an optical scanning device produces an image of the handwritten information, which is separated by a computer in single tokens and further segmentated into isolated patterns (details e.g. in Essink, 1986-b). The patterns are 'identified' using a statistical recognition model, and the resulting digit-output is compared with the OCR-b coding. If code and handwriting match, the document is marked 'okay'. If not, it enters a manual verification procedure similar to the one previously described. Figure 2 shows a box diagram of the complete automatic verification system, coined CODAL. Documents travel with a velocity of 1 m sec from left to right through the machine, which is indicated by the fat curved line in Figure 2. Five times per second a document is fed into the traject, and the documents move continuously to either an 'okay', or a 'reject' pocket. The verification system is build around a commercially available document sorting machine (NCR Company).

Omitting the details with respect to image segmentation and feature extraction (a description can be found in Essink, 1986-b), it suffices to state that the actual classification process starts with a 544 x 1 binary feature vector, representing the essential shape characteristics of the token read. It is a composition of 48 discrete features v_j, $j = 1,2,\ldots,48$, ranging between 1 and k (k=8, or 16). The feature values are represented by binary vectors of length k, having a one at element v_j, and zeroes otherwise. So, each binary sub-vector consists of a 1 out of k code. In extracting the features form the image instead of using the complete image itself, a great deal of data reduction is obtained, since the 'original' image is a 64 x 64 binary field (see Figure 3). The reduction in reliability of the classification as a result of the feature extraction is minimal.

Once the 48 discrete feature values of the unknown pattern are detected, the classification can be performed straightforward from the linear model

$$\begin{aligned} y &= A'x + \varepsilon, \\ d &= \max_i (y_i) \end{aligned} \qquad (6)$$

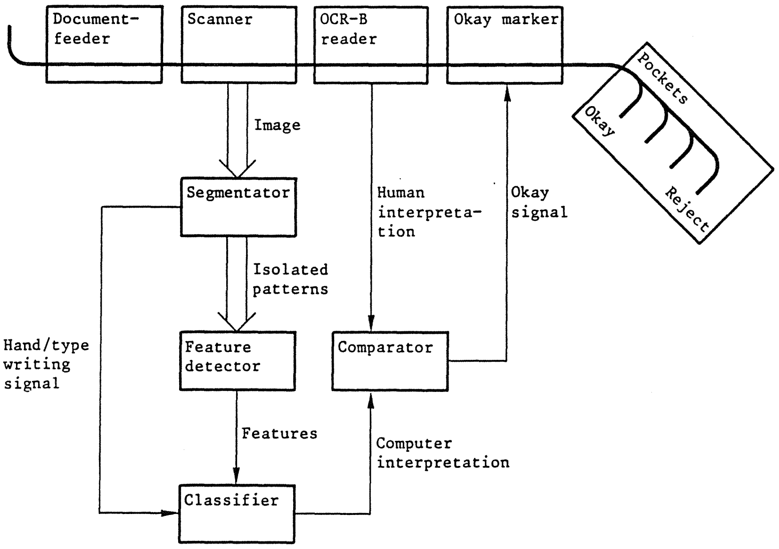
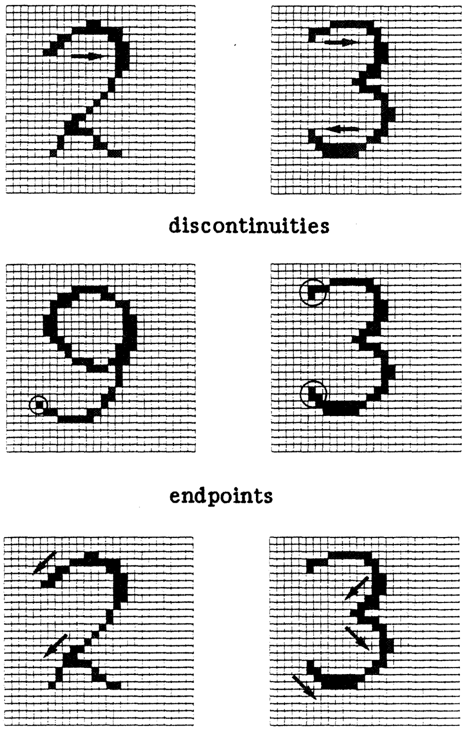


Fig. 2 System overview. Cheques are marked and sorted according to a match between computer interpretation and human interpretation in OCR-b.



discontinuities

endpoints

slopes

Fig. 3

Feature extraction. Example of three different features of the left faces of isolated patterns.

In (6) x denotes the 544 x 1 binary feature vector, A is a 544 x 11 matrix of feature weights called the recognition matrix, y is the 11 x 1 target vector of estimated probabilities y_i of class c_i, i = 0,1,..,9,10 (where c_{10} stands for the class of admissable non-numerals, as a stripe, star, ect.), ε is the 11 x 1 vector of residuals, and d is a maximum selector. Obviously, A has to be chosen so that the target vector y discriminates best (best in some statistical sense) between the classes. A may be a Bayesian classifier with (log) conditional densities log $p(x_k|c_i)$ as weights, k = 1,2,..,544, i = 0,1,..,10. Then y gives estimated a-posteriori (log) probabilities $p(c_i|x)$ that a pattern given in feature vector x belongs to class c_i. The conditional densities $p(x_k|c_i)$ are calculated from a training set of patterns with feature vector x and known class c_i (for details on the Bayesian classifier see Essink 1986-a). Here (6) will be considered as a regression model, i.e. with weights optimal in the OLS-sense. An unbiased and minimum variance estimate $\hat{A}$ of A can then be obtained by solving

$$\min_{A} \{tr(Y-A'X)'(Y-A'X)\}, \qquad (7)$$

with X a 544 x N matrix of feature vectors of N randomly sampled characters, and Y a 11 x N matrix of (known) target vectors (see for details, e.g., Duda & Hart, 1973, sections 5.8 and 5.12). Using well-known calculus (e.g. Lawson & Hanson, 1974) it follows that $\hat{A}$ is given by

$$\hat{A} = (XX')^{+}(XY'),$$

with $(XX')^{+}$ the Moore-Penrose inverse of XX'. Note that the inversion of the (possibly singular) 544 x 544 matrix XX' requires heavy computing. Once implemented and tested the recognition model (6) yields an acceptance rate of 97.2% on character level and 72% on document level. Further upgrading of A by adding 'filters' increases the rate to a current average of 98% accepted characters.

It is clear from the additive structure of (6) that the recognition model can easily be extended by adding independent features. Actually, this is the way in which (6) is optimized, namely by selecting the best discriminating shape features. Under the supposition that this selection process is optimal (i.e. in the MSE-sense), every new extension of the feature vector corresponds with a declining increase in acceptance rate, untill some break-even point is attained beyond which the processing 'parameters' discrimination, computer performance and processing speed cannot be improved without disproportional costs. Therefore, having an OLS-optimal parameter matrix A, further improvement of the classifier can only follow from using (6) in combination with other optimization criteria (i.e. looking for an estimate A with better first-order asymptotic properties), or developing a new, say non-linear, perception model. Concentrating on the first option, a possibility to improve (6) is to change from a mean squared residual criterium to a higher order power criterium. This in order to be able to capitalize on the bigger errors over the smaller ones. Then A is computed under the condition

$$\min \{(y - A'x)'(y - A'x)\}^p \quad , \quad p = 1,2,.. \qquad (8)$$

An unique weight matrix A exists for each value of p, but for p > 1 it cannot be expressed analytically in a closed form (see e.g. Devijver & Kittler, 1982, pg. 174 ff.). So in order to solve (8) one has to use some iterative procedure. Increasing p generally leads to less misclassifications in the training set of sampled characters. On the other hand, large values of p render the approach too sensitive to aberrant characters in the training set. Moderate values of p ($p \leqq 15$) usually give a good compromise.

A second possibility of improving (6) we want to mention here, is to allow for non-linearity in the feature relations by introducing quadratic terms in the binary feature vector. Then we have, for instance,
$x' = (x_1,x_2,..,x_{544},x_1^2,..,x_i x_j,..x_{544}^2)$ and optimization proceeds as in (7). But clearly the order of x can only be increased within certain limits, and so a selection of best discriminating quadratic elements must be made. Tests of some of these quadratic input classifiers showed promising results (Essink, 1986-a). And finally, combinations of these 'non-linear', non-MSE criteria showed promising results. Using a training set of 20,000 patterns sampled from the giro process, digit acceptance rates of 99.9%, and document acceptance rates of 87% were achieved (ibid., pg. 15). However, this extension is not as yet operational. In the present version of the CODAL machine model (6) is used with recognition matrix A optimal with respect to criterium (7).

A final variety of (6) we consider here more detailed, is the one using positional features in combination with shape features. Since both types of features may be expected to be independently distributed, extending (6) is simple. Positional features can be added to the feature vector without any restriction, the order of the recognition matrix A is changed accordingly, and for the optimization of A the OLS-criterium is used.

The question is, however, how many of these positional subvectors must be added and how. According to the results from the previous sections at least two positional features should be added, indicating the first guilder position (reading from left to right), and a first account number position (idem). For reasons of comparison it is also interesting to add at least second and third guilder/number position features to the model. Knowing that the class distributions of the second and following digits in the account numbers and transaction amounts rapidly converge to a uniform distribution, the contribution of the first digit features to the improvement of the model may be expected to be substantial, i.e. compared with that of the other digit features. To investigate this, we performed some experiments with various positional features. In the next section the model extension is formally denoted and the results of some tests on a set of characters classified accordingly, are presented.

4. A test of a combined positional pattern recognition model

Let x_1 and x_2 be two positional feature values, respectively, and A_1 and A_2 the corresponding recognition matrices. Then (6) can be written as

$$y = (A_1{}' \,|\, A_2{}') \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \varepsilon \;, \qquad (9)$$

$$d = \max_i \, (y_i).$$

Using criterium (7), optimal estimates of A_1 and A_2 were calculated from a training set of almost 40,000 characters. Concentrating on the first digit phenomenon primarily, it has been tested whether the phenomenon occurred in this set of some 4,000 combined transaction amounts and account numbers, and what the effect was of adding a first guilder feature to (6). Since in the training set (as in practice) non-numerals and non-siginificant zeroes appeared at the first guilder position, recognition rates were computed for the sets of first characters with and without the classes of non-numerals and zeroes. Finally, also the overall performance of the combined model (9) was tested.

Summarizing the results of the tests here, it is noted first that the training set applied to Benford's Law well enough to be representative of the actual collection of transaction amounts in the giro-process ($\chi^2_8 = 1.7$).

Further, comparing the recognition rates per class and the confusion matrices for the model both with and without a first guilder feature, all tests showed equal, or improved recognition rates for all character classes except for the class 7. To save space, only the results for the collection of 3958 classified characters read in the first guilder position are discussed here in some detail (see Table 2). A first observation is that the biggest increase in recognition rate is found for class 1 and, surprisingly, the class of non-numerals. In both cases this effect follows from the addition of the positional feature and can therefore be attributed uniquely to the dependance of the respective character classes on the position in the number.
Note in particular the improved discrimination in class 1 between the optically very similar digits 1 and 7 (from 12 to 9 incorrect classifications). The overall improvement in this set is about 0.25% on the digit level. Similar results, but with slightly minor overall improvement were obtained for the collection of first significant digits (i.e. without non-numerals and zeroes), and the total collection of characters read (resp. 0.1% and 0.04%). Details are available form the authors upon request.

Though seemingly marginal in absolute sense, the improvement in recognition rate indicates that applying the first digit feature twice (in account number and transaction amount) results in an expected 0.1% increase in the overall digit recognition rate, giving an average 1% increase in automatically accepted documents, or, complementary, a 4% decrease in manually checked documents [$(98,5\%)^{11} \doteq 85\%$, $(98,6\%)^{11} \doteq 86\%$]. Moreover, ranking a set of 750 selected feature components according to a successive maximum MSE-reduction criterium, it shows that the first guilder feature holds position 168 out of 650 linearly independent components. Thus positional features can be shown to perform well in a combined position/pattern recognition model.

Adding second and third guilder features to the positional binary vector, finally, showed that no further substantial improvement could be obtained. The separate and overall character recognition rates increased only marginally. Ranking features by maximum MSE-reduction again, the second and following guilder position features scored 450, 415, 518, 611 and 639, respectively, in a total set of 750 components.

Concluding, we state that the performance of pattern classifiers designed for the recognition of numerical (and related) characters, can be improved by implementing

Table 2. Recognition rates of first guilder characters, and confusion matrix of 3958 characters, combined model (between parentheses differences with model without positional feature)

	Class											
	0	1	2	3	4	5	6	7	8	9	NN*	all
recognition rate (%)	100	97.9 (+.5)	98.6 (+.4)	97.0 (+.2)	94.8	99.4	98.8	98.0 (-.5)	99.3	92.7	94.1 (+.5)	97.4 (+.2)
classified												
0	3	.	.	1	.	.	2	.	.	.	.	
1	.	995 (+6)	2	.	.	.	.	.	.	.	4	
2	.	4	703 (+2)	.	4	.	.	1 (+1)	.	.	3 (+1)	
3	.	1	.	396 (+1)	.	2	.	.	1	1	1	
4	.	5 (-2)	.	.	329	.	1	1	.	1	5 (-1)	
5	.	.	.	0 (-1)	.	349	.	.	.	3	2 (-1)	
6	.	.	0 (-1)	.	10	.	251	.	.	.	.	
7	.	9 (-3)	5	2	2	.	.	194 (-1)	.	1	7 (-1)	
8	.	0 (-1)	1	4	.	.	.	.	149	2	.	
9	.	1	2 (-1)	5	2	.	.	2	.	102	2	
NN*		.	1	.	.	.	.	.	.	.	384 (+2)	
												3855 (+10)
total	3	1016	713	408	347	351	254	198	150	110	408	3958

* non-numeral

statistically relevant positional information present in the data. The discussion has been restricted here to the first digit phenomenon in particular, but it is clear that similar research can be done with, for instance, transition distributions of digits. In certain types of the giro transactions discussed here, it has been established that the probability of a zero following any non-zero digit is almost 30%. Similarly, the probability of finding the digit 5 following any digit is more than 15%. Both results are statisticaly significant under the null hypothesis of having a uniform distribution of digit classes, and they indicate that further research on using contextual information in pattern classification is needed.

Literature

Adhikari, A.K. & Sarkar, B.P. (1968). Distribution of most significant digit in certain functions whose arguments are random variables. Sankya Ser. B., 30, 40, 47-58.

Benford, F. (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society, 78, 551-572.

Essink, H.P.M. (1986-a). Pattern recognition at the Dutch Postbank. In: E. Backer, et al. (Eds.). First Quinquennial Review 1981-1986. Dutch Society for Pattern Recognition & Image Processing. Pijnacker (The Netherlands): D.E.B. Publishers.

Devijver, P.A. & Kittler, J. (1982). Pattern recognition. London: Prentice Hall.

Duda, R.O. & Hart, P.E. (1973). Pattern classification and scene analysis. New York: Wiley.

Essink, H.P.M. (1986-b). Classificeren van hanepoten. Natuur & Techniek, 54 (10), 782-791.

Furlan, L.V. (1946). Das Harmoniegesetz der Statistik, eine Untersuchung über die metrische Interdependenz der socialen Erscheinungen. Basel: Verlag für Recht und Gesellschaft.

Johnson, N.L. & Kotz, S. (1970). Continuous univariate distributions I. New York: Wiley.

Lawson, C.L. & Hanson, R.J. (1974). Solving Least Squares Problems. Englewood Cliffs N.J.: Prentice Hall.

Lisman, J.H.C. (1986). Benford's law: Letter to the editor. Statistica, 46, 253-254.

Newcomb, S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers. American Journal of Mathematics, 4, 39-40.

Raimi, R.A. (1976). The first digit problem. The American Mathematical Monthly, 83, 521-538.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

A Stochastic Description of Copolymerisation and Network Formation in a Three-Stage Process

G.P.J.M. Tiemersma-Thoone, B.J.R. Scholtens and K. Dušek*

DSM Research, P.O. Box 18, 6160 MD Geleen, The Netherlands
*Institute of Macromolecular Chemistry,
Czechoslovak Academy of Sciences
16206 Prague, Czechoslovakia

ABSTRACT

With the theory of branching processes with cascade substitution a statistical description of a three-stage process of network formation is derived. End-functionalized prepolymers prepared in the first stage are modified in the second stage to prepolymers with a different type of end groups. These are subsequently crosslinked in the third stage with a mixture of hardeners. The distributions of units in the different reaction states are calculated with kinetic differential equations based on the mass action law. In these equations substitution effects can be taken into account.

1. INTRODUCTION

Crosslinking processes are technologically very important. They are used in a variety of fields in polymer technology, e.g. elastomers, adhesives, coatings and other thermosetting (e.g. construction) materials. By a crosslinking process the material is transformed from a monomeric or polymeric liquid into a permanent network, which is essentially a viscoelastic solid. Most crosslinking processes can be considered as multistage processes in which prepolymers are formed from monomers in one or several stages, upon which a network is obtained by crosslinking the functional prepolymers. Well known examples are, e.g., polyurethane networks, prepared with isocyanate-terminated polymers (macrodiisocyanates), which are synthesised from macrodiols with excess diisocyanate [1], or saturated polyester networks, prepared with carboxy-terminated polyesters, which are derived from macrodiols with excess dicarboxylic acid or anhydride [2]. Alternatively, a network formed in an intermediate stage can be further modified in a final stage.

In this contribution a theoretical scheme for a three-stage process of network formation is derived. It is based on the general scheme published recently [3], derived with the theory of branching processes with cascade substitution [4-6]. The Galton-Watson or universal consistency relation is assumed to be valid and cyclisation is postulated not to occur. Besides, substitution effects are allowed and can be incorporated in the probability generating functions (pgf's).

In the following paragraph a description of the three-stage process is given. Next the statistical method is applied for these three stages to obtain the pgf's and the relevant characteristics such as average molecular masses and functionalities in the pre-gel region, the conversion at the gel point and the network characteristics in the post-gel region. These pgf's contain the distributions of units calculated with a kinetic scheme described in the next paragraph (in the absence of substitution effects these distributions are more readily derived with statistical arguments). In that next paragraph a kinetic scheme is presented with which the fractions of units in the different reaction states are determined by kinetic differential equations based on the mass action law [7-11]. These distributions of units are thus a function of time or overall conversion and depend on the respective rate constants. It is in this scheme that substitution effects can be taken into account. Finally, some preliminary results are presented for a typical three-stage process.

2. DESCRIPTION OF THE PROCESS

The structure of the three-stage process is as follows.

Stage 1

Bifunctional monomers A, with functional endgroups called c, react exclusively with an excess mixture of difunctional monomers D and trifunctional monomers T, which have the same functional endgroups called h (and thus are equally reactive), to (mainly) h-terminated prepolymers.

Stage 2

The prepolymers of stage 1 are modified with an excess of difunctional C monomers, also with functionality c, into mainly c-terminated prepolymers. Unreacted functional groups of the A monomers are assumed not to react in this stage.

Stage 3

The mainly c-terminated prepolymers react in this last stage with a mixture of difunctional E and trifunctional F hardeners which are equally reactive. The h-endgroups are assumed not to react any further in this stage.

A scheme of the three stages is given in table 1. For these reactions substitution effects are allowed for monomers A and C in all three stages. In addition, monomer A is assumed to be completely insoluble in the reaction mixture and to be only reactive at the surface. As a result, the specific surface of A is an additional parameter.

3. STATISTICAL METHOD

Because of the complexity of the statistical characteristics of a three-stage process of network formation the three stages are treated separately. The statistical properties of the products are derived with the theory of branching processes with cascade substitution [3-8].

3.1 Stage 1

The process is started with N_A moles of monomer A with molecular mass M_A, N_D moles of monomer D with molecular mass M_D and N_T moles of monomer T with molecular mass M_T. Instead of moles it is more convenient to use mole or number fractions, denoted by n_A, n_D and n_T, respectively.

At any time, the conversion of A, α_A, which equals the ratio of the number of reacted to the number of initial c-groups in A, is related to the conversion of D and T, α_h, through the following balance equation:

$$2\alpha_A n_A = \alpha_h(2n_D + 3n_T). \qquad (3\text{-}1)$$

During the reaction a side product (e.g. water) may be produced and eliminated. A simple way to correct for this elimination is to correct the initial (molecular) masses with an amount which equals the molecular mass of the species eliminated times the conversion.

In the branching theory with cascade substitution, the distribution of units differing in number and type of bonds in which they are engaged is described by a vectorial pgf, $\underline{F}_0(\underline{z})$, where the subscript 0 refers to the root of the probability tree, see e.g. refs. [4-8]. The components of the dummy vector $\underline{z}$, which differ in their subscripts, denote the various types of bonds, their exponents indicate the numbers of those types of bonds. In the present treatment this idea is extended to a dummy vector $\underline{z}_f$ for the unreacted groups as well. For each monomer in the zeroth generation the pgf for the number of reacted groups (related to the dummy variables $\underline{z} = (z_A, z_D, z_T)^T$) and the groups once or twice unreacted, (related to the variables ${}_{(1)}\underline{z}_f = (z_{fA}, z_{fh}, z_{fAf})^T$) is formulated

$$
\begin{aligned}
{}_{(1)}F_{0A}(\underline{z}, {}_{(1)}\underline{z}_f) &= p_0 z_{fAf}^2 + p_1 z_{fA} z_h + p_2 z_h^2, \\
{}_{(1)}F_{0D}(\underline{z}, {}_{(1)}\underline{z}_f) &= [(1-\alpha_h) z_{fh} + \alpha_h z_A]^2, \qquad (3\text{-}2) \\
{}_{(1)}F_{0T}(\underline{z}, {}_{(1)}\underline{z}_f) &= [(1-\alpha_h) z_{fh} + \alpha_h z_A]^3,
\end{aligned}
$$

where p_0, p_1 and p_2 take into account the substitution effects (see also part 4 below) and $z_h = \phi_D z_D + \phi_T z_T$; $\phi_D = 2n_D/(2n_D + 3n_T)$; $\phi_T = 1 - \phi_D$ and α_A in (3-1) is equal to $(p_1 + 2p_2)/2$. As described in [3] the pgf's for each monomer in all next generations read

$$
\begin{aligned}
{}_{(1)}F_A(\underline{z}, {}_{(1)}z_f) &= [p_1 z_{fA} + 2p_2 z_h]/[p_1 + 2p_2], \\
{}_{(1)}F_D(\underline{z}, {}_{(1)}\underline{z}_f) &= [(1-\alpha_h) z_{fh} + \alpha_h z_A], \qquad (3\text{-}3) \\
{}_{(1)}F_T(\underline{z}, {}_{(1)}\underline{z}_f) &= [(1-\alpha_h) z_{fh} + \alpha_h z_A]^2,
\end{aligned}
$$

which is based on the Galton-Watson or universal consistency relationship [5]. Next the mass fraction generating function ${}_{(1)}W(z, {}_{(1)}\underline{z}_f)$ is derived as

$$
\begin{aligned}
{}_{(1)}W(z, \underline{z}_f) &= m_A\, {}_{(1)}W_A(z, {}_{(1)}\underline{z}_f) + m_D\, {}_{(1)}W_D(z, {}_{(1)}\underline{z}_f) + m_T\, {}_{(1)}W_T(z, {}_{(1)}\underline{z}_f) \\
&= m_A z^{M_A}\, {}_{(1)}F_{0A}({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f) + m_D z^{M_D}\, {}_{(1)}F_{0D}({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f) + \\
&\quad m_T z^{M_T}\, {}_{(1)}F_{0T}({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f), \qquad (3\text{-}4)
\end{aligned}
$$

where m_X is the mass fraction of X = A, D or T, ${}_{(1)}\underline{u} = (u_A(z), u_D(z), u_T(z))^T$ is the solution of the set of coupled nonlinear equations

$$\begin{aligned} u_A &= z^{M_A} \ {}_{(1)}F_A({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f), \\ u_D &= z^{M_D} \ {}_{(1)}F_D({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f), \\ u_T &= z^{M_T} \ {}_{(1)}F_T({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f). \end{aligned} \tag{3-5}$$

In the expressions (3-4) and (3-5) ${}_{(1)}F_{0X}$ and ${}_{(1)}F_X$ for component X are given by the systems (3-2) and (3-3) respectively. Notice that each component of the vector $\underline{z}$ in (3-4) is replaced by the scalar z, since we are not explicitly interested in the distribution according to the type of unit.

Before continuing we introduce the notations $\underline{m} = (m_A, m_D, m_T)^T$, $\underline{M} = (M_A, M_D, M_T)^T$, $\underline{n} = (n_A, n_D, n_T)^T$, ${}_{(1)}\underline{F}_0 = ({}_{(1)}F_{0A}, {}_{(1)}F_{0D}, {}_{(1)}F_{0T})^T$, ${}_{(1)}\underline{F} = ({}_{(1)}F_A, {}_{(1)}F_D, {}_{(1)}F_T)^T$ and $z^{\underline{M}} = (z^{M_A}, z^{M_D}, z^{M_T})^T$.
The expressions (3-4) and (3-5) read in shorthand notation [5]

$${}_{(1)}W(z,\underline{z}_f) = \underline{m} \cdot z^{\underline{M}} \cdot {}_{(1)}\underline{F}_0({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f), \tag{3-6}$$

$${}_{(1)}\underline{u} = z^{\underline{M}} \cdot {}_{(1)}F({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f), \tag{3-7}$$

where . means the inner product.

Subsequently ${}_{(1)}W(z, {}_{(1)}\underline{z}_f)$ is converted into the number fraction pgf ${}_{(1)}N(z, {}_{(1)}\underline{z}_f)$ by integrating ${}_{(1)}W(z, {}_{(1)}\underline{z}_f)/z$ as follows [3]

$${}_{(1)}N(z, {}_{(1)}\underline{z}_f) = {}_{(1)}\bar{M}_n \int_0^z \frac{{}_{(1)}W(z', {}_{(1)}\underline{z}_f)}{z'} \, dz', \tag{3-8}$$

where ${}_{(1)}\bar{M}_n$ is the number average molecular mass of the product of stage 1 (see 3-12 below). Substitution of (3-4) in (3-8), integration and another substitution of (3-5) in the result gives eventually

$$
\begin{aligned}
{}_{(1)}N(z,{}_{(1)}\underline{z}_f) = \frac{{}_{(1)}\bar{M}_n}{{}_{(1)}\bar{M}_{n,0}} \Bigg[& \underline{n}\cdot z^{\underline{M}}\cdot {}_{(1)}\underline{F}_0({}_{(1)}\underline{u},{}_{(1)}\underline{z}_f) - n_A(p_1+2p_2)u_A u_h \\
& + [n_A(p_1 + 2p_2) - (2n_D + 3n_T)\alpha_h] \int_0^z u_h \, du_A \Bigg],
\end{aligned}
\qquad (3\text{-}9)
$$

with ${}_{(1)}\bar{M}_{n,0} = \underline{n} \cdot \underline{M}$, the number average molecular mass before the start of stage 1, and $u_h = \phi_D u_D + \phi_T u_T$. Equation (3-9) can be simplified since the fraction of reacted c-groups equals the fraction of reacted h-groups, see (3-1). Thus the final expression for the number fraction pgf reads

$$
{}_{(1)}N(z,{}_{(1)}\underline{z}_f) = \frac{{}_{(1)}\bar{M}_n}{{}_{(1)}\bar{M}_{n,0}} \left[\underline{n}\cdot z^{\underline{M}}\cdot {}_{(1)}\underline{F}_0({}_{(1)}\underline{u},{}_{(1)}\underline{z}_f) - n_A(p_1+2p_2)u_A u_h \right], \qquad (3\text{-}10)
$$

where ${}_{(1)}\underline{u}$ is given implicitly by (3-7).

Next the condition for gelation is checked. The system is in the pre-gel region for

$$
D = \det\,(I - \partial\,{}_{(1)}\underline{F} \,/\, \partial \underline{z})\,\Big|_{\underline{z}=\underline{1}} > 0. \qquad (3\text{-}11)
$$

If the system is below the gel point the number average molecular mass of the product of stage 1, ${}_{(1)}\bar{M}_n$, is given by

$$
{}_{(1)}\bar{M}_n = \frac{{}_{(1)}\bar{M}_{n,0}}{1 - [n_A(p_1 + 2p_2) + (2n_D + 3n_T)\alpha_h]/2}\,, \qquad (3\text{-}12)
$$

and the mass average molecular mass, ${}_{(1)}\bar{M}_w$, is given by

$$ {}_{(1)}\bar{M}_w = \underline{m} \cdot \frac{\partial}{\partial z} \left(z^{\underline{M}} \cdot {}_{(1)}\underline{F}_0\left({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f\right)\right)\Big|_{z=1,\ {}_{(1)}\underline{z}_f=\underline{1}} = \tag{3-13} $$

$$ \underline{m} \cdot \underline{M} + \underline{m} . \left[\frac{\partial}{\partial\ {}_{(1)}\underline{u}}\ {}_{(1)}\underline{F}_0\left({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f\right) \left\{ I - \frac{\partial}{\partial_{(1)}\underline{u}}\ {}_{(1)}\underline{F}\ \left({}_{(1)}\underline{u}, {}_{(1)}\underline{z}_f\right) \right\}^{-1} \right] . \underline{M}. $$

The second term in the right hand side of (3-13) comes from the fact that ${}_{(1)}\underline{u}$ depends on z. The number average free functionality of type X, ${}_{(1)}\bar{\Phi}_{nX}$, can be derived from

$$ {}_{(1)}\bar{\Phi}_{nX} = \left[\frac{\partial}{\partial z_{fX}}\ {}_{(1)}N(z, {}_{(1)}\underline{z}_f) \right]_{z=1,\ {}_{(1)}\underline{z}_f=\underline{1}} . \tag{3-14} $$

We do not work out the right hand side of (3-14) here, because the resulting expression is very complex and does not give any further insight.

3.2 Stage 2

The product of stage 1 is subsequently mixed with newly added monomer C. As in stage 1 a side product may be produced and eliminated (a simple way to correct for this elimination is indicated in stage 1). The pgf's for the new monomer C in the second stage conform to the pgf's of monomer A in the first stage. The pgf of C in the zeroth generation is given by

$$ {}_{(2)}F_{0C}(\underline{z}, {}_{(2)}\underline{z}_f) = q_0 z_{fCf}^2 + q_1 z_{fC} z_P + q_2 z_P^2, \tag{3-15a} $$

and the pgf of C in all next generations is given by

$$ {}_{(2)}F_C(\underline{z}, {}_{(2)}\underline{z}_f) = [q_1 z_{fC} + 2q_2 z_P] \ /\ [q_1 + 2q_2], \tag{3-16a} $$

where $\underline{z} = (z_P, z_C)^T$ and ${}_{(2)}\underline{z}_f = (z_{fA}, z_{fC}, {}_{(2)}z_{fh}, z_{fAf}, z_{fCf})^T$. All information about the distribution of free functional groups in the prepolymers of stage 1 is collected in ${}_{(1)}N(z, {}_{(1)}\underline{z}_f)$. As a result this pgf is used to formulate the pgf for the prepolymers 1 in the zeroth generation in stage 2. Since this pgf is only dependent on number fractions and not on mass fractions z = 1. In addition the following cascade substitution is essential in formulating this pgf from ${}_{(1)}N(1, {}_{(1)}\underline{z}_f)$

$$z_{fh} = (1 - {}_{(2)}\alpha_h)\ {}_{(2)}z_{fh} + {}_{(2)}\alpha_h z_C. \qquad (3\text{-}17)$$

Thus the pgf for the prepolymer in the zeroth generation reads

$${}_{(2)}F_{0P}(\underline{z}, {}_{(2)}\underline{z}_f) = {}_{(1)}N(1, {}_{(1)}\underline{z}_f), \qquad (3\text{-}15b)$$

with (3-17) substituted in (3-15b). After applying the universal consisentency relationship, the pgf for the molecules in all next generations reads

$${}_{(2)}F_P(\underline{z}, {}_{(2)}\underline{z}_f) = \left.\frac{\partial\ {}_{(1)}N(1, {}_{(1)}\underline{z}_f)/\partial z_C}{\partial\ {}_{(1)}N(1, {}_{(1)}\underline{z}_f)/\partial z_C}\right|_{z_C=1,\ {}_{(1)}\underline{z}_f=\underline{1},\ {}_{(2)}\underline{z}_{fh}=\underline{1}} . \qquad (3\text{-}16b)$$

The mass fraction pgf is derived as

$${}_{(2)}W(z, {}_{(2)}\underline{z}_f) = m_{P1}\ {}_{(1)}W(z, {}_{(1)}\underline{z}_f({}_{(2)}\underline{u})) + m_C z^{M_C}\ {}_{(2)}F_{0C}({}_{(2)}\underline{u}, {}_{(2)}\underline{z}_f), \qquad (3\text{-}18)$$

with ${}_{(2)}\underline{u} = (u_{P1}, u_C)^T$ implicity given by the set of equations

$$u_{P1} = \frac{\partial_{(1)}N(z,{}_{(1)}\underline{z}_f({}_{(2)}\underline{u}))/\partial z_C}{\partial_{(1)}N(z,{}_{(1)}\underline{z}_f)/\partial z_C}\bigg|_{z=1,\ {}_{(1)}\underline{z}_f=\underline{1}} ,\tag{3-19}$$

$$u_C = z^{M_C}\ {}_{(2)}F_C({}_{(2)}\underline{u},{}_{(2)}\underline{z}_f),$$

and $m_{P1} + m_C = 1$. After a lot of elaborate calculations the first equation of system (3-19) can be simplified considerably to

$$u_{P1} = u_h\ (z_{fh} = (1-{}_{(2)}\alpha_h)\ {}_{(2)}z_{fh} + {}_{(2)}\alpha_h z_C),\tag{3-20}$$

where use has been made of the balance equation

$$n_{P1}\ \frac{{}_{(1)}\bar{M}_n}{{}_{(1)}\bar{M}_{n,0}}\ (2n_D + 3n_T)\ {}_{(2)}\alpha_h(1 - \alpha_h) = n_C(q_1 + 2q_2),\tag{3-21}$$

and $n_{P1} + n_C = 1$. Subsequently ${}_{(2)}W(z,{}_{(2)}\underline{z}_f)$ is converted into the number fraction pgf in a similar way as done in stage 1. As a result we find

$${}_{(2)}N(z,{}_{(2)}\underline{z}_f) = \frac{{}_{(2)}\bar{M}_n}{{}_{(2)}\bar{M}_{n,0}}\left[n_{P1}\ {}_{(1)}N(z,{}_{(1)}\underline{z}_f({}_{(2)}\underline{u})) + \right.\tag{3-22}$$

$$\left. + n_C z^{M_C}\ {}_{(2)}F_{0C}({}_{(2)}\underline{u},{}_{(2)}\underline{z}_f) - n_C(q_1 + 2q_2)u_C u_{P1}\right],$$

with ${}_{(2)}\bar{M}_n$ the number average molecular mass of the product of stage 2 (see 3-23 below) and ${}_{(2)}\bar{M}_{n,0} = n_{P1}\ {}_{(1)}\bar{M}_n + n_C\ M_C$. In the derivation of (3-22) the balance equation (3-21) is used again.

Next the condition for gelation is checked. The gel point is reached if the Perron-Frobenius eigenvalue of the Jacobian $\partial\ {}_{(2)}\underline{F}({}_{(2)}\underline{u},{}_{(2)}\underline{z}_f)/\partial\ {}_{(2)}\underline{u}\Big|_{{}_{(2)}\underline{u}=\underline{1};\ {}_{(2)}\underline{z}_f=\underline{1}}$ reaches the value 1, see refs. [12,13]. If the system is below the gel point the number average molecular mass of the prepolymer after stage 2 can be calculated with

$$ {}_{(2)}\bar{M}_n = {}_{(2)}\bar{M}_{n,0}/[1 - n_C(q_1 + 2q_2)]. \tag{3-23}$$

The mass average molecular mass of the products of stage 2, ${}_{(2)}\bar{M}_w$ and the number average free functionalities of type X, ${}_{(2)}\bar{\phi}_{nX}$, can be calculated with expressions similar to (3-13) and (3-14).

3.3 Stage 3

The product of stage 2 is mixed with a mixture of two hardeners, namely difunctional E and trifunctional F. In a similar way as in stage 1 the pgf's for these three components in the zeroth generation are given by the expressions

$$\begin{aligned} {}_{(3)}F_{0P2}(\underline{z}) &= {}_{(2)}N(1,{}_{(2)}\underline{z}_f), \\ {}_{(3)}F_{0E}(\underline{z}) &= (1 - \alpha_e + \alpha_e z_{P2})^2, \\ {}_{(3)}F_{0F}(\underline{z}) &= (1 - \alpha_e + \alpha_e z_{P2})^3, \end{aligned} \tag{3-24}$$

with $\underline{z} = (z_{P2}, z_E, z_F)^T$ and α_e the conversion of E and F in the third stage. Essential are the cascade substitutions applied in ${}_{(2)}\underline{z}_f$:

$$\begin{aligned} z_{fAf}^{2} &= k_0 + k_1 z_e + k_2 z_e^{2}, \\ z_{fA} &= k_3 + k_4 z_e, \\ z_{fCf}^{2} &= \ell_0 + \ell_1 z_e + \ell_2 z_e^{2}, \\ z_{fC} &= \ell_3 + \ell_4 z_e, \\ {}_{(2)}z_{fh} &= 1, \end{aligned} \tag{3-25}$$

where $z_e = \lambda_E z_E + \lambda_F z_F$; $\lambda_E = 2n_E/(2n_E + 3n_F)$; $\lambda_F = 1 - \lambda_E$. The pgf for molecules in all next generations is given by

$$\begin{aligned} {}_{(3)}F_{P2}(\underline{z}) &= \left.\frac{\partial\, {}_{(2)}N(1, {}_{(2)}\underline{z}_f)/\partial z_e}{\partial\, {}_{(2)}N(1, {}_{(2)}\underline{z}_f)/\partial z_e}\right|_{{}_{(2)}\underline{z}_f = \underline{1}}, \\ {}_{(3)}F_E(\underline{z}) &= (1 - \alpha_e + \alpha_e z_{P2}), \\ {}_{(3)}F_F(\underline{z}) &= (1 - \alpha_e + \alpha_e z_{P2})^{2}. \end{aligned} \tag{3-26}$$

The mass fraction pgf is derived as

$$\begin{aligned} {}_{(3)}W(z) = {} & m_{P2}\ {}_{(2)}W(z, {}_{(2)}\underline{z}_f({}_{(3)}\underline{u})) + m_E z^{M_E}\ {}_{(3)}F_{0E}({}_{(3)}\underline{u}) + \\ & m_F z^{M_F}\ {}_{(3)}F_{0F}({}_{(3)}\underline{u}), \end{aligned} \tag{3-27}$$

with ${}_{(3)}\underline{u} = (u_{P2}, u_E, u_F)^T$ implicitly given by the set of equations

$$\begin{aligned} u_{p2} &= \left.\frac{\partial\, {}_{(2)}N(z, {}_{(2)}\underline{z}_f({}_{(3)}\underline{u}))/\partial z_e}{\partial\, {}_{(2)}N(z, {}_{(2)}\underline{z}_f)/\partial z_e}\right|_{z=1;\ {}_{(2)}\underline{z}_f = \underline{1}}, \\ u_E &= z^{M_E}\ {}_{(3)}F_E({}_{(3)}\underline{u}), \\ u_F &= z^{M_F}\ {}_{(3)}F_F({}_{(3)}\underline{u}), \end{aligned} \tag{3-28}$$

and $m_{P2} + m_E + m_F = 1$. (The right hand side of the first equation of (3-28) can be worked out and in some special cases be simplified to a linear combination of u_A and u_C, given by (3-5) and (3-19) in which the cascade substitutions (3-25) are performed).

The number fraction pgf is derived as

$$ {}_{(3)}N(z) = \frac{{}_{(3)}\bar{M}_n}{{}_{(3)}\bar{M}_{n,0}} \left[n_{P2}\ {}_{(2)}N(z, {}_{(2)}\underline{z}_f({}_{(3)}\underline{u})) + n_E z^{M_E}\ {}_{(3)}F_{0E}({}_{(3)}\underline{u}) \right. $$
$$ \left. + n_F z^{M_F}\ {}_{(3)}F_{0F}({}_{(3)}\underline{u}) - (2n_E + 3n_F)\alpha_e u_e u_{P2} \right], \qquad (3\text{-}29) $$

with ${}_{(3)}\bar{M}_{n,0} = n_{P2}\ {}_{(2)}\bar{M}_n + n_E M_E + n_F M_F$, $n_{P1} + n_E + n_F = 1$, $u_e = \lambda_E u_E + \lambda_F u_F$ and ${}_{(3)}\bar{M}_n$ the number average molecular mass after stage 3 (see (3-31) below). In the derivation of (3-29) we have used that the fraction of reacted c-groups equals to fraction of reacted e-groups, or

$$ n_{P2} \frac{\partial}{\partial u_e}\ {}_{(2)}N(1, {}_{(2)}\underline{z}_f({}_{(3)}\underline{u})) \Big|_{{}_{(3)}\underline{u}=\underline{1}} = (2n_E + 3n_F)\alpha_e . \qquad (3\text{-}30) $$

Next we check the gel-condition. If the Perron Frobenius eigenvalue of the Jacobian $\partial\ {}_{(3)}\underline{F}({}_{(3)}\underline{u})/\partial {}_{(3)}\underline{u}|_{{}_{(3)}\underline{u}=\underline{1}} < 1$, the gel point is not yet reached. If the system is in the pre-gel state the number average molecular mass of the product 3 can be calculated with

$$ {}_{(3)}\bar{M}_n = {}_{(3)}\bar{M}_{n,0} / [1 - (2n_E + 3n_F)\alpha_e]. \qquad (3\text{-}31) $$

The mass average molecular mass of the product 3 can be calculated in a way similar to (3-13).

If the Perron-Frobenius eigenvalue is greater than one, the gel point is passed. In this case the extinction probabilities $\underline{v} = (v_{P2}, v_E, v_F)^T$ can be obtained by solving the set of coupled nonlinear equations

$$\underline{v} = {}_{(3)}\underline{F}(\underline{v}). \tag{3-32}$$

The solution $\underline{v} \neq \underline{1}$ of this set of equations exists and is unique for $0 \leqslant v_i \leqslant 1$ (i = P2, E, F). The sol fraction w_s is given by

$$w_s = m_{P2}\ {}_{(2)}W(1, {}_{(2)}\underline{z}_f(\underline{v})) + m_E\ {}_{(3)}F_{0E}(\underline{v}) + m_F\ {}_{(3)}F_{0F}(\underline{v}). \tag{3-33}$$

As an approximation for the fraction of elastically active network chains per monomer, N_{ea}, the following formula can be given [14]

$$N_{ea} = \frac{3[N_T\, \alpha_{P2}^3(1 - v_e)^3 + N_F\, \alpha_e^3(1 - v_{P2})^3]}{2(N_A + N_D + N_T + N_C + N_E + N_F)}\ , \tag{3-34}$$

with $v_e = \lambda_E v_E + \lambda_F v_F$. The conversion of the prepolymer of the second stage, α_{P2}, is obtained from

$$\alpha_{P2} = \frac{n_C[q_0(\ell_1+2\ell_2) + q_1\ell_4] + n_{P2}\ {}_{(1)}\bar{M}_n\ n_A[p_0(k_1+2k_2) + p_1k_1]/{}_{(1)}\bar{M}_{n,0}}{n_C(2q_0+q_1) + n_{P2}\ {}_{(1)}\bar{M}_n\ n_A(2p_0+p_1)/{}_{(1)}\bar{M}_{n,0}}. \tag{3-35}$$

4. KINETIC SCHEME

In this paragraph the distributions of units in the different reaction states are determined by kinetic differential equations [7-11], by means of which it is possible to take into account substitution and insolubility effects in an approximate way.

Monomer A is assumed to be completely insoluble in D and T and only reactive at the surface. Therefore the fraction of unreacted monomers A, p_A, is divided into a reactive fraction present at the surface, p_A^r , and an unreactive fraction p_A^u so that $p_A = p_A^r + p_A^u$. The ratio p_A^r/p_A^u is related to the specific surface area of the powder of monomer A.

Let p_A be the fraction of A with two unreacted functionalities, p_{AD} the fraction of A with one unreacted functionality and one endgroup reacted with D etcetera. In total there are 6 different states for D and E, 10 different states for T and F and 15 different states for A and C. Some states are shown schematically in figure 1.

Let k_{XY} be the rate constant for forming an X-Y-bond and K_{XY} be a substitution effect factor, multiplying the rate constant k_{XZ} for some Z if X bears already an X-Y-bond. If the X-group already bears two X-Y-bonds, the rate constant will accordingly be multiplied by K_{XY}^2. It follows from symmetry that $k_{XY} = k_{YX}$. The assumptions are made that $k_{XD} = k_{XT}$, $k_{XE} = k_{XF}$, $K_{XD} = K_{XT}$, $K_{XE} = K_{XF}$ for X = A, C.

Below some typical examples for the change of the 63 states as a function of time are derived.

_In_stage_1_:

$$\frac{d}{dt} p_A^r = -2\, p_A^r\, k_{surf} \left[k_{AD}\, N_D\, S_D + k_{AT}\, N_T\, S_T + k_{AE}\, N_E\, S_E + k_{AF}\, N_F\, S_F\right] +$$

$$+\, p_A^r \left[1 - \frac{2N_D+3N_T}{2N_A}\left\{\frac{N_D}{2N_D+3N_T}(p_{DA} + 2p_{DAA}) + \frac{N_T}{2N_D+3N_T}(p_{TA} + 2p_{TAA} + 3p_{TAAA})\right\}\right]^{2/3},$$

$$\frac{d}{dt}\, p_{TA} = - 2\, K_{TA}\, p_{TA} \left[k_{TA}\, N_A\, s_A + k_{TC}\, N_C\, s_C\right] + 3\, p_T\, k_{TA}\, N_A\, s_A,$$

In_stage_2:

$$\frac{d}{dt}\, p_{TAC} = - K_{TC}\, K_{TA}\, p_{TAC} \left[k_{TA}\, N_A\, s_A + k_{TC}\, N_C\, s_C\right] +$$

$$+ 2 \left[k_{TC}\, K_{TA}\, p_{TA}\, N_C\, s_C + k_{TA}\, K_{TC}\, p_{TC}\, N_A\, s_A\right],$$

In_stage_3:

$$\frac{d}{dt}\, p_{EAC} = k_{EC}\, K_{EA}\, p_{EA}\, N_C\, s_C + K_{EC}\, k_{EA}\, p_{EC}\, N_A\, s_A,$$

with N_X the number of molecules of monomer X, k_{surf} is a rate constant and

$$s_A = 2p_A^r + p_{AD}\, K_{AD} + p_{AT}\, K_{AT} + p_{AE}\, K_{AE} + p_{AF}\, K_{AF},$$

$$s_C = 2p_C + p_{CD}\, K_{CD} + p_{CT}\, K_{CT} + p_{CE}\, K_{CE} + p_{CF}\, K_{CF},$$

$$s_D = 2p_D + p_{DA}\, K_{DA} + p_{DC}\, K_{DC},$$

$$s_T = 3p_T + 2p_{TA}\, K_{TA} + 2p_{TC}\, K_{TC} + p_{TAA}\, K_{TA}^2 + p_{TAC}\, K_{TA}\, K_{TC} + p_{TCC}\, K_{TC}^2\, ,$$

$$s_E = 2p_E + p_{EA}\, K_{EA} + p_{EC}\, K_{EC},$$

$$s_F = 3p_F + 2p_{FA}\, K_{FA} + 2p_{FC}\, K_{FC} + p_{FAA}\, K_{FA}^2 + p_{FAC}\, K_{FA}\, K_{FC} + p_{FCC}\, K_{FC}^2.$$

To solve the set of 63 coupled differential equations in time or in conversion of for instance A and C, boundary conditions are needed for time or conversion zero.

At the beginning of stage 1, $k_{XC} = 0$ for X = D,T, $k_{YE} = k_{YF} = 0$ for Y = A,C, $r = p^r_A/p^u_A$ is known and $p_A = p_D = p_T = 1$. The conversion of A in this stage, ${}_{(1)}\alpha_C = \alpha_A$, is ${}_{(1)}\alpha_C = (p_{AD} + p_{AT} + 2p_{ADD} + 2p_{ADT} + 2p_{ATT})/2$.

At the beginning of stage 2, $k_{XA} = 0$ for X = D,T, $k_{YE} = k_{YF} = 0$ for Y = A,C and $p_C = 1$. The total conversion of A and C in this stage, ${}_{(2)}\alpha_C$, is ${}_{(2)}\alpha_C = [N_A\ (p_{AD} + p_{AT} + 2p_{ADD} + 2p_{ADT} + 2p_{ATT}) + N_C\ (p_{CD} + p_{CT} + 2p_{CDD} + 2p_{CDT} + 2p_{CTT})]\ /\ (2N_A + 2N_C)$.

At the beginning of stage 3, $k_{XD} = k_{XT} = 0$ for X = A,C and $p_E = p_F = 1$. The total conversion of A and C in this stage, ${}_{(3)}\alpha_C$, is ${}_{(3)}\alpha_C = (N_A\ \alpha_A + N_C\ \alpha_C)/(N_A + N_C)$, with $\alpha_X = p_{XD} + p_{XT} + p_{XE} + p_{XF} + 2p_{XDD} + 2p_{XDT} + 2p_{XDE} + 2p_{XDF} + 2p_{XTT} + 2p_{XTE} + 2p_{XTF} + 2p_{XEE} + 2p_{XEF} + 2p_{XFF}$ for X = A,C.

After the solution of the set of differential equations in conversion of A and C, the input parameters for the statistical method, described in the former paragraph are calculated with:

In_stage_1:

$p_0 = p^u_A + p^r_A + [p_{AE} + p_{AF} + p_{AEE} + p_{AEF} + p_{AFF}]$,

$p_1 = p_{AD} + p_{AT} + [p_{ADE} + p_{ADF} + p_{ATE} + p_{ATF}]$,

$p_2 = p_{ADD} + p_{ADT} + p_{ATT} = 1 - p_0 - p_1$,

In_stage_2:

$q_0 = p_C + [p_{CE} + p_{CF} + p_{CEE} + p_{CEF} + p_{CFF}],$

$q_1 = p_{CD} + p_{CT} + [p_{CDE} + p_{CDF} + p_{CTE} + p_{CTF}],$

$q_2 = p_{CDD} + p_{CDT} + p_{CTT} = 1 - q_0 - q_1,$

In_stage_3:

$k_0 = (p_A^u + p_A^r)/p_0,$

$k_1 = (p_{AE} + p_{AF})/p_0,$

$k_2 = (p_{AEE} + p_{AEF} + p_{AFF})/p_0 = 1 - k_0 - k_1,$

$k_3 = (p_{AD} + p_{AT})/p_1,$

$k_4 = 1 - k_3,$

$\ell_0 = p_C/q_0,$

$\ell_1 = (p_{CE} + p_{CF})/q_0,$

$\ell_2 = (p_{CEE} + p_{CEF} + p_{CFF})/q_0 = 1 - \ell_0 - \ell_1,$

$\ell_3 = (p_{CD} + p_{CT})/q_1,$

$\ell_4 = 1 - \ell_3.$

The terms between square brackets are zero in the first and second stage.

5. PROGRAM, RESULTS AND DISCUSSION

The computer programs for the statistical part, POLYM, and for the kinetic part, KINREL, are written in Fortran for the mainframe IBM 4381.

5.1 POLYM

General input values are the numbers of moles N_A, N_D, N_T, N_C, N_E, N_F and the molecular masses M_A, M_D, M_T, M_C, M_E and M_F. Input values for the first stage are p_1 and p_2, for the second stage q_1 and q_2 and for the third stage k_1, k_2, k_4, ℓ_1, ℓ_2 and ℓ_4 (all obtained with KINREL).
If the system is in the pre-gel region, output values in the first and second stage are the conversion of c and h and the relevant characteristics such as number and mass average molecular masses and number average functionalities. In the third stage POLYM calculates amongst others the conversion of c and e, the extinction probabilities, the sol fraction and the number fraction of elastically active network chains.

5.2 KINREL

Input values for this program are the numbers N_A, N_D, N_T, N_C, N_E and N_F, the rate constants k_{XY} and k_{surf}, the substitution effect factors K_{XY} for some monomers X and Y, the ratio p^r_A/p^u_A and the end conversions, $_{(i)}\alpha_C$, in every stage (i = 1, 2, 3).

Using Gear-stiff's method KINREL calculates the fractions of the 63 reaction states, described in paragraph 4, as a function of the conversion of c. Moreover the input values for the statistical part are determined.

5.3 RESULTS AND DISCUSSION

Some preliminary results are presented below, additional results are given in ref. [15].

The following input data were taken: $N_A = 8.98$, $N_D = 9.98$, $N_T = 0.00$, $N_C = 2.00$, $N_E = 0.00$, $N_F = 0.74$ mole, and $M_A = 0.166$, $M_D = 0.068$, $M_C = 0.166$, $M_F = 0.300$ kg/mol; $k_{XY} = 1\ (\text{mol.s})^{-1}$ for all possible combinations of X and Y, k_{surf} = 1; p_A^r/p_A^u = 1000. Substitution effects were only supposed to occur in C, so that K_{XY} = 1 for X ≠ C, and all K_{CY}, with Y = D, T, E, F, were taken equal and took the values 0.25, 0.50, 1.00, 2.00 and 4.00, respectively. Full conversion was assumed to occur in the first and second stage.

Since substitution effects play only a role in the last two stages, some typical results of KINREL for these stages are presented in figures 2 and 3. For a positive substitution effect ($K_{CY} > 1$) the ratio of q_2/q_1 is higher at each conversion than for a random reaction ($K_{CY} = 1$). The same holds for ℓ_2/ℓ_1 and ℓ_4/ℓ_3. As a result the gel point is shifted to lower conversion, see figures 4 and 5. The effects are more pronounced for positive than for negative substitution effects.

ACKNOWLEDGEMENT

The authors thank Dr. R. van der Linde, DSM Resins, for indicating the problem and for stimulating discussions, prof. M. Gordon for the discussions about KINREL, Mr. A.J.M. Sipers for this contribution to POLYM and Mr. E. Peters for his programming assistence.

REFERENCES

1. Dušek, K., 1982, Rubber Chem. Technol. 55, 1.

2. van der Linde, R., Scholtens, B.J.R. and Belder, E.G., 1985, Proceedings 11, XIth International Conference in Organic Coatings Science and Technology, Athens, p. 167; 1987, to be published in 'Organic Coatings Science and Technology', vol. 7, A.V. Patsis e.d., Marcel Dekker, New York.

3. Dušek, K., Scholtens, B.J.R and Tiemersma-Thoone, G.P.J.M., 1987, Polym. Bull., 17, in press.

4. Gordon, M., 1962, Proc. Roy. Soc. London, A268, 240.

5. Gordon, M. and Malcolm, G.N., 1966, Proc. Roy. Soc. London, A295, 29.

6. Dušek, K., 1986, Adv. Polym. Sci., 78, 1.

7. Gordon, M., and Scantlebury, G.R., 1964, Trans. Faraday Soc. 60, 604.

8. Gordon, M., and Scantlebury, G.R., 1967, J. Chem. Soc., B, 1.

9. Mikes, J. and Dušek, K., 1982, Macromolec., 15, 93.

10. Dušek, K. and Ilavsky, M., 1983, J. Polym. Sci. Polym. Phys. Ed., 21, 1323.

11. Riccardi, C.C. and Williams R.J.J., 1986, Polymer, 27, 913.

12. Seneta, E., 1973, 'Non-Negative Matrices and Markov Chains', Springer Verlag, New York.

13. Harris, T.E., 1963, 'The Theory of Branching Processes', Springer Verlag, Berlin.

14. Dobson, G.R. and Gordon, M., 1965, J. Chem. Phys., 43, 705.

15. Scholtens, B.J.R., Tiemersma-Thoone, G.P.J.M. and Dušek, K., 1987, Rolduc Polymer Meeting-2.

Table 1. Scheme of the three-stage process of network formation

stage	components	products
1	monomers A + (D + T) →	prepolymers 1, mainly h-functional, but with possibly some unreacted functional c-groups.
2	prepolymers 1 + monomer C →	prepolymers 2, mainly c-functional, but with possibly some unreacted functional h-groups.
3	prepolymers 2 + monomers (E + F) →	product 3, which may form a gel before the completion of the reaction.

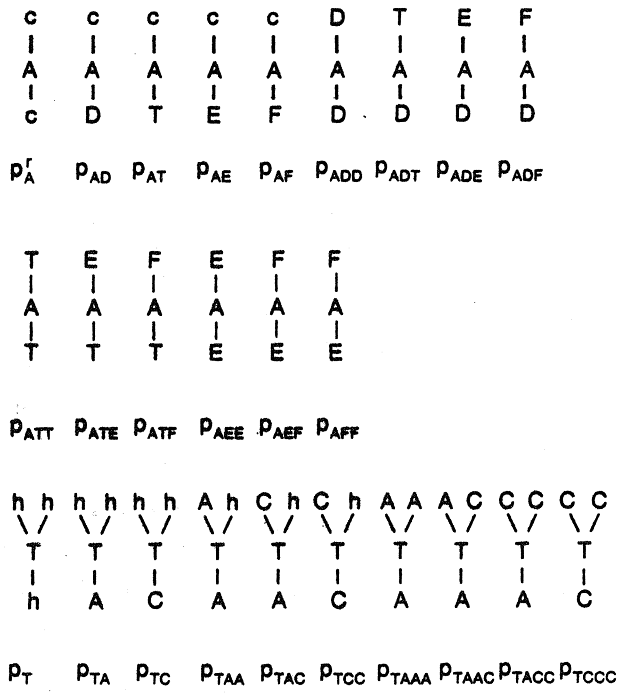


Fig. 1. Some examples of the 63 fractions of units in the different reaction states.

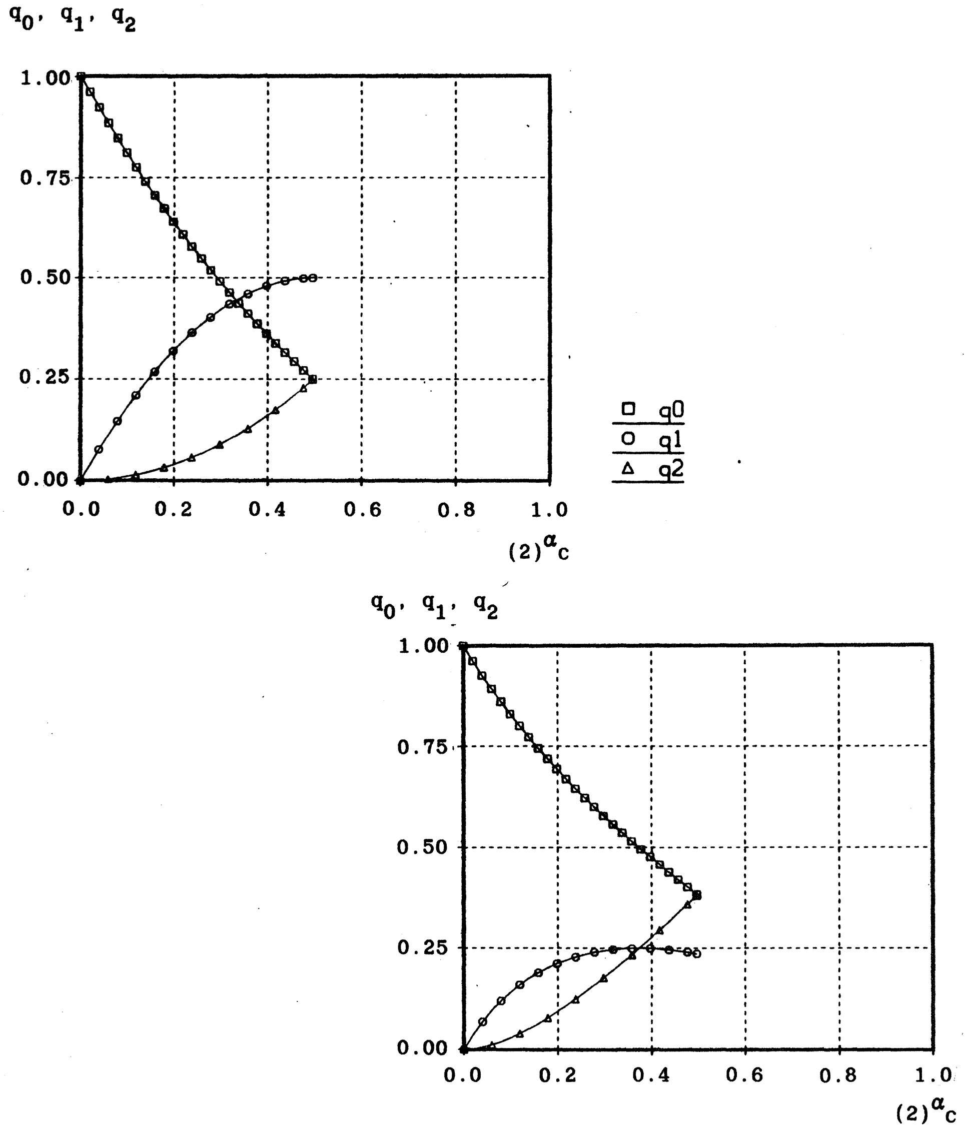


Fig. 2. Variation of q_0, q_1 and q_2 with $_{(2)}\alpha_c$ for K_{CY} = 1 (left) and K_{CY} = 4 (right), Y = D,T.

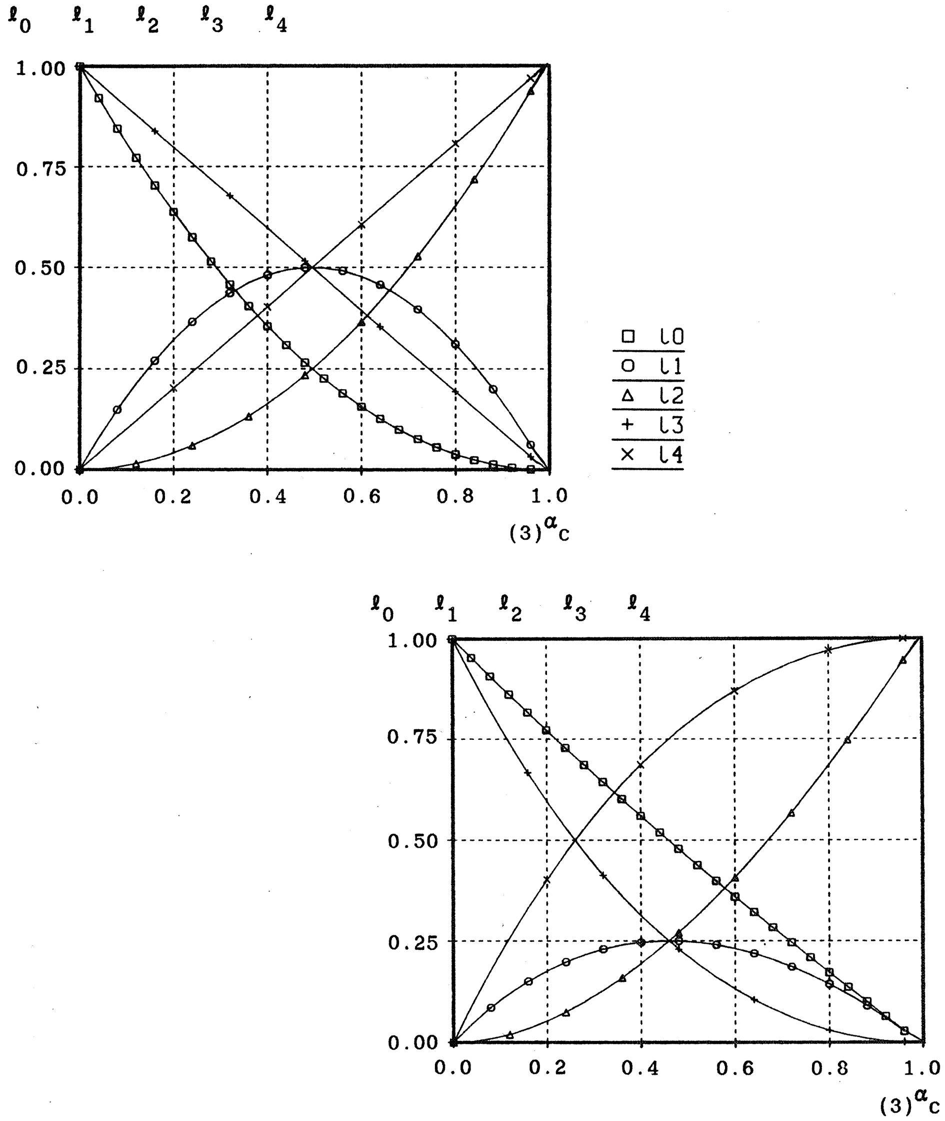


Fig. 3. Variation of ℓ_0, ℓ_1, ℓ_2, ℓ_3 and ℓ_4 with $_{(3)}\alpha_c$ for K_{CY} = 1 (left) and K_{CY} = 4 (right), Y = E, F.

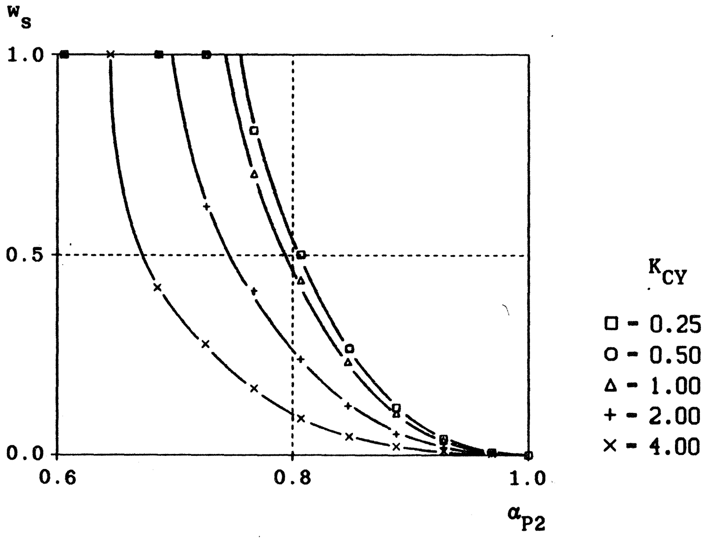


Fig. 4. Variation of the sol fraction, w_s, with $\alpha_{P2} = {}_{(3)}\alpha_c$ for K_{CY} = 0.25; 0.50; 1.00; 2.00 and 4.00, Y = E, F.

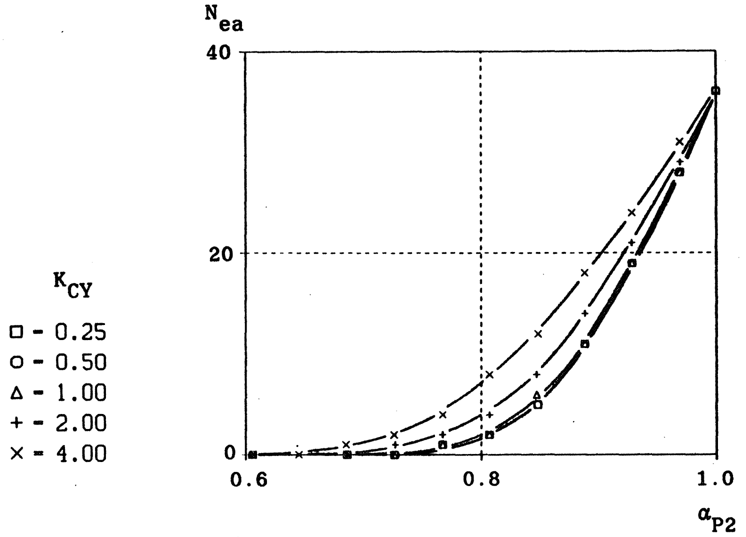


Fig. 5. Variation of N_{ea} with ${}_{(3)}\alpha_c$ for K_{CY} = 0.25; 0.50; 1.00; 2.00 and 4.00, Y = E, F.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Numerical Simulation of Hydraulic Crack Propagation

F.P.H. van Beckum
Department of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

A numerical method for hydraulic crack propagation is presented. It is designed for a one-dimensional model composed of the continuity equation, Darcy's momentum equation and England & Green's elasticity relation.

Essentials of the method are (1) scaling of the space coordinate with the fracture length, so that in the dimensionless coordinate the fracture tip has a fixed position; (2) elimination of flow rate and fracture width, giving a parabolic equation for the pressure; (3) discretization of the spatial operator conserving positive definiteness; (4) implicit time integration.

1. Introduction

In the oil industry the production of oil and gas wells can be increased by well stimulation through hydraulic fracturing of the hydrocarbon bearing formation. As the minimum total rock stress usually is horizontal, hydraulic fractures mostly propagate in the vertical plane. Assuming that such fractures grow lengthwise only, having constant height, the process can be described one-dimensionally.

In many hydraulic crack propagation studies additional assumptions are introduced to derive approximate relationships. (See for instance Geertsma & Haafkens 1979, Nordgren 1972 and Daneshy 1973).

However, to estimate the quality of these approximations and to be able to simulate more complex physical and chemical effects, the need for a robust numerical one-dimensional treatment was felt.

2. Mathematical problem

In a fracture, propagating from a wellbore in a vertical plane, a horizontal cross-section is considered. The coordinate in the propagating direction is named x; the x-domain of interest is $0 \leq x \leq L(t)$, where the fracture length L is changing with time. For all x and t a constant height is assumed.

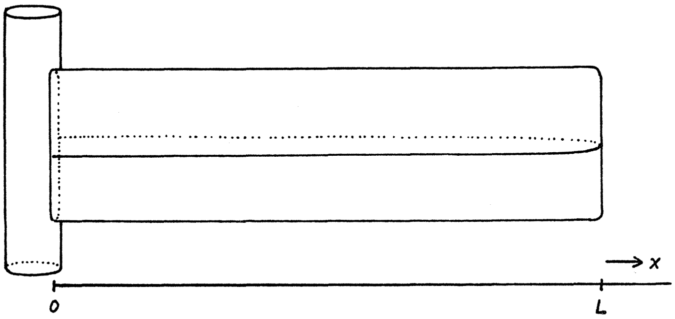


Figure 1.

The physical quantities in the model are the net pressure Δp (= the pressure in the fracture in excess of the total minimum horizontal principal in situ stress), the flow rate in x-direction q [m^2/s] and the fracture width w; of these functions only the variations in x and t are taken into account, consistent with the assumption of constant height.

Between Δp, q and w the following relations are adopted:

Continuity equation:
$$\frac{\partial w}{\partial t} + \frac{\partial q}{\partial x} + v_\ell = 0; \qquad (1)$$

Monentum equation:
$$\frac{\partial \Delta p}{\partial x} + 12\,\mu \frac{q}{w^3} = 0; \qquad (2)$$

Elasticity:
$$w(x) = \frac{1}{\sigma_0} \int_x^L \frac{\xi \, d\xi}{\sqrt{\xi^2 - x^2}} \int_0^\xi \frac{\Delta p(\eta)}{\sqrt{\xi^2 - \eta^2}} \, d\eta. \qquad (3)$$

Here v_ℓ is the leak off term, with dimension [m/s], representing the loss of fluid volume per second per area through the fracture walls, its explicit form being modelled later; μ is the effective viscosity of the fluid. The elasticity relation is adopted from England & Green 1963; of course Δp and w are time-dependent though it is not expressed in the arguments; $\sigma_0 = \pi\, G/4(1-\nu)$ is the isotropic elastic stiffness of the rock formation, where G is the shear modulus and ν is Poisson's ratio.

The boundary conditions are in terms of the flow rate q: at x = L(t) we have

$$q(L(t),t) = 0 \text{ for all } t, \tag{4}$$

and at x = 0 we assume a prescribed flow rate:

$$q(0,t) = q_{in}(t). \tag{5}$$

The initial conditions for L, Δp and w may be an arbitrary situation; a steady inflow solution taken from a simplified model may be convenient.

The condition for the crack propagation is chosen in the form

$$\frac{2}{\pi} \int_0^L \frac{\Delta p(x)}{\sqrt{L^2 - x^2}} \, dx = S_t \tag{6}$$

where S_t is called the tensile strength; see for instance Geertsma & de Klerk 1969.

3. Numerical treatment

Let y denote the dimensionless x-coordinate:

$$y = \frac{x}{L(t)} \, . \tag{7}$$

Let h = h(y) be a function on 0 < y ≤ 1, with h(y) ≤ y, and define A(y,t) to be the area of the fracture cross section between (y-h)L and yL:

$$A(y,t) \equiv \int_{yL-hL}^{yL} w(x,t) \, dx. \tag{8}$$

Differentiating with respect to t we find:

$$\frac{\partial A}{\partial t}(y,t) = L'(t) \left[y \, w(yL,t)\right]_{y-h}^{y} + \int_{yL-hL}^{yL} \frac{\partial w}{\partial t}(x,t) \, dx.$$

Upon substitution of the continuity equation (1) we get:

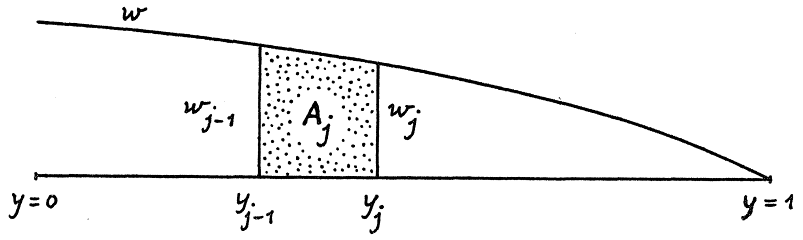
$$\frac{\partial A}{\partial t}(y,t) = L'(t) \left[y \, w(yL,t)\right]_{y-h}^{y} - q(yL,t) + q(yL-hL,t) - \int_{yL-hL}^{yL} v_\ell \, dx. \tag{9}$$

Now define a transformation $y = y(s)$, y monotonicly increasing from $y(0) = 0$ to $y(1) = 1$, and choose a discretization together with the function h such that $s_j = j\ \Delta s$, $(j=0,1,\dots,n)$, $\Delta s = 1/n$, $y_j = y(s_j)$ and $y_j - h(y_j) = y_{j-1}$. Further, denote $A(y_j,t)$ by A_j $(=A_j(t))$, $w(y_jL,t)$ by w_j, etcetera. Then for (9) we can write:

$$\frac{dA_j}{dt} = L'(t)\ [y_j w_j - y_{j-1} w_{j-1}] - q_j + q_{j-1} - \int_{y_{j-1}L}^{y_jL} v_\ell\ dx \qquad (10)$$

for $j=1,2,\dots,n$.

Figure 2.



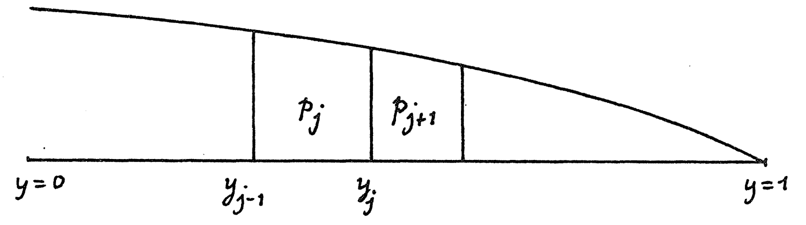
In the equation for j=n we can substitute boundary condition (4): $q_n = 0$, and the inflow condition (5) is incorporated in the equation for j=1. All the other q_j's are eliminated with the momentum equation (2):

$$q_j = -\frac{w_j^3}{12\mu}\left(\frac{\partial \Delta p}{\partial x}\right)_{y_j} \qquad (11)$$

in which $\frac{\partial \Delta p}{\partial x} = \frac{\partial \Delta p}{\partial s}\frac{ds}{dy}\frac{dy}{dx} = \frac{\partial \Delta p}{\partial s} / L\ \frac{dy}{ds}$ is discretised as

$$\left(\frac{\partial \Delta p}{\partial x}\right)_{y_j} = \frac{\Delta p_{j+1} - \Delta p_j}{L\Delta s\ \left(\frac{dy}{ds}\right)_{y_j}} \qquad (12)$$

Figure 3.



where Δp_j is associated with the j-th interval rather than with a point.

The time derivative is simply discretised as

$$\frac{dA_j}{dt} \approx \frac{A_j - \tilde{A}_j}{\Delta t} \tag{13}$$

where the superscript "~" denotes the last time level where all the quantities are known, and the symbols without superscripts refer to the new time level to be calculated. The pressure Δp in (12) will be taken at the new time level, while w in (11) and (10) are taken at the old level. Leaving the loss term undecided we have from (10):

$$A_j + \frac{\Delta t}{12\mu L\Delta s}\left[\tilde{w}^3_{j-1}\,\frac{\Delta p_j - \Delta p_{j-1}}{(\frac{dy}{ds})_{j-1}} + \tilde{w}^3_j\,\frac{\Delta p_j - \Delta p_{j+1}}{(\frac{dy}{ds})_j}\right] =$$

$$= \tilde{A}_j + \Delta L\,[y_j\tilde{w}_j - y_{j-1}\tilde{w}_{j-1}] - \Delta t \int_{y_{j-1}L}^{y_jL} v_\ell\, dx. \tag{14}$$

In this relation we want to express A_j in terms of Δp. To that end we integrate the elasticity relation (3), assuming Δp to be piecewise constant, according to the areas A_j. In the appendix it is shown that A_j can be expressed as

$$A_j = \sum_{m=1}^{n} c_1 K_{jm}\, \Delta p_m \tag{15}$$

with K_{jm} elements of a symmetric matrix. Altogether we now have:

$$(c_1K + c_2T)\Delta p = \tilde{A} + r_1 - r_2 \tag{16}$$

with c_1K: matrix defined by (15); see also (a3) and (a2) in the appendix;

c_2: constant in (14): $c_2 = \Delta t/(12\mu L\Delta s)$;

T: tridiagonal matrix in (14): elements of T are of the form $\tilde{w}^3_j/(\frac{dy}{ds})_j$;

$\Delta p = (\Delta p_1, \Delta p_2, \ldots, \Delta p_n)^T$;

$\tilde{A} = (\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_n)^T$;

r_1: convective term: $r_{1j} = \Delta L\,[y_j\tilde{w}_j - y_{j-1}\tilde{w}_{j-1}]$;

r_2: loss term: $r_{2j} = \Delta t \int_{y_{j-1}L}^{y_jL} v_\ell\, dx$.

Finally the crack propagation condition (6) is discretised as

$$\frac{2}{\pi}\sum_{j=1}^{n} \Delta p_j \ (\arcsin y_j - \arcsin y_{j-1}) = S_t. \qquad (17)$$

The time stepping procedure is:

(a) choose a time step Δt,

(b) guess the corresponding fracture length increase ΔL,

(c) construct the equations (16) and solve for Δp,

(d) check condition (17) and return to (b), i.e. iterate on ΔL,

(e) upon convergence, calculate the new width w from (15) and (8),

(f) if required, find q from (11).

Remarks

1. Instead of using L and $\tilde{w}$ in (14) at the old time level one may wish to have them at the new one; in this case these quantities should be updated within the iteration (b) - (d).

2. If the loss term v_ℓ in (1) is modelled linearly in Δp: $v_\ell = \alpha\Delta p + \beta$, with α and β functions of x and t (and of the history of the fracture) and $\alpha > 0$, then the corresponding term in (16) will be

$$r_{2j} = \Delta p_j \ \alpha_j + \beta_j \quad \text{with} \quad \alpha_j = \Delta t \int_{y_{j-1}L}^{y_jL} \alpha \, dx \quad \text{and} \quad \beta_j = \Delta t \int_{y_{j-1}L}^{y_jL} \beta \, dx .$$

The Δp_j-term can be moved over to the left-hand side in (16) where, by virtue of $\alpha_j > 0$, it reinforces the main diagonal of the matrix and the positive definiteness.

3. By a simple adaptation of the first row in (16) the program can also handle a given pressure as inlet condition.

4. Results

A few results will be discussed with the help of figures 4 to 6.

A first check on accuracy is possible by comparison with an analytical solution. In the steady state with $q_{in} = 0$ and $v_\ell = 0$ the situation is known to be: p = constant and $w(x) = \frac{\pi}{2}\frac{\Delta p}{\sigma_0}\sqrt{L^2 - x^2}$. Indeed for $q_{in} = 0$ the program simulates a transition to steady state with p = constant ($=S_t$) and with no further length increase. The accuracy in w is depicted in figure 4 where w^2 versus x^2 ought to show a straight line.

Figure 5 shows a typical pressure distribution in the fracture in the propagating condition. In view of condition (6), S_t is chosen as a reference level. The pressure gradient is negative everywhere in order to supply the fluid flow. The behaviour in the tip zone may indicate an integrable singularity, e.g. logarithmic, which is mild enough to handle numerically.

Figure 6 shows four stages in a violent fracture opening process: a small fracture ($L = 1m$, $w_{max} = .07mm$) in steady state is suddenly forced to take in 30 times its original volume within one second.

Figure 6a starts with a high pressure to widen the opening sharply; the rest of the fracture still has its original shape. The high pressure is balanced by a negative region; the computer output showed that along the positive pressure slope the flow rate is indeed negative and that the local width shrinks at first.

In figure 6b the process of widening is moving further into the fracture; the elliptical shape is created; the pressure dip deepens; and the last part of the fracture is still intact.

In figure 6c the final fracture shape has nearly been achieved. Over most of the fracture length the pressure distribution is settling like figure 5, although it still ends with a very deep peak. For physical interpretation this may cause some trouble, for the stability of the program it clearly is no problem.

Finally consider figure 6d. Until now the fracture length has shown no remarkable increase: all the inflow has been used to widen the fracture and to build up a combination of pressure and width that is capable of absorbing digest the given inflow rate. Indeed, from now on the simulation shows length increase; the phase of regular growing has begun.

Acknowledgement

This work has been supported by the Koninklijke/Shell Exploratie en Productie Laboratorium, Rijswijk, Netherlands, and benefited from discussions between the author and P.A. Choate, D. den Ottelander, A.D. van Kuyk and T.O.H. Kuiper, of the KSEPL.

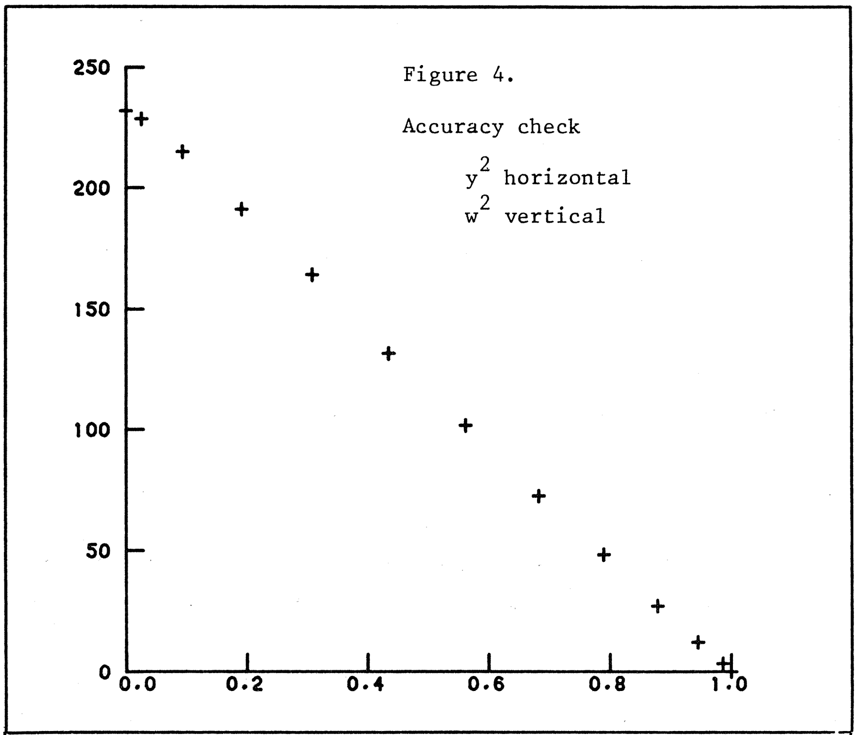


Figure 4.

Accuracy check

y^2 horizontal

w^2 vertical

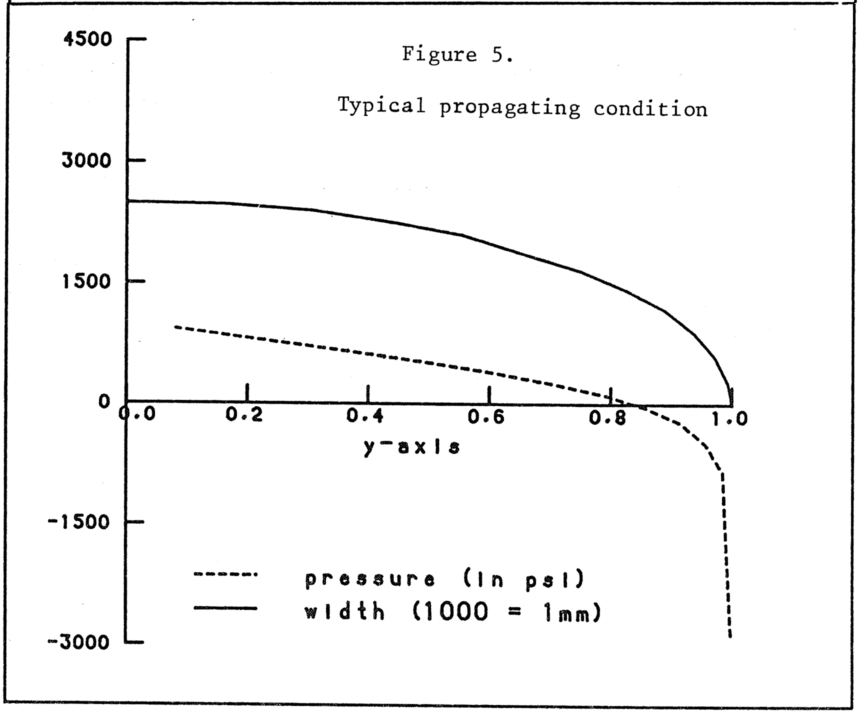
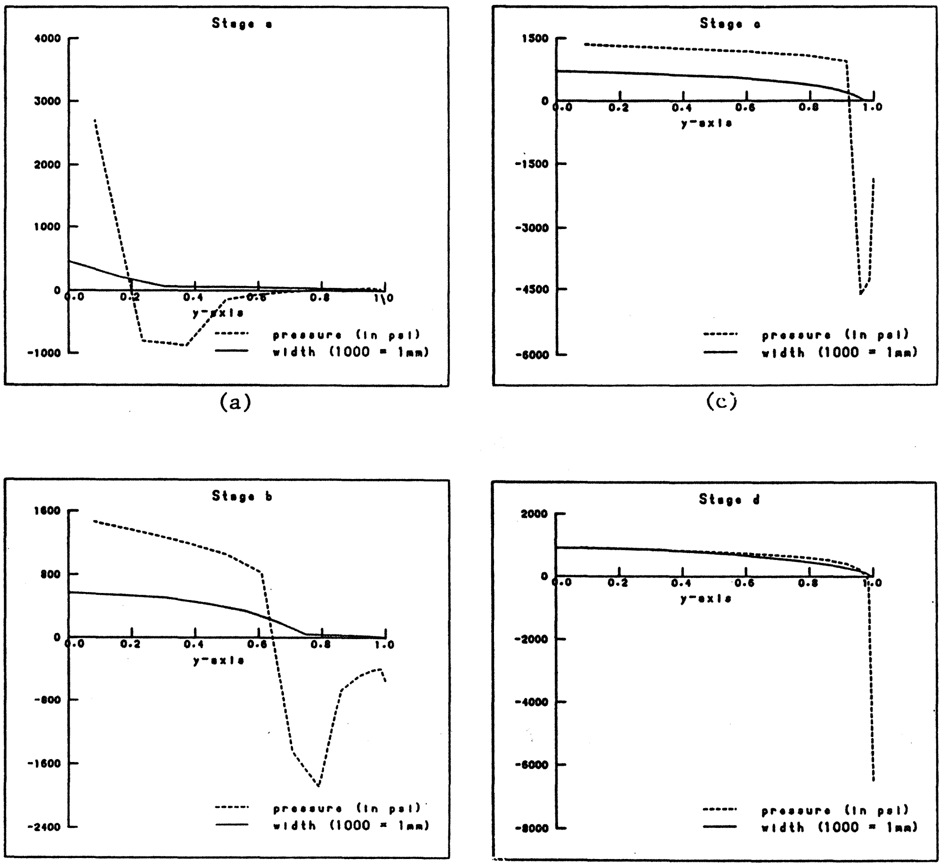


Figure 5.

Typical propagating condition



(a)

(c)

(b)

(d)

Figure 6. Width expansion

References

1. Daneshy, A.A., "On the design of vertical hydraulic fractures", Journal of Petroleum Technology, Jan 1973, p.83.

2. England, A.H. & A.E. Green, "Some two-dimensional punch and crack problems in classical elasticity", Proc. Camb.Phil.Soc. (1963), 59, p. 489.

3. Geertsma, J. & R. Haafkens, "A comparison of the theories for predicting width and extent of vertical hydraulically induced fractures", Trans. ASME, vol 101, March 1979, p. 8.

4. Geertsma, J. & F. de Klerk, "A rapid method of predicting width and extent of hydraulically induced fractures", Journal of Petroleum Technology, Dec. 1969, p. 1571.

5. Nordgren, R.P., "Propagation of a vertical hydraulic fracture", Society of Petroleum Engineering Journal, Aug 1972, p.306.

Appendix

Discretization of the elasticity relation

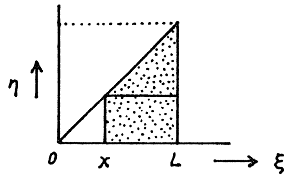
We start by reducing the double integral in (3) to a single integral. Then substitution of this integral into (8) will lead to the expression (15).

Expression (3), written as

$$w(x) = \frac{1}{\sigma_0} \int_{\xi=x}^{L} \int_{\eta=0}^{\xi} \frac{\xi}{\sqrt{\xi^2 - x^2}\,\sqrt{\xi^2 - \eta^2}}\, p(\eta)\, d\eta\, d\xi$$

is an integral over a trapezoidal domain in the $\xi\eta$-plane. Changing the order of integration we find

Figure 7.



$$w(x) = \frac{1}{\sigma_0} \left\{ \int_{\eta=0}^{x} \int_{\xi=x}^{L} \ldots\, d\xi d\eta + \int_{\eta=x}^{L} \int_{\xi=\eta}^{L} \ldots\, d\xi d\eta \right\}.$$

Substituting $\xi^2 = x^2\cosh^2 t - \eta^2\sinh^2 t$ into the first inner integral and $\xi^2 = \eta^2\cosh^2 t - x^2\sinh^2 t$ in the second, we find

$$w(x) = \frac{1}{\sigma_0} \{ \int_{\eta=0}^{x} p(\eta)\sinh^{-1} \surd \frac{L^2 - x^2}{x^2 - \eta^2}\, d\eta + \int_{\eta=x}^{L} p(\eta)\sinh^{-1} \surd \frac{L^2 - \eta^2}{\eta^2 - x^2}\, d\eta \}$$

$$= \frac{1}{\sigma_0} \int_{\eta=0}^{L} k(\eta,x)\, p(\eta)\, d\eta \qquad (a1)$$

where

$$k(\eta,x) = \log \frac{\surd\, L^2 - \eta^2 + \surd\, L^2 - x^2}{\surd\, |\eta^2 - x^2|} \; . \qquad (a2)$$

Now the mapping p → w is a symmetric positive definite operator, which property is conserved in our discretised formulation thanks to the introduction of the areas A. Indeed, with p piecewise constant:

$$p(x) = p_m \text{ for } \xi_{m-1} < x < \xi_m, \qquad m=1,2,\dots,n.$$

By (a1) we have:

$$w(x) = \frac{1}{\sigma_0} \sum_{m=1}^{n} p_m \int_{\eta=\xi_{m-1}}^{\xi_m} k(\eta,x)\, d\eta,$$

and thus

$$A_j \equiv \int_{x=\xi_{j-1}}^{\xi_j} w(x)\, dx = \frac{L^2}{\sigma_0} \sum_{m=1}^{n} K_{jm}\, p_m$$

$$\text{with} \quad K_{jm} = \frac{1}{L^2} \int_{x=\xi_{j-1}}^{\xi_j} \int_{\eta=\xi_{m-1}}^{\xi_m} k(\eta,x)\, d\eta\, dx. \qquad (a3)$$

Clearly K is a symmetric matrix. The positive definiteness requires a sufficiently fine quadrature and has been checked in the program. Finally, in dimensionless coordinates y = x/L and z = η/L:

$$K_{jm} = \int_{y=j_{j-1}}^{y_j} \int_{x=y_{m-1}}^{y_m} \log \frac{\surd\, 1-z^2 + \surd\, 1-y^2}{\surd|y^2 - z^2|}\, dz\, dy.$$

So the matrix K is independent of L and can be calculated and stored once for all time steps. The integrals K_{jm} are approximated by Gaussian quadrature with weight functions corresponding to the logarithmic behaviour.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Variational Approach to Bubble Deformation in Two-Phase Flow

J.A. Geurst and A.J.N. Vreenegoor
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands

ABSTRACT

Hamilton's variational principle for two-phase flow of a bubbly liquid/gas mixture including virtual mass (J.A. Geurst, Physica **129A** (1985) 233 and **135A** (1986) 455) is extended to include the effect of flow induced bubble deformation. The Weber number is introduced as an additional variable. The conservation of bubble number density is used as a constraint. The variational principle is brought in canonical form. The corresponding Lagrangian density is identified as the pressure (generalised Clebsch-Bateman principle). A stability analysis of the two-phase flow equations yields necessary and sufficient conditions for marginal stability. They take the form of PDE's, which admit an exact solution. An explicit expression is given for the virtual-mass coefficient.

1. INTRODUCTION

In classical hydrodynamics the Euler equations of motion of a perfect fluid may be derived from a generalised form of Hamilton's variational principle of least action (see e.g. [1]). The derivation is not presented in known textbooks, although it may be helpful in clarifying several aspects of the Euler equations. Instead the Euler equations are usually obtained from the laws of conservation of mass, momentum and energy.

In two-fluid hydrodynamics, which comprises the two-phase flow of a bubbly liquid/gas mixture in addition to the superfluid hydrodynamics of ^{4}He, variational methods seem to be indispensable. In a recent paper [2] it is demonstrated, how a generalised form of Hamilton's principle may be used to derive the correct form of the two-phase flow equations of a bubbly liquid/gas mixture in cases, where the virtual mass of the gas bubbles should be taken into account. The theory, however, does not allow deformations of the bubbles induced by their motion relative to the surrounding liquid. It is the aim of the present paper to extend the theory for a bubbly liquid/gas mixture by including those deformation effects in the two-phase flow equations.

Since flow induced bubble deformation is related to a non-vanishing Weber number, the dimensionless Weber number is introduced in the Lagrangian density as an additional variable in order to model the bubble deformation effects. In ref. [2] it was shown that, as a consequence of the theory, breakdown of bubbly flow should occur at the critical value 1/3 of the volume concentration of the gas bubbles. It is here investigated, how flow induced bubble deformation represented by a non-vanishing Weber number affects that critical value.

The analysis is confined to the non-dissipative behaviour of a bubbly liquid/gas mixture. Dissipative effects like viscosity may be included in a straightforward way (see [2]). An extensive review of two-phase flow theories including some variational methods is presented in [3].

2. VARIATIONAL PRINCIPLE AND TWO-PHASE FLOW EQUATIONS

A bubbly liquid/gas mixture may be characterised by the reduced densities ρ_1 and ρ_2 of the continuous liquid phase and the dispersed gas phase, respectively. The reduced densities are defined by

$$\rho_1 = (1 - \alpha)\rho_\ell, \qquad \rho_2 = \alpha\rho_g, \tag{2.1}$$

where ρ_ℓ denotes the constant mass density of the incompressible liquid, ρ_g is the mass density of the gas, while α represents the volume density of the gas phase usually called void fraction. The total mass density is given by

$$\rho = \rho_1 + \rho_2. \tag{2.2}$$

The gas satisfies the ideal-gas law

$$p_g = \frac{RT}{M}\rho_g, \tag{2.3}$$

where p_g represents the gas pressure, R the gas constant, T the absolute temperature, M the molecular mass of the gas. The flow is assumed to occur at isothermal conditions. In that case the thermodynamic properties of the liquid/gas mixture are obtained from the free energy density $F(\rho_1,\rho_2)$. The thermodynamic potentials μ_1 and μ_2 are defined by

$$dF = \mu_1 d\rho_1 + \mu_2 d\rho_2. \tag{2.4}$$

They satisfy

$$\rho_1\mu_1 + \rho_2\mu_2 = F + p_g. \tag{2.5}$$

Furthermore

$$d\mu_1 = \frac{RT}{M}\left\{\frac{\rho_2}{(\rho_\ell - \rho_1)^2}\, d\rho_1 + \frac{1}{\rho_\ell - \rho_1}\, d\rho_2\right\} \tag{2.6}$$

and

$$d\mu_2 = \frac{RT}{M}\left\{\frac{1}{\rho_\ell - \rho_1}\, d\rho_1 + \frac{1}{\rho_2}\, d\rho_2\right\}. \tag{2.7}$$

We refer to [2] for more details concerning the thermodynamic properties of a liquid/gas mixture.

The derivation of the two-phase flow equations starts from an extended form of Hamilton's principle of least action. The analysis will be confined to one-dimensional motion for the sake of convenience. The variation principle reads

$$\delta \int_{t_0}^{t_1} dt \int_{x_0}^{x_1} dx\; L = 0. \tag{2.8}$$

The Lagrangian density L is given by

$$L = K - F, \tag{2.9}$$

where K denotes the kinetic energy density of the liquid/gas mixture. The kinetic energy consists of the kinetic energy of the liquid phase, the kinetic energy of the gas phase and the kinetic energy associated with the motion of the gas bubbles relative to the liquid phase.

We write accordingly

$$K = \frac{1}{2}\rho_1 u_1^2 + \frac{1}{2}\rho_2 u_2^2 + \frac{1}{2}\rho_\ell m(\alpha,We)(u_2 - u_1)^2. \tag{2.10}$$

The local velocities of the liquid and the gas phase are denoted by u_1 and u_2, respectively. The virtual-mass effects associated with the relative motion are represented by the coefficient $m(\alpha,We)$. That coefficient is allowed to depend not only on the void fraction α but also on the Weber number We defined by

$$We = \frac{\rho_\ell(u_2 - u_1)^2}{(\gamma/2a)} \tag{2.11}$$

in order to model the flow induced bubble deformation of the gas bubbles in addition to their mutual interaction. The surface tension coefficient is denoted by γ, while 2a is the local average of the bubble diameter. It is one of the purposes of the investigation to obtain more information about the possible form of the function $m(\alpha,We)$. In view of the fact that the free energy density F as defined here does not include the energy associated with surface tension, it should be remarked that the last term at the right hand side of (2.10) may represent all surface tension effects.

The variation in (2.8) is restricted by the constraints

$$\frac{\partial\rho_1}{\partial t} + \frac{\partial}{\partial x}(\rho_1 u_1) = 0, \tag{2.12}$$

$$\frac{\partial\rho_2}{\partial t} + \frac{\partial}{\partial x}(\rho_2 u_2) = 0 \tag{2.13}$$

and

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x}(nu_2) = 0. \tag{2.14}$$

The constraints express, respectively, the conservation of mass of the liquid, the conservation of mass of the gas and the conservation of the number of bubbles. The number density of the gas bubbles is represented by n. Note that

$$\alpha = n\tau, \tag{2.15}$$

where τ is the local average of the bubble volume.

The local average 2a of the bubble diameter and the local average τ of the bubble volume are related by

$$\tau = \frac{4\pi}{3} a^3. \qquad (2.16)$$

When the constraints (2.12) to (2.14) are introduced in the variational principle (2.8) by means of Lagrange multipliers, the Lagrangian density is modified into

$$\hat{L} = L + \sum_{i=1}^{2} \phi_i \left[\frac{\partial \rho_i}{\partial t} + \frac{\partial}{\partial x} (\rho_i u_i)\right] + \lambda \left[\frac{\partial n}{\partial t} + \frac{\partial}{\partial x} (nu_2)\right]. \qquad (2.17)$$

Integrating by parts we obtain the equivalent Lagrangian density L* given by

$$L^* = L - \sum_{i=1}^{2} \rho_i \left(\frac{\partial}{\partial t} + u_i \frac{\partial}{\partial x}\right) \phi_i - n \left(\frac{\partial}{\partial t} + u_2 \frac{\partial}{\partial x}\right) \lambda. \qquad (2.18)$$

The variation applies to the independent variables ρ_1, ρ_2, n, u_1, u_2 and the Lagrange multipliers ϕ_1, ϕ_2, λ. The corresponding Euler-Lagrange equations read

$$\frac{1}{2} u_1^2 - \mu_1^* - \left(\frac{\partial}{\partial t} + u_1 \frac{\partial}{\partial x}\right) \phi_1 = 0, \qquad (2.19)$$

$$\frac{1}{2} u_2^2 - \mu_2 - \left(\frac{\partial}{\partial t} + u_2 \frac{\partial}{\partial x}\right) \phi_2 = 0, \qquad (2.20)$$

$$- \frac{1}{6} \rho_\ell \frac{We}{n} m_{We} (u_2 - u_1)^2 - \left(\frac{\partial}{\partial t} + u_2 \frac{\partial}{\partial x}\right) \lambda = 0, \qquad (2.21)$$

$$\rho_1 u_1 - \rho_\ell m^* (u_2 - u_1) - \rho_1 \frac{\partial \phi_1}{\partial x} = 0, \qquad (2.22)$$

$$\rho_2 u_2 + \rho_\ell m^* (u_2 - u_1) - \rho_2 \frac{\partial \phi_2}{\partial x} - n \frac{\partial \lambda}{\partial x} = 0, \qquad (2.23)$$

completed by the conservation equations (2.12) to (2.14).
The virtual-mass coefficient $m^*(\alpha, We)$ is given by

$$m^* = m + We m_{We}, \qquad (2.24)$$

where m_{We} denotes $(\partial/\partial We)m$, and the modified thermodynamic potential μ_1^* is defined according to

$$\mu_1^* = \mu_1 + \frac{1}{2}\left(m_\alpha + \frac{We}{3\alpha} m_{We}\right) (u_2 - u_1)^2, \tag{2.25}$$

where m_α denotes $(\partial/\partial\alpha)m$. The name of virtual-mass coefficient will be justified by the expression (3.12) for the kinetic energy density K^*. The term is furthermore suggested by the form of the equations of motion (2.29) and (2.30).

The total mass velocity is determined by

$$\rho u = \sum_{i=1}^{2} \rho_i u_i = \sum_{i=1}^{2} \rho_i \frac{\partial \phi_i}{\partial x} + n \frac{\partial \lambda}{\partial x}. \tag{2.26}$$

Combining (2.19) to (2.23) we obtain

$$\sum_{i=1}^{2} \rho_i u_i \frac{\partial \phi_i}{\partial x} + n u_2 \frac{\partial \lambda}{\partial x} = \sum_{i=1}^{2} \rho_i u_i^2 + \rho_\ell m^* (u_2 - u_1)^2 \tag{2.27}$$

and

$$\sum_{i=1}^{2} \rho_i \frac{\partial \phi_i}{\partial t} + n \frac{\partial \lambda}{\partial t} + K + F + p_g$$
$$+ \frac{1}{2} \rho_\ell \left[m + (1 - \alpha) m_\alpha + \left(\frac{1}{3\alpha} + 2\right) We m_{We}\right] (u_2 - u_1)^2 = 0. \tag{2.28}$$

Equation (2.28) constitutes a generalisation of the Bernoullian theorem, known from the classical hydrodynamics of one-phase fluids.

Differentiating (2.19) and (2.20) with respect to x and using (2.21) to (2.23) together with the conservation equations (2.12) to (2.14) we derive the equations of motion of the liquid and the gas phase:

$$\left(\frac{\partial}{\partial t} + u_1 \frac{\partial}{\partial x}\right) \left[u_1 - \frac{\rho_\ell}{\rho_1} m^* (u_2 - u_1)\right] - \frac{\rho_\ell}{\rho_1} m^* (u_2 - u_1) \frac{\partial u_1}{\partial x}$$
$$+ \frac{\partial}{\partial x} \mu_1^* = 0, \tag{2.29}$$

$$\left(\frac{\partial}{\partial t} + u_2 \frac{\partial}{\partial x}\right) \left[u_2 + \frac{\rho_\ell}{\rho_2} m^* (u_2 - u_1)\right] + \frac{\rho_\ell}{\rho_2} m^* (u_2 - u_1) \frac{\partial u_2}{\partial x}$$
$$+ \frac{\partial}{\partial x} \mu_2^* - \frac{1}{6} \frac{\rho_\ell}{n} We m_{We} (u_2 - u_1)^2 \frac{\partial}{\partial x}\left(\frac{n}{\rho_2}\right) = 0. \tag{2.30}$$

The modified thermodynamic potential μ_2^* is defined by

$$\mu_2^* = \mu_2 + \frac{1}{6}\frac{\rho_\ell}{\rho_2}\,\mathrm{Wem}_{\mathrm{We}}(u_2 - u_1)^2. \tag{2.31}$$

The two-phase flow of a bubbly liquid/gas mixture, which is characterised by giving ρ_1, ρ_2, n, u_1 and u_2 as functions of position and time, is determined by the evolution equations (2.12) to (2.14), (2.29) and (2.30).

3. THE CLEBSCH-BATEMAN PRINCIPLE

The conservation equations of energy and momentum may be derived by means of Noether's invariance theorem (see [2] and the references contained therein). The conservation of energy is expressed by

$$\frac{\partial H}{\partial t} + \frac{\partial Q}{\partial x} = 0, \tag{3.1}$$

where the Hamiltonian density H is given by

$$\begin{aligned} H &= -\sum_{i=1}^{2} \rho_i \frac{\partial \phi_i}{\partial t} - n\frac{\partial \lambda}{\partial t} - L^* \\ &= K + F + \rho_\ell \mathrm{Wem}_{\mathrm{We}}(u_2 - u_1)^2, \end{aligned} \tag{3.2}$$

while the energy flux Q is determined by

$$\begin{aligned} Q &= -\sum_{i=1}^{2} \rho_i u_i \frac{\partial \phi_i}{\partial t} - n u_2 \frac{\partial \lambda}{\partial t} \\ &= \rho_1 u_1 \left[\frac{1}{2} u_1^2 - \frac{\rho_\ell}{\rho_1} m^*(u_2 - u_1)u_1 + \mu_1^*\right] \\ &\quad + \rho_2 u_2 \left[\frac{1}{2} u_2^2 + \frac{\rho_\ell}{\rho_2} m^*(u_2 - u_1)u_2 + \mu_2^*\right]. \end{aligned} \tag{3.3}$$

The conservation of momentum is expressed by

$$\frac{\partial P}{\partial t} + \frac{\partial \Pi}{\partial x} = 0, \tag{3.4}$$

where the total momentum density P is determined according to

$$P = \sum_{i=1}^{2} \rho_i \frac{\partial \phi_i}{\partial x} + n\frac{\partial \lambda}{\partial x} = \sum_{i=1}^{2} \rho_i u_i = \rho u, \tag{3.5}$$

while the momentum flux Π is given by

$$\Pi = \sum_{i=1}^{2} \rho_i u_i \frac{\partial\phi_i}{\partial x} + nu_2 \frac{\partial\lambda}{\partial x} + L^*$$

$$= \sum_{i=1}^{2} \rho_i u_i^2 + \rho_\ell m^*(u_2 - u_1)^2 + p. \tag{3.6}$$

The pressure p is defined by

$$p = L^*. \tag{3.7}$$

For the definition of pressure we refer to [2]. It follows from (2.18), (2.27) and (2.28) that the pressure is determined by

$$p = p_g + \frac{1}{2}\rho_\ell \left[m + (1 - \alpha)m_\alpha + \frac{1}{3\alpha}\,\mathrm{Wem}_{\mathrm{We}}\right](u_2 - u_1)^2. \tag{3.8}$$

It is obvious from (3.2) that the Hamiltonian density H does not equal the sum of the kinetic and free energy densities K and F. Let us therefore introduce a new kinetic energy density K* and a new free energy density F* according to

$$H = K^* + F^*, \qquad L = K^* - F^*. \tag{3.9}$$

It follows that

$$K^* = K + \frac{1}{2}\rho_\ell \mathrm{Wem}_{\mathrm{We}}(u_2 - u_1)^2 \tag{3.10}$$

and

$$F^* = F + \frac{1}{2}\rho_\ell \mathrm{Wem}_{\mathrm{We}}(u_2 - u_1)^2. \tag{3.11}$$

Combining (2.10) and (3.10) we have

$$K^* = \frac{1}{2}\rho_1 u_1^2 + \frac{1}{2}\rho_2 u_2^2 + \frac{1}{2}\rho_\ell m^*(u_2 - u_1)^2. \tag{3.12}$$

It should be remarked, that the last term at the right hand side of (3.2), being invariant with respect to a Galilean transformation, may be attributed to the kinetic as well as to the free energy density.

It follows from (3.8), (3.10) and (3.11) that the Bernoullian theorem

(2.28) may be written in the form

$$\sum_{i=1}^{2} \rho_i \frac{\partial \phi_i}{\partial t} + n \frac{\partial \lambda}{\partial t} + K^* + F^* + p = 0. \tag{3.13}$$

According to (3.7) the modified Lagrangian density L* equals the pressure. Our variational principle therefore constitutes a generalisation of the Clebsch-Bateman principle valid for classical fluids (see [1]).

4. LINEAR MODES AND STABILITY

One of the interesting properties of a physical system is the behaviour of its linear modes. In order to investigate the linear modes the field quantities ρ_1, ρ_2, n, u_1, u_2 are taken in the form

$$u(x,t) = u_0 + \hat{u}e^{i(\omega t-kx)}, \tag{4.1}$$

where u_0 denotes the steady-state value and $\hat{u}$ represents the amplitude of a small perturbation. Neglecting products of perturbations in the evolution equations (2.12) to (2.14), (2.29) and (2.30) and eliminating the velocity perturbations $\hat{u}_1$ and $\hat{u}_2$ we arrive at the following system of linearised equations:

$$\begin{aligned} a_{11}\bar{\rho}_1 + a_{12}\bar{\rho}_2 &= 0, \\ a_{21}\bar{\rho}_1 + a_{22}\bar{\rho}_2 &= 0, \\ y\bar{\rho}_2 - y\bar{n} &= 0, \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} a_{11} = {} & \left\{1 + \frac{1}{1-\alpha} m + \frac{5}{1-\alpha} \mathrm{We} m_{\mathrm{We}} + \frac{2}{1-\alpha} \mathrm{We}^2 m_{\mathrm{WeWe}}\right\} y^2 \\ & + 2w_0 \left\{1 + \frac{2}{1-\alpha} m + m_\alpha + \frac{2+16\alpha}{3\alpha(1-\alpha)} \mathrm{We} m_{\mathrm{We}} + \mathrm{We} m_{\alpha \mathrm{We}} + \right. \\ & \qquad \left. + \frac{1+5\alpha}{3\alpha(1-\alpha)} \mathrm{We}^2 m_{\mathrm{WeWe}}\right\} y \\ & + w_0^2 \left\{1 + \frac{3}{1-\alpha} m + 2m_\alpha + \frac{-1+14\alpha+50\alpha^2}{9\alpha^2(1-\alpha)} \mathrm{We} m_{\mathrm{We}}\right. \end{aligned}$$

$$+ \frac{1-\alpha}{2} m_{\alpha\alpha} + \frac{1+5\alpha}{3\alpha} \mathrm{We}\, m_{\alpha \mathrm{We}} + \frac{1+10\alpha+25\alpha^2}{18\alpha^2(1-\alpha)} \mathrm{We}^2 m_{\mathrm{WeWe}} \Bigg\} +$$

$$- \frac{RT}{M} \beta \frac{1-\alpha}{\alpha}, \tag{4.3}$$

$$\begin{aligned} a_{12} = &- \left\{ \frac{1}{1-\alpha} m + \frac{5}{1-\alpha} \mathrm{We}\, m_{\mathrm{We}} + \frac{2}{1-\alpha} \mathrm{We}^2 m_{\mathrm{WeWe}} \right\} y^2 \\ &- w_0 \left\{ \frac{1}{1-\alpha} m + m_\alpha + \frac{2+11\alpha}{3\alpha(1-\alpha)} \mathrm{We}\, m_{\mathrm{We}} + \mathrm{We}\, m_{\alpha\mathrm{We}} + \right. \\ &\qquad \left. + \frac{1+4\alpha}{3\alpha(1-\alpha)} \mathrm{We}^2 m_{\mathrm{WeWe}} \right\} y \\ &+ w_0^2 \left\{ \frac{1+11\alpha}{18\alpha(1-\alpha)} \mathrm{We}\, m_{\mathrm{We}} + \frac{1}{6} \mathrm{We}\, m_{\alpha\mathrm{We}} + \frac{1+5\alpha}{18\alpha(1-\alpha)} \mathrm{We}^2 m_{\mathrm{WeWe}} \right\} - \frac{RT}{M} \beta, \end{aligned} \tag{4.4}$$

$$a_{21} = \frac{1-\alpha}{\beta\alpha} a_{12}, \tag{4.5}$$

$$\begin{aligned} a_{22} = &\frac{1}{\beta\alpha} \left\{ \beta\alpha + m + 5\mathrm{We}\, m_{\mathrm{We}} + 2\mathrm{We}^2 m_{\mathrm{WeWe}} \right\} y^2 \\ &- \frac{w_0}{\beta\alpha} \left\{ 2m + \frac{10}{3} \mathrm{We}\, m_{\mathrm{We}} + \frac{2}{3} \mathrm{We}^2 m_{\mathrm{WeWe}} \right\} y \\ &+ \frac{w_0^2}{\beta\alpha} \left\{ \frac{2}{9} \mathrm{We}\, m_{\mathrm{We}} + \frac{1}{18} \mathrm{We}^2 m_{\mathrm{WeWe}} \right\} - \frac{RT}{M} \end{aligned} \tag{4.6}$$

and

$$\bar{\rho}_1 = \frac{\hat{\rho}_1}{\rho_{1,0}}, \qquad \bar{\rho}_2 = \frac{\hat{\rho}_2}{\rho_{2,0}}, \qquad \bar{n} = \frac{\hat{n}}{n_0}. \tag{4.7}$$

The Doppler-shifted phase velocity y and the unperturbed relative velocity w_0 are given by

$$y = \frac{\omega}{k} - u_{2,0}, \tag{4.8}$$

$$w_0 = u_{2,0} - u_{1,0}. \tag{4.9}$$

Note that the zero suffix denoting unperturbed steady-state values has been deleted in the expressions (4.3) to (4.6) for the sake of convenience. The quantity β is defined by

$$\beta = \rho_{g,0}/\rho_{\ell}. \tag{4.10}$$

The velocity perturbations are determined by

$$\hat{u}_1 = (y + w_0)\bar{\rho}_1 \tag{4.11}$$

and

$$\hat{u}_2 = y\bar{\rho}_2. \tag{4.12}$$

The system (4.2) of linear equations admits a non-trivial solution if and only if the determinant of the coefficient matrix vanishes, i.e.,

$$y(a_{11}a_{22} - a_{12}a_{21}) = 0. \tag{4.13}$$

The equation (4.13), which is obviously a fifth degree algebraic equation in y, determines the phase velocities of the linear modes (dispersion equation). Considerations of stability require that the phase velocities possess real values. We therefore consider the discriminant D of the equation (4.13). It is known that D is negative when $m(\alpha,We)$ vanishes identically. It implies that the two-phase flow equations admit unstable steady-state solutions, when the virtual mass of the bubbles is neglected. A mathematical formulation is that the two-phase flow equations possess complex characteristics, when $m(\alpha,We) \equiv 0$. It is shown in [2] that D vanishes, when $m(\alpha,0) = (1/2)\alpha(1 - \alpha)(1 - 3\alpha)$. The vanishing of D entails marginal stability which corresponds to two coinciding (real) phase velocities. We investigate now, what form of $m(\alpha,We)$ might correspond to marginal stability (D = 0) in cases, where the Weber number does not vanish.

After some rather lengthy calculations it is found that the discriminant D vanishes independently of the value of β if and only if the following two partial differential equations for $f(\alpha,We)$ are *simultaneously* satisfied:

$$\frac{1}{2}\alpha(1 - \alpha)f_{\alpha\alpha} + (2\alpha - 1)Wef_{\alpha We} - 2We^2f_{WeWe} - 3Wef_{We} = 0, \tag{4.14}$$

$$(1 - \alpha)Wef_{\alpha We} + 2We^2f_{WeWe} + 5Wef_{We} + (1 - \alpha)f_{\alpha} + f + 1 = 0. \tag{4.15}$$

The function $f(\alpha,We)$ is related to $m(\alpha,We)$ according to

$$m(\alpha,We) = \alpha(1-\alpha)f(\alpha,We). \tag{4.16}$$

By introducing characteristic coordinates (η,ϕ) by means of

$$\begin{aligned} \eta &= We(1-\alpha)^2, \\ \phi &= 1-\alpha, \end{aligned} \tag{4.17}$$

the equations (4.14) and (4.15) are transformed into the following two partial differential equations:

$$\phi\eta\hat{f}_{\eta\phi} + \phi\hat{f}_{\phi} - \eta\hat{f}_{\eta} - \hat{f} - 1 = 0, \tag{4.18}$$

$$\phi\eta\hat{f}_{\eta\phi} - \eta\hat{f}_{\eta} + \frac{1}{2}\phi^2(1-\phi)\hat{f}_{\phi\phi} = 0, \tag{4.19}$$

where $\hat{f}(\eta,\phi)$ denotes the function f expressed in terms of characteristic coordinates. The equation (4.18) may be solved exactly according to

$$\hat{f}(\eta,\phi) = \phi C(\eta) + \frac{1}{\eta}D(\phi) - 1, \tag{4.20}$$

where $C(\eta)$ and $D(\phi)$ are unknown functions of the characteristic coordinates η and ϕ, respectively. Substitution of (4.20) in (4.19) yields an ordinary differential equation for $D(\phi)$, viz.,

$$\frac{1}{2}\phi(1-\phi)D'' - D' + \frac{1}{\phi}D = 0. \tag{4.21}$$

The general solution of (4.21) reads

$$D(\phi) = A\phi + B\frac{\phi}{1-\phi}, \tag{4.22}$$

where A and B are unknown constants.

It follows from (4.20) and (4.22) that the two partial differential equations (4.18) and (4.19) possess the *common* solution

$$\hat{f}(\eta,\phi) = \phi C(\eta) + \frac{1}{\eta}\left(A\phi + B\frac{\phi}{1-\phi}\right) - 1. \tag{4.23}$$

Combining (4.16), (4.17) and (4.23) we conclude that steady two-phase flow is marginally stable when the function $m(\alpha,We)$, which is related to the virtual-mass coefficient $m^*(\alpha,\mathrm{We})$ by means of (2.24), satisfies

$$m(\alpha,\mathrm{We}) = \alpha(1-\alpha)\left\{(1-\alpha)C(\mathrm{We}(1-\alpha)^2) + \frac{1}{\mathrm{We}(1-\alpha)}\left(A + \frac{B}{\alpha}\right) - 1\right\}. \tag{4.24}$$

5. DETERMINATION OF A, B AND C(η)

The form that might be taken by the unknown function $C(\eta)$, will be investigated by considering the limiting behaviour of the virtual-mass coefficient $m^*(\alpha,\mathrm{We})$ at small values of the void fraction α. That behaviour may be represented by

$$m^*(\alpha,\mathrm{We}) = \alpha\hat{m}(\mathrm{We}) + O(\alpha^2). \tag{5.1}$$

It follows from (2.24) and (4.16) that

$$m^*(\alpha,\mathrm{We}) = \alpha(1-\alpha)(f + \mathrm{We}f_{\mathrm{We}}). \tag{5.2}$$

Introducing the independent variables (η,ϕ) and using (4.23) we have

$$\begin{aligned} f + \mathrm{We}f_{\mathrm{We}} &= \hat{f} + \eta\hat{f}_\eta \\ &= \phi E(\eta) - 1, \end{aligned} \tag{5.3}$$

where

$$E(\eta) = \frac{d}{d\eta}(\eta C(\eta)). \tag{5.4}$$

By taking the limit $\alpha \to 0$ it is immediately inferred from the expressions (5.1), (5.2), (5.3) and the continuity of $E(\eta)$ that

$$E(\mathrm{We}) = 1 + \hat{m}(\mathrm{We}). \tag{5.5}$$

Note that, according to (5.1), the function $\hat{m}(\mathrm{We})$ represents the virtual-mass coefficient taken per unit volume of the gas in a low density dispersion of gas bubbles in liquid. Some information concerning the function $\hat{m}(\mathrm{We})$ might therefore be obtained by considering the inertial properties of the separate gas bubbles moving through the liquid.

It is shown in [4] that a gas bubble moving with velocity U relative to the surrounding liquid takes at small values of the Weber number $\widetilde{We}$ a near-spherical shape expressed in spherical polar coordinates by

$$r = a\left[1 + \frac{3}{64}\widetilde{We}\ (1 - 3\cos^2\theta)\right]. \tag{5.6}$$

The polar axis is assumed to coincide with the direction of relative motion, while the Weber number $\widetilde{We}$ is defined according to

$$\widetilde{We} = \frac{\rho_\ell U^2}{\gamma/2a}. \tag{5.7}$$

The near-spherical shape may be approximated by an oblate ellipsoid with eccentricity ε given by

$$\varepsilon^2 = \frac{9}{32}\widetilde{We} + O(\widetilde{We}^2). \tag{5.8}$$

According to [5] the virtual-mass coefficient $\hat{m}$ of an oblate ellipsoid is determined by

$$\hat{m} = \frac{\gamma_0}{2 - \gamma_0}, \tag{5.9}$$

where

$$\gamma_0 = \frac{2}{\varepsilon^2}\left(1 - \frac{\sqrt{1 - \varepsilon^2}}{\varepsilon}\ \arcsin\varepsilon\right). \tag{5.10}$$

It follows from (5.8), (5.9) and (5.10) that

$$\hat{m} = \frac{1}{2}\left[1 + \frac{27}{160}\widetilde{We} + O(\widetilde{We}^2)\right]. \tag{5.11}$$

In a low density dispersion of gas bubbles in liquid the kinetic energies associated with the motions of the separate gas bubbles may be added, because interaction effects between the bubbles are negligible. When it is assumed that the velocities of the gas bubbles relative to the liquid are nearly equal, the virtual masses of the separate gas bubbles may be added also. We therefore infer from (2.11), (5.7) and (5.11) that

$$\hat{m}(We) = \frac{1}{2}\left[1 + \frac{27}{160}We + O(We^2)\right]. \tag{5.12}$$

It has been assumed that the diameters of the gas bubbles are nearly equal.

It follows from (5.4), (5.5) and (5.12) that

$$c(\eta) = \frac{c_0}{\eta} + \frac{1}{2}\left(3 + \frac{27}{320}\eta\right) + O(\eta^2), \tag{5.13}$$

where c_0 is an unknown constant. According to (4.23) the constant c_0 may be taken equal to zero without affecting the generality of the expression for $\hat{f}(\eta,\phi)$.

The terms in the expression (4.24) for $m(\alpha,We)$ which contain the constants A and B, contribute to the free energy density F^*, but not to the kinetic energy density K^*. Those terms may be used to model the surface tension energy F_γ, which was not included in the free energy density F.

In the limit We → 0, the gas bubbles have a spherical shape. The surface tension energy F_γ is accordingly given by

$$F_\gamma = n\gamma 4\pi a^2 = \frac{3\gamma\alpha}{a}. \tag{5.14}$$

According to (3.11), (4.24) and (5.13) the difference of the free energy densities F^* and F is determined by

$$\begin{aligned} F^* - F &= \frac{1}{2}\rho_\ell We\, m_{We}(u_2 - u_1)^2 \\ &= \frac{3\gamma\alpha}{a}\left\{-\frac{A}{12} - \frac{B}{12\alpha} + \frac{3}{2560}(1-\alpha)^4 We^2 + O(We^3)\right\}. \end{aligned} \tag{5.15}$$

Taking the limit We → 0 and comparing with (5.14) we derive that

$$A = -12, \qquad B = 0. \tag{5.16}$$

The final expressions for $m(\alpha,We)$ and the virtual-mass coefficient $m^*(\alpha,We)$ read

$$m(\alpha,We) = \frac{1}{2}\alpha(1-\alpha)\left\{1 - 3\alpha + \frac{27}{320}(1-\alpha)^3 We\right\} - \frac{12\alpha}{We} + O(We^2), \tag{5.17}$$

$$m^*(\alpha,We) = \frac{1}{2}\alpha(1-\alpha)\left\{1 - 3\alpha + \frac{27}{160}(1-\alpha)^3 We\right\} + O(We^2). \tag{5.18}$$

The last term in the expression for $m(\alpha,We)$ gives rise to the well-known pressure difference $-2\gamma/a$ associated with the surface tension.

For physical reasons the virtual-mass coefficient $m^*(\alpha,We)$ should be non-negative. According to (5.18) the virtual-mass coefficient changes sign when $\alpha = \alpha_c = (1/3) + \delta_c$, where

$$\delta_c = \frac{1}{60}\,\mathrm{We} + O(\mathrm{We}^2). \tag{5.19}$$

The breakdown of bubbly flow (see [2]) is therefore shifted to larger values of α, in the case where the gas bubbles are deformed by the flow as a result of a finite value of the surface tension coefficient γ. In view of (5.19) the shift is relatively small. In any case it may be concluded that bubble deformation effects tend to stabilise two-phase bubbly flow in a first-order approximation.

REFERENCES

[1] R.L. Seliger and G.B. Witham, Proc. Roy. Soc. **A305** (1968) 1.
[2] J.A. Geurst, Physica **129A** (1985) 233.
[3] A. Bedford and D.S. Drumheller, Int. J. Engng. Sci. **21** (1983) 863.
[4] D.W. Moore, J. Fluid Mech. **6** (1959) 113.
[5] H. Lamb, Hydrodynamics, 6th ed., Dover, New York, 1945.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Low Order Spectral Models of the Atmospheric Circulation*

J. Grasman
Department of Mathematics
University of Utrecht
P.O. Box 80010, 3508 TA Utrecht, The Netherlands

and

H.E. de Swart
Department of Applied Mathematics
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

ABSTRACT

The spectral model consists of a system of coupled nonlinear ordinary differential equations. Low order barotropic models with 3 and 6 components have several stable solutions representing stream patterns with either a weak or a strong zonal component. With bifurcation theory these solutions are analyzed.
It is also shown that a 10 component model contains a strange attractor exhibiting alternately a weak and a strong zonal circulation pattern.
A comparable behaviour is found in 3 and 6 component models perturbed by noise.

* These investigations were supported by the Netherlands Foundation for the Technical Sciences (STW), future Science branch of the Netherlands Organization for the Advancement of Pure Research (ZWO).

1. INTRODUCTION

In this contribution low order spectral solutions of the barotropic potential vorticity equation are studied. A spectral model consists of a system of coupled nonlinear ordinary differential equations for the time dependent coefficients of the spectral expansion. Truncation of this expansion yields a finite dimensional approximation of the atmospheric flow described by the vorticity equation.

First we consider a 3-dimensional model having two stable equilibria and one unstable equilibrium at the separatrice. These equilibria can be seen as preferent states of the atmosphere. The irregular alternation of these preferent states, as observed in circulation patterns, is not reflected by the simple 3-dimensional model. One way to compensate the effects of the severe truncation in the spectral model is to add stochastic forcing to the system. In section 3 an analysis of this stochastic problem is presented. Special attention is given to the expected time of residence near a preferent state. Moreover, a discrete state Markov process is formulated; it describes the stochastic alternation of preferent states. In section 4 higher dimensional spectral models are discussed. A bifurcation analysis shows that the equilibria of the 3-dimensional deterministic system are unstable in the higher dimensional model and that for changing parameter values periodic solutions may branch off. Important is the occurrence of chaotic solutions (strange attractors) that visit, in an irregular way, different regular limit solutions, which are situated in different parts of state space (regimes).

2. DERIVATION OF THE SPECTRAL MODEL

For a large scale barotropic flow over a slowly varying topography in a midlatitude beta plane we assume the following: let H be the characteristic height, k^{-1} the horizontal length sale en σ^{-1} the time scale. The topography has a characteristic amplitude h_0. The meridional scale of the flow is assumed to be much smaller than the radius of the earth r_0. The potential vorticity equation for this circulation model reads in nondimensional form

$$\frac{\partial}{\partial t}\ \nabla^2\Psi + J(\Psi,\nabla^2\Psi) + \gamma J(\Psi,h) + \bar{\beta}\frac{\partial\Psi}{\partial x} + \bar{C}\nabla^2(\Psi-\Psi^*) = 0,$$

where $\Psi(x,y)$ is the stream function h the position of the earth's surface and Ψ^* a forcing stream function. Furthermore,

$$J(a,b) = \frac{\partial a}{\partial x}\frac{\partial b}{\partial y} - \frac{\partial a}{\partial y}\frac{\partial b}{\partial x},\quad \gamma = \frac{f_0 h_0}{\sigma H}\ ,\quad \bar{\beta} = \frac{\beta_0}{\sigma k}\ \text{ and } \bar{C} = \frac{f_0\delta_E}{2\sigma H}\ ,$$

where

$$f_0 = 2\Omega\sin\phi_0,\quad \beta_0 = \frac{2\Omega\cos\phi_0}{r_0}$$

with ϕ_0 the central latitude and Ω the angular speed of rotation of the earth. Finally, δ_E is the depth of the Ekman layer near the surface. We investigate the existence of travelling wave solutions in a rectangular channel with length L and width B = $\frac{1}{2}$bL. The nondimensional length and width are 2π and πb. The boundary conditions are

$$\Psi(x,y,t) = \Psi(x+2\pi,y,t),$$

$$\frac{\partial\Psi}{\partial x} = 0 \text{ and } \frac{\partial}{\partial t}\int_0^{2\pi}\frac{\partial\Psi}{\partial y}\,dx = 0 \qquad \text{at} \qquad y = 0 \text{ and } y = \pi b.$$

Let ϕ_i, i = 1,2,... be an orthonormal set of eigenfunctions of the Laplace operator for the domain of the channel:

$$\phi_1 = \sqrt{2}\cos(y/b),\qquad \phi_2 = 2\cos x\sin(y/b),$$
$$\phi_3 = 2\sin x\sin(y/b),\ \phi_4 = \sqrt{2}(\cos(2y/b)\ ,\ldots$$

Moreover, the functions Ψ^* and h are assumed to be of the form

$$\Psi^* = b(x_1^*\phi_1 + x_4^*\phi_4),\qquad h = \tfrac{1}{2}\phi_2.$$

Substitution of the expansion

$$\Psi(x,y,t) = b\sum_{n=1}^{\infty} x_n(t)\ \phi_n(x,y) \tag{1}$$

yields an infinite system of differential equations for $x_n(t)$, n = 1,2,... .

3. THE 3-DIMENSIONAL MODEL WITH STOCHASTIC FORCING

Taking $x_4^* = 0$ and $x_n(t) = 0$ for $n = 4,5,\ldots$, we obtain by substitution of (1) in the vorticity equation a system of differential equations for the remaining coefficients

$$\frac{dx_1}{dt} = bx_3 - C(x_1 - x_1^*), \tag{2a}$$

$$\frac{dx_2}{dt} = -ab(x_1 - \tfrac{1}{2}\beta)x_3 - Cx_2, \tag{2b}$$

$$\frac{dx_3}{dt} = ab(x_1 - \tfrac{1}{2}\beta)x_2 - \tfrac{1}{2}ax_1 - Cx_3 \tag{2c}$$

with

$$a = \frac{2b}{1+b^2}, \quad \beta = \frac{3\pi}{4\sqrt{2}}\bar{\beta} = 2.55, \quad C = \frac{3\pi}{4\sqrt{2}}\bar{C} = .2$$

or

$$\frac{dx_i}{dt} = f_i(x), \quad i = 1,2,3. \tag{3}$$

The stationary points $\bar{x}$ satisfy the equation $f(\bar{x}) = 0$. Depending on the parameter values either one or three real valued roots are found. In fig. 1 the first component of the equilibrium $\bar{x}$ is given as a function of x_1^*. Fig. 2 gives the three circulation patterns that correspond with the three equilibria for $x_1^* = 10$. The two stable equilibria with attraction domains Ω_i are denoted by $\bar{x}^{(i)}$, $i = 1,3$ $(x_1^{(1)} > x_3^{(1)})$ and the unstable one at the separatrice Γ by $\bar{x}^{(2)}$.

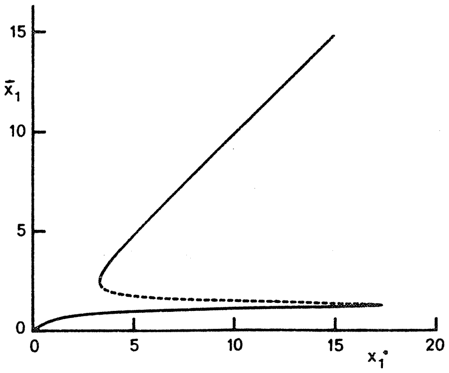


Fig. 1 Equilibrium solution $\bar{x}_1$ as a function of x_1^* for b = 1

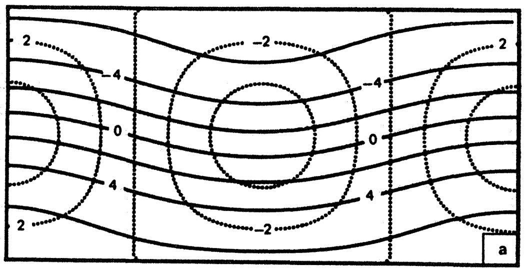
Next we consider the system (3) with each term forced by white noise of intensity ε:

$$dx_i = f_i(x)dt + \varepsilon dW_i(t), \quad = 1,2,3, \tag{4}$$

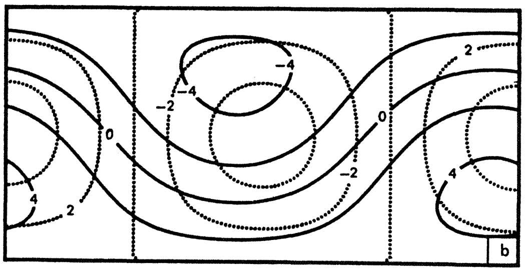
where $W_i(t)$, $i = 1,2,3$ are independent Wiener processes. This stochastic input compensates the absence of higher order spectral terms. The stochastic dynamical system (4) can be approximated by a diffusion process. Let $p(x,t)$ be the probability density distribution that the system is in state x at time t. Then $p(x,t)$ satisfies the so-called Fokker-Planck equation

$$\frac{\partial p}{\partial t} = \tfrac{1}{2}\varepsilon^2 \Delta p - \nabla .\ (pf(x)) \quad \text{or} \quad \frac{\partial p}{\partial t} = M_\varepsilon p. \tag{5}$$

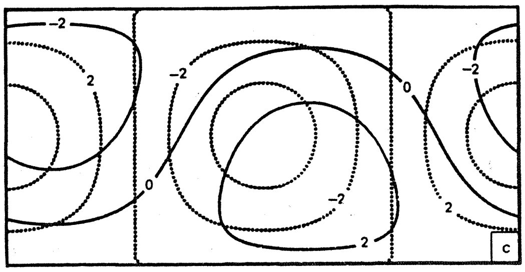
Let at time $t = 0$ the system be in x. Then $T(x)$ is defined as the first passage time of arriving at the separatrice $\Gamma = \partial\Omega_i$ of the deterministic system. Its expected value $T(x)$ satisfies Dynkin's equation



(a) the equilibrium $\overline{x}_1$



(b) the equilibrium $\overline{x}_2$



(c) the equilibrium $\bar{x}_3$

Fig. 2 Dimensional stream function patterns for the equilibrium states of the 3-dimensional spectral model. Dashed lines represent contours of the orography.

$$L_\varepsilon T = -1 \quad \text{in} \quad \Omega_i, \qquad (6a)$$

$$T = 0 \quad \text{at} \quad \Gamma, \qquad (6b)$$

where L_ε is the formal adjoint of M_ε:

$$L_\varepsilon \equiv \tfrac{1}{2}\varepsilon^2\Delta + f(x).\ \nabla\ .$$

The elliptic singular perturbation problem (6) has an asymptotic solution of the form

$$T(x) \approx C_i\ e^{K_i/\varepsilon^2} \quad \text{in } \Omega_i \text{ outside a neighborhood of } \Gamma, \qquad (7a)$$

$$T(x) \approx C_i\ e^{K_i/\varepsilon^2} \sqrt{\frac{2}{\pi}} \int_0^{s(x)} e^{-\frac{1}{2}t^2}\,dt \text{ in } \Omega_i \text{ near } \Gamma \qquad (7b)$$

with

$$s(x) = \frac{2}{\varepsilon}\ \{\int_0^{r(x)} \frac{\partial f}{\partial \nu}\ rdr\}^{\frac{1}{2}},$$

where ν is the normal at Γ and $r(x)$ the distance to Γ. The constants C_i and K_i are determined as follows. Let $p(x)$ satisfy the stationary Fokker-Planck equation in Ω_i and be of the form

$$p(x) = w(x)e^{-Q(x)/\varepsilon^2} \tag{8}$$

with

$$Q(\overline{x}^{(i)}) = 0 \quad \text{and} \quad Q(x) > 0 \quad \text{for} \quad x \neq \overline{x}^{(i)}.$$

The functions Q(x) and w(x) are determined by the ray method.
Substitution of (7) and (8) in the formula for the divergence theorem gives

$$\int_{\Omega_i} \{pL_\varepsilon T - TM_\varepsilon p\}dV = \int_\Gamma [\tfrac{1}{2}\varepsilon^2\{p\frac{\partial T}{\partial \nu} - T\frac{\partial p}{\partial \nu}\} + pTf(x).\nu]ds.$$

For $\varepsilon \to 0$ this equation must hold asymptotically which yields the values of K_i and C_i. We only give

$$K_i = \lim_{x \to \overline{x}_2} Q(x), \qquad K_1 = .23 \quad \text{and} \quad K_2 = .52.$$

It is concluded that most of the time the system is in an ε-neighborhood of one of the two stable equilibria. The attraction dom of these equilibria is most likely left through the separatrice Γ in an ε-neighborhood of the unstable equilibrium $\overline{x}_2$. The expected residence time in domain Ω_i is

$$T_i \approx C_i \; e^{K_i/\varepsilon^2} .$$

Near the unstable equilibrium the system remains a time of order

$$T_2 \approx \frac{1}{\lambda}\ln(\frac{1}{\varepsilon}),$$

where λ is the largest positive eigenvalue of the deterministic system linearized at $\overline{x}_2$. Estimates of ε for atmospheric models are given by Egger and Shilling (1983). They found $\varepsilon^2 \approx .2$.

A discrete Markov process is formulated as follows. Let Q_{ij} denote the transition probability per unit of time from state i to j (i,j = 1,2,3) and let $p_i(t)$ denote the probability of being in state i at time t. Then $p_i(t)$ satisfy

$$\frac{dp_1}{dt} = -(Q_{12}+Q_{21})p_1 - Q_{21}p_3 + Q_{21},$$

$$\frac{dp_3}{dt} = - Q_{23}p_1 - (Q_{32}+Q_{23})p_3 + Q_{23},$$

$$\Gamma_2 = 1 - p_1 - p_3 ,$$

where

$$Q_{23} = Q_{21} = \frac{1}{2T_2}, \qquad Q_{12} = \frac{1}{T_1} \quad \text{and} \quad Q_{32} = \frac{1}{T_3} .$$

In fig. 3 the probability functions $p_i(t)$ are given for a process that starts in state 1 with probability 1.

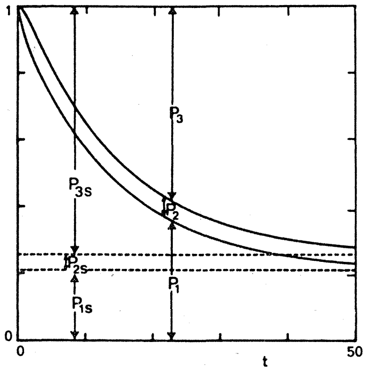
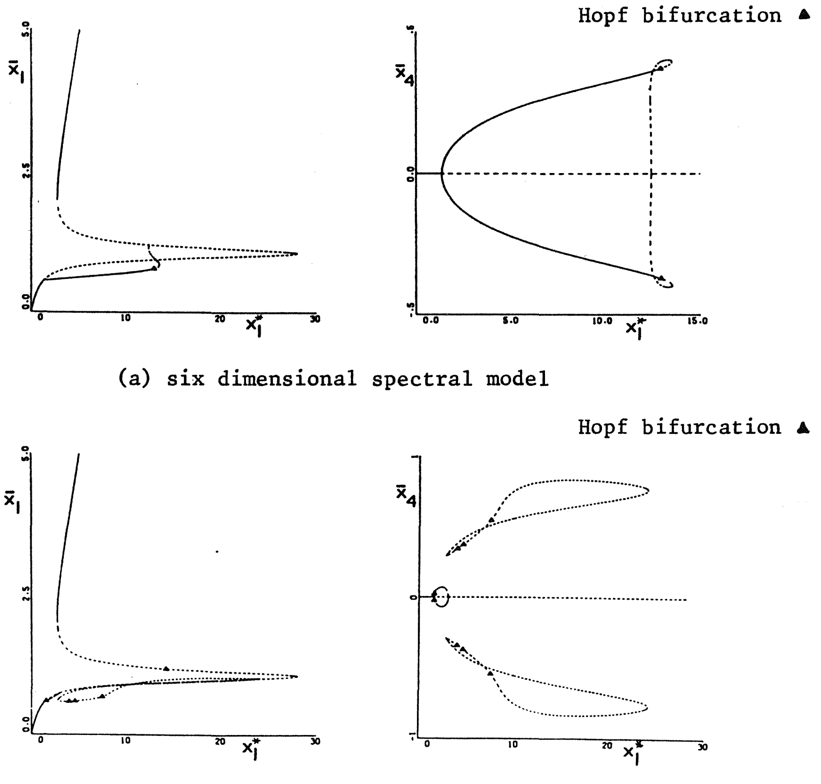


Fig. 3 Evolution of the probability distribution of the Markov process starting in state 1. The dotted lines represent the stationary distribution.

4. HIGHER DIMENSIONAL SPECTRAL MODELS

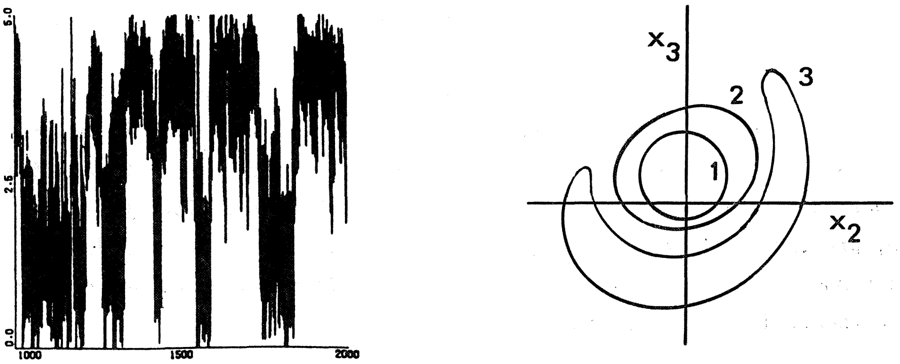
In higher dimensional spectral models the system will exhibit irregular dynamics from itself. No stochastic forcing is needed to obtain vacillation between states with a zonal flow of different intensities. In this section we summarize some results of De Swart (1987). The purpose is to formulate a spectral model with the lowest dimension that still has chaotic behavior with two clearly different scales of motion (planetary/synoptic) and that has a zonal component (x_1) that varies over a sufficiently large realistic range. In fig. 4 for a six and a ten dimensional model the bifurcation diagrams connected to the equilibria of the three dimensional model are given. One of the stable equilibria under goes a pitchfork bifurcation with the stable branches turning unstable through a Hopf bifurcation. However, the other equilibrium always remains stable. This is due to the symmetry in the forcing stream function. Therefore we take $x_4^* \neq 0$. It is verified that only in the ten dimensional model all equilibria get unstable for some x_4^*. Using physical arguments it is understood that 10 dimensions must be the minimum as only then energy exchange between wave triades is possible. In fig. 5a the x_1 component of a solution is given. Its largest Lyapunov exponent has a positive value, which indicates the presence of a strange attractor. Examining the course of a trajectory projected in the x_2,x_3-plane, we observe that this strange attractor remains from time to time close to three different periodic orbits. The behavior strongly resembles the discrete state Markov process of the preceding section, see fig. 5b. The deterministic chaotic model can be used to study the predictability of atmospheric flow from a theoretical point of view.



(a) six dimensional spectral model

(b) ten dimensional spectral model

Fig. 4 Bifurcation diagrams for higher dimensional spectral models. Solid (dotted) lines represent stable (unstable) stationary solutions.



(a) the x_1-component

(b) sketch of unstable periodic solutions

Fig. 5 A chaotic solution of the 10-dimensional model

LITERATURE

DE SWART, H.E., 1978, *Studies on low order spectral models of the atmospheric circulation: chaotic motion, predictability and vacillation behavior.* Thesis to appear.

DE SWART, H.E., and J. GRASMAN, 1987, *Effect of stochastic perturbations on a low-order spectral model of the atmospheric circulation,* to appear in Tellus A.

EGGER, J. and H.D. SCHILLING, 1984, *Stochastic forcing of planetary scale flow,* J. Atm. Sci. 41, p. 779-788.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

On the Variational Formulation of Hydrodynamic Lubrication Theory

E. van Groesen and R. Verstappen
Department of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

ABSTRACT

This contribution considers some aspects of the variational formulation for the creeping flow of a lubricant between rigid, moving bearings. Starting from a physical principle of virtual power, the restriction to creeping flow leads to a variational formulation for the differential equations and boundary conditions.
Performing the usual approximations based on the scaling of the problem directly into the functional provides an energy like functional that correctly produces the Reynolds equation and (free) boundary conditions (even for a pressure dependent viscosity). This approximated functional differs from the functional that is usually obtained in an ad hoc way by an additional term. This term depends on the velocity difference of the bearings and gives some new insight in the range of validity of the usual set of equations.

1. Introduction

Recently, many investigations into hydrodynamic lubrication problems deal with the variational formulation for the Reynolds equation. (see e.g. Capriz & Cimatti 1983). The variational approach has proved to be useful in several aspects: to derive existence and uniqueness results (cf. Bayada 1983, Oden & Wu 1985), for numerical purposes (cf. Wu 1986), and for the optimal construction of bearings (Mcallister & Rohde 1983). As in most of these references, we will consider the problem for rigid bearings. Although the variational structure for elastic deformations of bearings is also known (see e.g. Kalker 1977), a unified variational formulation for elasto-hydrodynamic lubrication problems is not yet completely understood. (Some quasi-variational formulations have been used by Wu 1986 and Strozzi 1986).

In the references quoted above, the variational formulation that is used is derived in an ad hoc way by simple presenting a functional for which the Euler-Lagrange equation is precisely the Reynolds equation.

The aim of this paper is to examine in detail how an (energy-like) functional can be derived from a physically well-understood and accepted principle of virtual power by performing approximations in the governing functional. In a certain approximation, the resulting functional will lead to the usual Reynolds equation, as required, but will also provide some new insight into additional restrictions that should be satisfied for this equation to be valid. An additional term that appears in this functional depends on the difference of the velocities of the bearings. This will probably effect the results for the optimal shape of bearings.

Following the description and notation of the basic problem, we present the principle of virtual power in section 2, and describe the approximations (which lead to the variational formulation of the problem) in section 3. In section 4 we analyse the resulting variational formulation, while section 5 is concerned with the modifications that are necessary to incorporate a viscosity-pressure relation.

Acknowledgement.

We would like to thank the tribology-group of the University of Twente for stimulating this research.

2. Basic formulation.

We consider the flow of a lubricant between two surfaces that are the rigid boundaries of two moving bearings. Restricting ourselves to the 2-D line contact problem, the extension to the 3-D point contact problem is obvious, the upper and lower boundary will be described in a x_1-x_2 plane by functions $x_2 = H_+(x_1)$ and $x_2 = H_-(x_1)$. The corresponding constant tangential velocities of the bearings will be denoted by w_+ and w_- respectively.

The stationary equilibrium equations for an imcompressible, Newtonian lubricant are described in terms of the velocity $\underline{v}$, the density ρ, the stress T, the pressure p, and the viscosity μ by the following set of equations:

$$\operatorname{div} \underline{v} = 0 \tag{1}$$

$$\operatorname{div} T = \rho(\underline{v}.\nabla)\underline{v} \tag{2}$$

$$T = -pI + \mu(\nabla\underline{v}+\nabla\underline{v}^T) \tag{3}$$

These equations describe the lubricant in a domain Ω in which no cavitation takes place. This domain Ω is unknown a priori, but defined by the requirement that p exceeds the cavitation pressure: $p > p_{cav}$, which value may be normalized by setting $p_{cav} = 0$.

The part of the boundary of Ω which coincides with part of the boundary of the bearings is denoted by $\partial\Omega_v$:

$$\partial\Omega_v = \partial\Omega_+ \cup \partial\Omega_- \text{ , with } \partial\Omega_{\underline{+}} = \{\underline{x} = (x_1,x_2) \mid x_2 = H_{\underline{+}}(x_1)\}$$

The no slip condition on $\partial\Omega_v$ can be described as follows:

$$\underline{v}_{\underline{+}} = W_{\underline{+}}(1,H_{\underline{+}}') \qquad \text{on } \partial\Omega_{\underline{+}}. \tag{4}$$

Here $\underline{v}_{\underline{+}}$ denotes the velocity on $\partial\Omega_{\underline{+}}$, the vector $(1, H_{\underline{+}}')$ is tangent to $\partial\Omega_{\underline{+}}$, and $W_{\underline{+}}$ is related to the constant tangential velocity $w_{\underline{+}}$ like

$$W_{\underline{+}} = w_{\underline{+}}(1+(H_{\underline{+}}')^2)^{-1} \tag{5}$$

Here, and in the following, a prime ' denotes differentiation with respect to x_1.

The remaining part of the boundary of Ω is defined by the isobar $p = 0$, and will be denoted by $\partial\Omega_p$. We suppose that on $\partial\Omega_p$ the surface stress is prescribed:

$$\underline{Tn} = \underline{t} \quad \text{on } \partial\Omega_p = \partial\Omega\backslash\partial\Omega_v. \tag{6}$$

In order to start with a well-posed problem the boundary condition (6) is necessary, although it is neglected in the literature. We will see later on that for a certain approximation to be valid, $\underline{t}$ cannot be prescribed arbitrarily.

To arrive at the principle of virtual power, it is to be noticed that eq. (2) with (6) follows from the vanishing of the expression

$$\int_\Omega (\text{div } T-\rho(\underline{v}.\nabla)\underline{v}).\delta\underline{v} \; d\Omega + \int_{\partial\Omega_p} (\underline{t}-T\underline{n}).\delta\underline{v} \; ds \tag{7}$$

for arbitrary functions $\delta\underline{v}$ with $\delta\underline{v} = \underline{0}$ on $\partial\Omega_v$. Integrating by parts gives

$$-\int_\Omega (T:\nabla\delta\underline{v} + \rho(\underline{v}.\nabla)\underline{v}.\delta\underline{v}) \; d\Omega + \int_{\partial\Omega_p} \underline{t}.\delta\underline{v} \; ds = 0. \tag{8}$$

When $\delta\underline{v}$ is considered as a virtual velocity, the volume integral represents the virtual mechanical work rate of the internal stresses and the virtual work rate of the inertia forces. The surface term represents the virtual work rate of the boundary forces.

Equation (8) subject to div $\underline{v}$ = 0 and the boundary condition (4) is an alternative statement of the problem, and is called the principle of virtual power (cf. Conner & Brebbia 1976).

3. Derivation of the approximate functional.

As it stands, equation (8) is not of the form of the vanishing of the first variation of a certain functional, due to the presence of the inertia term. Rather, (8) is a kind of quasi-variational principle. However, the usual restriction to creeping flow, i.e. assuming that the Reynolds number is very small compared to unity and omitting the inertia term, leads one to consider the following functional

$$E(p,\underline{v}) = \int_\Omega (p \text{ div } \underline{v} - \tfrac{1}{2}\mu(\nabla\underline{v}+\nabla\underline{v}^T) : \nabla\underline{v}) \; d\Omega + \int_{\partial\Omega_p} \underline{t}.\underline{v} \; ds \tag{9}$$

The notation $E(p,\underline{v})$ emphasises that E is considered as a functional of both p and $\underline{v}$. The critical points of this functional satisfy, in the required approximation, all the equations (1)-(2)-(3) and the boundary condition (6), provided the viscosity is constant. (In section 5 we will present the modification required to treat the general case.) Indeed, unrestricted variations of $p > 0$ in Ω lead to

$$\delta_p E = 0 \;\Rightarrow\; \mathrm{div}\,\underline{v} = 0 \quad \text{in } \Omega,$$

while arbitrary variations of $\underline{v}$, subject ot the boundary condition (4) only, i.e. subject to $\delta\underline{v} = \underline{0}$ on $\partial\Omega_v$, lead to

$$\delta_{\underline{v}} E = 0 \Rightarrow \begin{cases} -\nabla p + \mu(\Delta\underline{v} + \nabla\,\mathrm{div}\,\underline{v}) = \underline{0} & \text{in } \Omega \\ [-p + \mu(\nabla\underline{v} + \nabla\underline{v}^T)]\underline{n} = \underline{t} & \text{on } \partial\Omega_p. \end{cases}$$

Since $p = 0$ on $\partial\Omega_p$ (by definition), this set of equations reduces to

$$\begin{cases} \mathrm{div}\,\underline{v} = 0 \\ -\nabla p + \mu\Delta\underline{v} = \underline{0} \end{cases} \quad \text{in } \Omega \qquad\qquad (10)$$
$$\mu(\nabla\underline{v} + \nabla\underline{v}^T)\underline{n} = \underline{t} \qquad \text{on } \partial\Omega_p$$

It is to be noticed that the incompressibility of the lubricant is obtained from the variational formulation, and needs not to be imposed a priori (Stated differently, p can be considered as a Lagrange multiplier that is introduced to take account for this constraint).

A further approximation can be motivated by introducing dimensionless variables. With ε the quotient of characteristic lengths in the vertical (x_2) and horizontal (x_1) direction, the scaling of the lubrication problem is typically such that $\varepsilon \ll 1$.

Investigating the order of ε for the various terms in the functional (see Verstappen 1987 for details), the truncation up to second order leads to the approximate functional

$$E(p,\underline{v}) = \int_\Omega [p(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}) - \frac{1}{2}\mu(\frac{\partial v_1}{\partial x_2})^2 - \mu\frac{\partial v_1}{\partial x_2}\frac{\partial v_2}{\partial x_1}]\, d\Omega + \int_{\partial\Omega_p} \underline{t}\,.\,\underline{v}\, ds \qquad (11)$$

The term $\mu\int_\Omega \frac{\partial v_1}{\partial x_2}\frac{\partial v_2}{\partial x_1}\, d\Omega$ gives, in the approximation, rise to a boundary term only:

$$E(p,\underline{v}) = \int_\Omega [p(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}) - \frac{1}{2}\mu(\frac{\partial v_1}{\partial x_2})^2]\, d\Omega + \int_{\partial\Omega} \mu\,(v_1\frac{\partial v_2}{\partial x_2}\,,\,-v_1\frac{\partial v_2}{\partial x_1})^T.\,\underline{n}\, ds + \int_{\partial\Omega_p} \underline{t}\,.\,\underline{v}\, ds \qquad (12)$$

This functional, and the corresponding variational principle, will from now on be treated as the approximate basic formulation of the hydrodynamic lubrication problem. The analysis to follow consists in writing (12) as a functional in p by extremizing over $\underline{v}$ and evaluating it at the extremizing velocity.

We will reach this goal in several steps.

To start, we note that the variation with respect to v_2 leads to

$$\frac{\partial p}{\partial x_2} = 0 \qquad (13)$$

This implies that isobars are straight lines. In particular, the part of the boundary $\partial\Omega_p$, at which $p = 0$, is straight and we conclude that Ω can be described as

$$\Omega = \{\underline{x} \in \mathbb{R}^2 |\ H_-(x_1) \le x_2 \le H_+(x_1) \text{ and } x_a \le x_1 \le x_b\} \qquad (14)$$

for some, yet unknown, x_1-bounds x_a and x_b.

Exploiting (13) and (14), a part of the integration in (12) can be performed explicitly. Using the boundary condition (4) there results in a straightforward way

$$E(p,\underline{v}) = -\int_{x_a}^{x_b}\int_{H_-}^{H_+} [p'v_1 + \frac{1}{2}\mu(\frac{\partial v_1}{\partial x_2})^2]dx_2dx_1 + \mu \int_{H_-}^{H_+} v_1 \frac{\partial v_2}{\partial x_2}dx_2\Big|_{x_1=x_a}^{x_b}$$

$$+ \int_{\partial\Omega_p} \underline{t}.\underline{v}\ ds - \int_{x_a}^{x_b}[W_+(H_+'W_+)' - W_-(H_-'W_-)']dx_1 \qquad (15)$$

From now on we will neglect the last term in this expression, because the assumption that there exists characteristic lengths in the x_1- and x_2-direction implicitly implies that the curvatures of the surfaces $H_\pm$ are small. In that case, the velocities $W_\pm$ are approximately constant (independent of x_1).

Next, variations with respect to v_1 in (15) lead to the well known equation

$$p' = \mu \frac{\partial^2 v_1}{\partial x_2^2} \qquad \text{in } \Omega. \qquad (16)$$

From this equation and the boundary condition (4), v_1 can be expressed explicitly in terms of p'.

Introducing the filmthickness

$$h(x_1) = H_+(x_1) - H_-(x_1) \qquad (17)$$

and the quantity

$$k(x_1) = H_+(x_1) + H_-(x_1), \qquad (18)$$

the velocity v_1 is explicitly given by

$$v_1(x_1,x_2) = \frac{p'}{2\mu}(x_2^2-x_2k+H_-H_+) + \frac{x_2 - \frac{1}{2}k}{h}(W_+-W_-) + \frac{1}{2}(W_++W_-) \qquad (19)$$

This result was already derived by Berthe and Godet in 1973 using the differential formulation.

From the variations with respect to $\underline{v}$ there also result certain boundary conditions. Explicitly, variations of $\underline{v}$ on $\partial\Omega_p$ give

$$\delta_{v_1} E = 0 \quad \Rightarrow \quad t_1 = \begin{cases} -\mu\dfrac{\partial v_2}{\partial x_2} & (x_1=x_b) \\ \mu\dfrac{\partial v_2}{\partial x_2} & (x_1=x_a) \end{cases} \tag{20}$$

$$\delta_{v_2} E = 0 \quad \Rightarrow \quad t_2 = \begin{cases} \mu\dfrac{\partial v_1}{\partial x_2} & (x_1=x_b) \\ -\mu\dfrac{\partial v_1}{\partial x_2} & (x_1=x_a) \end{cases} \tag{21}$$

Of course, these conditions should be compatible with the expression (19) for the velocity. This implies that the approximation under consideration can only be valid provided that the surface stress $\underline{t}$, assumed to be prescribed so far, satisfies (20) and (21). In particular, t_1 should be of the order of ε. This point seems to be overlooked somewhat in the literature.

Resuming the results so far, (19), (20) and (21) hold for the velocity field which extremize the functional (15). Inserting these expressions in (15), the velocity field is eliminated in favour of the pressure. Performing the substitution and some explicit integrations there results

$$E(p) = \frac{1}{2}\int_{x_a}^{x_b} \left[\frac{1}{12}\,\frac{h^3}{\mu}(p')^2 - (W_+ + W_-)hp' - \mu\,\frac{(W_+ - W_-)^2}{h}\right]dx_1 . \tag{22}$$

We note that this functional depends on x_a, x_b, on the sum and the difference of the boundary velocities and on the filmthickness h (only; not on the boundaries $H_\pm$ themselves). When the flow of the lubricant is seen as a combination of a Couette flow

$$v_1 = \frac{x_2 - \frac{1}{2}k}{h}\,(W_+ - W_-) + \frac{1}{2}\,(W_+ + W_-)$$

and a Poisseuille flow

$$v_1 = \frac{p'}{2\mu}(x_2{}^2-x_2k+H_-H_+),$$

the terms in the functional (22) can be identified as the work rate of respectively the stress of the Poiseuille flow, and the normal and tangential stress of the Couette flow. Summarizing, we can say that the functional (22) has been derived from the basic variational principle for creeping flow (9) by approximating this functional up to second order of ε and extremizing over the velocity field.

4. Analysis of the variational formulation.

Having obtained the functional (22), let us now investigate the consequences of the corresponding variational principle. Variations with respect to p (satisfying p > 0), lead to the equation

$$\frac{1}{12\mu}(h^3p')' = \frac{1}{2}((W_++W_-)h)' \tag{23}$$

This is the well known Reynolds equation. The fact that this equation is obtained from (22) provides a formal, and partial, a posteriori justification for (22). However, (22) differs from the functional that is usually written down in an ad hoc way to provide the Reynolds equation (see e.g. Capriz & Cimatti 1983). The difference is the term

$$\mu \int_{x_a}^{x_b} \frac{(W_+-W_-)^2}{h}dx_1. \tag{24}$$

This term does not contribute when performing variations in p, i.e. does not alter the Reynolds equation. However, since (22) has the advantage of being derived directly from a physically sound principle, it is likely that the additional term should be incorporated. Therefore this additional term could effect the analysis for calculating the optimal shape of bearings (see Mcallister & Rohde 1983).

Next let us investigate what information can be derived for the boundary conditions at the free boundaries $x_1 = x_a$ and $x_1 = x_b$.

Assuming, without loss of generality, that $W_+ + W_- \geq 0$, x_a, the inlet boundary, is traditionally not treated as a free boundary (until now only Bayada (1983) has made an attempt to do so), but is taken to be fixed.

This is motivated by numerical results of Castle and Dowson (1972) who showed that the resulting pressure distribution is (almost) independent of the choice of x_a, provided $x_b - x_a$ is sufficiently large.

By constrast, the outlet condition at $x_1 = x_b$ has been studied intensively (cf. Dowson, 1975). Various arguments lead to the conclusion that p must satisfy

$$p'(x_b) = 0. \tag{25}$$

It is natural to ask if the Reynolds condition (25) can also be derived from the functional (22) by performing variations with respect to x_b (as is the usual way to treat free boundaries). Owing to the presence of the additional term in the functional this is not quite obvious, and the analysis provides some extra insight in the relevance of this term.

Variation with respect to x_b in (22) gives the following relation at x_b:

$$\frac{1}{12}\frac{h^3}{\mu}(p')^2 - (W_+ + W_-)hp' - \mu\frac{(W_+ - W_-)^2}{h} = 0. \tag{26}$$

Solving this for p' gives

$$p' = \frac{6\mu(W_+ + W_-)}{h^2}\left[1 \pm \left(1 + \frac{1}{3}\left(\frac{W_+ - W_-}{W_+ + W_-}\right)^2\right)^{\frac{1}{2}}\right] \text{ at } x_1 = x_b \tag{27}$$

On the other hand, since p satisfies $p > 0$ for $x_1 < x_b$ and $p(x_b) = 0$, we certainly must have $p' \leq 0$ at x_b. This shows that the solution (27) with the + sign is non realistic.

Moreover, requiring the Reynolds condition (25) implies that $W_+ = W_-$ at x_b and hence, approximately, $w_+ = w_-$.

Concluding, we can say that our results show that the usual free boundary condition (25) is obtained only if both bearings move with equal velocity; in case the difference $w_+ - w_-$ is small, the free boundary condition (27) at x_b reads approximately:

$$p' = \frac{6\mu(W_+ + W_-)}{h^2}\,[1-(1+\frac{1}{3}(\frac{W_+ - W_-}{W_+ + W_-})^2)^{\frac{1}{2}}] \approx -\frac{\mu}{h^2}\frac{(W_+ - W_-)^2}{(W_+ + W_-)} \tag{28}$$

We notice (see eq. (19)-(20)-(21) and (28)) that the cavitation boundary is approximately free of forces if $W_+ = W_-$ holds at $x_1 = x_b$; otherwise, for $W_+ - W_-$ small:

$$t_1 \approx 0,\ t_2 \approx -\frac{1}{2}(2x_2 - k)\,\frac{\mu}{h^2}\,\frac{(W_+ - W_-)^2}{W_+ + W_-} + \mu\,\frac{W_+ - W_-}{h}. \tag{29}$$

As far as we are aware of, no such findings have been reported yet.

5. <u>Pressure dependent viscosity.</u>

Let us now assume that μ is some given (smooth, monotone) function of p. In that case the functional (22) has to be modified in order to provide the Reynolds equation (23). The modification that is required can be described concisely by introducing an auxilliary variable q defined by $\frac{dq}{dp} = \frac{1}{\mu}$, i.e.

$$q = \int \frac{1}{\mu}\,dp \tag{30}$$

(This is a generalization of the Grubin transformation).
Then considering q, instead of p, as the basic variable in the variational principle:

$$\tilde{E}\ (q) = \frac{1}{2}\int_{x_a}^{x_b} [\frac{1}{12}h^3(q')^2 - (W_+ + W_-)\,hq' - \frac{(W_+ - W_-)^2}{h}]dx_1, \tag{31}$$

variations with respect to q lead to

$$\frac{1}{12}(h^3q')' = \frac{1}{2}((W_+ + W_-)h)', \tag{32}$$

which is the usual Reynolds equation, as can be seen by eliminating q in favour of p.

It may be noticed that the functional $\tilde{E}$ is definitely different from the functional E. Rewriting (31) in terms of p shows that the integrand contains an additional multiplicative factor $\frac{1}{\mu}$. (This point is often neglected; even in the standard monograph of Finlayson (1972), chapter 7). In particular, $\tilde{E}$ has a different physical meaning than the original functional E.

References.

Bayada, G., 1983, Variational formulation and associated algoritm for the starved finite journal bearing, ASME J. Lubr. Techn. 105, p. 453-457.

Berthe, D. & Godet, M., 1973, A more general form of Reynolds equation-applications to rough surfaces, Wear 27, p. 345-357.

Capriz, G. & Cimatti, G., 1983, Free boundary problems in the theory of hydrodynamic lubrication: a survey. In: Free boundary problems: theory and applications, vol. 2, A. Fasono & M. Primicerio (eds.), p.613-635.

Castle, P. & Dowson, D., 1972, A theoretical analysis of the starved elastohydrodynamic lubrication problem for cylinders in line contact, I. Mech. Engnrs., p. 131-137.

Conner, J.J. & Brebbia, C.A., 1976, Finite element techniques for fluid flow, Butterworth & Co., London.

Dowson, D. (a.o.), 1975, Cavitation and related phenomana in lubrication: proc. 1st Leeds-Lyon conf. on trib. held in Leeds 1974; publ.: Mech. Eng. Publ.

Finlayson, B.A., 1972, The method of weighted residuals and variational principles, Math. in Sci. and Engnrg. vol. 87, Academic Press, New York.

Kalker, J.J., 1977, Variational principles of contact elastostatics, J. Inst. Maths. Appl. 20, p. 199-219.

Mcallister, G.T. & Rohde, S.M., 1983, Optimum design of one-dimensional journal bearings, J. Opt. Th. Appl. 41, p. 599-617.

Oden, J.T. & Wu, S.R., 1985, Existence of solutions of the Reynolds equation of elastohydrodynamic lubrication, Inst. J. Engng. Sci 23, p. 207-215.

Strozzi, A., 1986, The elastohydrodynamic problem expressed in terms of extended variational formulation, ASME J. Trib. 108, p. 557-564.

Verstappen, R., 1987, On the variational formulation of hydrodynamic lubrication theory, Memorandum Un. of Twente, Dept. Appl. Math., to appear.

Wu, S.R., 1986, A penalty formulation and numerical approximation of the Reynolds-Hertz problem of elasto-hydrodynamic lubrication, Int. J. Engng. Sci. 24, p. 1001-1013.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

A Mathematical Model For a Falling Film Evaporator

by
R. van der Hout
Akzo Research
Corporate Research Department
P.O. Box 60, 6800 AB Arnhem, The Netherlands

1. SUMMARY

A falling film evaporator (FFE) is a vertically positioned tube, heated from the outside. Along the inner wall of the tube a liquid flows downward. Through the remaining space in the inside of the tube a gas is passed, in our case in the same direction as the liquid. Apart from flowing downward, the liquid evaporates partially into the gas flow.
Assuming turbulent fluid flow and turbulent gas flow, we present a mathematical model for a FFE, consisting of a system of ordinary differential equations.
FFE's are well known in the chemical engineering literature. As far as we know, the present model is new.

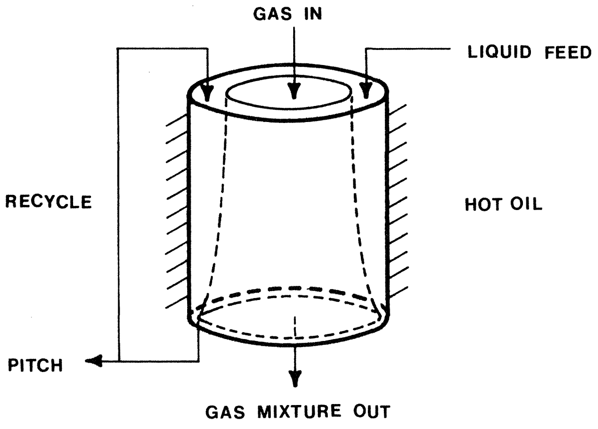


Figure 1. Sketch of a falling film evaporator.

List of symbols

symbol	meaning
ρ_1	density of the liquid
v_1	velocity of the liquid
ρ_n	density of the evaporated liquid
ρ_A	density of the "original" gas
$\rho_{A\circ}$	ρ_A at inflow
$v_\circ$	gas velocity at inflow
ρ	$\rho_n + \rho_A$
v	velocity of the gas
p	pressure of the gas
τ	temperature of the gas (°K)
T	temperature of the liquid (°K)
$M_\circ$	molar weight of original gas
M_1	molar weight of liquid
H	heat of evaporation
a, b, c	Antoine-constants, to be explained later
k	parameter in evaporation-formula to be explained later
$\emptyset_\circ$	flow of original gas
β	parameter in heat-flux formula liquid -> gas, to be explained later
Λ	heat conductivity of liquid
c_{vA}	specific heat original gas
c_{vn}	specific heat evaporated liquid (both at constant volume)
c_p	specific heat liquid
g	acceleration of gravity
η	viscosity of liquid
α	heat transfer coefficient oil —> liquid
R_g	gas constant

All quantities are of course supposed to be expressed in consistent units.
Many quantities (transfer coefficients, viscosity, ...) are dependent on the local conditions. Many dependencies are known approximately.

2. MATHEMATICAL MODEL

2.1. A mathematical model of a FFE must be

- a good quantitative description of the behaviour of the FFE under various process conditions.

- a good means for analysing and improving (and possibly optimizing) the performance of the FFE by means of simulation under various process conditions.

- a good means for optimizing the design of new units. Important questions to be answered are:

* What process conditions are optimal for given FFE-dimensions and specified performance?
* What are optimal FFE-dimensions and process conditions for specified process performance?

In what follows, we shall assume that, at inflow, all conditions are known (temperatures, fluxes, pressure).
Our model will have to consist of

- relations for fluid flow and heat transfer in the fluid
- relations for gas flow and heat transfer in the gas
- relations for interactions at the fluid/gas boundary
- relations for the free boundary
- relations for the heat transfer oil/liquid

We shall restrict ourselves to a quasi-stationary situation.

2.2. We now come to the model description proper.

We shall start from the following assumptions:

1) The gas flow is turbulent. All relevant quantities (pressure, temperature, ...) depend exclusively on the height in the tube.
2) The free surface fluid/gas is free of shear stresses. Viscous effects in the gas will be neglected.
3) For the fluid flow a Reynolds-number is defined by

$$Re_1 = \frac{2 * \text{mass flow}}{\pi\eta R}$$

The fluid flow is assumed to be turbulent whenever $Re_1 > 1600$. This criterion is somewhat arbitrary. Many authors prefer $Re_1 > 3200$ as a turbulence criterion. It is well known that, when the fluid flow is turbulent, there exist viscous boundary layers. These boundary layers will not be taken into consideration. The distinction between turbulent and viscous for the fluid flow is not only of physical importance, but has also mathematical consequences, resulting in different numerical procedures. Here, we shall almost exclusively consider turbulent fluid flow, because, for us, it is the only case of practical importance.

A sketch of the coordinate system is given in figure 2.

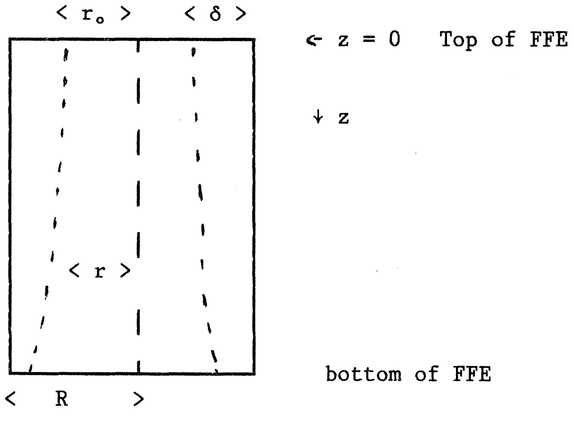


Figure 2. Cross-section of a FFE.

2.3. The fluid flow

We shall assume that the fluid velocity is pointing vertically downward and that the forces of gravity and the viscous force are in equilibrium throughout the liquid. The thickness of the fluid layer does not change dramatically over the length of the tube. Therefore we postulate that the fluid velocity is place-independent. In particular for turbulent flows this is not a very bad postulate. Moreover it is possible to get some feeling for the effect of this postulate by simply varying the fluid velocity in the computation.

For laminar flow we have

$$v_1 = \frac{\rho_1 g\ \delta^2}{3\eta} \text{ (= average value of } v_1\text{)}$$

$$\delta = {}^{3}\frac{3\eta * \text{mass flow}}{2\pi R\ g\ \rho_1{}^2} \qquad \text{(where } \delta << R\text{)}$$

These relations are easy to derive. Moreover, they can be found in the literature (for instance [1]).
For turbulent flow we use the following empirical relation:

$$\delta = .302\ \left(\frac{3\eta^2}{\rho_1{}^2 g}\right)^{1/3}\ \left(\frac{Re_1}{4}\right)^{8/15}$$

The corresponding value for v_1 can easily be derived from this relation together with the value of the flow.

A. For the sake of completeness we shall spend a few words on the laminar case. Here we have:

$$\rho_l c_p \frac{DT}{Dt} = \Lambda \nabla^2 T, \text{ where } \frac{D}{Dt} = \frac{\partial}{\partial t} + (v_l . \nabla\ )$$

Moreover, since only quasi-stationary flow is considered,

$$\frac{\partial T}{\partial t} = 0$$

As boundary conditions we have

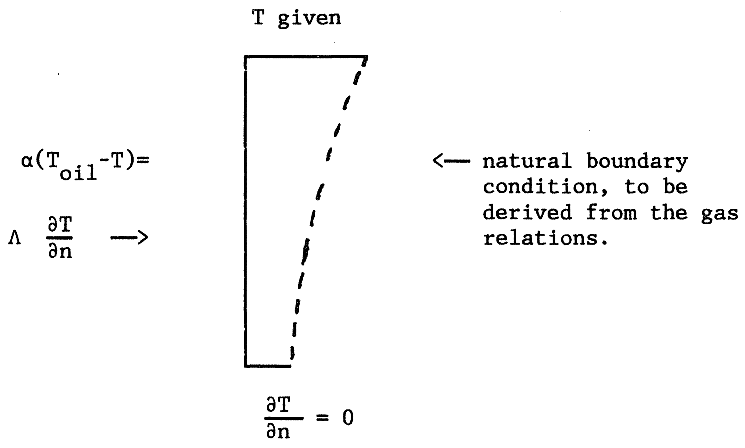
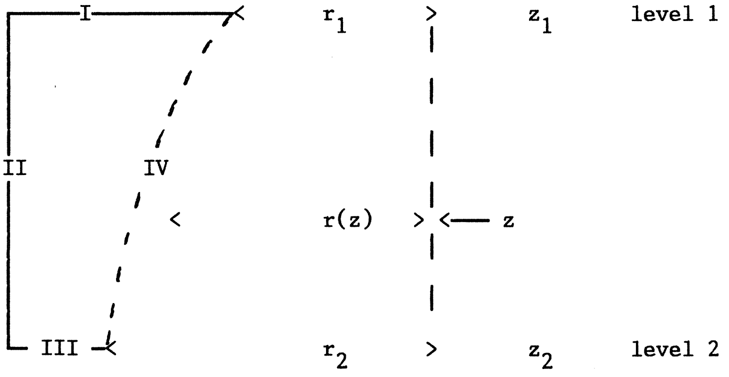

T given
$\alpha(T_{oil}-T)=$
$\Lambda \frac{\partial T}{\partial n}$ —>
<— natural boundary condition, to be derived from the gas relations.
$\frac{\partial T}{\partial n} = 0$

Figure 3. Boundary conditions.

In principle this heat problem is solvable as soon as the exact position of the free boundary is known. An additional relation is necessary in order to determine that position. An iterative procedure is needed to solve simultaneously the boundary position and the heat problem. Since turbulent flow is of more practical importance, we shall leave this laminar flow unconsidered.

B. The turbulent case.

Assume, for the time being, that the position of the free boundary is known. We have to construct an equation for the heat transfer in the liquid (the velocity being uniform, as was mentioned earlier). On the basis of the turbulence-assumption we postulate $T(r,z) = T(z)$



A simple formulation of the heat balance reads as follows: what flows in through I and II flows out through III and IV.

Expressed in formulas (where $T_1 = T(z_1)$; $T_2 = T(z_2)$) :

$$\begin{aligned}
&\text{Inflow through I} &:\quad & \pi(R^2 - r_1^2)\ \rho_1 v_1 c_p T_1 \\
&\text{Inflow through II} &:\quad & 2\pi R \int_{z_1}^{z_2} \alpha\ (T_{oil} - T(z))\ dz \\
&\text{Outflow through III} &:\quad & \pi(R^2 - r_2^2)\ \rho_1 v_1 c_p T_2 \qquad\qquad (1) \\
&\text{Outflow through IV} &:\quad & \pi H \rho_1 v_1 (r_2^2 - r_1^2) + \\
& & & 2\pi \int_{z_1}^{z_2} \beta r \sqrt{1 + (\frac{dr}{dz})^2}\ (T - \tau)\ dz + \\
& & & \int_{z_1}^{z_2} 2\pi \rho_1 v_1 c_p\ r \frac{dr}{dz}\ T\ dz
\end{aligned}$$

These terms describe successively: heat of evaporation, heating of the gas and the heat that is carried along with the evaporating liquid. Note that the enthalpy is described simply by $c_p T$.
We now have one relation with as unknowns $r(z)$, $T(z)$ and $\tau(z)$. Additional relations will follow in the next section.

2.4. The gas flow

1. Continuity equation

$$v = \rho_1 v_1 + \frac{r_o^2}{r^2}\ (\rho_{Ao} v_o - \rho_1 v_1) \qquad (2)$$

2. Equation of state

We assume (without good reason) that the Boyle-Gay Lussac relation holds:

$$p\ V = n\ R_g \tau \qquad \text{(n = number of moles)}$$

This formula can be interpreted as:

$$pr^2v = \left(\frac{r_o{}^2\rho_{A_o}\ v_o}{M_o} + \frac{(r^2 - r_o{}^2)\rho_1 v_1}{M_1}\right) R_g\tau \qquad (3)$$

3. Relation of evaporation

The driving force for the evaporation is supposed to be $P^* - P_{part}$ where

P^* = saturation - pressure of the evaporated liquid in the gas.

P_{part} = the actual partial pressure of the evaporated liquid.

For P^* we use a so called Antoine-relation:

$$P^* = \exp\ (a - \frac{b}{c + \tau})$$

Furthermore P_{part} = total pressure * (mol fraction of the evaporated liquid)

The relation of evaporation reads:
evaporation flux/unit surface = k $(P^* - P)$.

In formula:

$$\exp(a - \frac{b}{c+\tau}) = p\ \{1 - \frac{\frac{r_\circ^{2}\ \rho_{A\circ}v_\circ}{M_\circ}}{\frac{r_\circ^{2}\rho_{A\circ}v_\circ}{M_\circ} \quad \frac{(r^{2} - r_\circ^{2})\rho_1 v_1}{M_1}}\} + \frac{\rho_1 v_1 \frac{dr}{dz}}{k\sqrt{1 + (\frac{dr}{dz})^{2}}} \tag{4}$$

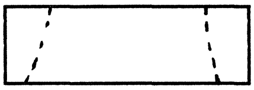
Note that we use τ in the formula for P^*. The use of τ in stead of T is not trivial, since, in our model, the jump between T and τ is partly due to the construction of the model. If we want to be sure which choice has to be made, we have to carry out an analysis of the boundary layer between gas and liquid. Although this might be important, no attempts were made to do so. As it is, it seems to be logical to try both T and τ and perhaps $\frac{T + \tau}{2}$.

4. Equation of motion

Since we postulated that the fluid velocity remains unchanged throughout the tube, we assume that the net force on every fluid particle vanishes. This force is composed of

- gravitation
- viscous force
- pressure force

Consider a segment R(t) of the tube, or rather the matter contained in R(t): liquid and gas



We have

$$\frac{d}{dt}\iiint\limits_{R(t)} \rho\ v\ d\ \mathrm{vol} = \text{sum of the forces, } \underline{\text{acting on the gas}}.$$

Those forces are: - gravitation
- pressure-forces

The total force amounts to:

$$\iiint\limits_{\text{gas}} \rho\ g\ d\ \mathrm{vol} + \iint\limits_{\substack{\text{boundary}\\ \text{of gas}}} -\ p\ n_z\ ds,$$

where n_z denotes the component in z - direction of the outward unit normal and ds the surface measure.

Consequently:

$$\iiint\limits_{\text{gas}} \rho\ g\ d\ \mathrm{vol} + \iint\limits_{\substack{\text{boundary}\\ \text{of gas}}} -\ pn_z d\ s = \frac{d}{dt}\iiint\limits_{R(t)} \rho\ v\ d\ \mathrm{vol} =$$

$$= \iiint\limits_{R(t)} \frac{\partial}{\partial t}(\rho\ v)\ d\ \mathrm{vol} + \iint\limits_{\substack{\text{boundary}\\ \text{of } R(t)}} \rho\ v\ (\vec{v}.\vec{n})\ ds.$$

Note that, due to the quasi-stationary character, we have

$$\iiint \frac{\partial}{\partial t} (\rho\ v)\ d\ \text{vol} = 0.$$

Working out this equation and differentiating it with respect to z_2 we get

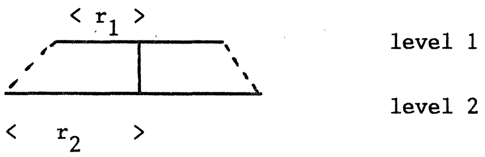
$$\frac{d}{dz}\ (r^2 P_{tot}) = r^2\ \rho\ g + 2r\frac{dr}{dz}\ (\rho_1 v_1{}^2 + p), \tag{5}$$

where $P_{tot} = p + \rho\ v^2$.

5. Energy - equation

The energy-equation is specified for turbulent fluid flow. For laminar flow we have a somewhat different boundary condition and the difference can be found in the energy equation.

We consider the energy in the gas only.



The energy equation reads as follows (in words):
the energy that flows per unit time through the lower surface equals the total energy that flows per unit time through the other surfaces plus the energy, supplied by forces.

The energy is built up of kinetic energy and heat. For the enthalpy we write $c_v\tau$. Note that a gap in energy seems to arise at the transition liquid ——> gas. There is no danger, however, if no comparisons are made on the basis of the enthalpy assumptions. Within each phase, the formulas are reasonably correct. The energy equation reads as follows:

$$\pi \frac{r_2^2 \rho_2 v_2^3}{2} + (\emptyset_o c_{vA} + \pi (r_2^2 - r_o^2) \rho_1 v_1 c_{vn}) \tau_2 =$$

$$= \pi \frac{r_1^2 \rho_1 v_1^3}{2} + (\emptyset_o c_{vA} + \pi (r_1^2 - r_o^2) \rho_1 v_1 c_{vn}) \tau_1$$

$$+ \int_{z_1}^{z_2} 2\pi r \frac{dr}{dz} \rho_1 \frac{v_1^3}{2} dz + \int_{z_1}^{z_2} 2\pi r \frac{dr}{dz} v_1 \rho_1 c_{vn} T dz$$

$$+ \int_{z_1}^{z_2} \beta . 2\pi r \sqrt{1 + (\frac{dr}{dz})^2} (T - \tau) dz + \int_{z_1}^{z_2} \pi r^2 \rho g v dz +$$

$$+ \iint_{\substack{\text{boundary}\\ \text{of gas}}} (-p\vec{n}, \vec{v}) ds,$$

where $\vec{n}$ is the outward unit normal.

These terms describe succesively

1. The kinetic energy, flowing per unit time through level 2
2. The heat - " " " " " " " 2
3. The kinetic " " " " " " " 1
4. The heat - " " " " " " " 1
5. The kinetic " " " " " " the free surface
6. The heat, carried along per unit time by the evaporating liquid.
7. The heat, used per unit time for warming up of the gas.
8. The energy, due to gravitation per unit time.
9. The energy, due to pressure per unit time.

The last term incorporates a troublesome detail: at the free boundary, the velocity is discontinuous (in our model). As a consequence, the surface integral cannot simply be transformed into a volume-integral. The energy equation looks somewhat friendlier when differentiated with respect to z_2:

$$\begin{aligned}
&\frac{d}{dz}\left(\frac{1}{2}\, r^2\rho\, v^3\right) + \left\{\frac{\emptyset_\circ}{\pi}\, c_{vA} + (r^2 - r_\circ^2)\, \rho_1 v_1 c_{vn}\right\} \frac{d\tau}{dz} + \\
&2r\, \frac{dz}{dz}\, \rho_1 v_1 c_{vn}\, (\tau - T) = \qquad\qquad (6)\\
= \quad & r\, \rho_1 v_1{}^3\, \frac{dr}{adz} + 2r\, \beta \sqrt{1 + \left(\frac{dr}{dz}\right)^2}\, (T - \tau) + r^2\rho\, g\, v \\
- \quad & r^2 p\, \frac{dv}{dz} - r^2 v\, \frac{dp}{dz} + 2\, pr\, \frac{dr}{dz}\, (v_1 - v)
\end{aligned}$$

(The last term represents the discontinuity)

The five equations (2), ..., (6) for the gas, together with the equation (1) for the liquid are sufficient to compute all unknowns:

$T(z)$, $r(z)$, $\tau(z)$, $p(z)$, $\rho(z)$ and $v(z)$.
Since at inflow all is known, we are left with an initial value problem for a system of ordinary differential equations. The computational results are within a 10% marge of the measurements.

Literature:

1. Bird/Stewart/Lightfoot - Transport Phenomena
   Wiley 1960.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Clustering and Nesting of Energy Spectra

P. van Mouche
Mathematisch Instituut Rijksuniversiteit Utrecht
P.O. Box 80010, 3508 TA Utrecht, The Netherlands

ABSTRACT: In some models of solid state physics, mysterious phenomena can be observed: The clustering and nesting of energy spectra. This happens in particular with models leading to the discrete Mathieu equation

$$g(n+1)+(2A\cos(2\pi n\alpha-\nu)-\varepsilon)g(n)+g(n-1) = 0 \quad (n \in \mathbb{Z}).$$

Hofstadter (1976) gave an empirical description of the combinatorics of both phenomena. In this paper we shall give a natural explanation of the combinatorics by introducing and applying the concept of infinitesimal clustering.

1. INTRODUCTION

Studying the quantum mechanical spectral problem of a Bloch electron in a magnetic field, Hofstadter (1976) obtained a miraculous picture; see figure 1a. It is a plot of the set spect$(1,\alpha)\subseteq\mathbb{R}$ as a function of α. For $A,\alpha\in\mathbb{R}$ spect(A,α) is defined by

$$\text{spect}(A,\alpha) = \bigcup_{\nu\in\mathbb{R}} \text{spec}(A,\alpha;\nu)$$

where

$$\text{spec}(A,\alpha;\nu) = \sigma(H_{A,\alpha,\nu}/l^2(\mathbb{Z}))$$

the spectrum of the discrete Mathieu operator

$$H_{A,\alpha,\nu}: \mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}}$$

$$(H_{A,\alpha,\nu}g)(n) = g(n+1)+(2A\cos(2\pi n\alpha-\nu))g(n)+g(n-1)$$

restricted to the Hilbert space $l^2(\mathbb{Z})$.

For $\alpha\in\mathbb{Q}$ the set spect(A,α) has a band structure (Hochstadt 1975;

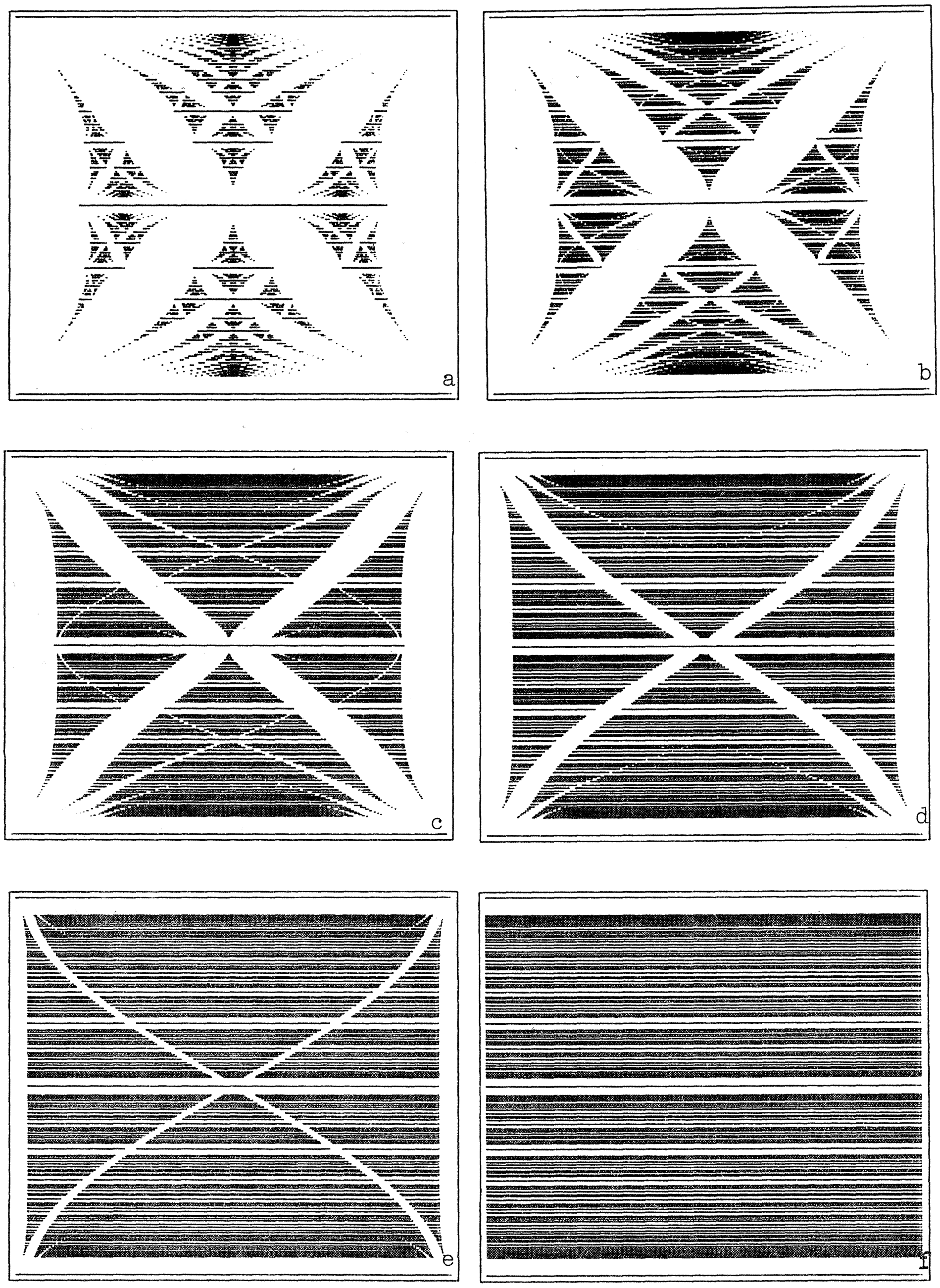


Figure 1: Hofstadter pictures. a A=1.0 b A=0.6 c A=0.3 d A=0.15 e A=0.06 f A=0.0. Abscissa: spect(A,α). Ordinate: 0≤α≤1.

Hofstadter 1976; van Moerbeke 1976), while for $A \neq 0$ and α irrational there is a Cantor-like structure (Simon 1982; Bellisard and Simon 1982; Bougerol and Lacroix 1985; Sokoloff 1985). Of course figure 1a consists only of the spect$(1,\alpha)$ for some selected rational values of α. The global picture exhibits a recursive structure (nesting) while at each height α the set spect$(1,\alpha)$ shows clustering. Such phenomena have geometrical and combinatorial aspects. Hofstadter gave empirical descriptions of the combinatorial aspects. We shall call these descriptions the nesting- and clustering hypotheses respectively. They can be represented as trees of numbers $H(\alpha)$ (§3). The main purpose of this paper is to give a natural explanation of the origin of these trees by introducing and applying the concept of infinitesimal clustering (§5). The idea of infinitesimal clustering occurred to us while watching "Hofstadter pictures" for values of A near 0, as shown in figures 1a-1f. These figures show "gap hierarchies" which somehow induce the combinatorial trees $H(\alpha)$.

The model of a Bloch electron in a magnetic field is but one source of "Hofstadter pictures" (Azbel 1964; Butler and Brown 1968; Hofstadter 1976; Claro and Wannier 1979; van Mouche 1983; Thouless 1984; Sokoloff 1985). Similar pictures can be derived from the modulated spring model (de Lange and Janssen 1983; Sokoloff 1985). In this case the underlying operator is the generalized discrete Mathieu operator. It would be interesting to see if the concept of infinitesimal clustering could also be used to explain the origin of these combinatorics in a natural way.

This paper sketches only the main lines. The technical details will be presented elsewhere.

2. THE DISCRETE MATHIEU OPERATOR

The discrete Mathieu operator $H_{A,\alpha,\nu}$, which derives its name from its resemblance to the differential equation of Mathieu (Magnus and Winkler 1979), is a special case of the second-order linear almost periodic recursion operator H_b: $\mathbb{C}^Z \to \mathbb{C}^Z$

$$(H_b g)(n) = g(n+1)+b_n g(n)+g(n-1) \quad (n\in\mathbb{Z})$$

where b: $\mathbb{Z}\to\mathbb{R}$ is an almost periodic function. For our purposes, we only need rational α's and thus periodic b's. Let us briefly review some elementary spectral theory for the periodic H_b, fixing terminology as we go along.

Firstly one has the following relations between the spectrum σ, the point spectrum σ_p and the continuous spectrum σ_c

$$\sigma(H_{b/l^2(\mathbb{Z})})=\sigma_c(H_{b/l^2(\mathbb{Z})})=\sigma_p(H_{b/l^\infty(\mathbb{Z})})=\sigma(H_{b/l^\infty(\mathbb{Z})}).$$

In particular there are no non-trivial localized solutions g for each $\varepsilon\in\mathbb{C}$ of the recursion relation

$$g(n+1)+(b_n-\varepsilon)g(n)+g(n-1) = 0 \ (n\in\mathbb{Z}) \qquad (b)_\varepsilon .$$

Moreover, we see that, with $\mathrm{spec}(b)=\sigma(H_{b/l^2(\mathbb{Z})})$,

$$\mathrm{spec}(b)=\{\varepsilon\in\mathbb{C}\,|\,(b)_\varepsilon \text{ has a non-trivial bounded solution}\}$$

and, because $H_{b/l^2(\mathbb{Z})}$ is self-adjoint, $\mathrm{spec}(b)\subseteq\mathbb{R}$.

In the following, $q\in\mathbb{N}$ is a fixed period of b. Using Floquet-theory and Bloch-wave analysis one can prove the following

1) Let
$$\Delta_q(\varepsilon) = \mathrm{Tr}\begin{pmatrix}\varepsilon-b_q & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}\varepsilon-b_{q-1} & -1\\ 1 & 0\end{pmatrix}\ldots\ldots\begin{pmatrix}\varepsilon-b_1 & -1\\ 1 & 0\end{pmatrix}$$
Then $\mathrm{spec}(b)=\{\varepsilon\in\mathbb{R}\,|\,-2\leqslant\Delta_q(\varepsilon)\leqslant 2\}$

2) The polynomial $\Delta_q-c : \mathbb{R}\to\mathbb{R}$ obeys the oscillation-theorem, i.e. for all $-2\leqslant c\leqslant 2$ the polynomial has q real zeros (counted with multiplicities); if $-2<c<2$ each zero is simple and if $c=-2$ or $c=2$ each zero has multiplicity $\leqslant 2$, the smallest and largest zero of $(\Delta_q-2)(\Delta_q+2)$ are simple. The typical form of Δ_q is as in figure 2.

3) The bands $E_1\leqslant E_2\leqslant\ldots\leqslant E_q$ and the gaps $G_1<G_2 \ldots <G_{q-1}$ of H_b (we write $A\leqslant B$ if $a\leqslant b$ for each $a\in A$ and $b\in B$, we write $A<B$ if $a<b$ for each $a\in A$ and $b\in B$) are the closures of the q connected components of $\Delta_q^{-1}(-2,2)$ and the q-1 closed intervals or touching points between these bands respectively. If a gap consist of one only point, we call the gap degenerate. To find the bands and gaps, one has to solve the algebraic equations $\Delta_q(\varepsilon)=2$ (giving the periodic spectrum)

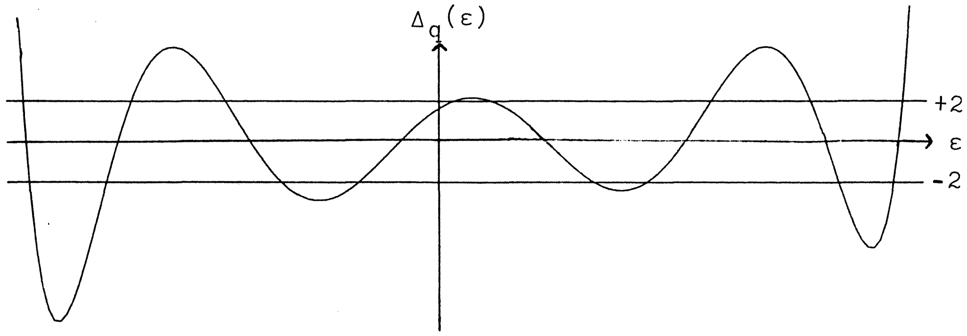


Figure 2: The typical form of Δ_q (here q=8).

and $\Delta_q(\varepsilon)=-2$ (giving the anti-periodic spectrum). Now we have (co means convex hull)

$$\mathrm{spec}(b)=E_1\cup E_2\cup\ldots\cup E_q$$
$$\mathrm{co}(\mathrm{spec}(b))= E_1\cup G_1\cup E_2\cup\ldots\cup G_{q-1}\cup E_q$$

Note that the definition of bands and gaps depends on the choice of the period q of b; spec(b), of course, does not depend on this choice. All these statements are similar to the corresponding ones for Hill's equation (Magnus and Winkler 1979).

Let us return to the discrete Mathieu operator. Hereafter we also write $\frac{p}{q}$ (with $p\in\mathbb{Z}$, $q\in\mathbb{N}$ relatively prime) instead of $\alpha\in\mathbb{Q}$, and take q as a period of the discrete Mathieu potential $b^{(\alpha,\nu)}=2\cos(2\pi n\alpha-\nu)$. One has the following simple relations

$$\mathrm{spect}(A,\alpha+k)=\mathrm{spect}(A,\alpha)\ (k\in\mathbb{Z}) \qquad (2.1)$$

$$\mathrm{spect}(A,\alpha)=\mathrm{spect}(A,-\alpha)=\mathrm{spect}(-A,\alpha)=-\mathrm{spect}(A,\alpha) \qquad (2.2)$$

and the deeper relations

$$\mathrm{spect}(A,\alpha)=\{\varepsilon\in\mathbb{R}\,|\,|\Delta_q(A,\alpha,\tfrac{\pi}{2q};\varepsilon)|\leqslant 2+2|A|^q\} \qquad (2.3)$$

$$G_{q/2}(A,\alpha;0)=\{0\}\quad \text{if } q=0\ (\mathrm{mod}4) \qquad (2.4)$$

$$G_{q/2}(A,\alpha;\tfrac{\pi}{q})=\{0\}\quad \text{if } q=2\ (\mathrm{mod}4) \qquad (2.5)$$

$$\mathrm{co}(\mathrm{spect}(A,\alpha))=\mathrm{spect}(A,\alpha)\sqcup\bigsqcup_{\substack{N=1\\N+q\ \text{even}}}^{q-1}\overset{\circ}{G}_N(A,\alpha;0)\sqcup\bigsqcup_{\substack{N=1\\N+q\ \text{odd}}}^{q-1}\overset{\circ}{G}_N(A,\alpha;\tfrac{\pi}{q})\ (A\geqslant 0) \qquad (2.6)$$

where ⊔ denotes disjoint union. It follows that

$$0\in\mathrm{spect}(A,\alpha) \qquad (2.7)$$

(2.3) is a consequence of the Butler-Brown phase-relation (Butler and Brown 1968 , Hofstadter 1976). A very important conjecture is the following version of the discrete Ince conjecture (Claro and Wannier 1978 ,Bellisard and Simon 1982)

$$\text{spect}(A,\alpha) \text{ consists of } \begin{matrix} q \\ q-1 \end{matrix} \text{ connected components if } q \begin{matrix} \text{odd} \\ \text{even} \end{matrix} \ (A\neq 0)$$

This conjecture is in fact a problem of coexistence type which has been solved completely by Ince (Magnus and Winkler 1979) for the differential equation of Mathieu.

Recently, considerable progress has been made in the spectral theory of almost-periodic Schrödinger operators. The discrete Mathieu operator has therein played an important rôle.

3. HOFSTADTER'S CLUSTERING AND NESTING HYPOTHESES

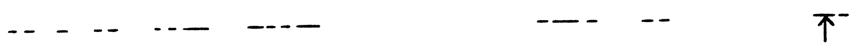
Before stating Hofstadter's hypotheses we need some definitions concerning clustering.

A band is a subset of $\mathbb{R}$ of the form $\{a<x<b \mid x\in\mathbb{R}\}$ with $a<b$. A clustering input is a non-empty finite collection U of bands which at most touch, that is

$$I_1, I_2 \in U \Rightarrow \mathring{I}_1 \cap \mathring{I}_2 = \emptyset$$

The closed intervals or touching points between the given bands of U are called gaps. So if $\#U=q$, then there are q bands and q-1 gaps. In case of a touching point we speak of a degenerate gap (cf. with §2). When the clustering input is pictured, the touching cannot be seen, so we will indicate touching by a ↑ below each touching point.

Consider a picture of a clustering input U. It may happen that the eye sees that the bands are organized into groups. Consider now such a group then it may happen that the eye sees again the same phenomenon, etc. One can in an obvious way present the combinatorics of the entire clustering process of U by a finite tree of positive integers. Denote this tree by CC(U). Here is an exemplaric example



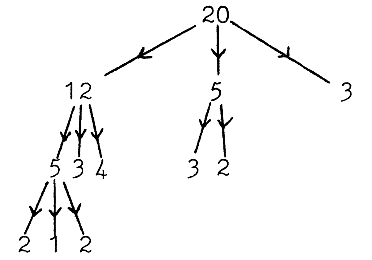


Figure 3: Representation of the combinatorics of the clustering process by a tree.

In this paper we shall not attempt to define clustering: the eye has to decide if a given clustering input clusters and in what ways. Beside reasonable trivial conditions which must be satisfied in order that a clustering input U cluster in $U_1, U_2, \dots, U_L$ (where $\{U_1, U_2, \dots, U_L\}$ is a partition of U) such as

$$\forall i,j\Big((\forall I\in U_i \ \forall J\in U_j \ I\leqslant J) \vee (\forall I\in U_i \ \forall J\in U_j \ I\geqslant J)\Big),$$

we shall require a reasonable but non-trivial one, namely

$$L\geqslant 2 \wedge \ \#U_i\geqslant 2 \text{ for at least two i's} \tag{3.1}$$

We will now specify precisely the clustering inputs $U(\alpha)$ ($\alpha\in\mathbb{Q}$) for the clustering hypothesis. An obvious first idea might be to simply take the collection of the connected components of spect$(1,\alpha)$. This however is not the right thought because then one possibly neglects "natural" degenerated spectral gaps. Remember that we have, by (2.6):

$$\mathrm{co}(\mathrm{spect}(1,\alpha))=\mathrm{spect}(1,\alpha)\sqcup\bigsqcup_{\substack{N=1\\ N+q \text{ even}}}^{q-1}\overset{\circ}{G}_N(1,\alpha;0)\sqcup\bigsqcup_{\substack{N=1\\ N+q \text{ odd}}}^{q-1}\overset{\circ}{G}_N(1,\alpha;\tfrac{\pi}{q}) \tag{3.2}$$

The gaps appearing on the right-hand side of (3.2) are the natural gaps for spect$(1,\alpha)$. (3.2) implies that the set

$$\mathrm{co}(\mathrm{spect}(1,\alpha))\Big\backslash\bigsqcup_{\substack{N=1\\ N+q \text{ even}}}^{q-1} G_N(1,\alpha;0)\sqcup\bigsqcup_{\substack{N=1\\ N+q \text{ odd}}}^{q-1}G_N(1,\alpha;\tfrac{\pi}{q})$$

consists of q connected components, and that the collection of closures of these components is a non-empty finite collection of bands. Take this collection as $U(\alpha)$. The gaps of $U(\alpha)$ are then the $G_N(1,\alpha;0)$ (N+q even) and the $G_N(1,\alpha;\frac{\pi}{q})$ (N+q odd). If the discrete Ince conjecture is true this somewhat technical definition can be restated in a more manageable form: If q is odd $U(\alpha)$ is the collection of the q connected components of spect$(1,\alpha)$. If q is even $U(\alpha)$ consists of the collection of all the connected components of spect$(1,\alpha)$ that do not contain 0 (there are q-2 of them) and the left and right section of the connected component which contains 0.

Because of (2.1) we may assume $0\leqslant\alpha<1$. Hofstadter's clustering hypothesis is now

> For all $0\leqslant\alpha<1$ the tree $CC(U(\alpha))$ equals the tree $DEN(H(\alpha))$, defined as follows

Let $[\ \ ]$ denote the entire function, and let the functions

$$\Lambda:\ (0,1)\to[0,1)\quad\text{and}\quad\Gamma:\ (0,1)\setminus\{\tfrac{1}{2}\}\to[0,1)$$

be defined by

$$\Lambda(x)=\frac{1}{x}-\left[\frac{1}{x}\right]\ \text{if}\ 0<x\leqslant\tfrac{1}{2}\ ,\ \Lambda(x)=\Lambda(1-x)\ \text{if}\ \tfrac{1}{2}<x<1$$

$$\Gamma(x)=\frac{1}{\frac{1}{x}-2}-\left[\frac{1}{\frac{1}{x}-2}\right]\ \text{if}\ 0<x<\tfrac{1}{2}\ ,\ \Gamma(x)=\Gamma(1-x)\ \text{if}\ \tfrac{1}{2}<x<1$$

Following Hofstadter we call the elements of $\{\frac{1}{n},\frac{n-1}{n}\mid n\geqslant1\}$ ’pure cases’ and the elements of $\{\frac{n}{2n+1},\frac{n+1}{2n+1}\mid n\geqslant2\}$ ’special cases’. Now define the tree $H(\alpha)$ of rational numbers as in figure 4. The formation of a branch stops as soon as we meet a pure case. Because Λ,Γ reduce the denominators of rational numbers, the tree $H(\alpha)$ is finite. By $DEN(H(\alpha))$ we mean now the tree of positive integers obtained from $H(\alpha)$ by replacing each rational number by its denominator.

Remarks: 1 Each point of the trees $H(\alpha)$ has out-degree 0 or 3; that is: we have uniform clustering. 2 The function Λ appears also in connection with continued fractions. 3 As far as I can check, the first one who noticed the clustering phenomenon was Azbel (Azbel 1964). He used a semi-classical approach to the spectral problem of a Bloch electron in a magnetic field. Hofstadter's clustering description looks like that of Azbel.

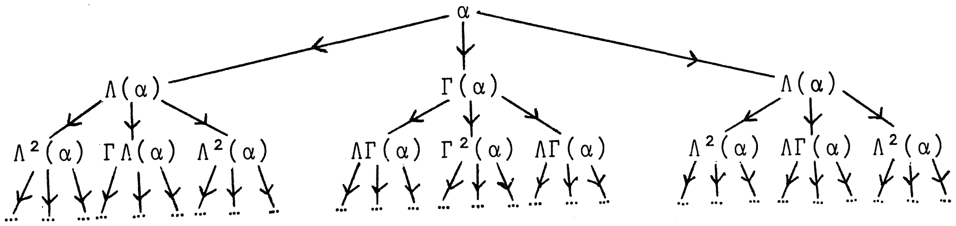
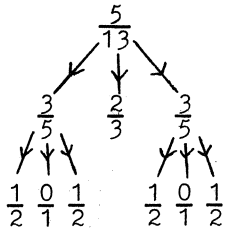
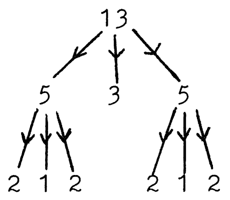


Figure 4: The tree H(α).

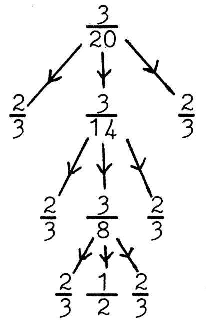
Example 1



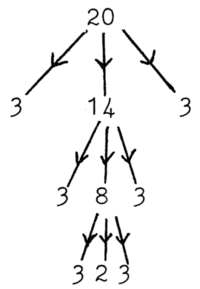
$H(\frac{5}{13})$



$DEN(H(\frac{5}{13}))$



$H(\frac{3}{20})$



$DEN(H(\frac{3}{20}))$

Figure 5 presents a selection of U(α)'s. The reader might want to check now the validity of the clustering-hypothesis for these cases.

Now we turn to the nesting-hypothesis. Consider figure 1a. We see a recursive structure, that is we see a motif (in this

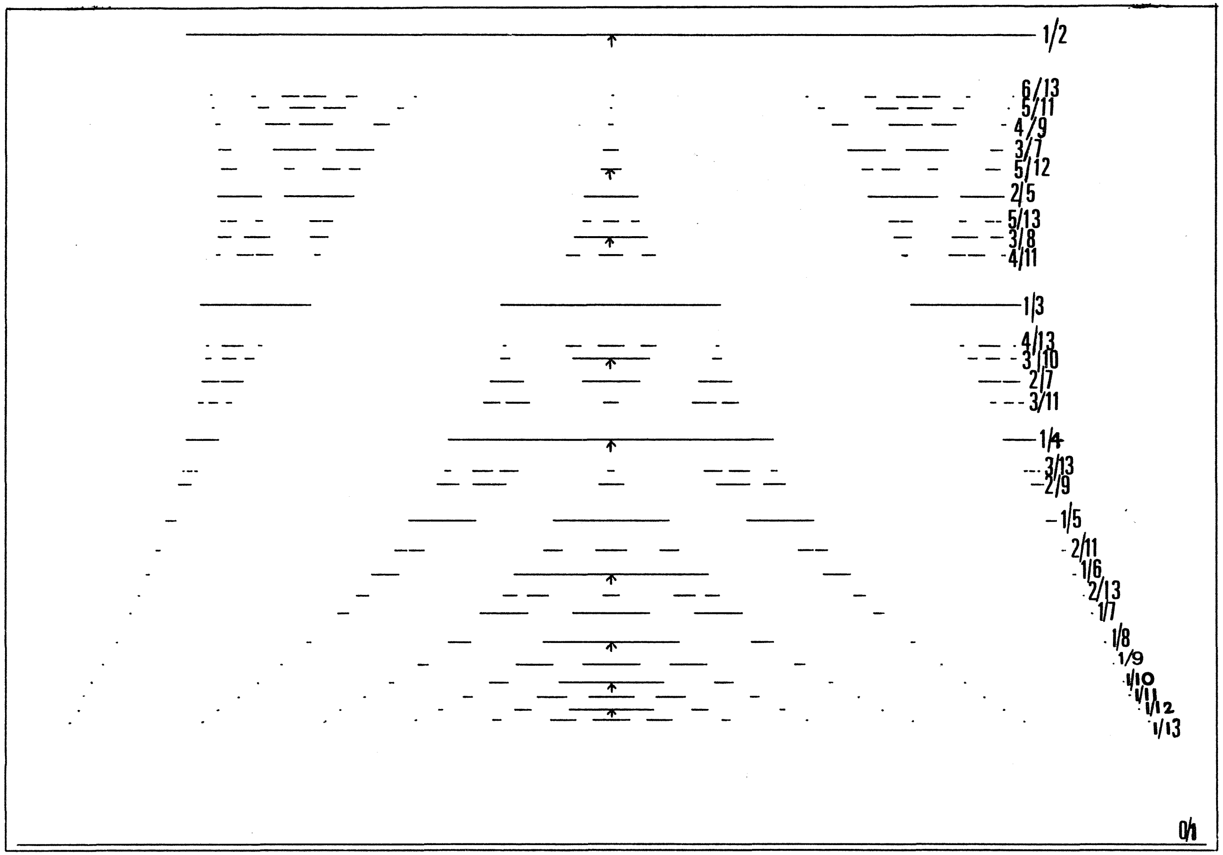


Figure 5: A magnification of a certain part of figure 1a.

case a butterfly) from which the picture is built up recursively. Hofstadter(1976) gives a description of how the building-up process proceeds (nesting process) from a basic skeleton (the butterfly), which essentially consists of the spect(1,α) for pure and special cases. Without going into details we notice that this process has a distorting aspect: At every step one has to compress the skeleton down to a certain small fraction of its size and then distort its vertical and horizontal scales before inserting it (in many places) in the picture obtained one step before. The combinatorics of the nesting process is described by the full trees H(α) in the following way. Consider a tree H(α), see figure 4. Then the meaning is that U(α) clusters into U_1,U_2,U_3, where U_1,U_2,U_3 is a distorted $U(\Lambda(\alpha)),U(\Gamma(\alpha)),U(\Lambda(\alpha))$ version respectively,etc. Again, the the reader may check this statement (nesting hypothesis) for some of the given U(α)'s in figure 5. The tree H(α)contains more information than the tree DEN(H(α)). This observation agrees with the fact that nesting is in general a mechanism leading to clustering.

In §5 we will give natural explanations of the origins of both hypotheses. Before doing this we must introduce "gap opening powers",which are needed for the concept of infinitesimal clustering.

4.GAP OPENING POWERS

Suppose we have a real analytic family of clustering inputs $U_A(A\in W)$, where W is an interval containing 0. This means

1) There is a $q\in\mathbb{N}$ such that $\#U_A=q$ for all $A\in W$.
2) All gaps of U_0 are degenerated.
3) There exist a 2q-tuple of real analytic functions $W\to\mathbb{R}$ that exactly describes the band boundaries of the bands of U_A as a function of A, the so-called band boundary functions

Because of the definition of bands and gaps we can number (with 1,2,...,q-1) and label (with e,o) the band boundary functions such that

$$\lambda_0^{e}(A)<\lambda_1^{\substack{e\\o}}(A)<\lambda_2^{\substack{e\\o}}(A)<\ldots\ldots<\lambda_{q-1}^{\substack{e\\o}}(A)<\lambda_q^{e}(A)$$

The numbering is uniquely determined, the only freedom lies in the labeling e,o for the numbers 1,2,...,q-1. We have

$$|\lambda_N^{o}(A)-\lambda_N^{e}(A)| \text{ is the length of gap } G_N(A)$$

An important case of such an family $U_A(A\in\mathbb{R})$ is given by the collection of bands $\{E_1(A),E_2(A),\ldots,E_q(A)\}$ of a q-periodic recursion operator H_{Ab}. Indeed, all gaps of H_0 are closed (see figure 6 and the intermezzo) and the real-analyticity is guaranteed by a deep theorem of Rellich (Baumgartel 1985).

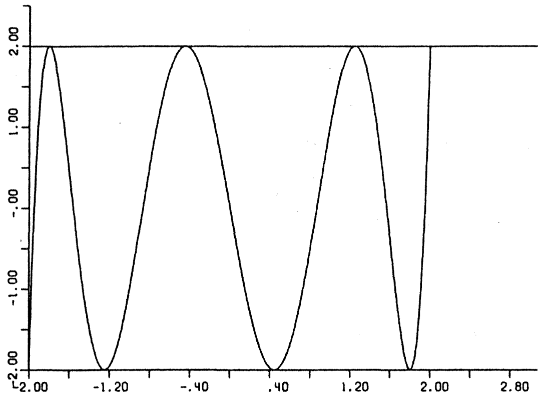


Figure 6: Δ_q for b=0 (here q=7).

Intermezzo:

For b=0 we have the equation

$$g(n+1)-\varepsilon g(n)+g(n-1) = 0 \quad (n\in\mathbb{Z})$$

Then every $q\in\mathbb{N}$ is a period of b and $\Delta_q(\varepsilon)=2T_q(\frac{\varepsilon}{2})$, where T_q is the q^{th} Chebyshev polynomial of the first kind. It follows that

- spec(0)= $[-2,2]$ (see also figure 1f)
- all gaps are closed
- the band boundaries are $-2\cos(\frac{N}{q}\pi)$ $(0\leqslant N\leqslant q)$.

We call the real analytic functions $\lambda_N^o - \lambda_N^e$ $(1 \leq N \leq q-1)$ the signed gap lengths. Denote their power series expansion about A=0 by

$$\sum_{j \geq 0} (\alpha^o_{N,j} - \alpha^e_{N,j}) A^j$$

Define the gap opening power o(N) of gap N by

$$o(N) = \inf\{j \mid j \in \mathbb{N}_0,\ \alpha^o_{N,j} \neq \alpha^e_{N,j}\}$$

Because for A=0 all gaps are closed we have with $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ that $o(N) \in \overline{\mathbb{N}}$. We call the vector

$$\vec{O} = (o(1), o(2), \ldots, o(q-1)) \in \overline{\mathbb{N}}^{q-1}$$

the gap opening power vector of U_A $(A \in W)$. Denote by POT(q) the collection of all q-periodic functions b: $\mathbb{Z} \to \mathbb{R}$. Then we have, via the H_{Ab} and U_A $(A \in \mathbb{R})$, the mapping

$$\vec{O}\colon \mathrm{POT}(q) \to \overline{\mathbb{N}}^{q-1}$$

A simple but important property is that, for every $b \in \mathrm{POT}(q)$, $\vec{O}(b)$ is symmetric: we call in general an $f \in \overline{\mathbb{N}}^M$ (with $M \in \mathbb{N}_0$) symmetric if $f(n) = f(M+1-n)$ for all $1 \leq n \leq M$.

It is no small feat to calculate $\vec{O}(b)$ for a given $b \in \mathrm{POT}(q)$. The way to do this is to use Rayleigh-Schrödinger perturbation theory (Baumgartel 1985). Because for the unperturbed situation A=0 there are (for q>1) roots with multiplicity 2 one needs degenerate perturbation theory which is highly nontrivial. Fortunately, if b is symmetric with respect to a $t \in \frac{1}{2}\mathbb{Z}$ one can split the periodic and anti-periodic eigenvalue problem, by restricting H_{Ab} to even and odd functions with respect to t, into non-degenerate problems, which are more easy to handle.

For generic b one has $\vec{O}(b) = (1,1,\ldots,1)$. So the occurrence of high gap opening powers is a non generic phenomenon. Very interesting gap opening powers appear for the discrete Mathieu potential $b^{(\alpha,\nu)}$. Calculation of its gap opening powers $o(N;\alpha\ ,\nu)$, using a discrete version of a procedure of Levy and Keller(1963), yields

$$o(N;\alpha,\nu+\frac{2\pi}{q}) = o(N;\alpha,\nu) = o(N;\alpha,-\nu) \quad (4.1)$$
$$o(N;\alpha,\nu) = \min(\tau_\alpha^{-1}(N),\ q-\tau_\alpha^{-1}(N)) \quad (N\neq\frac{q}{2}) \quad (4.2)$$
$$o(\frac{q}{2};\alpha,\nu)\geqslant\frac{q}{2},\ \text{even} \quad (4.3)$$
$$o(\frac{q}{2};\alpha,0) = \begin{matrix}\infty\\ q/2\end{matrix} \quad \text{if } q= \begin{matrix}0(\text{mod}4)\\ 2(\text{mod}4)\end{matrix} \quad (4.4)$$
$$o(\frac{q}{2};\alpha,\frac{\pi}{q}) = \begin{matrix}\infty\\ q/2\end{matrix} \quad \text{if } q = \begin{matrix}2(\text{mod}4)\\ 0(\text{mod}4)\end{matrix} \quad (4.5)$$
$$o(\frac{q}{2};\alpha,\nu) = q/2 \quad \text{if } 0<\nu<\pi/q \quad (4.6)$$

where τ_α is the permutation of $\{0,1,\dots,q-1\}$ defined by

$$\tau_\alpha(N) = (pN)(\text{mod}q)$$

Note that $o(N;\alpha,\nu)<\frac{q}{2}$ if $N<\frac{q}{2}$ and that for each $q\in\mathbb{N}$ the collection $\{\tau_\alpha | p\in\mathbb{Z},\ p,q \text{ relatively prime}\}$ is an Abelian group under the composition and that $o(N;\frac{p'p}{q},\nu) = o(\tau^{-1}_{\frac{p'}{q}}(N);\ \frac{p}{q},\nu)$.

(4.1)-(4.5) are extremely important. We will show in the next section that somehow all the information needed for a natural explanation of the origin of both hypotheses of Hofstadter is contained in these results. Let us here mention the following result confirming that the gap opening powers of the discrete Mathieu potential are extraordinarily high

For all $b'\in\text{POT}(q)$, $b^{(\alpha,\nu)}$:
$o(N;b') \geqslant o(N;\alpha,\nu)$ for all $1\leqslant N\leqslant q-1 \implies$
there are $c,d,\nu'\in\mathbb{R}$ such that $b'=cb^{(\alpha,\nu')}+d$.

In figure 7 some stability diagrams, i.e. the collection of bands as a function of A, are given in a neighbourhood of A=0.

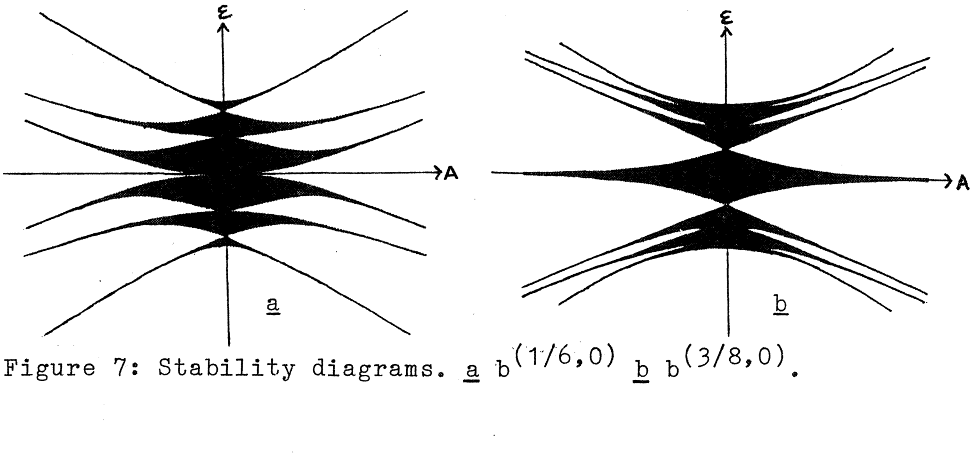


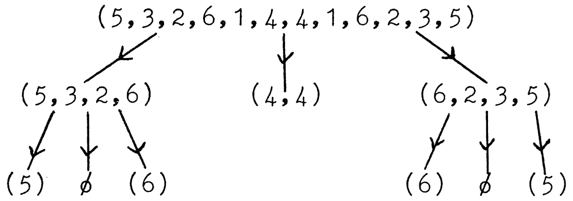
Figure 7: Stability diagrams. a $b^{(1/6,0)}$ b $b^{(3/8,0)}$.

5. A NATURAL EXPLANATION OF THE ORIGIN OF THE CLUSTERING- AND NESTING HYPOTHESIS

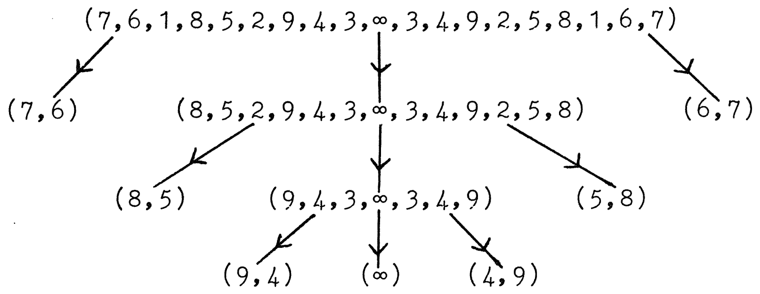
First some definitions.

1) In what follows we interpret a function $f\in\overline{\mathbb{N}}^M$ ($M\in\mathbb{N}_0$) as a row $(f(1),f(2),\dots,f(M))$; $\dim(f)=M$, the dimension of f. $\overline{\mathbb{N}}^0$ consists of the void row $\emptyset$.
2) Let $f\in\overline{\mathbb{N}}^M$ and $n\in\mathbb{N}$. If there are disjoint subsets R,T of $\{1,2,\dots,M\}$ with $R\cup T=\{1,2,\dots,M\}$ such that $f(r)<f(t)$ for all $r\in R$, $t\in T$ and $\#R=n$, then we call the set R the n smallest points of f.
3) Given $f\in\overline{\mathbb{N}}^M$ and $s\in\mathbb{N}$ with $s\geq 2$, we can construct a finite tree of rows, $IC_s(f)$, as follows: Let (if it exists) S be the set of s-1 smallest points of f. Removing the set S from the domain of f, we obtain s new rows (possibly some of whichmay be void). Repeat this process with these new rows,etc. We stop locally with a row, say g, if we meet one of the following obstructions
   1 $s>\dim(g)+1$ (dimension obstruction)
   2 g has not s-1 smallest points (s-section obstruction)
   3 the number of new rows with dimension ≥ 1 which g gives is less than 2 (clustering obstruction) (cf (3.1)).
   This process can be represented in an obvious way by a finite tree, $IC_s(f)$.

Example 2



$IC_3(5,3,2,6,1,4,4,1,6,2,3,5)$



$IC_3(7,6,1,8,5,2,9,4,3,\infty,3,4,9,2,5,8,1,6,7)$

Now we can introduce the concept of infinitesimal clustering. Suppose we have a real analytic family of clustering inputs $U_A (A \in W)$. Let $\vec{O}$ be the gap opening power vector of this family. Then we can construct, for each $s \geqslant 2$, the tree $IC_s(\vec{O})$. Now if, for some $A_0 \in W$ $(A_0 \neq 0)$ and $s \geqslant 2$ we have

$$CC(U_{A_0}) = \dim(IC_s(\vec{O})) + 1 \qquad (5.1)$$

where $\dim(IC_s(\vec{O})) + 1$ is the tree obtained from $IC_s(\vec{O})$ by first replacing each row by its dimension and then adding +1 everywhere, then it is reasonable to see this as a natural explanation for the origin of the clustering combinatorics of $U(A_0)$.

Remarks:

1 The +1 comes from the fact that an M-dimensional row of gap opening powers is associated with a clustering input consisting of M+1 bands.

2 $o(N_1) < o(N_2)$ implies that for A in a sufficiently small puntured neighbourhood of 0, the length of gap $G_{N_1}(A)$ is larger than the length of gap $G_{N_2}(A)$.

3 s specifies the choice of the infinitesimal clustering criterion.

4 In words (5.1) means: The infinitesimal clustering combinatorics born at A=0 survive until $A=A_0$.

We apply this infinitesimal clustering concept to $U(\alpha)$. Consider the identity (2.6). In the same way as we defined $U(\alpha)$, define the clustering inputs $U_A(\alpha)$ for $A\geqslant 0$. One can prove that this is a real analytic family of clustering inputs. Of course $U_1(\alpha)=U(\alpha)$. Using (4.2)-(4.5), we find for the gap opening powers of $U_A(\alpha)$ ($A\geqslant 0$), which we denote by $\vec{O}(\alpha)$, that

$$o(N;\alpha) = \min(\tau_\alpha^{-1}(N), q-\tau_\alpha^{-1}(N)) \qquad (1\leqslant N\leqslant q-1,\ N\neq\tfrac{q}{2})$$
$$o(\tfrac{q}{2};\alpha) = \infty$$

Note that $\vec{O}(\alpha)$ is symmetric, $\vec{O}(\alpha)=\vec{O}(1-\alpha)$ and that

$$o(N;\frac{p'p}{q}) = o(\tau^{-1}_{\frac{p'}{q}}(N);\frac{p}{q}).$$

Since the following theorem can be proved,

Theorem 1

For all $0\leqslant\alpha<1$: $\quad DEN(H(\alpha)) = \dim(IC_3(\vec{O}(\alpha))) + 1$

we have a natural explanation for the origin of the clustering hypothesis. Let us check this statement for some special cases in example 4.

Example 3

$\vec{O}(0/1) = \emptyset$
$\vec{O}(1/2) = (\infty)$
$\vec{O}(1/3) = (1,1)$
$\vec{O}(1/4) = (1,\infty,1)$
$\vec{O}(1/5) = (1,2,2,1)$
$\vec{O}(2/5) = (2,1,1,2)$
$\vec{O}(3/8) = (3,2,1,\infty,1,2,3)$
$\vec{O}(5/13) = (5,3,2,6,1,4,4,1,6,2,3,5)$
$\vec{O}(3/14) = (5,4,1,6,3,2,\infty,2,3,6,1,4,5)$
$\vec{O}(3/20) = (7,6,1,8,5,2,9,4,3,\infty,3,4,9,2,5,8,1,6,7)$

Example 4 (use examples 1,2 and 3 !)

$\dim(IC_3(\vec{O}(5/13))) + 1 = DEN(H(5/13))$
$\dim(IC_3(\vec{O}(3/20))) + 1 = DEN(H(3/20))$

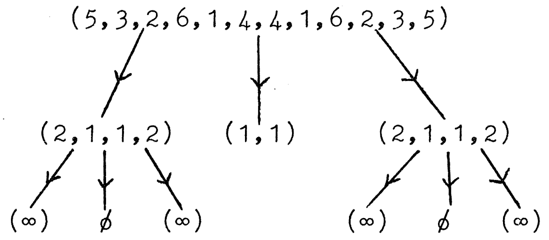
Next we turn to the nesting hypothesis. We ought to give a natural explanation for the origin of the full tree $H(\alpha)$. We shall succeed in doing something less, namely in explaining the tree min(.,1-.)H(α). This is the tree obtained from H(α) by replacing all rational numbers x in the tree H(α) by min(x,1-x). Because of spect(1,α)=spect(1,1-α), however,this is not a serious failure. The following observations will be useful:

1) The elements of the tree $IC_3(\vec{O}(\alpha))$ are not in general symmetric.
2) $\mathcal{O}: [0,\frac{1}{2}]\cap\mathbb{Q}\longrightarrow\bigcup_{M=0}^{\infty}\overline{\mathbb{N}}^M$ defined by $\alpha\mapsto\vec{O}(\alpha)$ is an injective mapping.

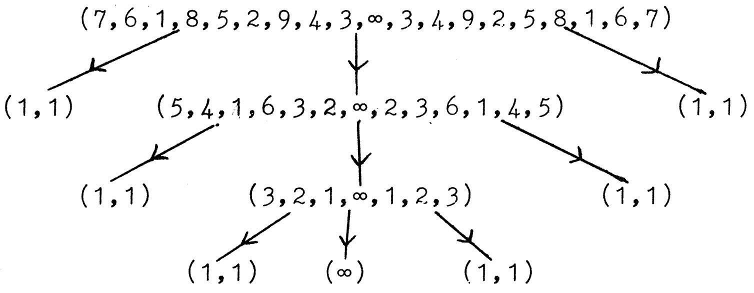
Note that the lack of symmetry in the trees $IC_3(\vec{O}(\alpha))$ agrees with the distorti on aspect mentioned in §3. We will repair this lack of symmetry as follows.

Take a row f of the tree $IC_3(\vec{O}(\alpha))$. Take the two smallest points of f and replace their corresponding coefficients by 1. Then take the next two smallest points of f and replace their corresponding coefficients by 2, etc. In the final replacement we will meet a single point if dim(f) is odd, in which case we replace its corresponding coefficient by ∞. Denote the resulting row by Sym(f). It can be proved that this process is well defined.

Example 5 (use examples 2 and 3 !)



$Sym(IC_3(\vec{O}(5/13)))$



$\mathrm{Sym}(\mathrm{IC}_3(\vec{O}(3/20)))$

One can prove that the image of the injective mapping $\mathcal{O}$ contains the rows occurring in the trees $\mathrm{Sym}(\mathrm{IC}_3(\vec{O}(\alpha)))$. Therefore we can define for each α the tree of rational numbers (in $[0,\frac{1}{2}]$) $\mathcal{Q}^{-1}(\mathrm{Sym}(\mathrm{IC}_3(\vec{O}(\alpha))))$. Finally one can prove that

Theorem 2

For all $0 \leqslant \alpha < 1$: $\min(.,1-.)H(\alpha) = \mathcal{Q}^{-1}(\mathrm{Sym}(\mathrm{IC}_3(\vec{O}(\alpha))))$.

which can be considered as a natural explanation for the origin of the nesting hypothesis. Note that theorem 2 implies theorem 1. The reader is invited to check the validity of this theorem for some α.

ACKNOWLEDGMENTS

I would like to thank Prof. Duistermaat for the concept of infinitesimal clustering, and Dr. Stienstra for useful comments on the presentation.

REFERENCES

1) M.Azbel, 1964, Energy spectrum of a conduction electron in a magnetic field, Sov.Phys.Jetp., 19, 634.

2) H.Baumgartel, 1985, Operator theory and its applications, OT 15.

3) J.Bellisard and B.Simon, 1982, Cantor spectrum for the almost Mathieu potential, J.of Funct.Analysis, 48, 408.

4) P.Bougerol and J.Lacroix, 1985, Products of random matrices with applications to Schrödinger operators, Birkhauser vol8.

5) F.Butler and E.Brown, 1968, Model calculations of magnetic band structure, Phys.Rev., 166, 630.

6) F.Claro and G.Wannier, 1978, Closure of bands for Bloch electrons in a magnetic field, Phys.Stat.Sol.(b), 88, K147.

7) F.Claro and G.Wannier, 1979, Magnetic subband structure of electrons in hexagonal lattices, Phys.Rev.B, 19, 12, 6068.

8) H.Hochstadt, 1975, On the theory of Hill's matrices and related inverse problems, Lin.Alg&Appl., 11, 41.

9) D.Hofstadter, 1976, The energy-levels of Bloch electrons in rational and irrational magnetic fields, Phys.Rev.B, 14, 2239.

10) C.de Lange, 1983 Phonons and electrons in modulated crystals, Thesis, Katholieke Universiteit Nijmegen.

11) D.Levy and J.Keller, 1963, Instability intervals of Hill's equation, Commun.on Pure and Appl.Math., 16, 469.

12) W.Magnus and S.Winkler, 1979, Hill's equation, Dover Publ.

13) P.van Moerbeke, 1976, The spectrum of Jacobi-matrices, Inventiones Math., 37, 45.

14) P.van Mouche, 1983, Sur un électron de Bloch dans un champ magnétique et sur l'équation presque-Mathieu, Essay, Katholieke Universiteit Nijmegen.

15) B.Simon, 1982, Almost periodic Schrödinger operators: a review, Adv.in Appl.Math., 3, 4, 463.

16) J.Sokoloff, 1985, Unusual band structure, wave functions and electrical conductance in crystals with incommensurate periodic potentials, Phys.Reports, 126, 189.

17) D.Thouless, 1984, Quantized Hall-effect in two-dimensional periodic potentials, Phys.Reports, 110, 279.

Proceedings ICIAM 87, Paris-La Villette, June 29-July 3 1987

Tools for the Development and Usage of Industrial Mathematical Software

F.J. Heerema, W. Loeve and J.J.P. van Hulzen
National Aerospace Laboratory NLR
P.O. Box 90502, 1006 BM Amsterdam, The Netherlands

ABSTRACT

Cost effectiveness of many applied scientific research activities can be improved considerably when they are directed towards the development of digital simulation methods. The kernel of such methods is a mathematical model that describes the aspects that have to be simulated. Valuable simulation methods are those that can be used for analysis processes in the engineering phase of technical products. As a consequence the developed so-called mathematical software can serve as a means to transfer knowledge from institutes for applied scientific research to industry. However, in that case it is required that the software is integrated in systems for computer-aided engineering.

To facilitate this integration, industry and institutes need an adequate infrastructure for information processing, comprising hardware and software components to be used in various disciplines.
The requirements imposed on such an infrastructure are:

- sufficient computer power shall be available to guarantee the fast development and the usage of mathematical software of increasing complexity due to increasing competition on the market of the resulting industrial products;
- it shall be possible to treat the information generated at different locations in the organization as one single source of information.

As a basis of the infrastructure mentioned above the National Aerospace Laboratory NLR has developed a computer and terminal network with a

general purpose central mainframe that serves mainly database applications and a supercomputer for fast computations. The network serves both geographically separated parts of NLR and it allows access from virtually any location at NLR and from outside to software and information from experiments, and digital simulations.

The kernel of the software infrastructure that is implemented on the NLR network consists of an engineering data management system with 4th generation characteristics. This system EDIPAS that has been developed by NLR, enables engineers to store and to use data according to various engineering views and to facilitate data exchange between various computational processes, and between experimental and theoretical investigations.

As a means for user interaction a standardized command language system COLAS has been developed by NLR. Use of this system results in uniform interfaces for a wide variety of applications.

Finally, in analogy with databases a methodbase system MEBAS is being developed which supports the management, assemblage and use of software components. Essential in the methodbase is the administrative information concerning the function, interfaces, limitations, etc. of the methods.

It has become clear that the described infrastructure, comprising integrated hardware and software components, results in higher efficiency in research and in support of engineering activities of industry.

1. INTRODUCTION

In many industrial organizations the need increases to improve the engineering process. This is due to quicker changes in specific customer requirements and to more severe requirements with respect to the price-performance ratio of the products. As a consequence of competition on the national and international market, the time for innovation and design of the product and the time for the production process itself is to be decreased. Application of computers is a necessaty for simulations that are used to analyse the characteristics of products or product parts before they will be realized.

The software to perform the simulations is based on applied mathematics to a great extent and is developed in industry and in institutes for applied scientific research. In order to be applicable in industry this mathematical simulation software has to be integrated in Computer-Aided Engineering (CAE) systems. Such a CAE-system has to be built in such a way, that the various disciplines involved in the development of a product are assisted with state-of-the-art simulation tools and are able to exchange information. To enable this, the CAE-system comprises tools for the generation of output of mathematical software, for the analysis of the output, and for the treatment of all information that is relevant for the concerning design and production process as one single source of information.

The National Aerospace Laboratory NLR is the central institute in the Netherlands for aerospace research. Its principle mission is to render scientific support and technical assistance under contract to Dutch and foreign aerospace industries and organizations, civil and military aircraft operators, and governmental agencies concerned with aviation and spaceflight. In principle the activities of institutes for applied scientific research, such as NLR, consist of supply of information and not the actual design or operation of systems like aircraft and spacecraft. Experience at NLR learns that the effectivity of applied theoretical and experimental scientific research can be improved if it is focussed on development or improvement of computer based information systems. This is especially the case if the results of the research are to be used by contractors. As such, information systems support technology and information transfer from NLR to industry and operators of high tech industrial products.

Although many of the information systems developed by NLR for industry are transferred to other organizations separately, it is considered essential that they are developed in such a way that they fit into the industrial infrastructure for multi-disciplinary design and analysis. As a result, NLR decided to develop an infrastructure, that reflects the infrastructure for information processing needed in industry. This infrastructure serves primarily as a test bed for the development of computer based information systems.

Moreover access had to be made possible by NLR to recent simulation tools based on mathematical software to potential users outside the organization, for familiarization and validation before the decision is made to transfer the software to the own organization.

The requirements that are the result of the above mentioned considerations are presented in the next chapters. The technical characteristics of the standard components of the infrastructure of NLR that assist in the development and usage of the mathematical software are outlined in the paper.

2 THE DEVELOPMENT OF SIMULATION METHODS

The nucleus of any digital simulation method is a mathematical description of the aspects of the real world that have to be simulated. As a result of limited knowledge of the aspects or in view of required simplicity of the mathematical description, the resulting mathematical model contains modeling errors.

Discretization of a mathematical model in general is required when a digital computer has to be used for the simulation according to the model. The resulting numerical model contains discretization errors. To minimize these errors, in many cases it is required to derive the discretization method from the nature of the aspects that have to be simulated. So it is well known in computational fluid dynamics that discretization schemes have to be based on physical considerations, such as the need to guarantee conservation of mass, momentum, and energy.

Finally, a solution strategy has to be defined for the equations that are the result of the discretization. Many sources of errors exist in the solution techniques. In principal accuracy and computational efficiency have to be balanced. Many of the concerning aspects are discussed in Boerstoel et al., 1986, for computational fluid mechanics.

For analysis of the results of digital simulation with respect to errors, the best strategy is to compare the results with information from other sources. These can be more accurate simulation methods. However, in applied technical research the purpose much often is to develop a better method than available at that time. The only way to obtain reference material in that case is to make use of physical experiments. In physical experiments there are sources of errors as well as in digital simulation methods. Measuring errors exist in all cases. When use is made in the experiments of scale models, which very often is the case in aircraft and ship design, scaling errors are present too. The various sources of errors mentioned so far are summarized in Fig. 1.

Validation of simulation methods requires the capability to define and execute special experiments for the many different errors that will be present in digital simulation methods. Physical experiments require special facilities. Digital simulation experiments with more accurate models,

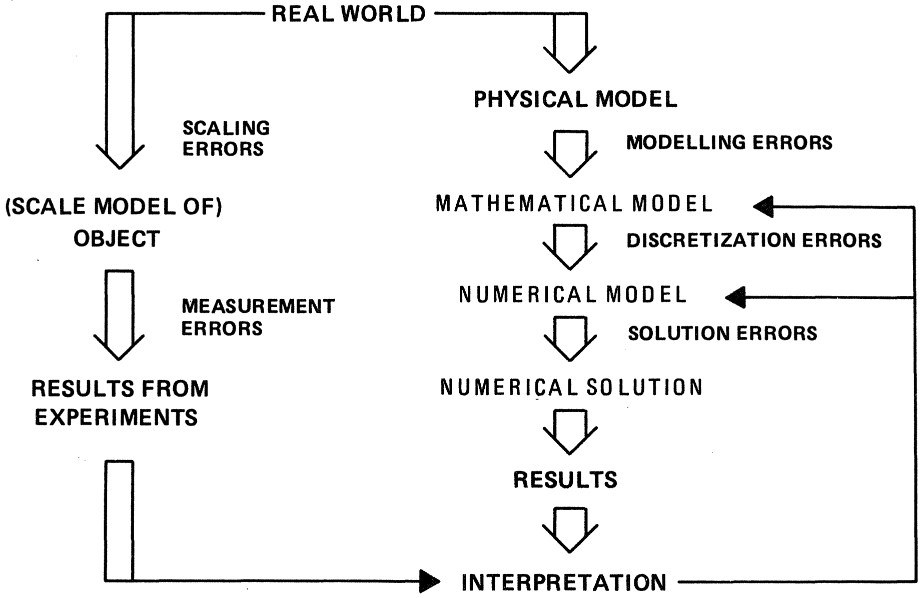


Fig. 1 Development of a mathematical simulation model

finer discretization or better solution methods require as facility in principle only more powerful computers than are necessary for the more simplified methods under development.

Apart from the facilities for the development of advanced digital simulation methods and the validation of these methods, various disciplines are required. As far as the digital simulation method itself is concerned, it is necessary that extensive knowledge is available in the fields of mathematical modeling, numerical analysis, and software engineering. For physical experimentalizing, experimenters are required who can define instrumentation and who are able to execute experiments. Experimenters as well as mathematicians must be enabled to apply the digital simulation methods for interpretation of the results from the physical experiments and from the digital simulation experiments. This means that already during the development of the digital simulation methods it must be possible for representatives of various disciplines to apply available parts of the methods.

The considerations mentioned above lead to the conclusion that digital simulation methods have to be part of an information system from which the required output can be obtained and for which the related specialist is able to generate the proper input. For the concerning organization it is advisable to organize the processing of information from physical experiments and from digital simulation in such a way that the results can be considered as one single source of information for interpretation purposes. In many cases this affects different disciplines that are housed in different departments of the organization.

It is not only the handling of information that requires multi-disciplinary co-ordination in the organizations that develop advanced simulation methods. Also the computers that are required, very often are too expensive to be cost effective when only available for one of the disciplines in the organization. Integral solutions for the various processing requirements have to be realized for the organization as a whole as part of the infrastructure of the organization. This requires the consciousness that local interests have to be made subordinate to the interests of the organization as a whole, and adequate centralized management.

3 REQUIREMENTS FOR TOOLS IN THE INFRASTRUCTURE FOR INFORMATION PROCESSING

The infrastructure for information processing consists of hardware, software and information. Essential for the creation and operation of such an infrastructure, are specialists in the various branches of computer sciences, mathematics and in the various engineering disciplines.

For purposes of overall efficiency of the organization, in the development of the infrastructure in a CAE environment a number of technical/organizational requirements must be fulfilled. The most important of these are:

- the infrastructure shall be developed in such a way that incremental growth is possible;
- re-use of software in successive phases of incremental development and in a large range of disciplines shall be made possible;
- standards shall be applied to software development to guarantee consistency in information and in engineering methods used in different disciplines. Maintaining standards also avoids culture shocks to the users, when confronted with the next generation of elements in the infrastructure and results of applications. Moreover, when imposing standards, the transfer of knowledge is facilitated between various user groups.
- the infrastructure shall be accessible from virtually any location in the organization;

- the infrastructure and the organizational conditions shall make it possible that time to start up new applications is minimal.

From a data management point of view, in the product development process various sources of information can be recognized: designers, experiments and digital simulations. For evaluation of data from various sources the users need means for "post processing", such as the determination of differences between data from different sources or of isobar patterns on a wing from results of flow calculations or wind tunnel experiments. In the product development processes it is very important to facilitate access to relevant data such as the parameters that define the current geometry and the physical quantities that are related to that geometry. The data not only have to be made available to representatives of the discipline that has generated the data. In aircraft development, for example, structural analysis requires aerodynamic loadings. Also for analysis of the dynamic behaviour of an aircraft tail in flight, structural stiffness characteristics and dynamic aerodynamic loads are needed by the flutter annalist. Based on these observations the requirements for data management are:

- the complete collection of information related to a design process, functionally shall be treated as one single source of information;
- it must be possible for any engineer who is involved in the design process to store and retrieve data in a structured way, related to the specific aspects of his application. As a consequence of changing views on the data during the design process, re-structuring of the data for the various engineering views must be possible in a short time without laborous re-design of the database.
- facilities must be available for maintenance and quality assurance of the information to end users of the CAE infrastructure. End users of various engineering disciplines shall be able to operate the data management tools.

Industrial competition gives rise to an ever increasing demand for new methods for processing of data, and for analyses and simulation purposes. Amongst others mathematical software for analysis of design characteristics has to be improved continuously. For example the accuracy of flow simulations for determination of aerodynamic characteristics of aircraft are improved continuously. Older methods have to be kept available, because in general they require less processing power so they are more suitable for parameter studies in the preliminary design phase. Furthermore, in a number of cases results of simple methods are used as starting conditions for more complicated methods. The reason is, that convergence of complicated and expensive iterative computational procedures can be improved considerably in that way. The continuous development of methods leads to the following requirement:

- the software infrastructure for series of simulation methods with increasing degree of complexity shall be structured in such a way that the user can switch simply between the methods. As a consequence, tools have to be made available for the management of the various computational methods.

In order to avoid an ever changing type of interaction with the end user of simulation methods:

- standards have to be imposed on the user interface. At least the style of interaction has to be the same.

The above mentioned requirements mean that the infrastructure for CAE comprises facilities for the management of data, for the management of methods, and for user interaction. These facilities have to be implemented in such a way that access is possible from any point in the organization.

The software components mentioned so far have to be implemented on computers. In aerospace research, methods for computational fluid mechanics and computational solid mechanics require large processing power used for batch processing type of work. For the interactive applications and batch processing together also through-put requirements have to be fulfilled. In technical applications such as CAE a continuous growth of required processing performance is apparent. The growth rate partly depends on development speed of new application software and on the dispersion rate of computer applications in the concerning organization.

The various components of the infrastructure for development and usage of industrial mathematical software available at NLR are discussed in the next sections into some more detail.

4 DATA MANAGEMENT

With as main reason the quicker realization of technical applications, a commercially available database management system (DBMS) was introduced in the mid seventies. Utilization of this DBMS and using the same type of abstraction as in data management technology, led to the development of an application software system, that meets the requirements mentioned. In 1983 an operational version of this system came available, called EDIPAS (Engineering Data Integrated Processing and Analysis System). EDIPAS comprises facilities for engineering data management and for engineering data evaluation. The data management functions enable the creation and maintenance of a database, with data from computersimulations and experiments. Computer programs can easily be interfaced to the database in order to exchange data and to supply data for data evaluation purposes. A network of computer programs for engineering purposes can be built, resembling the architecture presented in Fig. 2, given as an example in the area of robotics (Groothuizen et al., 1986).

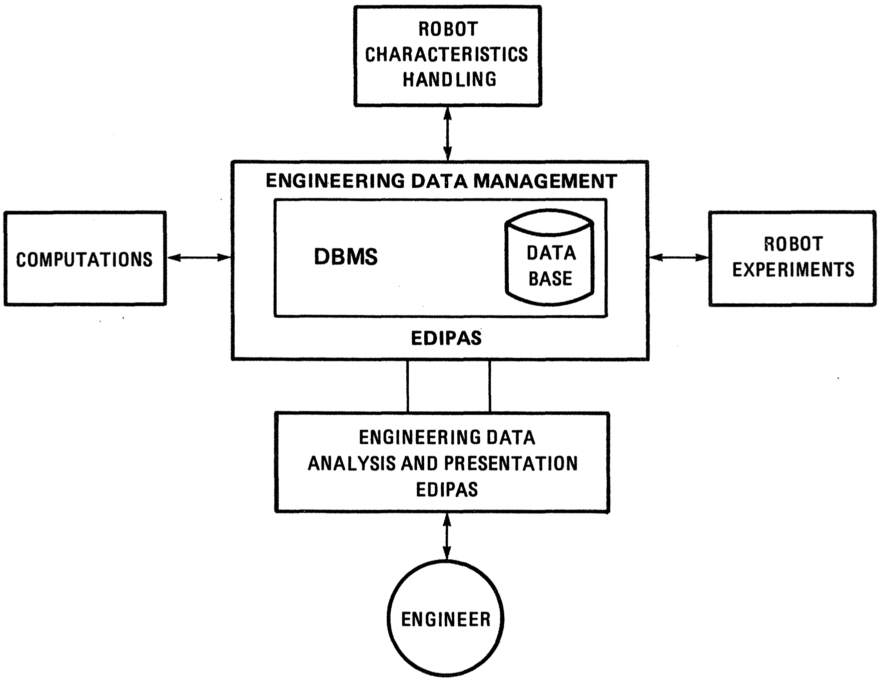

ROBOT CHARACTERISTICS HANDLING
ENGINEERING DATA MANAGEMENT
COMPUTATIONS
DBMS
DATA BASE
ROBOT EXPERIMENTS
EDIPAS
ENGINEERING DATA ANALYSIS AND PRESENTATION EDIPAS
ENGINEER

Fig. 2 Architecture of system for the analysis of robot dynamic properties

In order to meet requirements with respect to flexibility of application and ease of use, EDIPAS employs a data model with characteristics that are easy to understand for engineers using computers. In contrast to specific application of a general purpose DBMS the user is allowed to work with the terminology as applied in his own discipline. In fact the data management facilities of EDIPAS provide a layer, that translates the instructions of the user to data manipulation commands for the general purpose DBMS that is used as data management kernel. The data model as applied in EDIPAS (Kreijkamp et al., 1985), allows the user to manipulate entities for basic engineering data, for derived data, and for data items that support system operation. The basis of the data model is an entity called "data block". A data block is identified by user defined characteristics and contains user defined (names, types, number and length) scalars and matrices (like Fortran arrays). The identifying characteristics of data blocks are used to define relations between data blocks: the structures.

Within one database several structures can exist at the same time. In order to cope with the dynamic character of the view of engineers on the database contents, it is possible during the operational phase of a database to modify or to add to the identification of existing data blocks and to define new structures. The kernel DBMS guarantees the internal consistency of the database, when such operations are performed.

In large project groups the function of database administration is required, in analogy with the situation in business applications. This function will be performed by the prime user or by one of the most experienced users. The tasks of the project database administrator are to advise and to assist the project team, to initialize and to maintain the project database, and to grant access to the database for users within the project team. The facilities required to support the project database administrator are provided by EDIPAS (Steenbergen et al., 1985).

The data analyses facilities of EDIPAS provide a standard set of functions, that are common to various engineering activities. The facilities comprise functions to interrogate the project database, to perform arithmetic operations on and with selected data, to perform curve fitting and interpolation, and to present results in graphical or tabular form. The presentation form may be in a quick and mainly system defined layout, or in a layout that is completely defined by the user, with all descriptive information added, so the graphs can be communicated to others.

The analysis functions are command driven and can be executed interactively or in batch. Help information for interactive usage is available. An important feature is the possibility to create and maintain procedures of commands. These procedures may be build during execution, and can be used later on with an actual parameter setting. This procedure facility is used to build complete jobs that perform on a routine basis the required data evaluation and presentation of results. In this way the system acts as a 4th generation language for engineers, with which ad hoc queries can turn into routine operations.

Except for these standard analysis tools, the various engineering disciplines need more specific and specialized tools. Such tools can be easily interfaced to the database by means of a set of routines in the data management facilities. These Fortran callable routines give access to the database at the level of data entity and the data items contained therein.

For introductory purposes, it appeared essential that an EDIPAS workshop was developed, where potential users get insight in the concepts of engineering data management and analysis, and get their first hands-on experience. This workshop had to be composed and given by experienced users and members of the development team. Experience has learned, that engineers with some support are able to operate within one week the basics of the data analysis functions on a database that is fed from one data source. Once they have learned to operate the system and are acquainted to the level of abstraction, they find more engineering problems that can be solved by applying EDIPAS than ever was foreseen. The effectiveness of such engineers increases considerably, where at the same time they consume more computer resources. The impact of this fact has to be dealt with in a sound plan for the further development of the computer hardware and data-communications layers in the infrastructure.

5 METHOD MANAGEMENT

The infrastructure for CAE support at NLR focusses on research type of applications, that are continuously evolving and more and more are to support multi-disciplinary co-operation. Separately developed application subsystems and algorithms are gradually integrated into larger information systems.

The evolution in user requirements and in analysis techniques results in an information system development, in which several versions of algorithms are tested in different combinations, in order to select a combination with favourable characteristics. This selection is supported by comparison of results with experimental data.

The results of algorithmic research are made available as information systems, ready for use by the customer. Therefore development of algorithms is only a part of total system development, be it the driving part.

More and more attention is paid to data management and organizational aspects and to the production of information systems for operational environments.

In information system development the need arises for re-use of application components within and across projects, and for easy exchange of algorithms in information systems. Within projects, this need leads to standardization of functional and technical component interfaces. Across projects, the same type of standards are needed as well, as support for management and selection of software components. However, existing object and source library mechanisms do not store the formalized information on the functionality of the software components needed for this support. Therefore the method base system MEBAS is being developed as part of the infrastructure for CAE.

Strong analogies exist between database and methodbase systems, the most important being that they derive their right to exist from the integrity and access control on data and methods respectively. They store not only the data or methods on their own, but also information on the functionality and on technical aspects (place of storage, ways of access and usage).

The methodbase system addresses three types of users, not necessarily impersoned by different persons. The user types are:

- the application programmer, who adds new methods to the methodbase;
- the system designer, who assembles methods creating stand alone application programs;
- the end user, who uses stand alone programs to solve his specific problem. An important aspect is the ability for the end user to use his own language.

As one person can act in different user types, the user interaction of the methodbase system should provide a consistent users view for all user types.

A methodbase system shall support the following functions:

- addition, replacement and deletion of methods, including the associated information. It should be possible to define a structure of the methods for a specific application, before actually adding method sources. Method exchange across projects requires operations for copying methods from other method bases or even merging complete methodbases.
- selection of methods by users, based on keywords and attributes. The structure of the methodbase is also important to support the selection

process.

- assemblage of methods to new methods and finally to application programs for end-users. Implementation details should be handled by the methodbase system (e.g. calling sequences, type conversions). Method assemblage is the prime support function of the methodbase system. Method assemblage should support
  - . the need to define an assembled method in which certain aspects can be left open e.g. the specific algorithm to be used for a defined purpose and storage place for inputs and results. In this case, these have to be filled in by the end-user before execution.
  - . incorporation of database access methods, which leads to the specific requirement to check the integrity of using the database, e.g. to check that results of a computation are stored with identifications consistent with those of the inputs. Moreover means to ensure the correct state of the database (open or closed) are needed.
  - . incorporation of user interaction methods. Attention should be paid to standard ways of data entry by users, allowing short specifications, and to the requirement that certain programs have to be executed in batch as well as interactively. This might lead to two different versions generated for each program or to a systematic subdivision of programs in interactive and batch sections.

An important requirement for methods assemblage is that versions of assembled methods can be identified, with a precise specification of the methods used in a particular version. It shall always be possible to trace the program version that produced certain results.

- checking of consistency of assembled methods at construction or at runtime. Aside from the checks mentioned for database access methods, checks on dataflow and pre- and postconditions on methods can be performed. These checks cannot be exhaustive, but contribute to faster development of correct programs.
- execution of methods, controlled in a language with the end-users terminology. Specific support for monitoring execution and collection of execution information (again in end-users terminology!) is asked for.

The technical concept of the methodbase system is depicted in Fig. 3. The different user types will access the components as follows:

- the application programmer will access the methodbase via the method manager to store new methods;

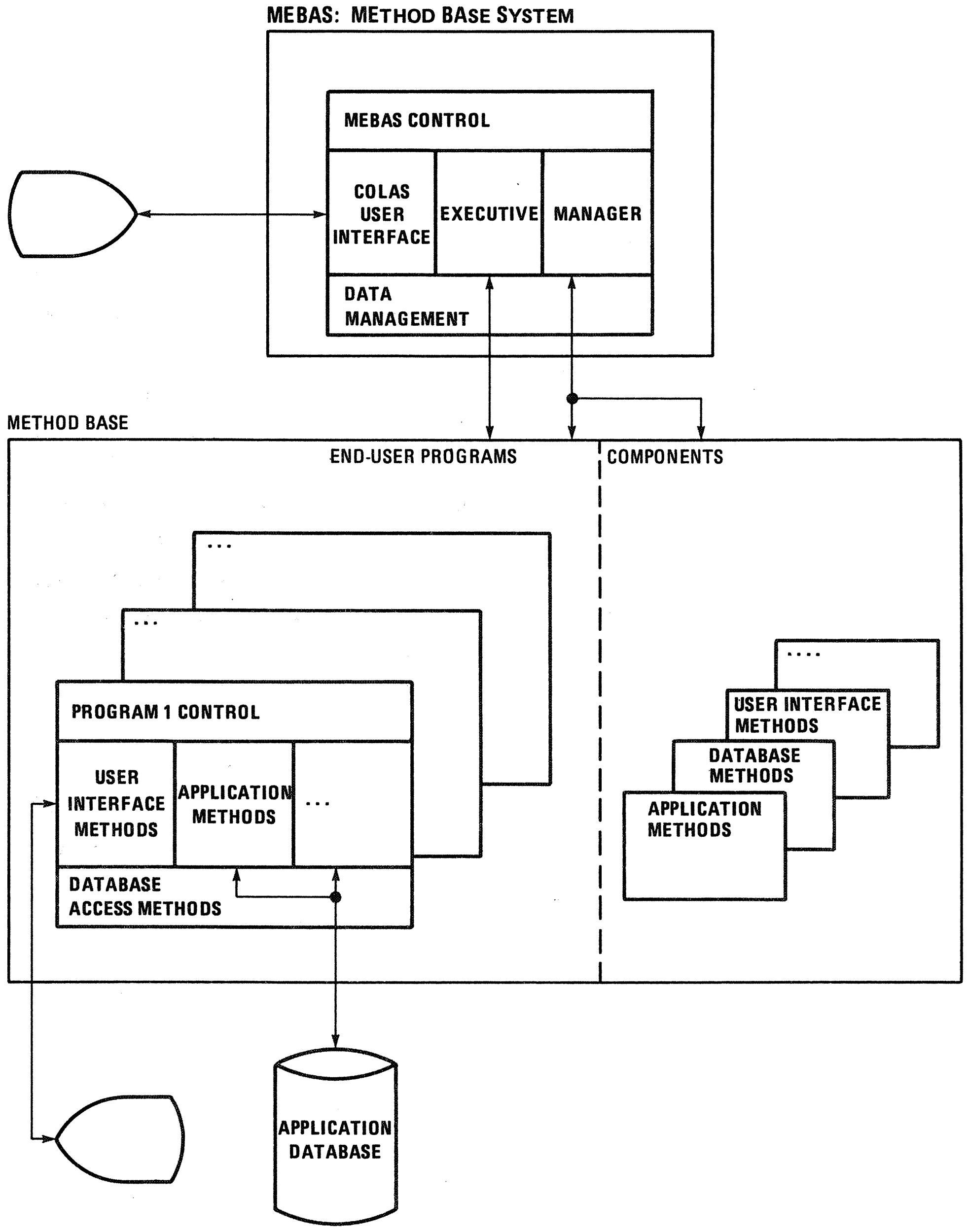


Fig. 3 Concept of MEBAS

- the system designer will access the methodbase via the method manager to assemble methods, creating programs for end-user execution;

- the end user will access the programs via the executive in order to bring them into execution.

For the implementation of a methodbase system in Fortran use is made of the characteristics of promising new programming languages such as Ada. Availability of the concerning compilers will decrease development time considerably.

In a distributed computer- and terminal network methodbases will be needed in different computers (workstations, mainframes,etc). In view of the expected effort, implementation of a distributed methodbase is not considered for the near future by NLR. However, co-existence of methodbases, in which a methodbase administrates information on methods available (in method bases) on other computers in the network, will enable local consultation of remote methods. The first implementation is a stand-alone methodbase. Support for distributed computer facilities will be one of the directions of further development.

Although MEBAS is in its initial stage of operation, at NLR a wide range of application groups have accepted the approach and are contributing in the further evolution of the facilities (Boerstoel,1986; Van den Dam,1985).

6. USER INTERFACE

The user interface of an information system determines to a large extent the acceptation of the system by its users. A user interface component in the infrastructure for CAE shall satisfy the following requirements:

- support the application developers in the development of the user interface for applications by handling syntax checking and error recovery;
- support the user by providing a flexible style of interaction and by additional facilities as help information and abbreviation of input;
- provide uniformity of interaction over a range of applications by standardizing the structure of the input syntax and the special commands to request help information or to read alternate input.

At NLR a user interaction facility, called COLAS, has been taken in operational use. The properties and the technical concept of COLAS will be explained in the sequel.

Basic properties of a user interface are the master/slave relationship between user and information system and the style of interaction, both discussed in the following paragraphs.

The master in an interaction is the one determining the sequence of actions. With the information system as the master, the user supplies requested data. The sequence of actions and data requests are determined by the information system. With the user as master, the user selects an action and provides additional data. The information system performs the indicated action. An information system designed to have the user as master, provides better control for the user, possibly leading to faster interaction. However, it may require more user experience with the information system.

The interaction between the user and the COLAS interaction System is based on the principle that the user serves as the master in the interaction, but COLAS allows the developer to structure the interaction to provide more guidance to the users. In COLAS allowed actions (commands) are grouped in command sets, of which several may be defined for one application program. The application program selects the command set for which the user can supply an entry. The user selects the command with its parameters. The two extreme uses of this concept are:

- only one command is defined in each command set, making the user the slave in the interaction;
- only one command set with all possible commands is defined, making the user the master in the interaction.

The developer will try to find a compromise between these two extremes, in order to deliver a user interaction offering reasonable guidance and flexibility for its users.

There are a number of basic interaction styles, which generally are mixed in a particular interaction. The three styles are:

- question/answer style. The information system presents a question which is to be answered by the user. This style is appropriate for an interaction with the information system as master or for recovery of erroneous or missing input.
- menu style. The information system presents a list of allowed answers and the user has to indicate a selection. In this way the user can select actions to be performed or (sets of) input items to be entered or changed subsequently.
- command style. The information system presents a prompt. The user answer with an input sentence, indicating the required action as well as input data. The syntax of the input sentence could be formalized or based on

natural language.

For the different interaction styles the command style may allow the largest speed of interaction, again requiring more user experience. Users regularly using an information system generally prefer this style for frequently used actions, but do require more support for less frequently used actions.

The basic interaction style in COLAS is command driven. The application developer defines names for commands and parameters and types of parameters. The remainder of the syntax is standardized for all applications to provide a uniform view for users of these applications. The command set approach makes a basic kind of menu interaction already possible. In the evolution of COLAS, an extension to a menu based interaction is defined. It is found that the command sets used in COLAS, can be mapped to a menu style interaction as well as to a command style interaction, without a need to change the way application programs access the interaction system.

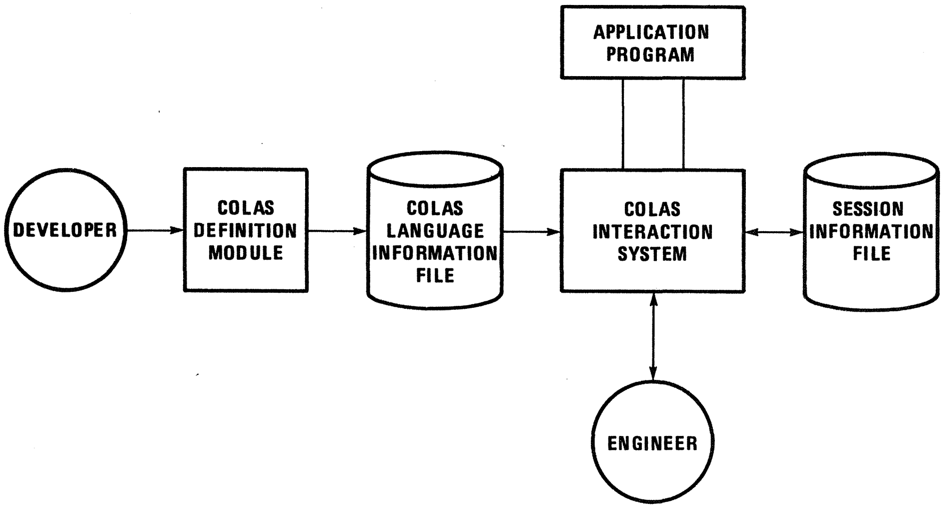


Fig. 4 Technical concept of the user interface facility COLAS

The technical concept of COLAS (Fig. 4) is based on the philosophy, to keep the interaction separated from the processing functions in the application program. Therefore the interaction is defined in the COLAS Definition Module (CDM). As a consequence changes are allowed that do not affect the application program, such as changes in command and parameter names, defaults, prompts, help information and message formats. In this way changing an entire interaction from English to Dutch typically requires a one day effort. The interaction definition for a specific application is stored in the COLAS Language Information File (CLIF). It is possible to retain a Session Information File (SIF) for each user after leaving the program. This file contains most recently entered parameter and message items and can be used for subsequent program executions.

To serve the separation of the user interface development from the application program development an extension to COLAS is currently used as a prototype system. The prototype system usage generates an interaction routine for each command set, using the interaction description given by the developer. This routine handles the interaction and directly calls appropriate application routines with the command parameters provided as actual arguments with the call. In this way the application routines do not handle any aspects of the interaction, and the interface between the user interaction routines and the application program is application specific. In this concept the master/slave relationship is directly reflected in the program structure. The prototype system also generates stubs for all required application routines, enabling immediate evaluation of the user interaction.

7. COMPUTER AND TERMINAL NETWORK

NLR is located at two geographically separated sites at a distance of 100 km. In the beginning of the seventies it was decided to serve both sites with one mainframe computer with a communication front-end in one site and a remote communication controller in the other site. The computer and terminal network as evolved to its current state is depicted in Fig. 5. The alternative in the decision process was a set of large minicomputers for specialized applications. Both alternatives were set up in such a way that investment costs where the same (Loeve, 1976).

The main reason for the choice made, is that only with this configuration NLR could afford a mainframe with sufficient processing power for advanced computational fluid mechanics and extensive data processing for wind tunnel tests, flight tests and integrated theoretical/experimental research. At the moment the reasoning still remains the same. Although computers have become cheaper, for competitive reasons requirements for power of processing have increased accordingly.

Moreover the growth potential in computer capacity needed for increase of the number of applications at NLR could also be provided by a

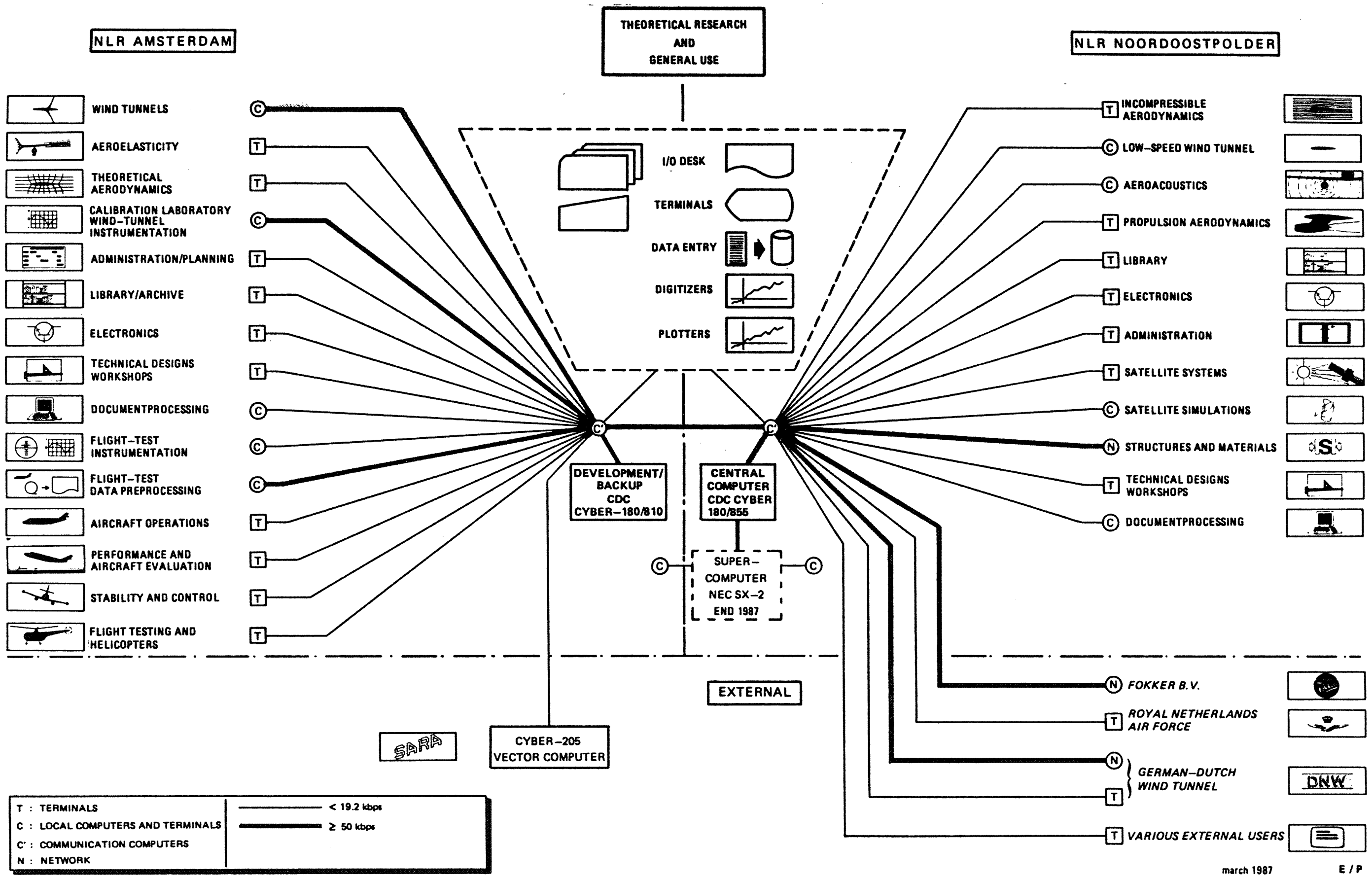


Fig. 5 NLR computer and terminal network

mainframe. For example the number of terminals for interactive applications of the central computer is increasing continuously. Storage capacity of the computer, the transfer capacity between computer and on-line storage devices, and central memory size still can be adapted continuously, to guarantee acceptable response times under these circumstances. The selection criteria for the mainframe stress this growth potential.

In addition to economic aspects of hardware usage, there also exists an urgent need for re-use of software. This not only concerns re-use in successive phases of incremental system development, but also re-use in applications of other disciplines whenever possible. An example is the use of information management. The development of EDIPAS described in chapter 3 and the large number of applications at NLR, would have been impossible to organize without one central mainframe for all large scale computations and dataprocessing. A physical centralized information management also facilitates multi-disciplinary design activities and the development of CAE applications to support those activities. As a result of these considerations the policy of NLR for the time being is to maintain central computer facilities as powerful as possible from the point of view of invested capital. Into line with this policy, NLR decided to add to the existing mainframe computer power a supercomputer, in order to obtain sufficient power to support new simulation methods in the area of computational physics. The performance of existing supercomputers related to the mainframe at NLR is shown in Fig. 6, which was modified after Fernbach, 1985.

The supercomputer selected by NLR (NEC SX2) will be installed end 1987, as integrated part of the NLR computer facilities. In that situation, the supercomputer is foreseen as the extended mainframe computer for fast simulations in computational physics. The general purpose mainframe itself is used for data management and data evaluation, interactive applications, and the moderate batch activities. Access to the

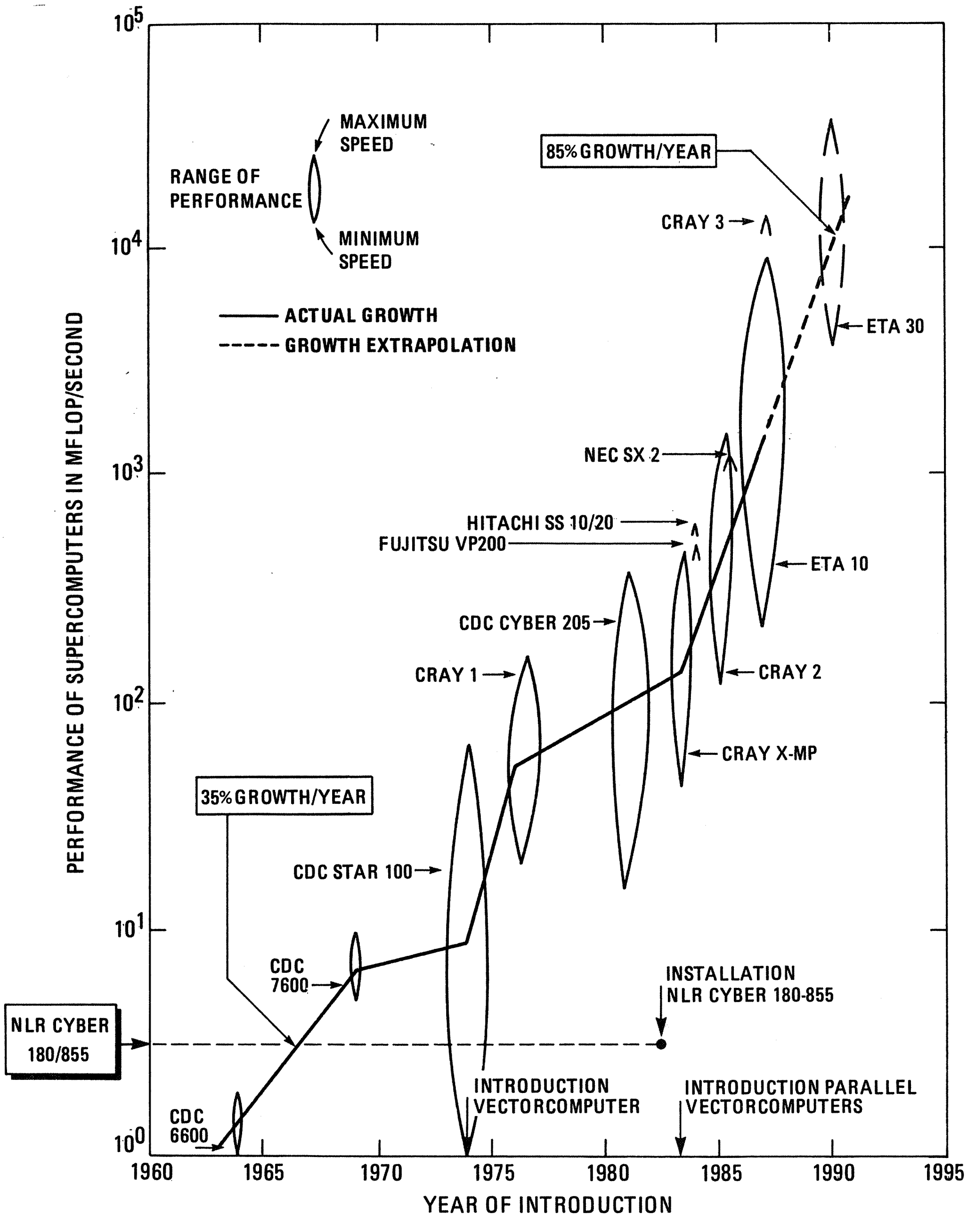


Fig. 6 Performance of supercomputers in the years 1960-1990

supercomputer is possible via a high speed data link from the general purpose mainframe, and for specific applications from minicomputers or workstations.

As was already mentioned in Loeve, 1976, local computing facilities shall be installed for all processing with real time aspects such as data acquisition in experimental facilities. In the NLR computer and terminal network a large number of local minis is integrated in facilities for experiments that can be regarded as information sources for the centralized information management system.

In recent years personal computers have been introduced as part of the network mainly for administrative applications (document processing, project planning activities). Communication especially between both NLR sites is realized via the network.

The strong centralization of computer power imposes requirements with respect to availability of the central system to the end users. To increase this availability a computer is implemented in the network for development of system software (operating system upgrading and communication software development). The development system of course had to be compatible with the central computer. In view of this, one of the criteria for selection of the mainframe at NLR always is that the computer has to form part of a series of computers with a long compatible growth path. The development computer also serves as a back up for a small number of critical applications such as flight test programs of the Dutch national industry. Special measures are taken for these applications to make sure that up to date information for processing of data is available for back up purposes. Back-up facilities for the supercomputer except for the general purpose mainframe, is not foreseen.

Introduction of workstations based on powerful microprocessors is one of the means considered to increase throughput of the network for activities with a high degree of interaction. However, it appears that making use of growth potential of the central mainframe (main memory, disc controllers, discs) is much more cost effective than introduction of the workstations for the organization as a whole.

As far as general purpose software is concerned, use is made of commercially available products as much as possible. In the past, however, severe limitations in the applicability of commercial communication hardware and software have forced NLR to develop dedicated software for the communication controllers in the network. The software is based on the finite state approach. Each connection of terminals and computers in the network is considered as a user-server relation. Between the systems for which this relation is initiated a virtual path is determined in the communication concept (V.d. Bosch et al., 1985).

The usage of a supercomputer requires high volumes of data to be transfered, due to the amount of data that is necessary to feed the computational methods or that are produced as result. As a consequence the datacommunications need a considerable upgrade in possible transfer speed. Investigation of possible solutions learned, that standard facilities in the multi-supplier situation of NLR is still not possible. The planned situation comprises a proprietary Local Area Network for high volume data transmission (file transfer) in combination with the existing network for data transfer for interactive usage.

Summarizing it can be stated that when technical and psychological equal access from all parts of an organization to a central computer is guaranteed for information processing and computations, a sound basis is formed for an infrastructure that can serve as the central nerve system of a multi-disciplinary organization. In practice it appears to lead to a manageable system with maximal possibilities for re-use of software. The benefits of a centralized approach can be maintained also when local computers are introduced for real time aspects if these are connected to the central system to guarantee that generated information can be made available to all whom it may concern. For economic and organizational reasons introduction of intelligent workstations has to be treated carefully.

8. CONCLUSIONS

An infrastructure for information processing is in operation at the National Aerospace Laboratory NLR. This evolutionary developed infrastructure is based on general applicable requirements. As a consequence the applications built with this infrastructure, such as simulation methods based on mathematical software, can be used to support industrial research and product innovation, and can be introduced in other organizations.

The main components of the infrastructure are facilities for data management, method management, and user interfaces, all implemented on a computer- and terminal network. The facilities for data management and user interfaces have already gained wide spread acceptance by application system developers and end users. The first version of the facilities for method management is available and will be developed evolutionary. The computer and terminal network, in which a supercomputer gives access to data and methods from virtually any location at both sites of NLR. This network and the software components are upgraded continuously with evolving user requirements and increasing usage.

9. REFERENCES

Boerstoel, J.W., Veldman, A.E.P., Van der Vooren, J., Van der Wees, A.J., 1986, "Trends in CFD for aeronautical 3-D steady applications: the Dutch situation", NLR MP 86074 U.

van den Bosch, F.J., Nelis, W.J.M., 1985, "Conceptual design of communication software of the NLR computer network", NLR TR 85153.

van den Dam, R.F., 1985, "A perspective of mathematical simulation and optimization techniques in computer-aided design.", Paper presented at CAPE'86, May 20-23 1986, Copenhagen, NLR MP 85088.

Fernbach, S., 1985, "Supercomputers - Past, present, prospects", Supercomputers, FGCS 124, North Holland Publ. Comp.

Groothuizen, R.J.P., van den Berg, J.I., van Leeuwen, W., 1986, "A software design tool for the analysis of robot dynamic properties", NLR MP 86064 U.

Hameetman, G.J., 1982, "The creation of the NLR bench mark package and the performance of two new CDC computers with it", NLR MP 82015.

Kreijkamp, H.A., Van Hedel, H., Heerema, F.J., 1985, "Flexible and Dynamic Data modeling aspects of the Engineering Data management System EDIPAS", The Internal Congress Intelligencia 85, May 21-24, 1985, Paris, France, NLR MP 85026 U.

Loeve, W., 1976, "A hierarchical network linking two research laboratories", Computer networks 1 (1976) 119-129.

Steenbergen, H., Heerema, F.J., 1985, "Database Administrator Facilities for Engineering Data Management", Fourth International Conference and Exhibition on Engineering Software, London, United Kingdom, June 18-20, 1985, NLR MP 85010 U.

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print; replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservation of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statement.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Emperical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallman spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *Go-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewal theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequences.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.

2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.

3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.

4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.

5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.

6 A.P.W. Böhm. *Dataflow computation.* 1984.

7 A. Blokhuis. *Few-distance sets.* 1984.

8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.

9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.

10 C.G. van der Laan, N.M. Temme. *Calculation of special functions: the gamma function, the exponential integrals and error-like functions.* 1984.

11 N.M. van Dijk. *Controlled Markov processes; time-discretization.* 1984.

12 W.H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods.* 1985.

13 D. Grune. *On the design of ALEPH.* 1985.

14 J.G.F. Thiemann. *Analytic spaces and dynamic programming: a measure theoretic approach.* 1985.

15 F.J. van der Linden. *Euclidean rings with two infinite primes.* 1985.

16 R.J.P. Groothuizen. *Mixed elliptic-hyperbolic partial differential operators: a case-study in Fourier integral operators.* 1985.

17 H.M.M. ten Eikelder. *Symmetries for dynamical and Hamiltonian systems.* 1985.

18 A.D.M. Kester. *Some large deviation results in statistics.* 1985.

19 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 1: Philosophy, framework, computer science.* 1986.

20 B.F. Schriever. *Order dependence.* 1986.

21 D.P. van der Vecht. *Inequalities for stopped Brownian motion.* 1986.

22 J.C.S.P. van der Woude. *Topological dynamix.* 1986.

23 A.F. Monna. *Methods, concepts and ideas in mathematics: aspects of an evolution.* 1986.

24 J.C.M. Baeten. *Filters and ultrafilters over definable subsets of admissible ordinals.* 1986.

25 A.W.J. Kolen. *Tree network and planar rectilinear location theory.* 1986.

26 A.H. Veen. *The misconstrued semicolon: Reconciling imperative languages and dataflow machines.* 1986.

27 A.J.M. van Engelen. *Homogeneous zero-dimensional absolute Borel sets.* 1986.

28 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 2: Applications to natural language.* 1986.

29 H.L. Trentelman. *Almost invariant subspaces and high gain feedback.* 1986.

30 A.G. de Kok. *Production-inventory control models: approximations and algorithms.* 1987.

31 E.E.M. van Berkum. *Optimal paired comparison designs for factorial experiments.* 1987.

32 J.H.J. Einmahl. *Multivariate empirical processes.* 1987.

33 O.J. Vrieze. *Stochastic games with finite state and action spaces.* 1987.

34 P.H.M. Kersten. *Infinitesimal symmetries: a computational approach.* 1987.

35 M.L. Eaton. *Lectures on topics in probability inequalities.* 1987.

36 A.H.P. van der Burgh, R.M.M. Mattheij (eds.). *Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87).* 1987.