Kuwait University
Department of Mathematics &
Computer Science

Kuwait Foundation
for the Advancement
of Sciences (KFAS)

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON MATHEMATICS AND ITS APPLICATIONS

(ICMA 2004)

April 5-7, 2004

State of Kuwait

ORGANIZING COMMITTEE

Dr. Mansour A. Al-Zanaidi, Chairman

Prof. Bader N. Al-Saqabi, Treasurer
Prof. Adnan H. Al-Aqeel
Prof. Reyadh R. Khazal
Prof. Ali Al-Zamel
Prof. Shyam L. Kalla
Prof. Man M. Chawla, Secretary

FOREWORD

This volume contains papers and invited talks presented at the *"International Conference on Mathematics and Its Applications"* (ICMA04-Kuwait), which was held at Kuwait University from April 5 to 7, 2004. All papers contained in this volume are refereed/peer reviewed.

The Conference covered active research fields of the Department of Mathematics and Computer Science:

(A) Computational Differential Equations and Linear Algebra.

(B) Integral Transforms and Special Functions with Fractional Calculus.

(C) Groups, Rings & Categories, and Differential Geometry.

The ICMA04-Kuwait was jointly sponsored and organized by the *Department of Mathematics and Computer Science, Kuwait University, and the Kuwait Foundation for the Advancement of Sciences (KFAS)*. The Conference was attended by sixty three participants from twenty two countries.

Keynote speakers at the conference were: G. V. Berghe (Belgium), S. Caenepeel (Belgium), L. Debnath (USA), D. J. Evans (UK), V. D. Mazurov (Russia), A. J. Scholl (UK) and B. Wegner (Germany).

The organizers would like to thank all the participants for their cooperation in preparing their contributions for this Proceedings and their active interest in the peer reviewing process for the volume. Thanks are also due to all colleagues who helped in the reviewing process.

All papers in this volume are arranged alphabetically according to the name of the first author.

We would like to take this opportunity to express our gratitude to the administration at Kuwait University and Kuwait Foundation for the Advancement of Sciences, for sponsoring this Conference and publication of the Proceedings.

Finally we offer our sincere thanks to the faculty and staff of the Department of Mathematics and Computer Science at Kuwait University for taking active interest, shouldering several responsibilities during the conference and publication of these proceedings.

Mansour A. Al-Zanaidi, Chairman

Shyam L. Kalla and Man M. Chawla
Editors

Kuwait, December 2004.

CONTENTS

ON AN INTEGRAL TRANSFORM INVOLVING BESSEL FUNCTIONS

M. Al-Hajri and S. L. Kalla
Department of Mathematics & Computer Science
Kuwait university, P. O. Box. 5969, Safat 13060, KUWAIT
email: kalla@mcs.sci.kuniv.edu.kw

Abstract

This paper deals with a new integral transform, involving a combination of Bessel functions as a kernel. The inversion formula is established and some properties are given. This transform can be used to solve some mixed boundary value problems. We consider here a problem of heat conduction in an infinite and semi-infinite cylinder $a \leq r \leq b$, with radiation-type boundary conditions.

AMS Subj. classification: 44A20, 35K20, 80A20.

Keywords: Integral transform, Bessel Functions, Differential equations, Boundary conditions.

1. INTRODUCTION

Let $f(t)$ be a given function defined on an interval $[a, b]$, that belongs to a certain class of functions. An integral transform of $f(t)$ is a mapping of the form,

$$T[f(t); s] = \bar{f}(s) = \int_a^b K(s,t) f(t) dt,$$

provided that the integral exists. $K(s,t)$ is a prescribed function, called the kernel of the transform [2,5,6,11]. Among the well known transforms are the Laplace, Fourier, Hankel, Stieltjes and Mellin transforms. The most versatile of these, the Laplace transform has been widely used to solve differential equations, and particularly problems related to heat transfer and electrical circuits. On the other hand for problems in which there is an axial symmetry, the Hankel transforms are found to be most appropriate. The Mellin transform being closely related to the Fourier transform, has its own peculiar uses, as for deriving expansion and solving problems with wedge shape boundaries. In general, the use of an integral transform often reduces a partial differential equation in n independent variables to $(n-1)$ variables, that provides a simplification of the problem.

The success of the use of integral transforms to solve boundary value problems and to exclude a variable with range $(0,\infty)$ or $(-\infty,\infty)$ led investigators to consider finite integral transforms. Doetsch considered finite Fourier transforms, and Sneddon [11] extended the idea of Bessel function kernel, called 'Finite Hankel Transforms'. Using the Sturm-Liouville theory [3,], a number of integral transforms can be implemented, according to the prescribed boundary conditions.

Recently Khajah [9] has considered a modified Hankel transform in the form,

$$J_\mu[f(z);s,\lambda]=\int_0^b z^\lambda f(z)J_\mu(zs)dz$$

where $f(z)$ satisfies Dirichlet's conditions on the interval $[a,b]$. He has derived the inversion formula, Parseval-type identities, transform of derivatives, as well as transforms of products of the form $z^\lambda f(z)$.

Using the Sturm-Liouville theory Kalla and Villalobos [7,8] have defined and studied an integral transform defined as,

$$T[f(x),a,b,\nu;\lambda_i]=\bar{f}_\nu(\lambda_i)=\int_a^b xf(x)C_\nu(\lambda_i x)dx,$$

where

$$\begin{aligned}C_\nu(\lambda_i x) &= \{Y_\nu(\lambda_i a)+B_\nu(\lambda_i b)\}J_\nu(\lambda_i x)\\ &\quad -\{J_\nu(\lambda_i a)+A_\nu(\lambda_i b)\}Y_\nu(\lambda_i x)\end{aligned}$$

and

$$\begin{aligned}A_\nu(\lambda x) &= J_\nu(\lambda x)+h\lambda J_\nu^{'}(\lambda x)\\ B_\nu(\lambda x) &= Y_\nu(\lambda x)+h\lambda Y_\nu^{'}(\lambda x)\end{aligned}$$

and λ_i are the positive roots of equation,

$$J_\nu(\lambda a)B_\nu(\lambda b)-Y_\nu(\lambda a)A_\nu(\lambda b)=0$$

This transform has been used to solve a heat conduction problem in an infinite cylinder bounded by the surface $r=a$, $r=b$ $(b>a)$.

In this paper we define and study a new integral transform involving Bessel functions of first and second kind, by invoking the Sturm-Liouville theory. Inversion formula is established and some properties are mentioned. The transform has been

used to solve a heat conduction problem in an infinite and a semi- infinite circular cylinder, bounded by surfaces $r = a$ and $r = b$ $(b > a)$, with radiation-type boundary conditions on both surfaces.

2. DEFINITION AND INVERSION FORMULA

Consider Bessel's differential equation

$$x^2 y'' + x y' + (\lambda^2 x^2 - \nu^2) y = 0\,, \qquad x \in [a, b] \tag{1}$$

with homogeneous boundary conditions:

$$y(a) + h_1 y'(a) = y(b) + h_2 y'(b) = 0 \tag{2}$$

The general solution of (1) is given by:

$$y(x) = c_1 J_\nu(\lambda x) + c_2 Y_\nu(\lambda x) \tag{3}$$

where $c_1\,, c_2$ are arbitrary constants, and $J_\nu(x)$, $Y_\nu(x)$ are the Bessel functions of first and second kind respectively. To obtain a solution of (1) that satisfies conditions (2), we have

$$c_1 \left[J_\nu(\lambda a) + h_1 \lambda J'_\nu(\lambda a) \right] + c_2 \left[Y_\nu(\lambda a) + h_1 \lambda Y'_\nu(\lambda a) \right] = 0 \tag{4}$$

$$c_1 \left[J_\nu(\lambda b) + h_2 \lambda J'_\nu(\lambda b) \right] + c_2 \left[Y_\nu(\lambda b) + h_2 \lambda Y'_\nu(\lambda b) \right] = 0 \tag{5}$$

from which we deduce

$$\frac{c_1}{c_2} = -\frac{Y_\nu(\lambda a) + h_1 \lambda Y'_\nu(\lambda a)}{J_\nu(\lambda a) + h_1 \lambda J'_\nu(\lambda a)} = -\frac{Y_\nu(\lambda b) + h_2 \lambda Y'_\nu(\lambda b)}{J_\nu(\lambda b) + h_2 \lambda J'_\nu(\lambda b)} \tag{6}$$

Let

$$\begin{aligned} A_\nu(\lambda x, h_k) &= J_\nu(\lambda x) + h_k \lambda J'_\nu(\lambda x)\,, \qquad k = 1, 2 \\ B_\nu(\lambda x, h_k) &= Y_\nu(\lambda x) + h_k \lambda Y'_\nu(\lambda x)\,, \qquad k = 1, 2 \end{aligned}$$

Then, the function given by (3) is a solution of equation (1), subject to the conditions (2), if λ is a root of the transcendental equation,

$$B_\nu(\lambda a, h_1) A_\nu(\lambda b, h_2) - A_\nu(\lambda a, h_1) B_\nu(\lambda b, h_2) = 0 \tag{7}$$

Henceforth, we take λ_i $(i = 1, 2, \ldots)$ to be the positive roots of equation (7). Then, from (4–5), we have

$$y_i(x) = \frac{c_1}{B_\nu(\lambda_i a, h_1)} \left[J_\nu(\lambda_i x) B_\nu(\lambda_i a, h_1) - A_\nu(\lambda_i a, h_1) Y_\nu(\lambda_i x) \right] \tag{8}$$

$$= \frac{c_1}{B_\nu(\lambda_i b, h_2)} \left[J_\nu(\lambda_i x) B_\nu(\lambda_i b, h_2) - A_\nu(\lambda_i b, h_2) Y_\nu(\lambda_i x) \right] \tag{9}$$

If we define

$$\mathsf{Z}_i = B_\nu(\lambda_i a, h_1) + B_\nu(\lambda_i b, h_2)\,, \qquad \mathsf{W}_i = A_\nu(\lambda_i a, h_1) + A_\nu(\lambda_i b, h_2)$$

then the following functions are taken to be solutions of (1–2):

$$M_\nu(\lambda_i x) = \mathsf{Z}_i\, J_\nu(\lambda_i x) - \mathsf{W}_i\, Y_\nu(\lambda_i x) \tag{10}$$

By Sturm-Liouville theory [3], the functions of the system (10) are orthogonal on the interval $[a, b]$ with weight function x, that is

$$\int_a^b x M_\nu(\lambda_i x)\, M_\nu(\lambda_j x) dx = \begin{cases} 0 & , \quad i \neq j \\ \\ \mathcal{M}_\nu(\lambda_i) & , \quad i = j \end{cases} \tag{11}$$

where $\mathcal{M}_\nu(\lambda_i) = \|\sqrt{x} M_\nu(\lambda_i x)\|_2^2$ — the weighted L^2 norm. If a function $f(x)$ and its first derivative are piecewise continuous on the interval $[a, b]$, then the relation

$$T[f(x), a, b, \nu; \lambda_i] = \bar{f}_\nu(\lambda_i) = \int_a^b x f(x) M_\nu(\lambda_i x) dx \tag{12}$$

defines a linear integral transform. To derive the inversion formula for this transform, given the series expansion,

$$f(x) = \sum_{i=1}^{\infty} a_i M_\nu(\lambda_i x) \tag{13}$$

we multiply (13) by $x M_\nu(\lambda_j x)$ and integrate both sides with respect to x to get the coefficients:

$$a_i = \frac{1}{\mathcal{M}_\nu(\lambda_i)} \int_a^b x\, f(x) M_\nu(\lambda_i x)\, dx = \frac{\bar{f}_\nu(\lambda_i)}{\mathcal{M}_\nu(\lambda_i)}\,, \qquad i = 1, 2, \ldots \tag{14}$$

and the inversion formula becomes

$$f(x) = \sum_{i=1}^{\infty} \frac{\bar{f}_\nu(\lambda_i)}{\mathcal{M}_\nu(\lambda_i)}\, M_\nu(\lambda_i x) \tag{15}$$

Using some well known properties of Bessel functions [12] we can easily derive the following relation:

$$\begin{aligned} 2\,\mathcal{M}_\nu(\lambda_i) = \mathsf{Z}_i^2 \left[b^2 P(\lambda_i, b, \nu) - a^2 P(\lambda_i, a, \nu)\right] - 2\mathsf{Z}_i\mathsf{W}_i \left[b^2 Q(\lambda_i, b, \nu) - a^2 Q(\lambda_i, a, \nu)\right] \\ + \mathsf{W}_i^2 \left[b^2 R(\lambda_i, b, \nu) - a^2 R(\lambda_i, a, \nu)\right] \end{aligned} \tag{16}$$

in which

$$P(\lambda_i,\mu,\nu) = J_\nu^2(\lambda_i\mu) - J_{\nu-1}(\lambda_i\mu)J_{\nu+1}(\lambda_i\mu)$$
$$R(\lambda_i,\mu,\nu) = Y_\nu^2(\lambda_i\mu) - Y_{\nu-1}(\lambda_i\mu)Y_{\nu+1}(\lambda_i\mu)$$

and

$$Q(\lambda_i,\mu,\nu) = J_\nu^{'}(\lambda_i\mu)Y_{\nu-1}(\lambda_i\mu) - \frac{1}{\lambda_i\mu}J_{\nu-1}(\lambda_i\mu)Y_\nu(\lambda_i\mu) - J_{\nu-1}^{'}(\lambda_i\mu)Y_\nu(\lambda_i\mu)$$

and μ stands for a or b. It is not difficult to verify some properties of the transform from definition (12). For example,

$$T[\alpha f(x) + \beta g(x), a, b, \nu; \lambda_i] = \alpha\bar{f}(\lambda_i) + \beta\bar{g}(\lambda_i) \tag{17}$$

$$T[f(px), a, b, \nu; \lambda_i] = \int_a^b xf(px)M_\nu(\lambda_i x)dx = \frac{1}{p^2}T[f(x), pa, pb, \nu; \lambda_i/p] \tag{18}$$

3. TRANSFORM OF A DIFFERENTIAL OPERATOR

We derive the transform of the following operator

$$Df(x) = \frac{d^2}{dx^2}f(x) + \frac{1}{x}\frac{d}{dx}f(x) - \frac{\nu^2}{x^2}f(x), \qquad a \le x \le b \tag{19}$$

Let I be the transform of the first two terms of D, that is

$$I = \int_a^b x\left[f''(x) + \frac{1}{x}f'(x)\right]M_\nu(\lambda_i x)\,dx = \int_a^b xf''(x)M_\nu(\lambda_i x)dx + \int_a^b f'(x)M_\nu(\lambda_i x)dx$$

Integration by parts of the first integral leads to,

$$\int_a^b xf''(x)M_\nu(\lambda_i x)dx = xM_\nu(\lambda_i x)f'(x)\Big|_a^b - \int_a^b f'(x)\left[x\lambda_i M_\nu'(\lambda_i x) + M_\nu(\lambda_i x)\right]dx,$$

and hence,

$$I = xM_\nu(\lambda_i x)f'(x)\Big|_a^b - \lambda_i\int_a^b xf'(x)\,M_\nu'(\lambda_i x)\,dx$$

Integrating by parts once again leads to,

$$I = x\left[f'(x)M_\nu(\lambda_i x) - \lambda_i f(x)M_\nu'(\lambda_i x)\right]\Big|_a^b + \int_a^b x^{-1}\left[\lambda_i^2x^2M_\nu''(\lambda_i x) + \lambda_i xM_\nu'(\lambda_i x)\right]f(x)\,dx$$

Since M_ν satisfies (1)we have

$$\lambda_i^2x^2M_\nu''(\lambda_i x) + \lambda_i xM_\nu'(\lambda_i x) = (\nu^2 - \lambda_i^2x^2)\,M_\nu(\lambda_i x)$$

and

$$\int_a^b x^{-1}\left[\lambda_i^2 x^2 M_\nu''(\lambda_i x)+\lambda_i x M_\nu'(\lambda_i x)\right] f(x)\,dx=\int_a^b x\left[\frac{\nu^2}{x^2}-\lambda^2\right] f(x)\,M_\nu(\lambda_i x)\,dx$$

Furthermore, it follows from the boundary conditions (2) that

$$\lambda_i M_\nu'(\lambda_i a)=\frac{1}{h_1}\,M_\nu(\lambda_i a)\,,\qquad \lambda_i M_\nu'(\lambda_i b)=\frac{1}{h_2}\,M_\nu(\lambda_i b)$$

Hence

$$I=\frac{b}{h_2}\,M_\nu(\lambda_i b)\left[f(b)+h_2 f'(b)\right]-\frac{a}{h_1}\,M_\nu(\lambda_i a)\left[f(a)+h_1 f'(a)\right]-\lambda_i^2\bar{f}(\lambda_i)+T\left[\frac{\nu^2}{x^2}f(x)\right]$$

and the transform of the operator D in (19) becomes

$$T\left[Df(x)\right]=\frac{b}{h_2}\,M_\nu(\lambda_i b)\left[f(b)+h_2 f'(b)\right]-\frac{a}{h_1}\,M_\nu(\lambda_i a)\left[f(a)+h_1 f'(a)\right]-\lambda_i^2\bar{f}(\lambda_i) \tag{20}$$

Transform of x^{ν}

From definition (12) we have

$$T[x^\nu,a,b,\nu;\lambda_i]=\int_a^b x^{\nu+1}M_\nu(\lambda_i x)\,dx$$

Using a result of [12], namely

$$\int x^{\rho+1}\mathcal{Z}_\rho(x)dx=x^{\rho+1}\mathcal{Z}_{\rho+1}(x)$$

where $\mathcal{Z}_\rho(x)$ is any of the Bessel functions, we obtain

$$T[x^\nu,a,b,\nu;\lambda_i]=\frac{1}{\lambda_i}\left[b^{\nu+1}\,M_{\nu+1}(\lambda_i b)-a^{\nu+1}\,M_{\nu+1}(\lambda_i a)\right]$$

Since

$$M_{\nu+1}(cz)=\frac{\nu}{cz}\,M_\nu(cz)-M_\nu'(cz)$$

the transform becomes,

$$T[x^\nu,a,b,\nu;\lambda_i]=\frac{b^{\nu+1}}{\lambda_i}\left[\frac{\nu}{\lambda_i b}\,M_\nu(\lambda_i b)-M_\nu'(\lambda_i b)\right]-\frac{a^{\nu+1}}{\lambda_i}\left[\frac{\nu}{\lambda_i a}\,M_\nu(\lambda_i a)-M_\nu'(\lambda_i a)\right]$$

Then, from the boundary conditions (2) this reduces to,

$$T[x^\nu,a,b,\nu;\lambda_i]=\frac{b^{\nu+1}}{\lambda_i^2}\left[\frac{\nu}{b}+\frac{1}{h_2}\right]\,M_\nu(\lambda_i b)-\frac{a^{\nu+1}}{\lambda_i^2}\left[\frac{\nu}{a}+\frac{1}{h_1}\right]\,M_\nu(\lambda_i a) \tag{21}$$

In particular, the transform of a constant (where $\nu = 0$) is found to be

$$T[c, a, b, 0; \lambda_i] = \frac{c}{\lambda_i^2}\left[\frac{b}{h_2}\,M_0(\lambda_i b) - \frac{a}{h_1}\,M_0(\lambda_i a)\right] = c\,T[1, a, b, 0; \lambda_i] \tag{22}$$

4. HEAT CONDUCTION IN AN INFINITE CYLINDER

Consider a long hollow cylinder of inner radius a and outer radius b, with radiation type boundary conditions in both outer and inner surface, and a prescribed initial temperature. The differential equation of the phenomena is:

$$\frac{1}{K}\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r}\,, \tag{23}$$

where $a < r < b\,,\ t > 0$ and $U(r,t)$ denotes the temperature at any radial position r at time t; K is a constant that depends on the material of the cylinder. The initial and boundary conditions are as follows:

$$U(r,0) = I(r)\,, \qquad a < r < b \tag{24}$$

$$U + h_1\frac{\partial U}{\partial r}\bigg|_{r=a} = f(t)\,, \qquad U + h_2\frac{\partial U}{\partial r}\bigg|_{r=b} = g(t)\,, \qquad t > 0 \tag{25}$$

Taking $\nu = 0$, we consider the transform of U with respect to the radial variable, that is

$$\overline{U}(\lambda_i, t) = \int_a^b r\,U(r,t)\,M_0(\lambda_i r)\,dr \tag{26}$$

Referring to (20) and (23), we obtain

$$\frac{1}{K}\frac{\partial \overline{U}}{\partial t} = \frac{b}{h_2}\,M_0(\lambda_i b)\left[U + h_2\frac{\partial U}{\partial r}\right]_{r=b} - \frac{a}{h_1}\,M_0(\lambda_i a)\left[U + h_1\frac{\partial U}{\partial r}\right]_{r=a} - \lambda_i^2\overline{U}$$

From the boundary conditions (25) we have this reduced to

$$\frac{1}{K}\frac{\partial \overline{U}}{\partial t} = \frac{b}{h_2}\,M_0(\lambda_i b)\,g(t) - \frac{a}{h_1}\,M_0(\lambda_i a)\,f(t) - \lambda_i^2\overline{U}$$

and the following ODE is obtained

$$\frac{\partial \overline{U}}{\partial t} + K\lambda_i^2\overline{U} = K\left[\frac{b}{h_2}M_0(\lambda_i b)\,g(t) - \frac{a}{h_1}M_0(\lambda_i a)\,f(t)\right] \tag{27}$$

whose solution is given by

$$\overline{U}(\lambda_i, t) = \exp(-K\lambda_i^2 t)\,\left[K\int_0^t \exp(K\lambda_i^2 s)\Big[\frac{b}{h_2}M_0(\lambda_i b)\,g(s) - \frac{a}{h_1}M_0(\lambda_i a)\,f(s)\Big]\,ds + C\right]$$

Taking the transform of the initial condition (24), namely

$$\overline{U}(\lambda_i, 0) = \overline{I}(\lambda_i)$$

leads to $C = \overline{I}(\lambda_i)$, hence

$$\overline{U}_s(\lambda_i; t) = \exp(-K\lambda_i^2 t)$$
$$\Big[K \int_0^t \exp(-K\lambda s\big[\frac{b}{h_2} M_0(\lambda_i b)\, g(s) - \frac{a}{h_1} M_0(\lambda_i a)\, f(s)\big]\, ds + \overline{I}(\lambda_i)\Big] \tag{28}$$

The solution of (23) follows after applying the inversion formula to the above, thus

$$U(r,t) = \sum_{i=1}^{\infty} \frac{\overline{U}(\lambda_i, t)}{\mathcal{M}_\nu(\lambda_i)}\, M_0(\lambda_i r) \tag{29}$$

with the understanding that the summation is taken over all the positive roots of the equation:

$$B_0(\lambda a, h_1) A_0(\lambda b, h_2) - A_0(\lambda a, h_1) B_0(\lambda b, h_2) = 0 \tag{30}$$

Special cases:

(i) Let us consider the previous problem with the following initial and boundary conditions;

$$U(r,0) = 0\,, \qquad a < r < b \tag{31}$$

$$U + h_1 \frac{\partial U}{\partial r}\bigg|_{r=a} = U_0 \text{ (constant)}, \quad U + h_2 \frac{\partial U}{\partial r}\bigg|_{r=b} = U_1 \text{(constant)}, \quad t > 0 \tag{32}$$

Then equation (28) will become

$$\overline{U}(\lambda_i; t) = \exp(-K\lambda_i^2 t)$$
$$\left[K \int_0^t \exp(K\lambda_i^2 x) \left\{\frac{b}{h_2} M_0(\lambda_i b) U_1 - \frac{a}{h_1} M_0(\lambda_i a) U_0\right\} dx\right] \tag{33}$$
$$= \frac{1 - \exp(-K\lambda_i^2 t)}{\lambda_i^2} \left[\frac{bU_1}{h_2} M_0(\lambda_i b) - \frac{aU_0}{h_1} M_0(\lambda_i a)\right]$$

And according to (29) the solution will become

$$U(r,t) = \sum_{i=1}^{\infty} \frac{1 - \exp(-K\lambda_i^2 t)}{\lambda_i^2 \mathcal{M}_\nu(\lambda_i)} \left[\frac{bU_1}{h_2} M_0(\lambda_i b) - \frac{aU_0}{h_1} M_0(\lambda_i a)\right] M_0(\lambda_i r), i = 1, 2, 3, ... \tag{34}$$

where the summation is taken over all positive roots of (30).

(ii) If $h_1 \longrightarrow 0$ in (25), that is $U\,|_{r=a} = f(t)$, our result (29) reduces to a result of Kalla and Villalobos [8, p.41, eq.(20].

5. HEAT CONDUCTION IN A SEMI-INFINITE CYLINDER

Let us consider the problem of finding the temperature distribution $U(r, z, t)$ in a hollow semi-infinite cylinder with outer radius b and inner radius a. This problem is expressed by the differential equation,

$$\frac{1}{K}\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r} + \frac{\partial^2 U}{\partial z^2} \tag{35}$$

where $a < r < b\,;\ z\,,\ t > 0\,$, and the initial/boundary conditions are taken to be:

$$U(r, z, 0) = I(r, z) \tag{36}$$

$$U(r, 0, t) = V(r, t)\,, \qquad \lim_{z\to\infty} U(r, z, t) = 0 \tag{37}$$

$$U + h_1\frac{\partial U}{\partial r}\bigg|_{r=a} = f(z, t)\,, \qquad U + h_2\frac{\partial U}{\partial r}\bigg|_{r=b} = g(z, t) \tag{38}$$

Consider the integral transform,

$$\overline{U}(\lambda_i, z, t) = \int_a^b r\,U(r, z, t)M_0(\lambda_i r)\,dr \tag{39}$$

and the Fourier sine transform

$$\overline{U}_s(\lambda_i, p, t) = \int_0^\infty \overline{U}(\lambda_i, z, t)\,\sin(p\,z)\,dz \tag{40}$$

Following a similar argument as in the previous section, the transformed equation of (35) becomes

$$\frac{1}{K}\frac{\partial\overline{U}}{\partial t} = \frac{b}{h_2}M_0(\lambda_i b)g(z, t) - \frac{a}{h_1}M_0(\lambda_i a)f(z, t) - \lambda_i^2\bar{U} + \frac{\partial^2\overline{U}}{\partial z^2}$$

whose Fourier sine transform is found to be,

$$\frac{1}{K}\frac{d\overline{U}_s}{dt} = \frac{b}{h_2}M_0(\lambda_i b)g_s(p, t) - \frac{a}{h_1}M_0(\lambda_i a)f_s(p, t) - \lambda_i^2\overline{U}_s - p^2\overline{U}_s + p\overline{V}(\lambda_i, t)$$

and which is expressed as,

$$\frac{d\overline{U}_s}{dt} + K(\lambda_i^2 + p^2)\overline{U}_s$$
$$= K\left[\frac{b}{h_2}M_0(\lambda_i b)g_s(z, t) - \frac{a}{h_1}M_0(\lambda_i a)f_s(z, t) + p\overline{V}(\lambda_i, t)\right] \tag{41}$$

Now we have the solution of the above ODE given by

$$\overline{U}_s(\lambda_i,p,t) = e^{-K(\lambda_i^2+p^2)t}\left[K\int_0^t e^{K(\lambda_i^2+p^2)\,\tau}\left[\frac{b}{h_2}M_0(\lambda_i b)g_s(z,\tau) - \frac{a}{h_1}M_0(\lambda_i a)f_s(z,\tau)+p\overline{V}(\lambda_i,\tau)\right]d\tau + C\right]$$

and from the initial condition, we have $C=\overline{I}_s(\lambda_i,p)$, so that

$$\overline{U}_s(\lambda_i,p,t) = e^{-K(\lambda_i^2+p^2)t}\left[K\int_0^t e^{K(\lambda_i^2+p^2)\,\tau}\left[\frac{b}{h_2}M_0(\lambda_i b)g_s(z,\tau) - \frac{a}{h_1}M_0(\lambda_i a)f_s(z,\tau)+p\overline{V}(\lambda_i,\tau)\right]d\tau + \overline{I}_s(\lambda_i,p)\right] \tag{42}$$

Finally, taking respective inverse transforms will lead to the solution:

$$U(r,z,t)=\frac{2}{\pi}\sum_{i=1}^{\infty}\frac{1}{\mathcal{M}_\nu(\lambda_i)}\left[\int_0^\infty \overline{U}_s(\lambda_i,p,t)\,\sin(pt)\,dp\right]M_0(\lambda_i r). \tag{43}$$

Special cases:

(i) We consider the previous partial differential equation with the following conditions

$$\left(U+h_1\frac{\partial U}{\partial r}\right)_{r=a}=0, \qquad z>0,\ t>0$$
$$\left(U+h_2\frac{\partial U}{\partial r}\right)_{r=b}=\frac{1}{z}, \qquad z>0, t>0$$
$$U(r,0,t)=U_0, \quad \text{(constant)} \qquad a<r<b, t>0$$
$$U(r,z,0)=0 \qquad a<r<b, z>0$$
$$U(r,z,t)\to 0 \qquad \text{as} \qquad z\to\infty$$

then (42) will become

$$\bar{U}_s(\lambda_i,p,t) = e^{-K(\lambda_i^2+p^2)t}\left[K\int_0^t e^{-K(\lambda_i^2+p^2)x}\left\{\frac{b}{h_2}M_0(\lambda_i b)\frac{\pi}{2} + p\frac{U_0}{\lambda_i^2}\left[\frac{b}{h_2}M_0(\lambda_i b)-\frac{a}{h_1}M_0(\lambda_i a)\right]\right\}dx\right]$$
$$= \frac{1-e^{-K(\lambda_i^2+p^2)t}}{\lambda_i^2+p^2}\left\{\frac{p}{h_2}\left[\frac{\pi}{2}+\frac{PU_0}{\lambda_i^2}\right]M_0(\lambda_i b)-\frac{aPU_0}{\lambda_i^2 h_1}M_0(\lambda_i a)\right\}$$

and according to (43)

$$U(r,z,t)=\sum_{i=1}^{\infty}\left\{\frac{2}{\pi}\int_0^{\infty}\frac{1-e^{-K(\lambda_i^2+p^2)t}}{\lambda_i^2+p^2}\left\{\frac{p}{h_2}\left[\frac{\pi}{2}+\frac{PU_0}{\lambda_i^2}\right]M_0(\lambda_i b)-\frac{aPU_0}{\lambda_i^2 h_1}M_0(\lambda_i a)\right\}\right.$$

$$\left.\sin(pt)dz\right\}\times\frac{M_0(\lambda_i r)}{\mathcal{M}_\nu(\lambda_i)}.$$

(ii) For $h_1 \longrightarrow 0$ in (38), that $U\,|_{r=a}=f(z,t)$, our result (43) reduces to one considered in [8, p.43; eq. (36)].

6. CONCLUSIONS

Here we have introduced a new finite integral transform (Hankel-type) involving product of Bessel functions as the Kernel. This transform can be used to solve certain class of mixed boundary value problems. As indicated in previous sections 4 and 5, this transform is suitable to solve heat conduction problems in hollow cylinders with radiation (mixed) conditions on both surfaces ($r=a$ and $r=b$). The Hankel-type finite transforms considered earlier in [5, 7, 8, 11] were able to solve problems with both surfaces of the cylinder kept at prescribed temperature or radiation condition on only one surface, $r=b$ [8].

Numerical treatment of the results obtained here can be done by using *"Matematica"* and [S. L. Kalla and S. Conde: Tables of Bessel Functions and Roots of Transcendental Equations, Univ. Zulia, Venezuela, 1978].

ACKNOWLEDGEMENT

The authors are thankful to the referee for some useful comments and suggestions for a better presentation of this paper.

REFERENCES

[1] Ali, I. and Kalla, S.L., A generalized Hankel transform and its use for solving certain partial differential equations. *Jour. Australian Math.* Soc. 41 B (1999), 105-117.

[2] Andrews, L. C. and Shivamoggi, B. K., *Integral Transforms for Engineers,* SPIE, Bellingham, Washington, (1999).

[3] Birkhoff, G. and Rota, G. C., *Ordinary Differential Equations*, Wiley, New York, (1989).

[4] Churchill, R. V., *Operational Mathematics,* McGraw-Hill, New York, (1972).

[5] Debnath, L. *Integral Transforms and their Applications.* CRC Press, Boca Raton, FL. , (1995).

[6] A. Erdèlyi, (Ed.), *Tables of Integral Transforms,* Vol. 1 &2., McGraw-Hill, New York. (1954).

[7] Kalla, S.L. and Villalobos, A., On a new integral transform I., *Jnanabha,* 9-10 (1980), 149-154.

[8] Kalla, S. L. and Villalobos, A., On a new integral transform II., *Rev. Tec. Ing. ,* Univ. Zulia, 5 (1982), 40-44.

[9] Khajah, H. G. A modified finite Hankel transform. *Integral Transforms Spec. Func.,* 14(2003), 403-412.

[10] Prudnikov, A. P., Brychkov, Yu. A., and Marichev, O. I., *Series and Integrals,* Gordon and Breach, New York.,(1993).

[11] Sneddon, I. N. *The Use of Integral Transforms,* McGraw-Hill, New Yourk, (1973).

[12] Watson, G. N. *Theory of Bessel Functions,* Cambridge Univ. Press, (1966).

PVM-BASED IMPLEMENTATION OF A NEW ADI TECHNIQUE (IADEI) IN SOLVING ONE DIMENSIONAL PARABOLIC EQUATION USING PC CLUSTER SYSTEM

N. Alias[1], M. S. S. Mohamed[2], A. R. Abdullah[3]
[1]Department of Mathematics, Science Faculty,
Universiti Teknologi Malaysia, Skudai, Johor.
email: norm_ally@hotmail.com

[2]Department of Engineering
Universiti Tenaga Nasional , Kajang,
email: Sallehs@uniten.edu.my

[3] Department of Industrial Computing,
FTSM, UKM, Bangi, SEL.
email: ara@ftsm.ukm.my

Abstract

In Evans & Sahimi (1989) the discretization of parabolic partial differential equation is derived from Iterative Alternating Decomposition Explicit Method (IADE). Six strategies of parallel algorithms for IADE were found to be more effective using a cluster of workstations (Alias, Sahimi & Abdullah ,2002). In this review paper, we consider some important developments and trends in algorithm design for IADEI (Sahimi, et. al 2003), concentrating on aspects that involve the modification of ADI scheme. IADEI was found to be more convergent and accurate compared to IADE. The absorption of the several parallel strategies for IADEI has been developed to be run on PC cluster systems based on Parallel Virtual Machine environment (Geist .el, 1994). This paper surveys how the cost communication affected the parallel strategies. The analysis of the proposed strategies demonstrates that parallelism is limited by using explicit blocks technique. The elements of explicit blocks were expressed as sub-blocks. These schemes can be effective in reducing data storage accesses on a distributed parallel computer systems. The experiments were run on the homogeneous PC cluster system, which contains of 20 Intel Pentium IV CPUs, each with a storage of 20GB and speed of 1.6Mhz, connected with internal network Intel 10/100 NIC under RedHat Linux 7.2 operation and using message-passing libraries, PVM 3.4. The results of some computational experiments and the parallel performance measurements of the parallel strategies will be discussed.

Keywords. Iterative Alternating Decomposition Explicit Method (IADE), A New

Iterative Alternating Decomposition Explicit Method (IADEI), Parabolic equation, Parallel Virtual Machine (PVM).

1. INTRODUCTION

The model problem under consideration is one dimensional Parabolic equation (Smith, 1979).

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 \leq x \leq 1, \quad 0 < t, \tag{1}$$

with initial condition, $U(x,0) = f(x), \quad 0 \leq x \leq 1,$.
Boundary condition, $U(0,t) = g(t), \quad 0 < t \leq T$ and $U(1,t) = h(t), \quad 0 < t \leq T$.
Where $f(x) = sin(\pi x), \quad 0 \leq x \leq 1$, and $g(t) = h(t) = 0, \quad 0 < t \leq 1,$.
and subject to the exact solution of equation, $U(x,t) = e^{-\frac{2t}{\pi}} sin(\pi x)$
Lets Ω be the domain of $0 \leq x \leq 1$ and $0 < t \leq 1$ with mesh $\Delta x = \frac{1}{m} = h$.
Equation 1 is discretized by the finite difference formula and leads to the generalised approximation stencil as,

$$\begin{aligned} & -\lambda\theta u_{i-1,j+1} + (1 + 2\lambda\theta) u_{i,j+1} - \lambda\theta u_{i+1,j+1} \\ = \; & \lambda(1-\theta) u_{i-1,j} + [1 - 2\lambda(1-\theta)] u_{i,j} \\ + \; & \lambda(1-\theta) u_{i+1,j} \end{aligned} \tag{2}$$

where $i = 1, 2, 3, ..., m$, $j = 1, 2, 3, ..., T$ and $\lambda = \frac{\Delta t}{(\Delta x)^2}$ leads to the a large system of equation,

$$\mathbf{A}u = \mathbf{f} \tag{3}$$

The ADI technique known as a new of Iterative Alternating Decomposition Explicit Method (IADEI) is applied to linear system in equation (3). The sequential and the parallel strategies of IADEI algorithms are described in detail and the numerical results are presented.

2. A NEW ITERATIVE ALTERNATING DECOMPOSITION EXPLICIT METHOD

The New Iterative Alternating Decomposition Explicit Method (IADEI) is based on Iterative Alternating Decomposition Explicit (Sahimi, 1989) and the modification of matrix **A**. Another stable and (4,2) accurate difference replacement of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{4}$$

is as follows,

$$\begin{aligned}
&(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_x^2)(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2)u_{i,j,k+1}\\
&=\frac{2}{3}\lambda(\delta_x^2+\delta_y^2+\frac{1}{2}\delta_x^2\delta_y^2)u_{i,j,k}+(1+(\frac{1}{12}+\frac{2}{3}\lambda)\\
&\delta_x^2+\delta_y^2))u_{i,j,k-1}+(\frac{1}{12}-\frac{2}{3}\lambda)^2\delta_x^2\delta_y^2(2u_{i,j,k}-u_{i,j,k-1})
\end{aligned} \tag{5}$$

whose ADI analogue is given by

$$\begin{aligned}
&(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_x^2)u_{i,j,k+1/2}\\
=\;&-(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2(2u_{i,j,k}-u_{i,j,k-1})+\frac{2}{3}\lambda(\delta_x^2+\delta_y^2+\frac{1}{2}\delta_x^2\delta_y^2)u_{i,j,k}\\
&+(1+\frac{1}{12}+\frac{2}{3}\lambda)\delta_x^2+\delta_y^2))u_{i,j,k-1}
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
&(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2)u_{i,j,k+1}\\
=\;&u_{i,j,k+1/2}+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2(2u_{i,j,k}-u_{i,j,k-1})
\end{aligned} \tag{7}$$

As the temperature reaches steady state over time, the parabolic equation 4 reduce to elliptic equation (Laplace's equation). Whose numerical solution can be obtained iteratively using ADI technique,

$$\begin{aligned}
(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_x^2)u_{i,j}^* &= (-(\frac{1}{12}-\frac{2}{3}r)\delta_y^2+\frac{2}{3}r(\delta_x^2+\delta_y^2+\frac{1}{2}\delta_x^2\delta_y^2)\\
&+ (1+(\frac{1}{12}+\frac{2}{3}\lambda)\delta_x^2+\delta_y^2))u_{i,j}^p
\end{aligned} \tag{8}$$

and

$$(1+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2)u_{i,j}^{(p+1)}=u_{i,j}^*+(\frac{1}{12}-\frac{2}{3}\lambda)\delta_y^2u_{i,j}^p \tag{9}$$

Where r is the acceleration parameter. By considering the two-step iterates corresponding to equations 8 and 9, IADEI formulas are as follows,

at time level $(k+\frac{1}{2})$

$$(I+\alpha\mathbf{G}_1)u^{(k+\frac{1}{2})}=(I+(\alpha+2r)\mathbf{G}_1)(I+2r\mathbf{G}_2)+\beta\mathbf{G}_1\mathbf{G}_2)u^{(k)}-2r\mathbf{f} \tag{10}$$

at time level $(k+1)$

$$(I+\alpha\mathbf{G}_2)u^{(k+1)}=u^{(k+\frac{1}{2})}+\alpha\mathbf{G}_2u^{(k)} \tag{11}$$

with $\alpha=\dfrac{1}{12}-\dfrac{2}{3}r$ and $\beta=\dfrac{2r(3\nu-2r)}{3}$

The coefficient matrix **A** in equation (3) however is decomposed into,

$$\mathbf{A} = \mathbf{G}_1 + \mathbf{G}_2 + \frac{1}{6}\mathbf{G}_1\mathbf{G}_2 \tag{12}$$

The discretization of IADEI method is obtained in the implicit and explicit forms as follows,

i. at time level $(k+\frac{1}{2})$

$$\begin{aligned} u_i^{(k+\frac{1}{2})} &= \frac{1}{A}(s_{i-1}u_{i-1}^{(k)} + v_i u_i^{(k)} + ssu_{i+1}^{(k)} - w_{i-1}u_{i-1}^{(k+\frac{1}{2})} - DDf_i), \\ & i = 1, 2, ..., m \end{aligned}$$

$$s_0 = v_0 = w_0 = 0 \text{ and } A \neq 0, \quad \forall i \in [1, m]. \tag{13}$$

ii. At time level $(k+1)$

$$\begin{aligned} u_{m+1-i}^{(k+1)} &= \frac{1}{1+d_{m+1-i}}(u_{m+1-i}^{(k+\frac{1}{2})} + d_{m+1-i}u_{m+1-i}^{(k)} + ddu_{m+2-i}^{(k)} - ddu_{m+2-i}^{(k+1)}) \\ &\text{with } d_i \neq 0 \text{ and } \forall i \in [1, m] \end{aligned} \tag{14}$$

3. PARALLEL STRATEGIES

The sequential algorithm for IADEI shown that the approximation solution for $u_i^{(k+\frac{1}{2})}$ is dependent on $u_{i-1}^{(k+\frac{1}{2})}$ and the approximation solution for $u_{m+1-i}^{(k+1)}$ is dependent on $u_{m+2-i}^{(k+1)}$. To avoid dependently situation, some parallel strategies are developed to create the non-overlapping subdomains for domain Ω.

3.1. IADEI_SUB

The strategy of Incomplete Block LU preconditioners is slightly non-overlapping subdomains, the domain Ω is decomposed into p processors with incomplete subdomain $\overline{\Omega}$. This strategy implemented the incomplete factorization with parameter β of algebraic boundary condition as follows,

i. at time level $(k+\frac{1}{2})$

$$\begin{aligned} &Au_{i-1}^{(k+\frac{1}{2})} - s_{i-1}u_{i-1}^{(k)} - v_{i-1}u_{i-1}^{(k)} - ssu_i^{(k)} + \beta w_{i-2}u_{i-2}^{(k+\frac{1}{2})} - \beta w_{i-1}u_{i-1}^{(k+\frac{1}{2})} \\ &= -DDf_{i-1}),\ i\ \epsilon\ \overline{\Omega}. \end{aligned}$$

ii. at time level $(k+1)$

$$\begin{aligned} &(1+d_{i+1})u_{i+1}^{(k+1)} - (u_i^{(k+\frac{1}{2})} + d_{i+1}u_{i+1}^{(k+1)} + ddu_{i+2}^{(k)} - ddu_{i+2}^{(k+1)}) = 0 i\ \epsilon\ \overline{\Omega}. \\ &i\ \epsilon\ \overline{\Omega}. \end{aligned}$$

3.2. IADEI_RB

On IADEI_RB strategy, the domain Ω is decomposed into two different subdomains Ω^H and Ω^M. Ω^H is the approximate solution on the odd grids and Ω^M is the approximate solution on the even grids. Computation on Ω^H is executed followed by Ω^M. These two subdomains are not dependent on each other. Ω^H is decomposed into groups, $H_1, H_2, ..., H_{\frac{m}{p}}$ and Ω^M is decomposed into groups, $M_1, M_2, ..., M_{\frac{m}{p}}$. Every group of H_i or M_i, $i = 1, 2, ..., \frac{m}{p}$ is assigned to processors p . IADEI_RB is run in parallel for each subdomain in alternating way on time steps $(k + \frac{1}{2})$ and $(k + 1)$. The parallel strategy of IADEI_RB for equations 13 and 14 are as follows, i) at time level $(k + \frac{1}{2})$

$$\begin{aligned} u_i^{(k+\frac{1}{2})} &= u_i^{(k+\frac{1}{2})} + \frac{\omega}{A}(-Au_i^{(k+\frac{1}{2})} + s_{i-1}u_{i-1}^{(k)} + v_i u_i^{(k)} + ssu_{i+1}^{(k)} - w_{i-1}u_{i-1}^{(k+\frac{1}{2})} \\ &\quad -DDf_i), i \in \Omega^R \\ u_i^{(k+\frac{1}{2})} &= u_i^{(k+\frac{1}{2})} + \frac{\omega}{A}(-Au_i^{(k+\frac{1}{2})} + s_{i-1}u_{i-1}^{(k)} + v_i u_i^{(k)} + ssu_{i+1}^{(k)} - w_{i-1}u_{i-1}^{(k+\frac{1}{2})} \\ &\quad -DDf_i), i \in \Omega^H \end{aligned}$$

ii) at time level $(k + 1)$

$$\begin{aligned} u_{m+1-i}^{(k+1)} &= \frac{\omega}{1 + d_{m+1-i}}(-(1 + d_{m+1-i})u_{m+1-i}^{(k+1)} + u_{m+1-i}^{(k+\frac{1}{2})} + d_{m+1-i}u_{m+1-i}^{(k)} + ddu_{m+2-i}^{(k)} \\ &\quad -ddu_{m+2-i}^{(k+1)}), i \in \Omega^R \\ u_{m+1-i}^{(k+1)} &= \frac{\omega}{1 + d_{m+1-i}}(-(1 + d_{m+1-i})u_{m+1-i}^{(k+1)} + u_{m+1-i}^{(k+\frac{1}{2})} + d_{m+1-i}u_{m+1-i}^{(k)} + ddu_{m+2-i}^{(k)} \\ &\quad -ddu_{m+2-i}^{(k+1)}), i \in \Omega^H \end{aligned}$$

3.3 IADEI_SOR

Using the well-known fact of the IADEI_RB, the parallel algorithm for IADE_CG_SOR takes the form similar to IADEI_RB. The acceleration parameter ω was chosen to provide the most rapid convergence.

3.4 IADEI_MULTI

Multicoloring technique has been used extensively for the solution of the a large-scale problems of linear system of equations $\mathbf{Ax} = \mathbf{b}$ on parallel and vector computer (Ortega, 1987). By the definition of multidomain, domain Ω is decomposed into w different groups. IADEI_MULTI is an advanced concept of IADEI_RB. Typically, one chooses the minimum number of colors w so that the coefficient matrix takes

the block form. In particular. If $w = 2$, then IADEI_MULTI is the IADEI_RB. The Domains for colors $1, 2, 3, ..., w$ are noted as $\Omega^{w_1}, \Omega^{w_2}, ..., \Omega^{w_w}$. The subdomains Ω^{w_i} are distributed into different groups of grid $W_{i1}, W_{i2}, ..., W_{i\frac{m}{wp}}$, where $i = 1, 2, ..., w$. In the process of assignment, $W_{ij}, i = 1, 2, ..., w$ and $j = 1, 2, ..., \frac{m}{wp}$ are mapped into the processors p in the alternating way.

At each time step, the computational grid for domain Ω started its execution with Ω^{w_1}, followed by Ω^{w_2} and ends with Ω^{w_w}. The IADEI_MULTI allows the possibility of a high degree of parallelism and vectorization. However, IADEI_MULTI, as opposed to the natural ordering, may have a deleterious effect on the rate of convergence.

3.5. IADEI_VECTOR

The parallel strategy of IADEI_VECTOR is implemented in two convergence sections. The first section is at time level $(k+\frac{1}{2})$ and the second section is time level $(k+1)$. This method converges if the inner convergence criterion is achieved for each section. The inner convergence criterions $\varepsilon^{(k+\frac{1}{2})}$ and $\varepsilon^{(k+1)}$ are definite global convergence criterion ε.

3.6 IADEI_Michell-Fairweather

The IADEI_Michell-Fairweather (IADEI_MF) which is fully explicit is derived to produce the approximation of grid-i and which is not totally dependent on the grid $(i-1)$ and $(i+1)$. The approximation at the first and second intermediate levels are computed directly by inverting $(rI + \mathbf{G}_1)$ and $(rI + \mathbf{G}_2)$. The explicit form of equation (10) and (11) are given by,
i. at time level $(k+\frac{1}{2})$,

$$
\begin{aligned}
u_i^{(k+\frac{1}{2})} &= \sum_{l=1}^{i-1} \frac{(-1)^T v^{i-l-1} \prod_{j=l}^{i-2} k_j (Ek_{i-1} + Ge_{i-1}k_{i-1})}{\prod_l^{i-1} A} u_l^{(k)} \\
&\quad \sum_{l=1}^{i} \frac{(-1)^H v^{i-l} \prod_{j=l}^{i-1} k_j (J + He_i + G(k_{i-1}h + e_i)}{\prod_l^{i} A} u_l^{(k)} \\
&\quad \sum_{l=1}^{i+1} \frac{(-1)^T v^{i-k+l} \prod_{j=l}^{i} k_j (Hh + Gh)}{\prod_l^{i+1} A} u_l^{(k)} + \sum_{l=1}^{i} \frac{(-1)^H v^{i-k} \prod_{j=l}^{i-1} k_j f_k}{\prod_l^{i} A} u_l^{(k)}
\end{aligned}
$$

$$
H = \begin{cases} l+1, & i = 1, 3, 5, ..., m-1 \\ l, & i = 2, 4, 6, ..., m \end{cases} \quad T = \begin{cases} l, & i = 1, 3, 5, ..., m-1 \\ l+1, & i = 2, 4, 6, ..., m \end{cases}
$$

with $L = 2r$, $E = \alpha + L$, $G = L(\alpha + L) + \beta$ dan $J = 1 + \alpha + L$.

at time level $(k+1)$,

$$
\begin{aligned}
u_m^{(k+1)} &= \frac{\alpha e_m}{t_m} u_m^{(k)} + \frac{u_m^{(k+\frac{1}{2})}}{t_m} \\
u_i^{(k+1)} &= \frac{\alpha e_i}{d_i} u_i^{(k)} + \alpha \sum_{l=i}^{m-1} \frac{(-1)^H \prod_{j=i}^{l} h}{\prod_{j=i}^{l} t_j} u_{l+1}^{(k)} + \alpha \sum_{l=i}^{m-1} \frac{(-1)^T \prod_{j=i}^{l} h e_{l+1}}{\prod_{j=i}^{l+1} t_j} u_{l+1}^{(k)} \\
&+ \sum_{l=i}^{m-1} \frac{(-1)^T h_j^{l-i+1}}{\prod_{j=i}^{l+1} t_j} u_{l+1}^{(k+\frac{1}{2})}
\end{aligned}
$$

$$
H = \begin{cases} l+1, & i = 1,3,5,...,m-1 \\ l, & i = 2,4,6,...,m \end{cases} \quad T = \begin{cases} l, & i = 1,3,5,...,m-1 \\ l+1, & i = 2,4,6,...,m \end{cases}
$$

4. IMPLEMENTATION ON MESSAGE PASSING SYSTEMS

All the parallel strategies are based on the non-overlapping subdomain. There are no data exchange between the neighboring processors at the iteration (k) but there are inter-processor communications between the iteration (k) and the next iteration $(k+1)$. A typical parallel implementation of a parallel IADEI assigns several mesh points to each processor p such that each processor only communicates with its two nearest neighbors.The computations of the approximation solutions in subdomain Ω^p are executed independently.

The stopping criteria in the processors p is investigated by measuring the size of the inner residuals. Let us define the residual computed in the processors p as $R^p(k) = max\{\mid u_{i,j}^{(k+1)} - u_{i,j}^{(k)} \mid, (i,j) \in p\}$. This quantity is kept in the processor's memory between successive iterations and it is checked if the residual is reduced by the convergence criterion $\epsilon^p = 1.0 \times 10^{-15}$. The master processor checked the maximum of $R^p(k)$ and the iteration stopped when the global convergence criterion is met.

5. NUMERICAL RESULTS

Table 1 shows that the sequential performance of IADEI is better than IADE in terms of time execution and number of iterations for all cases. The results obtained for the various parallel strategies of IADEI in table 2. The worst performances are shown by IADEI_MF and IADEI_VECTOR. The sequential IADEI is better in accuracy and convergence than all the parallel strategies of IADEI. In comparison with the parallel strategies of IADEI, these results also show that the time execution for IADEI_SUB was about 2 times shorter than other parallel strategies. Furthermore, IADEI_SUB is the best in terms of convergence and accuracy.

Table 1: Sequential performance of IADEI and IADE are based on three cases

method	CASE 1			CASE 2			CASE 3		
	IADEI-IMP	IADEI-CG	IADEI-DG	IADEI-IMP	IADEI-CG	IADEI-DG	IADEI-IMP	IADEI-CG	IADEI-DG
time	487	442	435	671	589	570	692	596	609
iteration	111	103	100	147	131	127	159	136	134
rmse	2.89E-03	1.81E-03	3.76E-06	5.00E-03	1.22E-03	2.85E-05	1.58E-03	1.95E-04	1.99E-05
r.maxs	9.08E-06	3.57E-06	2.21E-11	2.82E-05	1.68E-06	9.37E-10	2.82E-06	4.29E-08	4.52E-10
ave.rmx	3.76E-03	2.36E-03	8.79E-06	7.12E-03	1.74E-03	4.11E-05	2.25E-03	2.78E-04	2.86E-05
|r|	2.51E-05	2.83E-05	5.16E-05	5.09E-05	3.52E-05	2.99E-05	1.86E-05	8.38E-06	7.64E-06
lam.	0.1	0.1	0.1	0.5	0.5	0.5	1.0	1.0	1.0
t	0.05	0.05	0.05	0.25	0.25	0.25	0.5	0.5	0.5
del.t	0.001	0.001	0.001	0.005	0.005	0.005	0.01	0.01	0.01
del.x	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
r	0.820	0.813	0.818	0.830	0.830	0.830	0.800	0.820	0.830
exp	1.00E-04	1.00E-04	1.00E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04

method	CASE 1			CASE 2			CASE 3		
	IADE-IMP	IADE-CG	IADE-DG	IADE-IMP	IADE-CG	IADE-DG	IADE-IMP	IADE-CG	IADE-DG
time	590	477	438	678	621	606	900	598	544
iteration	150	107	103	165	150	149	231	150	138
rmse	2.89E-03	1.80E-03	1.43E-05	4.99E-03	1.22E-03	2.82E-05	1.53E-03	1.92E-04	1.24E-05
r.maxs	9.07E-06	3.53E-06	2.87E-10	2.81E-05	1.68E-06	9.35E-10	2.62E-06	4.19E-08	1.72E-10
ave.rmx	3.76E-03	2.35E-03	2.67E-05	7.10E-03	1.74E-03	4.15E-05	2.16E-03	2.74E-04	1.75E-05
|r|	1.46E-05	5.15E-05	4.01E-05	1.96E-05	5.68E-06	9.99E-05	4.40E-05	2.72E-05	1.84E-05
lam.	0.1	0.1	0.1	0.5	0.5	0.5	1.0	1.0	1.0
t	0.05	0.05	0.05	0.25	0.25	0.25	0.5	0.5	0.5
del.t	0.001	0.001	0.001	0.005	0.005	0.005	0.01	0.01	0.01
del.x	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
r	0.960	1.140	1.240	0.720	0.920	0.980	2.700	0.770	0.700
exp	1.00E-04	1.00E-04	1.00E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04

rmse= root means square error, |r|= absolute error, r.maxs = maximum error and ave_rmx = average of rmse

This paper presents the numerical properties of the parallel solver on the homogeneous architecture which contains of 20 Intel Pentium IV CPUs, each with a storage of 20GB and speed of 1.6Mhz, connected with internal network Intel 10/100 NIC under RedHat Linux 7.2 operation and using message-passing libraries, PVM 3.4. The following definitions are used to measure the parallel strategies, speedup $S_p = \frac{T_1}{T_p}$, efficiency $C_p = \frac{S_p}{p}$ effectiveness $F_p = \frac{S_p}{C_p}$ and temporal performance $L_p = T_p^{-1}$. Where T_1 is the execution time on one processor, T_p is the execution time on p processors and the unit of L_p is work done per micro second. Parallel Gauss Seidel Red Black is chosen as the control scheme. The graph of the execution time, speedup, efficiency, effectiveness and the temporal performance versus number of the processors were plotted in Figures 1, 2, 3, 4 and 5. The algorithm of parallel strategies with the highest performance executed in the least time and is therefore the best algorithm. As expected, Figure 1 shows the execution time decreases with the increasing p. IADEI_SUB strategy is found to give the best performance because of the minimum memory access and data sharing. Figure 2 illustrates that

Table 2: Performance measurements of the parallel strategies of IADEI methods

$m = 720003$, $\Delta x = 1.3889E^{-6}$, $\Delta t = 9.6450E^{-13}$, level=50 $t = 4.8225E^{-11}$ $\lambda = 0.5$, $\theta = 1.00$, $\epsilon = 1.0E^{-15}$

IADEI	SQENT	SUB	SOR	RB	MULTI	VEKT0R	MF	GBRB
t_{para}	43.0972	45.5205	103.873	114.8103	198.0183	291.0482	2106.659	156.4432
iterat.	261	301	358	360	360	261	261	600
rmse	$1.5921E^{-9}$	$1.5921E^{-9}$	$1.5567E^{-9}$	$1.5921E^{-9}$	$1.5921E^{-9}$	$1.5921E^{-9}$	$1.5921E^{-9}$	$1.5921E^{-9}$
|r|	$6.6613E^{-16}$	$9.9920E^{-16}$	$6.1062E^{-16}$	$6.1062E^{-16}$	$8.8818E^{-12}$	$9.9920E^{-16}$	$6.6613E^{-16}$	$1.1123E^{-16}$
ave_rmse	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$	$1.9846E^{-7}$
r.maxs	$5.3374E^{-17}$	$5.3374E^{-17}$	$5.3374E^{-17}$	$5.3314E^{-17}$	$5.3374E^{-17}$	$5.3374E^{-17}$	$5.3374E^{-17}$	$5.3374E^{-17}$
r	0.74	0.7	0.54	0.55	0.56	0.73	0.74	–
ω_y	–	–	1.01	1.0	1.02	–	–	–
ω_z	–	–	1.0	1.0	0.94	–	–	–
L_y	–	–	–	–	–	1036	–	–
L_z	–	–	–	–	–	987	–	–
ϵ_y	–	–	–	–	–	$1.0E^{-14}$	–	–
ϵ_z	–	–	–	–	–	$1.0E^{-15}$	–	–

rmse= root means square error, |r|= absolute error, r.maxs = maximum error and ave_rmse = average of rmse

at $p = 20$ processors, all the parallel strategies of IADEI yield approximately equal performance in speedup.

Figure 3 shows that the efficiency of IADEI_MF and IADEI_VECTOR strategies are decreased drastically. This is the result of the additional overhead imposed by having communications routed though the PVM daemon with high number of iterations.

From table 2 and figure 4, the results have shown that the effectiveness of IADEI_SUB is superior than the IADEI_SOR, IADEI_RB and IADEI_MULTI for all numbers of processors. Usually, the temporal performance is used to compare the performance of different parallel algorithms. Figure 5 has shown that the temporal performance of the parallel strategies as in the following order,
1. IADEI_LU 2. IADEI_SOR 3. IADEI_RB 4. IADEI_MULTI 5. IADEI_VECTOR 6. IADEI_MF

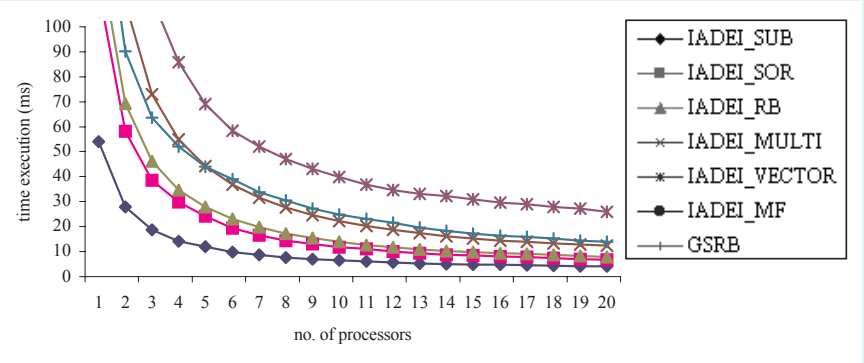


Figure 1: The execution time vs. number of processors

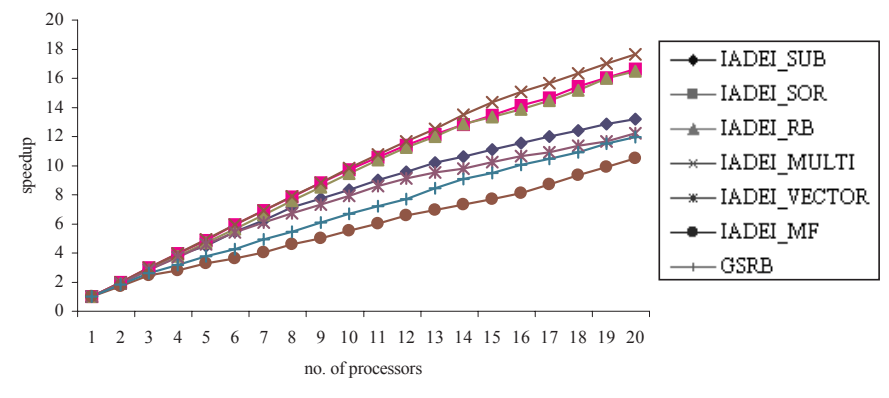


Figure 2: The speedup vs. number of processors

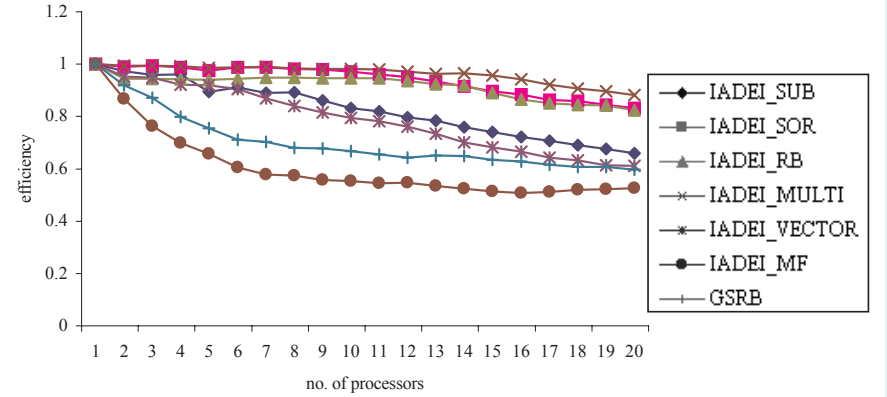


Figure 3: The efficiency vs. number of processors

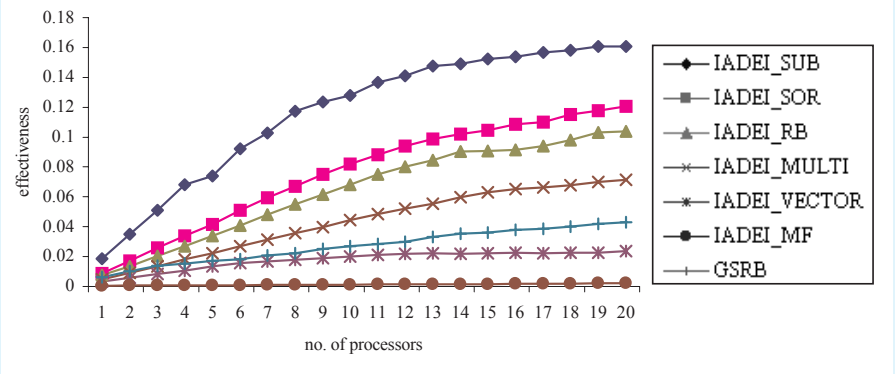


Figure 4: The effectiveness vs. number of processors

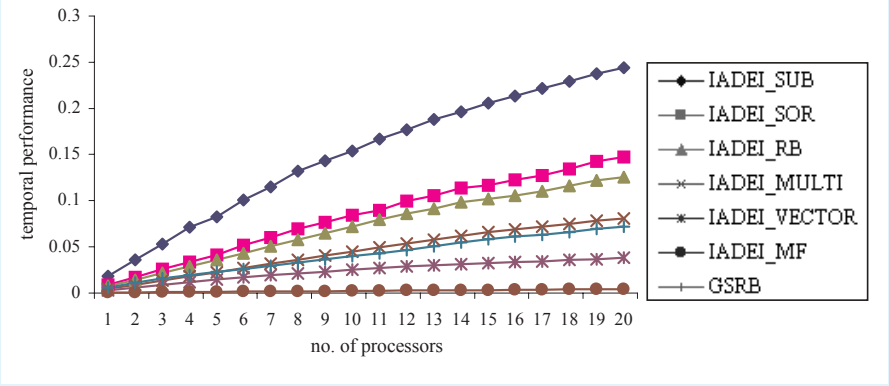


Figure 5: Temporal performance vs. number of processors

Table 3: Computational complexity per iteration for the parallel strategies of IADEI methods

method	c.computation	c.communication
SUB	$(\frac{2408m}{p} + 4214)T + (\frac{3010m}{p} + 4816)D$	$3010t_{data} + 2408(t_{start} + t_{idle})$
SOR	$\frac{3580mT}{p} + \frac{5012mD}{p}$	$5012t_{data} + 2864(t_{start} + t_{idle})$
RB	$\frac{3600mT}{p} + \frac{5040mD}{p}$	$5040t_{data} + 2880(t_{start} + t_{idle})$
MULTI	$\frac{3600mT}{p} + \frac{4680mD}{p}$	$5762t_{data} + 3600(t_{start} + t_{idle})$
VECTOR	$\frac{2112012mT}{p} + \frac{2652804mD}{p}$	$2113584t_{data} + 1057050(t_{start} + t_{idle})$
MF	$261(\frac{pD}{\sum_{i=1}^{p} i} + 2m - \frac{3m}{p}$ $+\frac{pT}{\sum_{i=1}^{p} i} + 14p - 14)$	$3136t_{data} + 2610(t_{start} + t_{idle})$
GSRB	$(\frac{1800m}{p} + 1200)T + (\frac{2400m}{p} + 1800)D$	$8400t_{data} + 4800(t_{start} + t_{idle})$

c.computational=computational complexity , c.communication= communications cos, D=multiplications, T=additions

6. CONCLUDING REMARKS

This paper has outlined the parallel strategies of a numerical schemes in class of iterative and explicit two-steps methods to solve one dimensional he equations. As the basic of derivation is the unconditionally stable (4,2) accurate IADEI scheme, the parallel strategies of IADEI are convergent and computationally stable. A comparison with the parallel strategies of IADEI scheme, shows that the IADEI_SUB has extended range of efficiency, speedup and effectiveness. Furthermore, the superiority of the IADEI_SUB is also indicated by the highest value of the temporal performance, accuracy and convergence in solving a large-scale linear algebraic equations on a PC cluster system. Because of message latency, load balancing, data-sending and number of iteration, the communication times of IADEI_SUB is lowest than other strategies.

ACKNOWLEDGMENT

We wish to express our gratitude and indebtedness to the Universiti Kebangsaan Malaysia, Universiti Tenaga Nasional and Malaysian Government for providing the moral and financial support under IRPA grant for the successful completion of this project.

REFERENCES

[1] Alias, N., Sahimi, M.S., and Abdullah, A.R., The AGEB Algorithm for Solving the Heat Equation in Two Space Dimensions and Its Paralleization on a Distributed Memory Machine. Proceedings of the 10^{th} European PVM/ - MPI User's Group Meeting: *Recent Advances In Parallel Virtual Machine and Message Passing Interface.* 7 (2003) 214–221

[2] Alias, Sahimi., and Abdullah, A., Parallel Algorithms On Some Numerical Techniques Using PVM Plat form On A Cluster Of Workstations. *Proceedings of the* 7^{th} *Asian Technology Conference in Mathematics* 7 (2002) 390–397.

[3] Al Geist, Al., Beguelin, A., Dongarra, J., Jiang, W., Manchek, R. and Sunderam, V., *PVM : Parallel Virtual Machine, A Users' Guide and Tutorial for Networked Parallel Computing.* Cambridge: MIT Press 1994.

[4] Chalmer, A. and Tidmus, J. *Practical Parallel Processing - An Introduction to Problem Solving in Parallel.* International Thomson Computer Press 1996.

[5] Evans, D. J. and Sahimi, M. S., The alternating Group Explicit Iterative method (AGE) to Solve Parabolic and Hyperbolic Partial Differential Equation. *In Annual Review of Numerical Fluid Mechanics and Heat Transfer* 2 (1989) 283–389.

[6] Evans, D. J. and Sahimi, M. S., The alternating Group Explicit (AGE) Iterative method for Solving Parabolic Equations II: 3 Space Dimensional Problems. *International Journal Computer Mathematic* 26 (1989) 117–142.

[7] Evans, D. J. and Sahimi, M. S. The alternating Group Explicit (AGE) Iterative method for Solving Parabolic Equations I. *International Journal Computer Mathematic* 24 (1988) 127–145.

[8] Foster, I. *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering.* Inc: Addison- Weslley Publishing Company 1996.

[9] Hageman, L. A. and Young, D. M. *Applied Iterative Methods.* Academic Press, New York 1981.

[10] Lewis, T.G. and El-Rewini, H., *Distributed and Parallel Computing.* New York: Manning Publication 1998.

[11] Ortega, J. M. and Harrar, D. I., Multicoloring With Lots Of Colors. Proceedings Of The 3rd International Conference On Supercompting New York. ACM Press: 1–6 (1986) 1987.

[12] Sahimi, M. S., Alias, N., Mansor, N.A, and Nor, M. N., Parabolic-Elliptic Correspondence of a Three Level Finite Difference Approximation to Heat Equation. *Bulletin of the Malaysian Mathematical Sciences Society.* 26 (2003) 79-85.

[13] Wilkinson, B. and Allen M., *Parallel Programming: Techniques and Applications Using Networked workstations and Parallel Computers.* New Jersey : Prentice Hall, 1999.

[14] Zamoya, A. Y. *Parallel and Distribution Computing Handbook.* New York: Mc Graw-Hill, 1996.

A NEW NONLINEAR DIFFERENTIAL EQUATION FOR DESCRIBING ION-ACOUSTIC WAVES IN PLASMA

H. Alinejad[1], J. Mahmoodi[2], S. Sobhanian[1], M. Momeni[1]
[1]Atomic & Molecular Physics, Tabriz University, Tabriz, IRAN
[2]Physics Department, Qom university, Qom, IRAN

Abstract

Using the standard reductive perturbation method, the two dimensional dynamics of nonlinear ion-acoustic waves in plasma comprising cold ions and nonisothermal electrons has been considered. A new nonlinear equation is found which is valid for unmagnetized and magnetized plasmas. For exactly vanishing magnetic fields the Modified Kadomtsev-Petviashvili (MKP) equation is recovered. For weak magnetic fields, however, the dynamics is mainly different from the MKP equation, depending on amplitude. For increasing magnetic field, the new equation is similar (but not identical) to the Modified Zakharov-Kuznetsov (MZK) equation which is fulfilled for very strong magnetic fields.

1. INTRODUCTION

Now a days, the study of nonlinear waves in plasma, especially ion-acoustic waves, has become one of the most important subject of plasma physics, Washimi and Taniuti[1] derived the Korteweg-deVries (KdV) equation in their study of ion-acoustic solitary waves in a cold plasma, using the reductive perturbation method. Without external magnetic field, the behavior of the two dimensional small-amplitude weakly nonlinear ion acoustic wave in a plasma comprising cold ions and isothermal electrons is described by the Kadomtsev-Petviashvili (KP) equation [2]. On the other hand in the presence of a strong external magnetic field and under the same conditions, Zakharov-Kuznetsov (ZK) equation describes ion acoustic waves in plasma [2]. In the presence of resonant electrons, the plasma behaves nonisothermally. Resonant electrons strongly interact with the wave during its evolution and therefore cannot be treated with a Boltzmann distribution $n_e = \exp(\Phi)$ as considered in an isothermal plasma. There are interesting results in the case of a plasma consisting of non-isothermal electrons. Schamel [3,4] was first who considered the effects of nonisothermality of electrons in a plasma and derived Modified Korteweg-deVries (MKdV) equation. So far only in the extreme values of $\Omega_c(\Omega_c = 0, \quad \Omega_c \to \infty)$ with small but finite amplititudes the models have been derived, where $\Omega_c = \frac{B_0}{\sqrt{4\pi n_i m_i c}}$, MKP equation for $\Omega_c = 0$, and MKZ equation for $\Omega_c \to \infty$ [5,6]. In this article we

tried to fill this gap and to derive a new model equation, which is valid for finite Ω_c, which occur in most experiments. This paper is organized as follows. The basic equation governing the non-isothermal plasma model under investigation are given in Section 2. Derivation of MKP and MZK equation are briefly studied by reductive perturbation method in Section 3. The new scaling leading to a new model equation is given in Section 4. Finally, a brief discussion is presented in Section 5.

2. BASIC EQUATIONS

We consider a collisionless magnetized plasma comprising cold ions and hot electrons. In dimensionless form, the ion dynamics are governed by the following system of equations:

$$\partial_t n + \nabla.(nV) = 0, \tag{1}$$

$$\partial_t V + (V.\nabla)V + \nabla\Phi + \Omega_c \hat{x} \wedge V = 0, \tag{2}$$

$$\nabla^2\Phi = n_e - n. \tag{3}$$

where, $n, n_e, u\Phi$ are respectively the non-dimensional ion number density, the electron number density, the ion fluid velocity and the electrostatic potential. The reference density, speed, time, length and electrostatic potential are respectively the unperturbed number density n_0, the ion sound speed $c_s = \left(\frac{K_B T_{ef}}{m_i}\right)^{\frac{1}{2}}, (\omega_{pi})^{-1}, \lambda_{De}$ and $\frac{K_B T_{ef}}{e}$. Here K_B is Boltzmann constant, T_{ef} is the constant temperature of the free electrons, m_i is the ion mass, $\lambda_{De} = \left(\frac{\varepsilon_o K_B T_e}{n_{0e} e^2}\right)^{\frac{1}{2}}$ is the electron Debye length and ε_0 is the vaccum permittivity. We consider the more realistic situation in which the electrons are non-isothermal. In this case the electron number density is replaced by [3,4],

$$n_e = \exp(\Phi) erf\left(\Phi^{\frac{1}{2}}\right) + \beta^{-\frac{1}{2}} \times \left\{ \begin{array}{ll} \exp(\beta\Phi) erf((\beta\phi)^{\frac{1}{2}}) & \beta \geq 0 \\ (2/\pi^{\frac{1}{2}}) W((-\beta\Phi)^{\frac{1}{2}}) & \beta < 0 \end{array} \right\}$$

Where W is the Dawson integral. $\beta = \frac{T_{ef}}{T_{et}} \neq 1$ and T_{et} is the constant temperature of the trapped electrons. For weekly nonlinear waves the electron number density becomes

$$n_e = 1 + \Phi - \frac{4}{3} b \Phi^{\frac{3}{2}} + \frac{1}{2}(\Phi^2) \tag{4}$$

where $b = (1-\beta)\pi^{-\frac{1}{2}}$, measures the deviation from isothermality. We assume that $b > 0$, which is suggested by experiment [3], and this term is the contribution of

the resonant electrons to the electron density. $b = 0$ if electrons are isothermal and resonant effects are absent.

3. DERIVATION OF MKP AND MZK EQUATION

To derive the Modified Kadomtsev-Petviashvili equation, we use the standard reductive perturbation method and in order to find a suitable choice of scaling for the independent variable, we use a linear dispersion argument, similar to the one used by Infeld and Rowlands (1990, Appendix 1) in their derivation of the ZK equation. According, we choose the following scaling for the independent variables:

$$\xi = \varepsilon^{\frac{1}{4}}(x - t) \qquad , \qquad \sigma = \varepsilon^{\frac{1}{2}} y \qquad , \qquad \tau = \varepsilon^{\frac{3}{4}} t \tag{5}$$

From the basic Eqs. (1)-(3) and (4), we expand the density, fluid velocities and electrical potential asymptotically by a smallness parameter ε as

$$\begin{aligned} n &= 1 + \varepsilon n^{(1)} + \varepsilon^{\frac{3}{2}} n^2 + ... \\ \Phi &= \varepsilon \Phi^{(1)} + \varepsilon^{\frac{3}{2}} \Phi^{(2)} + ... \\ \nu_x &= \varepsilon \nu_x^{(1)} + \varepsilon^{\frac{3}{2}} \nu^{(2)} + ... \\ \nu_y &= \varepsilon^{\frac{1}{4}} (\varepsilon \nu_y^{(1)} + \varepsilon^{\frac{3}{2}} \nu_y^{(2)} + ...) \end{aligned} \tag{6}$$

For $\Omega_c = 0$ and using the scaling above, the basic equation is simplified to the Modified Kadomtsev-Petviashili equations [5].

$$\left[\partial_\tau n + b n^{\frac{1}{2}} \partial_\xi n + \frac{1}{2} \partial_\xi^3 n \right]_\xi + \frac{1}{2} \partial^2 \sigma n = 0 \tag{7}$$

Here, we have set $n^1 = n$. One dimensionally ($\partial_\sigma = 0$) this equation is identical to the Modified Korteweg-deVries (MKdV) equation. It was shown that the MKP equation has the stationary plane soliton solutions [5].

On the other hand, for $\Omega_c \sim 1$, the MZK equation can be derived. In the presence of an external magnetic field we replace in (5) and (6) the following variables

$$\sigma = \varepsilon^{\frac{1}{4}} y \qquad , \qquad \eta = \varepsilon^{\frac{1}{4}} z \qquad , \qquad v_\perp = \varepsilon^{\frac{5}{4}} \nu_\perp^{(1)} + \varepsilon^{\frac{3}{2}} v_\perp^{(2)} + ... \tag{8}$$

The longitudinal dynamics and therefore the longitudinal scaling is unchanged compared with (5),(6). Straightforward but lengthy calculations lead to the MZK equation

$$2\partial_\tau n + 2b n^{\frac{1}{2}} \partial_\xi n + \partial_\xi^3 n + (1 + \Omega^{-2} c) \partial_\xi (\partial_\sigma^2 + \partial_\eta^2) n = 0 \tag{9}$$

As discussed in reference [6], the soliton solutions are stable with respect to transverse perturbations when $\Omega_c = 0$, on the other hand, for $\Omega_c \sim 1$ soliton solutions are unstable.

4. SCALING FOR WEAK MAGNETIC FIELDS

In Sec. 3 we have shown that in the two limits $\Omega_c = 0$ and $\Omega_c \sim 1$ qualitatively different results for the two-dimensional dynamics are obtaining most particle applications are in the region where $\Omega_c \ll 1$; the behavior in that region is described correctly neither by the MKP equation nor by the MZK equation. Therefore, we derive a new equation which is valid in this intermediate region. To derive the new equation, we use the following stretched variables

$$\xi = \varepsilon^{\frac{1}{4}}(x-t) \qquad \tau = \varepsilon^{\frac{3}{4}}t$$

$$\sigma = \varepsilon^{\frac{1}{2}}y \qquad \eta = \varepsilon^{\frac{1}{2}}z \tag{10}$$

together the additional condition, $\Omega_c \sim \varepsilon^{\frac{1}{4}}$. We expect the dependent variables to take the form

$$\begin{aligned} n &= 1 + \varepsilon^{r_{n1}} n^{(1)} + \varepsilon^{r_{n2}} + ... \\ \Phi &= \varepsilon^{r_{\Phi 1}} \Phi^{(1)} + \varepsilon^{r_{\Phi 2}} \Phi^{(2)} ... \\ \nu_x &= \varepsilon^{r_{x1}} \nu_x^{(1)} + \varepsilon^{r_{x2}} \nu_x^{(2)} + ... \\ \nu_\perp &= \varepsilon^{r_{\perp 1}} \nu_\perp^{(1)} + \varepsilon^{r_{\perp 2}} \nu_\perp^{(2)} + ... \end{aligned} \tag{11}$$

From the basic equation and Eqs.(10) and (11), we obtain all powers of variables in Eq.(11). Comparing the various order in ε, we obtain the lowest order

$$\partial_\xi n^{(1)} = \partial_\xi \nu_x^{(1)}, \qquad r_{n1} = r_{x1} \tag{12}$$

$$\partial_\xi v_x^{(1)} = \partial_\xi \Phi^{(1)}, \qquad r_{x1} = r_{\Phi 1}, \tag{13}$$

$$\partial_\xi \nu_y^{(1)} = \partial_\sigma \Phi^{(1)} - \Omega_c \nu_z^{(1)} \quad , \qquad \nu_{\perp 1} = r_{\Phi 1} + \frac{1}{4} \tag{14}$$

$$\partial_\xi \nu_z^{(1)} = \partial_\tau \Phi^{(1)} - \Omega_c \nu_y^{(1)}, \tag{15}$$

$$\Phi^{(1)} = n^{(1)} \qquad r_{\Phi 1} = r_{n1}. \tag{16}$$

the higher-order equations then yield

$$\partial_\tau n^{(1)} - \partial_\xi n^{(2)} + \partial_\xi \nu_x^{(2)} + \partial_\sigma \nu_y^{(1)} + \partial_\eta \nu_z^{(1)} = 0 \qquad r_{x2} = r_{n2} \tag{17}$$

$$\partial_\tau \nu_x^{(1)} + \partial_\xi \Phi^{(2)} - \partial_\xi \nu_x^{(2)} = 0 \tag{18}$$

$$\partial_\xi \nu_y^{(2)} = \partial_\tau \nu_y^{(1)} + \partial_\sigma \Phi^{(2)} - \Omega_c \nu_z^{(2)} \qquad r_{\perp 2} = r_{\perp 1} + \frac{1}{2} \tag{19}$$

$$\partial_\xi \nu_z^{(2)} = \partial_\tau \nu_z^{(1)} + \partial_\eta \Phi^{(2)} + \Omega_c \nu_y^{(2)} \tag{20}$$

$$\partial_\xi^2 \Phi^{(1)} = \Phi^{(2)} - \frac{4}{3} b \Phi^{(1)^{\frac{3}{2}}} - n^{(2)} \quad , \quad r_{\Phi 2} = r_{n2} \tag{21}$$

From Eqs. (12), (13) and (16) we get

$$n^{(1)} = \Phi^{(1)} = \nu_x^{(1)} \tag{22}$$

Taking the derivative of Eq. (21) with respect to

$$\partial_\xi n^{(2)} = \partial_\xi \Phi^{(2)} - 2b\Phi^{(1)^{\frac{1}{2}}} \partial_\xi \Phi^{(1)} - \partial_\xi^3 \Phi^{(1)},$$

and using equation (16) and (17), we obtain from Eq. (18)

$$2\partial_\tau n^{(1)} + 2b\Phi^{(1)^{\frac{1}{2}}} \partial_\xi \Phi^{(1)} + \partial_\xi^3 \Phi^{(1)} + \partial_\sigma \nu_y^{(1)} + \partial_\eta \nu_z^{(1)} = 0 \tag{23}$$

whereas Eqs. (14) and (15) yield.

$$\left(\partial_\xi^2 + \Omega_c^2\right) \partial_\sigma \nu_y^{(1)} = \partial_{\sigma\sigma\xi} \Phi^{(1)} - \Omega_c \partial_{\eta\sigma} \Phi^{(1)} \tag{24}$$

$$\left(\partial_\xi^2 + \Omega_c^2\right) \partial_\eta \nu_z^{(1)} = \partial_{\eta\eta\xi} \Phi^{(1)} + \Omega_c \partial_{\sigma\eta} \Phi^{(1)}$$

We can combine Eqs. (23) and (24) to get

$$2\partial_\tau n + 2bn^{\frac{1}{2}} \partial_\xi^3 n + \partial_\xi^3 n + \left(\partial_\xi^2 + \Omega_c^2\right)^{-1} \partial_\xi \left(\partial_\sigma^2 + \partial_\eta^2\right) n = 0. \tag{25}$$

where we have set $n^1 = n$. Equation (25) is the new nonlinear equation for ion acoustic waves in weak magnetic fields. First for $\partial_\sigma = \partial_\eta = 0$ we recover the MKVdV equation. Secondly for $\Omega_c = 0$, the MKP equation (7) is obtained. Finally for $\Omega_c^2 \gg \partial_\xi^2$, an equation similar to the MZK equation, (9) arises. However it should not be expected that Eq.(9) and (25) have a common region of applicability since they have been derived for complementary Ω_c regions.

5. CONCLUSION

In this paper we have derived a new nonlinear equation using the well-known reductive perturbation method. This nonlinear equation describes ion-acoustic waves in a plasma consisting of cold ions and nonisothermal electrons with weak magnetic fields. The new equation is transformed into the MKP from for $\Omega_c = 0$, and the MZK equation for Ω_c 1. It covers the most important region, where the transition from stable to unstable behavior occurs. The stability of this equation is not investigated in this paper but is under investigation.

REFERENCES

[1] Washimi, H. and Taniuti, T., Phys. Rev. Lett. 17, 996 (1966).

[2] Infeld, E. and Rowlands, G., *"Nonlinear wave, soliton and chaos"*(Cambridge University Press, Cambridge 1990).

[3] Schamel, H., J. Plasma Phys. 14, 905 (1972).

[4] Schamel, H., J. Plasma Phys. 9, 377 (1973).

[5] Okeir, S. and Parkes, J., Physica Scripta, 55, 135 (1997).

[6] Munro, S. and Parkes, J., J. Plasma Phys. 62, 305 (1999).

USING MONTE CARLO METHODS TO EVALUATE SUB-OPTIMAL EXERCISE POLICIES FOR AMERICAN OPTIONS

G. Alobaidi[1] and R. Mallier[2]
[1]Department of Mathematics and Statistics,
American University of Sharjah, Sharjah, United Arab Emirates
e-mail: galobaidi@yahoo.ca

[2]Department of Applied Mathematics
University of Western Ontario, London ON N6A 5B7 Canada
e-mail: rolandmallier@hotmail.com

1. INTRODUCTION

Derivatives are financial instruments which *derive* their value from some other underlying asset; examples include options and futures on equities and equity indices. This paper is concerned with options, which as their name suggests give the holder a choice, carrying the right but not the obligation to buy (or sell) the underlying asset. They have numerous uses, such as speculation, hedging, generating income, and they contribute to market completeness. Although options have existed for far longer, their use has exploded since the occurrence in 1973 of two seminal events in the history of options: the publication of the Black-Scholes option pricing formula [1, 2], enabling investors to price certain options, and the opening of the Chicago Board Options Exchange (CBOE), the first secondary market for options, which gave an investor holding options a marketplace for reselling those options to another investor, in addition to the choices of holding the option to expiry, or exercising early if that was permitted.

Options can be categorized in several ways, one method being by the exercise characteristics. It is fairly easy to value European options, which can be exercised only at expiry, a pre-determined date specified in the option contract. However, American options, which can be exercised at the holder's discretion at or before expiry, are much harder to price, since the early exercise feature necessitates a decision by the holder as to whether and when to exercise such an option, and this is one of the best-known problems in mathematical finance, leading to an optimal exercise boundary and an optimal exercise policy, which if followed will maximize the expected return to the holder and thereby the value of the option. Ideally, an investor would be able to constantly calculate the expected return from continuing to hold the option, and if that is less than the return from immediate exercise, he

should exercise the option. This process would tell the investor the location of the optimal exercise boundary. However, except for one or two very special cases, no closed form solutions are known for the location of the optimal exercise boundary, and in general either numerical solutions or approximations must be used to locate it. Both of these approaches are fairly well-developed, and for a review, the reader is referred to the monographs [3, 4]. Both of these approaches can also be difficult and time-consuming, and whereas an institution can perform those calculations and thereby optimize their return, an individual may well be unable to do this, and instead have his own elementary exercise policy, choosing to exercise the option when certain conditions are met, for example when the value of the option reaches some multiple of the exercise price. We will refer to such an individual as an uninformed investor. The expected return from such sub-optimal strategies will be less than or equal to that when the optimal exercise policy is pursued, indeed hence the term *sub-optimal*.

2. MONTE CARLO SCHEME

In this study, we use a Monte Carlo scheme to look at several such strategies that an amateur investor might follow, and calculate the expected return using each of these strategies. We considered both call options, which give the holder the right to buy the underlying stock at the strike price E, and put options, which carry the right to sell the underlying at the price E. In what follows, S denotes the price of the underlying at time t, while $S_0 = S(t_0)$ is the initial value of S at the time t_0 the option was purchased. In terms of S and S_0, the 8 strategies we used for the call option to exercise the option when:
(1): Never (*i.e.* treat the option like a European).
(2): If S is 110% or more of S_0 (put: $S \leq 0.9\, S_0$).
(3): If S is 115% or more of S_0 and in money (put: $S \leq 0.85\, S_0$).
(4): If S is greater than S_0 and at or in money (put: $S < S_0$).
(5): If S goes down by 10% and still in money (put: $S \geq 1.1 S_0$).
(6): If S goes down by 5% (put: $S \geq 1.05 S_0$).
(7): If S goes down by 10% from its peak and in money (put: S up by 10% from trough).
(8): If S goes up on 5 successive time-steps and is in money (put: down).

In the above, the corresponding strategy used for the put is given in parentheses where it differs from that for the call. We should recall that for the call with no dividends, it is never optimal to exercise, so we would expect strategy (1) to be the best for the call. In strategies (2)-(6), we are treating the option as a barrier

option. The motivation for strategies (2)-(3) was that some investors will exercise when they feel they have made sufficient profit, while that for strategies (5)-(6) was that other investors will be spooked by losses and pull out of the market.

In addition to evaluating the expected return to an investor if he were to follow one of these elementary strategies, we will also look at how the expected return is affected by how often the investor checks to see if his exercise criteria have been met. As we mentioned above, we will tackle this problem with Monte Carlo simulation. This approach is well-suited for this particular problem, since the underlying stock price S is assumed to follow a random walk. The use of Monte Carlo methods for option pricing was pioneered by Boyle [5], and these methods have since become extremely popular because they are both powerful and extremely flexible. Although the use of Monte Carlo methods to value American options is still a nebulous problem, with for example several researchers pursuing Malliavin calculus while others are attempting different approaches, the difficulties with American options stem from the need to locate the optimal exercise boundary, and for the problem studied here, that is not an issue: rather, we are calculating what an option is worth if one of several elementary strategies is followed, and so the location of our (sub-optimal) exercise boundary is fairly simple. Returning to option pricing in general, in this context, Monte Carlo methods involve the direct stochastic integration of the underlying Langevin equation for the stock price, which is assumed to follow a log-normal random walk or geometric Brownian motion. The heart of any Monte Carlo method is the random number generator, and our code employed the Netlib routine RANLIB, which produces random numbers which are uniformly distributed on the range $(0, 1)$ and which were then converted to normally distributed random numbers. This routine was itself based on the article by L'Ecuyer and Cote [6]. Antithetic variables were used to speed convergence, and a large number of realizations were performed to ensure accurate results. Our simulations, including other runs not presented here, required about a month's CPU time on a DEC Alpha and were performed on the Beowulf cluster at the University of Western Ontario.

The starting point of our analysis is the risk-neutral random walk for the price of the underlying S in the absence of dividends,

$$dS/S = rdt + \sigma dX, \tag{1}$$

where dX describes the random walk, dt is the step size, taken to be 0.01 in our simulations, r is the risk free rate, σ the volatility, and dS is the change in S in time dt. If we assume that the simulation is started at time t_0 and ends at expiry T, then the other parameters which affect the simulations are the initial stock price

$S_0 = S(t_0)$, the exercise price E and the tenor or time to expiry, $\tau = T - t_0$. For each value of the parameters, a separate set of runs was done for each of the exercise strategies. For each realization, at each time step, we first check to see if the exercise criteria has been satisfied, and either exercise at that step or continue to the next time step, and repeat this procedure either the option has been exercised or we reach expiry, at which time the option is either exercised or expires worthless. For each realization, we calculate the payoff, which is $\max\left[S\left(T_E\right) - E, 0\right]$ for the call and $\max\left[E - S\left(T_E\right), 0\right]$ for the put, where T_E is the time at which the option was exercised. We then discount this value back to the starting time to find its present value, using the discount factor $\exp\left[r(T_E - t_0)\right]$. The value of the option is the average over all realizations of this present value.

3. RESULTS

In this section, we present the results of some of our simulations, and in particular examine the effects of varying the various parameters. In figures 1 (for the call) and 2 (for the put), we look at the effect of varying the strike price E for the call while holding the other parameters fixed.

For the call, strategy (1) (holding) is best, which is to be expected given that it is never optimal to exercise a call with no dividends. By contrast, for the put, no one strategy is best, and in actuality, they are all bad. Holding is no longer optimal and is sometimes the worst strategy amongst those studied. While for the call, the value always increased with time to expiry, for the put sometimes the value decreased and sometimes it increased. Presumably, this happens because some of the strategies for the put are especially bad, and increasing the tenor increases the possibility of inopportune exercise. The results from figures 1 and 2 are collapsed onto single curves in figure 3, where we see that as we increased the exercise price, the value of the call decreased while that of the put increased. This dependence on exercise price is of course to be expected from our knowledge of the greeks. Similarly, in figure 4, we looked at the effect of varying the initial stock price, finding as expected that as we increased S_0 the option value increased for the call but decreased for the put. As expected, the dependence on S_0 is in the opposite direction to that on E.

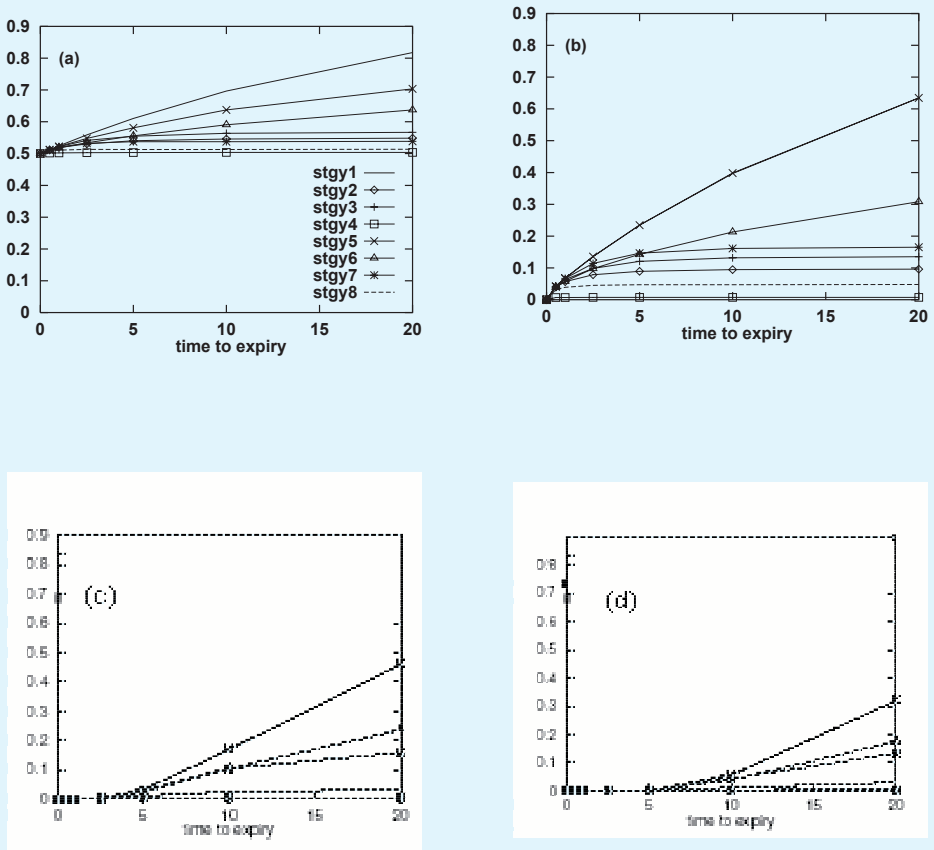
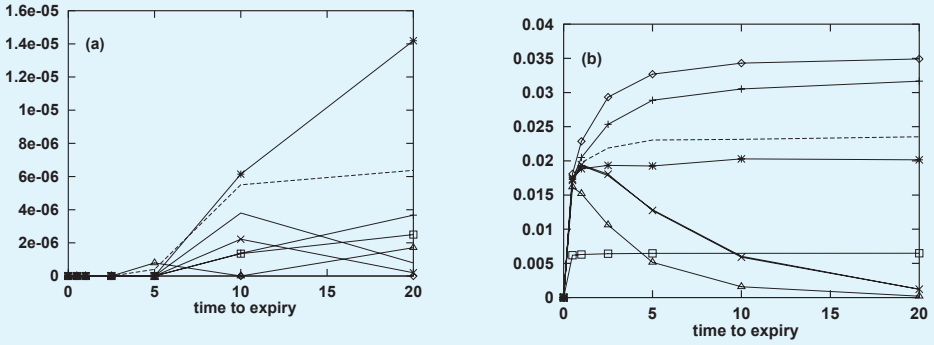


Fig1: Effect of E: call. $S_0 = 1$, $r = .05$, $\sigma = .1$. (a) E=.5, (b) 1, (c) 1.5, (d) 2.



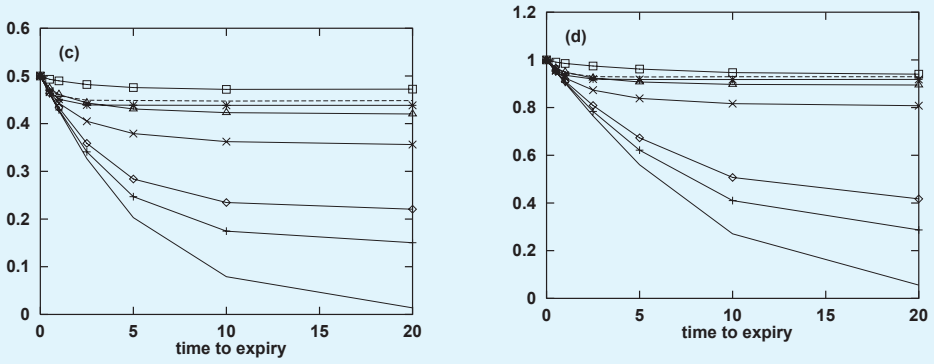


Figure 2: As for figure 1 but for put.

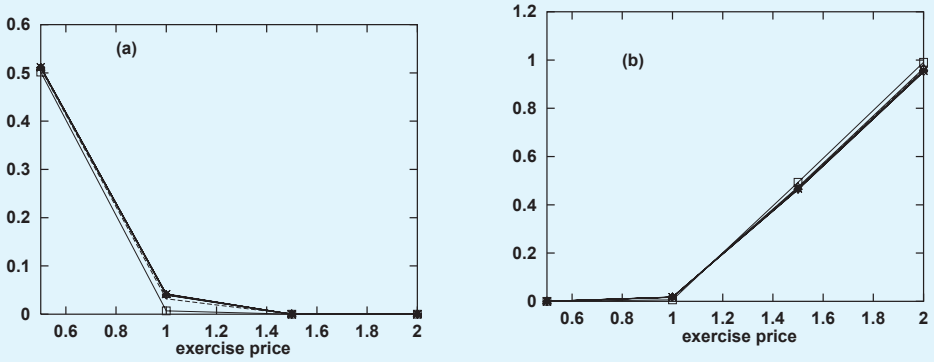


Figure 3: Effect of E. $S_0 = 1$, $r = .05$, $\sigma = .1$, $\tau = .5$. (a) call, (b) put.

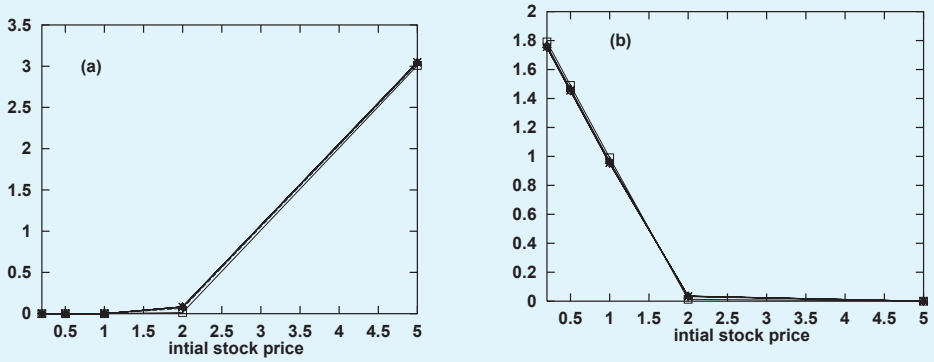


Figure 4: Effect of S_0. $E = 2$, $r = .05$, $\sigma = .1$, $\tau = .5$. (a) call, (b) put.

In figure 5, we examine the effects of varying the volatility, and find that for both the put and call, increasing σ leads to an increase in the value of the option, again as expected from our knowledge of vega.

In figure 6, we look at the effect of varying the risk-free rate, and find that increasing r increases the option value for the call but decreases it for the put, and again, this was as expected from our knowledge of theta.

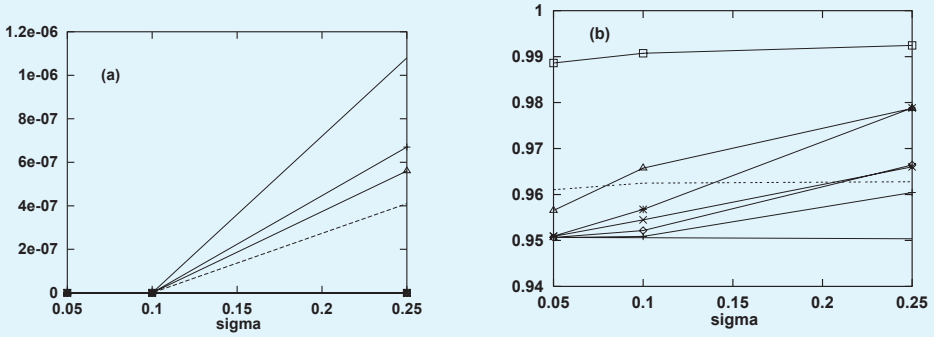


Figure 5: Effect of σ. c$E = 2$, $r = .05$, $S_0 = 1$, $\tau = .5$. (a) call, (b) put.

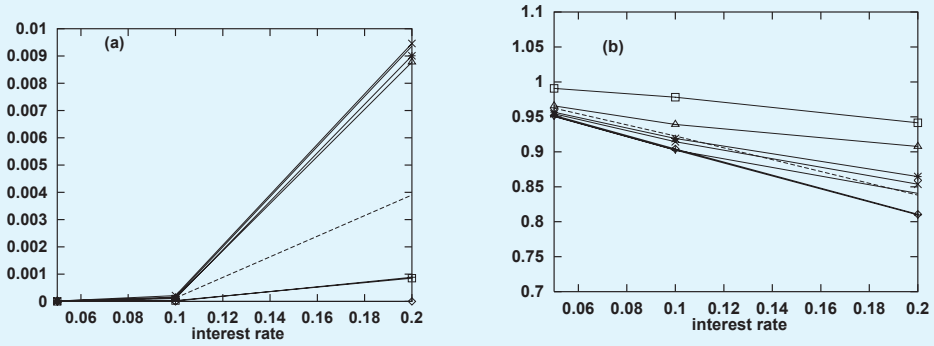


Figure 6: Effect of r. $E = 2$, $\sigma = .1$, $S_0 = 1$. (a) call $\tau = 2.5$, (b) put $\tau = .5$.

We also studied the effect that the frequency of application of the strategy had on the expected returns from the option. Our results are shown in figure 7, with a logarithmic scale used solely for ease of viewing. The time-step used in our simulations was $dt = 0.01$, and to examine the effects of frequency we applied the strategy initially every step or 0.01 time units, and then (in different runs) every 10 steps (0.1 units), 100 steps (1 units), 500 steps (5 units) and 1000 steps (10 units). The motivation for this was an attempt to model the real world behavior of different classes of investor, ranging from institutions using computer trading through a day trader who is constantly checking prices, and an average investor who might check prices daily of weekly, to a pension fund investor gets report once a month. Here, we are essentially treating the option like a Bermudan (which is sometimes known as a semi-American option as it can be exercised on a finite number of dates prior to expiry), as indeed we have in this entire study since we are using a finite time-step and therefore checking for exercise on a finite (if fairly large) number of occasions.

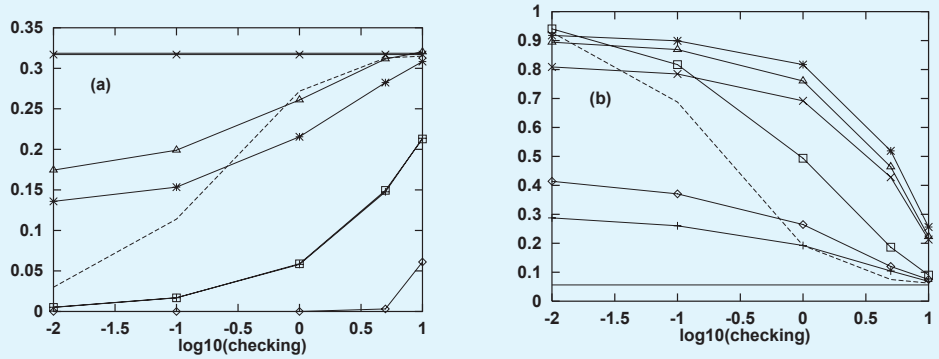


Figure 7: Effect of checking. $E = 2$, $\sigma = .1$, $r = .05$, $S_0 = 1$, $\tau = 20$. (a) call, (b) put.

We see that for the call, strategy (1), which was holding, is unaffected by the frequency of checking and while (5), a down-and-out barrier at $S = 0.9S_0$, is little affected by the frequency since exercise is infrequent for these particular parameter value, but that amongst the other strategies increasing the interval between checks leads to an increase in value. We should recall that it is never optimal to exercise the call without dividends, so that increasing the interval reduces the likelihood of inopportune exercise. For the put, strategy (1), which was holding, is again unaffected by the frequency, while for the other strategies, increasing the interval leads to a decrease in value. We should recall that it is sometimes optimal to exercise the put even without dividends, so that increasing the interval reduces exercise possibilities.

4. CONCLUSION

In this paper, we have looked at a number of elementary exercise strategies for American options, and used a Monte Carlo scheme to find the returns that an investor would expect if he followed one of these strategies, looking at the effects of varying each parameter while holding the others fixed. The variation of the expected returns with these parameters was largely as expected from the greeks such as vega and rho. As expected, for a call without dividends, holding was the best strategy. For the put, no single strategy amongst those studied was best, with different strategies being better in different areas of parameter space; in fact, all of the strategies for the put and all apart from holding for the call were fairly bad strategies from the point of view of the returns that an investor would expect if he pursued one of those strategies, and so our advice to an unsophisticated investor would be to steer clear of American options.

In closing, we would like to point out some related work which we recently presented. In [7], rather than consider the elementary exercise strategies presented here, we calculated the value of an American option if the holder employed an exercise strategy based on the series solutions for the optimal exercise boundary presented in [8, 9].

ACKNOWLEDGEMENTS

The work presented here formed part of G.A.'s PhD dissertation [10], funded in part by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Black, F. and Scholes, M., The pricing of options and corporate liabilities, *J. Pol. Econ.*, 81 (1973), 637-659.

[2] Merton, R.C., The theory of rational option pricing, *Bell J. Econ. Manag. Sci.*, 4 (1973), 141-183.

[3] Kwok, Y.K., *Mathematical Models of Financial Derivatives*, Springer, Singapore, 1998.

[4] Wilmott, P., *Paul Wilmott on Quantitative Finance*, Wiley, Chichester, 2000

[5] Boyle, P., Options: a Monte-Carlo approach, *J. Fin. Econ.*, 4 (1979), 323-338.

[6] L'Ecuyer, P. and Cote, S., Implementing a Random Number Package with Splitting Facilities, *ACM Trans. Math. Soft.*, 17 (1991), 98-111.

[7] Mallier, R., Evaluating approximations to the optimal exercise boundary for American options, *J. Appl. Math.*, 2 (2002), 71-92.

[8] Alobaidi, G. and Mallier, R., Asymptotic analysis of American call options, *Int. J. Math. Math. Sci.*, 27 (2001), 177-188.

[9] Alobaidi, G. and Mallier, R., On the optimal exercise boundary for an American put option, *J. App. Math.*, 1 (2001), 39-45.

[10] Alobaidi, G., *American options and their strategies*, Ph.D. Thesis, University of Western Ontario, Canada, 2000.

GROUP THEORY AND THE CAPITULATION PROBLEM FOR CERTAIN BIQUADRATIC FIELDS

A. Azizi
Department of Mathematics and Computer Sciences
Faculty of Sciences
University Mohamed 1, Oujda - MOROCCO

1. INTRODUCTION

Let G be a finite group, H a subgroup of G and H' the subgroup of H generated by the commutators of H. If $H\tau_1$, $H\tau_2 \cdots H\tau_n$ are the distinct cosets of H in G, then for every $\sigma \in G$ and each index i there is an element $\phi_i(\sigma) \in G$ and an index j such that

$$\tau_i \sigma = \phi_i(\sigma)\tau_j.$$

The transfert from G to H is a homomorphism V defined on G with values in H/H' by

$$V(\sigma) = H'\phi_1(\sigma) \cdots \phi_n(\sigma).$$

The principal ideal theorem of group theory is an application of the transfert map from G to G'.

Theorem 1.1 (Principal ideal theorem of group theory). *Let G be a finite group, then the transfert from G to G' is the trivial homomorphism.*

Let $\mathbf{k}$ be a number field of finite degree over $\mathbf{Q}$, $O_{\mathbf{k}}$ be its ring of integers and $C_{\mathbf{k}}$ the class group of $\mathbf{k}$, i.e. the quotient of the set of fractional ideals of $\mathbf{k}$ by the set of principal fractional ideals of $\mathbf{k}$.

Let $\mathbf{F}$ be an unramified extension of $\mathbf{k}$ of finite degree and let $O_{\mathbf{F}}$ be its ring of integers. We say that an ideal $\mathcal{A}$ of $\mathbf{k}$ (or the ideal class of $\mathcal{A}$) capitulates in $\mathbf{F}$ if it becomes principal in $\mathbf{F}$, i.e., if $\mathcal{A}O_{\mathbf{F}}$ is principal in $\mathbf{F}$.

The Hilbert class field $\mathbf{k}_1$ of $\mathbf{k}$ is by definition the maximal abelian unramified extension of $\mathbf{k}$. Let p be a prime number; the Hilbert p-class field $\mathbf{k}_1^{(p)}$ of $\mathbf{k}$ is the maximal abelian unramified extension of $\mathbf{k}$ such that $[\mathbf{k}_1^{(p)} : \mathbf{k}] = p^n$ for some integer n.

The first important result on capitulation was conjectured by D. Hilbert and proved by E. Artin and P. Furtwängler by the transfer theory of groups (theorem 1.1). It

deals with the case $\mathbf{F} = \mathbf{k}_1$.

Theorem 1.2 (Principal ideal theorem). *Let $\mathbf{k}_1$ be the Hilbert class field of $\mathbf{k}$, then every ideal of $\mathbf{k}$ capitulates in $\mathbf{k}_1$.*

The case where $\mathbf{F}/\mathbf{k}$ is a cyclic extension of prime degree was studied by D. Hilbert in his Theorem 94:

Theorem 1.3 (Theorem 94). *Let $\mathbf{F}/\mathbf{k}$ be an unramified cyclic extension of prime degree, then there exists at least one non trivial class in $\mathbf{k}$ which capitulates in $\mathbf{F}$.*

We find in the proof of Theorem 94 this result:

Let σ be a generator of the Galois group of $\mathbf{F}/\mathbf{k}$ and $N_{\mathbf{F}/\mathbf{k}}$ be the norm of $\mathbf{F}/\mathbf{k}$. Let $E_{\mathbf{L}}$ be the unit group of the field $\mathbf{L}$. Let $E_{\mathbf{F}}^*$ be the group of units of norm 1 in $\mathbf{F}/\mathbf{k}$. Then the group of classes of $\mathbf{k}$ which capitulates in $\mathbf{F}$ is isomorphic to the quotient group $E_{\mathbf{F}}^*/E_{\mathbf{F}}^{1-\sigma} = H^1(E_{\mathbf{F}})$, the cohomology group of $G = \langle\sigma\rangle$ acting on the group $E_{\mathbf{F}}$.

With this result and other results on cohomology, we have:

Theorem 1.4 *Let $\mathbf{F}/\mathbf{k}$ be an unramified cyclic extension of prime degree, then the number of classes which capitulate in $\mathbf{F}/\mathbf{k}$ is equal to $[\mathbf{F} : \mathbf{k}][E_{\mathbf{k}} : N_{\mathbf{F}/\mathbf{k}}(E_{\mathbf{F}})]$.*

The case where $\mathbf{F}/\mathbf{k}$ is an unramified abelian extension was treated by H. Suzuki who has proved Miyake's conjecture: *In an unramified abelian extension $\mathbf{F}/\mathbf{k}$ the number of classes of $\mathbf{k}$ which capitulate in $\mathbf{F}$ is a multiple of $[\mathbf{F} : \mathbf{k}]$*. Moreover, H. Suzuki has proved the next theorem which is generalization of the principal ideal theorem, the Hilbert theorem 94 and Tannaka-Terada's principal ideal theorem:

Theorem 1.5 *Let $\mathbf{k}$ be a finite cyclic extension of an algebraic number field $\mathbf{k}_0$ of finite degree, and let $\mathbf{K}$ be an unramified extension of $\mathbf{k}$ which is abelian over $\mathbf{k}_0$. Then the number of the $G(\mathbf{k}/\mathbf{k}_0)$-invariant ideal classes of $\mathbf{k}$ which become principal in $\mathbf{K}$ is divisible by the degree $[\mathbf{K} : \mathbf{k}]$ of the extension $\mathbf{K}/\mathbf{k}$.*

The group theoretical endomorphism version of this theorem is

Theorem 1.6 *Let g be an endomorphism of finite group G, and H be a normal subgroup of G containing the commutator subgroup G'. If $g(H) \subset H$ and g induces*

the identity map on G/H, then the order of the subgroup

$$\{hG' \in KerV_{G\longrightarrow H} : g(h)h^{-1} \in G'\}$$

of the transfer kernel is divisible by $[G : H]$. Here $V_{G\longrightarrow H}$ is the group transfer from G to H.

For more details, see [23], [22], [18], [7], [5], [10] and [12].

2. CAPITULATION OF THE 2-IDEAL CLASSES OF SOME BIQUADRATIC FIELDS

2.1 The General Case

Let $\mathbf{k}$ be a number field such that the 2-component $C_{\mathbf{k},2}$ of $C_{\mathbf{k}}$ is isomorphic to $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$. Let G_2 be the Galois group of $\mathbf{k}_2^{(2)}/\mathbf{k}$ and $\mathbf{k}^*$ be the genus field of $\mathbf{k}$
(the maximal unramified extension of $\mathbf{k}$ of the form $\mathbf{k}Ł$ where Ł is an abelian extension of $\mathbf{Q}$). By class field theory, $Gal(\mathbf{k}_1^{(2)}/\mathbf{k}) \simeq C_{\mathbf{k},2} \simeq \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$. Then $\mathbf{k}_1^{(2)}$ contains three quadratic extensions of $\mathbf{k}$ denoted by $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$.

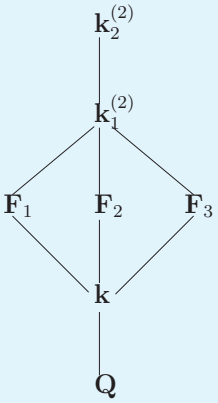


Diagram 1

Under these conditions, from the results of Kisilevsky [12], we have the following

Theorem 2.1 *Let $\mathbf{k}$ be such that $C_{\mathbf{k},2} \simeq \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$. Then we have three types of capitulation:*

Type 1: The four classes of $C_{\mathbf{k},2}$ capitulate in each extension $\mathbf{F}_i$, $i = 1, 2, 3$. This is possible if and only if $\mathbf{k}_1^{(2)} = \mathbf{k}_2^{(2)}$. In this case $G_2 \simeq \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$.

Type 2: The four classes of $C_{\mathbf{k},2}$ capitulate only in one extension among the three extensions $\mathbf{F}_i$, $i = 1, 2, 3$. In this case the group G_2 is dihedral.

Type 3: Only two classes capitulate in each extension $\mathbf{F}_i$, $i = 1, 2, 3$. In this case the group G_2 is semi-dihedral or quaternionic. It is the quaternionic group if and only if there exists an i such that the non trivial class which capitulates in $\mathbf{K}_i$ is norm in $\mathbf{K}_i/\mathbf{k}$.

2.2 The Case Of Biquadratic Fields

In this paragraph, we suppose that $\mathbf{k}$ is a biquadratic field such that the 2-group $C_{\mathbf{k},2} \simeq \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$, and we study the capitulation problem in the extensions $\mathbf{F}_i/\mathbf{k}$, $i = 1, 2, 3$ and the structure of the group G_2.

In particular, we found the next theorem

Theorem 2.2 *Let $\mathbf{k} = \mathbf{Q}(\sqrt{2q_1q_2}, i)$ where q_1 and q_2 are primes such that, $(\frac{q_1}{q_2}) = -(\frac{q_2}{q_1}) = (\frac{2}{q_1}) = -(\frac{2}{q_2}) = 1$, then $\mathbf{k}_2^{(1)} = \mathbf{Q}(\sqrt{2}, \sqrt{q_1}, \sqrt{q_2}, i)$. Let $\mathbf{K}_1 = \mathbf{Q}(\sqrt{q_1}, \sqrt{2q_2}, i)$, $\mathbf{K}_2 = \mathbf{Q}(\sqrt{q_2}, \sqrt{2q_1}, i)$ and $\mathbf{K}_3 = \mathbf{Q}(\sqrt{2}, \sqrt{q_1q_2}, i)$, then two classes of $C_{\mathbf{k},2}$ capitulate in each extension $\mathbf{K}_i$, $i = 1, 2, 3$. In this case the group G_2 is semi-dihedral or quaternionic.*

Proof.

By [2], we have that the 2-class group of $\mathbf{k}$ is the Klien group, and by computing the degree of the genus field of $\mathbf{k}$ we found that $\mathbf{k}_2^{(1)} = \mathbf{Q}(\sqrt{2}, \sqrt{q_1}, \sqrt{q_2}, i)$. We have to study the capitulation problem in the three sub-extensions of $\mathbf{k}_2^{(1)}/\mathbf{k}$. Using the fundamental system of units of the fields $\mathbf{K}_i$, $i = 1, 2, 3$ (Theorem 9 and 12 of [5] and [3]), we have:

Let ϵ_1 (resp. ϵ_2, ϵ_3) be the fundamental units of $\mathbf{k}_1' = \mathbf{Q}(\sqrt{q_1})$ (resp. $\mathbf{k}_2' = \mathbf{Q}(\sqrt{2q_2})$, $\mathbf{k}_3' = \mathbf{Q}(\sqrt{2q_1q_2})$. In our case $2\epsilon_3$ is not a square in $\mathbf{k}_3'$; then the fundamental system of units of $\mathbf{K}$ is $\{\epsilon_3\}$ and the fundamental system of units of $\mathbf{K}_1$ is

$$\{\sqrt{i\epsilon_1}, \sqrt{i\epsilon_2}, \sqrt{i\epsilon_3}\} \text{ or } \{\sqrt{i\epsilon_1}, \sqrt{i\epsilon_2}, \sqrt{\epsilon_3}\}.$$

The torsion subgroup of the units group of $\mathbf{K}$ is generated by $\sqrt{-1} = i$ and the

torsion subgroup of the units group of $\mathbf{K}_1$ is generated by $\sqrt{-1} = i$. So, we have $[E_{\mathbf{k}} : N_{K_1/\mathbf{k}}(E_{K_1})] = 1$.

Let ϵ_1 (resp. ϵ_2, ϵ_3) be the fundamental units of $\mathbf{k}'_1 = \mathbf{Q}(\sqrt{q_2})$ (resp. $\mathbf{k}'_2 = \mathbf{Q}(\sqrt{2q_1})$, $\mathbf{k}'_3 = \mathbf{Q}(\sqrt{2q_1q_2})$. In our case $2\epsilon_3$ is not a square in $\mathbf{k}'_3$; then the fundamental system of units of $\mathbf{K}$ is $\{\epsilon_3\}$ and and the fundamental system of units of $\mathbf{K}_2$ is

$$\{\sqrt{i\epsilon_1}, \sqrt{i\epsilon_2}, \sqrt{i\epsilon_3}\} \text{ or } \{\sqrt{i\epsilon_1}, \sqrt{i\epsilon_2}, \sqrt{\epsilon_3}\}.$$

The torsion subgroup of the units group of $\mathbf{K}$ is generated by $\sqrt{-1} = i$ and the torsion subgroup of the units group of $\mathbf{K}_2$ is generated by $\sqrt{-1} = i$. So, we have $[E_{\mathbf{k}} : N_{K_2/\mathbf{k}}(E_{K_2})] = 1$.

Let ϵ_1 (resp. ϵ_2, ϵ_3) be the fundamental units of $\mathbf{k}'_1 = \mathbf{Q}(\sqrt{2})$ (resp. $\mathbf{k}'_2 = \mathbf{Q}(\sqrt{q_1q_2})$, $\mathbf{k}'_3 = \mathbf{Q}(\sqrt{2q_1q_2})$. In our case $2\epsilon_3$ is not a square in $\mathbf{k}'_3$; then the fundamental system of units of $\mathbf{K}$ is $\{\epsilon_3\}$ and and the fundamental system of units of $\mathbf{K}_3$ is

$$\{\epsilon_1, \epsilon_2, \sqrt{\epsilon_2\epsilon_3}\}.$$

The torsion subgroup of the units group of $\mathbf{K}$ is generated by $\sqrt{-1} = i$ and the torsion subgroup of the units group of $\mathbf{K}_3$ is generated by $\sqrt{i}$. So, we have $[E_{\mathbf{k}} : N_{K_3/\mathbf{k}}(E_{K_3})] = 1$.

Finally, the number of classes of $C_{\mathbf{k},2}$ which capitulate in each extension $\mathbf{K}_i$, $i = 1, 2, 3$ is equal to $[K_i : \mathbf{k}][E_{\mathbf{k}} : N_{K_i/\mathbf{k}}(E_{K_i})] = 2$.

The 2-class group of $\mathbf{k}$ can be generated by classes of prime ideals of $\mathbf{k}$ laying above the ramified primes in $\mathbf{k}/\mathbf{Q}$. But in our case this is not possible, we have only one non trivial class generated by the prime ideal I_0 above the prime 2 ($2O_{\mathbf{k}} = I_0^4$). We choose the second generator as the following:
Let l be a prime such that $(\frac{q_1}{l}) = (\frac{q_2}{l}) = -1$ and $(\frac{2}{l}) = (\frac{-1}{l}) = 1$. The prime l exists
(see [sim- 95]) and splits completely in $\mathbf{k}/\mathbf{Q}$, then there exists some ideals I_1, I_2, I_3 and I_4 of $\mathbf{k}$ such that $lO_{\mathbf{k}} = I_1I_2I_3I_4$. We prove that I_1 is not principal and if m is the odd part of the class number of $\mathbf{k}$ then $C_{\mathbf{k},2}$ is generated by the class of I_0 and the class of I_1^m.

The ideal I_1 is inert in $\mathbf{K}_i/\mathbf{k}$, $i = 1, 2$ and splits in $\mathbf{K}_3/\mathbf{k}$. Also the ideal I_0 is inert in $\mathbf{K}_i/\mathbf{k}$, $i = 2, 3$ and splits in $\mathbf{K}_1/\mathbf{k}$. So we have
- the ideal class of I_1^m isn't norm in $\mathbf{K}_2/\mathbf{k}$.
- the ideal class of I_0 isn't norm in $\mathbf{K}_2/\mathbf{k}$.

- the ideal class of $I_0 I_1^m$ isn't norm in $\mathbf{K}_2/\mathbf{k}$.
- the ideal class of I_1^m is norm in $\mathbf{K}_3/\mathbf{k}$.
- the ideal class of I_0 isn't norm in $\mathbf{K}_3/\mathbf{k}$.
- the ideal class of $I_0 I_1^m$ isn't norm in $\mathbf{K}_3/\mathbf{k}$.
- the ideal class of I_1^m isn't norm in $\mathbf{K}_1/\mathbf{k}$.
- the ideal class of I_0 is norm in $\mathbf{K}_1/\mathbf{k}$.
- the ideal class of $I_0 I_1^m$ isn't norm in $\mathbf{K}_1/\mathbf{k}$.
- The ideal I_0 capitulates in $\mathbf{K}_3$:

We have $(\sqrt{2}\sqrt{i})O_{\mathbf{K}_3} = I_0^2$ and $\sqrt{2}\sqrt{i}\epsilon_1 = \alpha^2$ where

$$\alpha = \frac{1+\sqrt{2}+i}{\sqrt{2}} \in \mathbf{K}_3.$$

Then

$$I_0 = \alpha O_{\mathbf{K}_3} \text{ with } \alpha \in \mathbf{K}_3.$$

From this remarks and using theorem 2.1, we obtain that

- One class from the classes of I_0, I_1^m or $I_0 I_1^m$ capitulates in $\mathbf{K}_2/\mathbf{k}$ and it isn't norm from $\mathbf{K}_2$.
- If I_0 capitulates in $\mathbf{K}_1/\mathbf{k}$, then the group G_2 is quaternionic.
- If I_1 capitulates in $\mathbf{K}_3/\mathbf{k}$, then the group G_2 is quaternionic; but I_0 capitulates in $\mathbf{K}_3/\mathbf{k}$ and the number of classes which capitulate in $\mathbf{K}_3/\mathbf{k}$ is equal to 2. So this case can't occur.
- In the other cases the group G_2 is semi-dihedral.

We conclude that G_2 is quaternionic if and only if I_0 capitulates in $\mathbf{K}_1/\mathbf{k}$. □

More generally, we have the following theorem

Theorem 2.3 *Let* $\mathbf{k}$ *be a biquadratic field such that* $C_{\mathbf{k},2} \simeq \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$*, and* G_2 *be the Galois group of* $\mathbf{k}_2^{(2)}/\mathbf{k}$*.*
i) If $\mathbf{k} = \mathbf{Q}(\sqrt{d}, i)$*, then the group* G_2 *is dihedral, semi-dihedral, quaternionic or* $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$*.*
ii) If $\mathbf{k} = \mathbf{Q}(\sqrt{d}, \sqrt{-2})$*, then the group* G_2 *is dihedral, semi-dihedral, quaternionic or* $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$*.*
iii) If $\mathbf{k} = \mathbf{Q}(\sqrt{d}, \sqrt{p})$ *where* $d \in \mathbf{N}$*,* $p \equiv 1 \bmod 4$ *is prime, then the group* G_2 *is dihedral, quaternionic or* $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$*.*

iv) If $\mathbf{k} = \mathbf{Q}(\sqrt{d}, \sqrt{2})$*, where* $d \in \mathbf{N}$ *then the group* G_2 *is dihedral, quaternionic or* $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$.
v) If $K = \mathbf{Q}(\sqrt{2p'}, \sqrt{qp})$ *where* p, p' *and* q *are primes such that* $(\frac{p}{p'}) = (\frac{q}{p}) = -(\frac{2}{p}) = -(\frac{2}{p'}) = -(\frac{q}{p'}) = -(\frac{2}{q}) = 1$ *then the group* G_2 *is semi-dihedral.*

As the last theorem, to prove this theorem we must
1) study the structure of $C_{\mathbf{k},2}$. Using the Genus theory, the class number formula for biquadratic fields, Kaplan's results on the 2-part of the class number for quadratic number fields and other results.
2) determine the number of ideal classes which capitulate in $\mathbf{F}_i/\mathbf{k}$, $i = 1, 2, 3$. So we have to determine the unit group of each $\mathbf{F}_i$, $i = 1, 2, 3$, where $\mathbf{F}_i$ is a composite of three quadratic fields. 3) determine the conditions such that $\mathbf{k}_1^{(2)} = \mathbf{k}_2^{(2)}$.
4) study the structure of the 2-class group of some $\mathbf{F}_i$.
5) determine some classes of $C_{\mathbf{k},2}$ and study their capitulation in every $\mathbf{F}_i$.

For more details on this theorem, see [4], [5], [18], [7] and [10].

Numerical examples

Let be $\mathbf{k} = \mathbf{Q}(\sqrt{d}, (\sqrt{d'})$.

d	d'	G_2
119	-1	$\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$
287	-1	dihedral
-2	102	dihedral
-2	493	quaternionic
2	24947	semi-dihedral
113	10	quaternionic

ACKNOWLEDGMENTS

I want to thank the Organizing Committee, of the International Conference on Mathematics and it's applications held in Kuwait University, for their courtesy and their support.

REFERENCES

[1] Azizi, A. *Sur la capitulation des 2-classes d'idéaux de* $\mathbf{Q}(\sqrt{d}, i)$. C. R. Acad. Sci. Paris, t. 325, serie I, (1997), p. 127-130.

[2] Azizi, A. *Sur le groupe de classes d'idéaux de* $\mathbf{Q}(\sqrt{d}, i)$. Rendiconti del circolo matematico di Palermo, vol. 48 (1999).

[3] Azizi, A. *Unités de certains corps de nombres imaginaires et abéliens sur* $\mathbf{Q}$. Annales des Sciences Mathématiques du Québec, 23 (1999), 87-93.

[4] Azizi, A. *Capitulation of the* 2*-Ideal Classes of* $\mathbf{Q}(\sqrt{p_1p_2}, i)$ *Where* p_1 *and* p_2 *Are Primes Such That* $p_1 \equiv 1 \pmod{8}$, $p_2 \equiv 5 \pmod{8}$ *and* $(\frac{p_1}{p_2}) = -1$. Lecture notes in pure and applied mathematics, volume 208 (1999), p.13-19.

[5] Azizi, A. *Capitulation des 2-classes d'idéaux de* $\mathbf{Q}(\sqrt{2pq}, i)$. Acta arithmetica XCIV.4 (2000), 383-399.

[6] Azizi, A. and Mouhib, et A., *Sur le rang du* 2*-groupe de classes de* $\mathbf{Q}(\sqrt{m}, \sqrt{d})$ *où* $m = 2$ *ou un premier* $p \equiv 1 \pmod 4$. Trans. Amer. Math. Soc. Volume 353, Number 7 (2001), 2741-2752.

[7] Azizi, A. Ayadi M. and Ismaili, et M. C., *Capitulation in certain number Fields.* Advanced Studies in Pure Mathematics 30 (2001), pp. 467-482.

[8] Azizi, A. *Sur une question de Capitulation.* Proc. Amer. Math. Soc. 130 (2002), 2197-2202.

[9] Azizi, A. *Construction de la tour des 2-corps de classes de Hilbert de certains corps biquadratiques.* Pacific Journal of Mathematics, vol. 208, N.1, (2003), 1-10.

[10] Azizi, A. and Mouhib, et A., *Capitulation des 2-classes d'idéaux de certains corps biquadratiques réels de la forme* $\mathbf{Q}(\sqrt{d}, \sqrt{2})$. Acta Arithmetica, 109.1 (2003).

[11] Kaplan, P. *Sur le 2-groupe des classes d'idéaux des corps quadratiques.* J. reine. angew. Math. 283/284. (1976), 313-363.

[12] Kisilevesky, H. *Number fields with class number congruent to 4 modulo 8 and Hilbert's theorem 94.* J. Number theory 8 (1976), 271-279.

[13] Kubota, T., *Über die beziehung der Klassenzahlen der Unterkörper des bizyklischen Zahlkörpers.* Nagoya Math. J, 6 (1953), 119-127.

[14] Kubota, T. *Über den bizyklischen biquadratischen zahlkörper.* Nagoya Math. J, 10 (1956), 65-85.

[15] Kucera, R. *On the parity of the class number of a biquadratic field.* J.Number theory 52 (1995), 43-52.

[16] Kuroda, S. *Über den Dirichletschen Zahlkörper.* J. Fac. sc. Imp. Univ. Tokyo sec. I, vol IV part. (1943), 383-406.

[17] Thomas M. McCall, Charles J. Parry, and Ramona R. Ranalli, *Imaginary Bicyclic Biquadratic Fields With Cyclic* 2- *Class Group.* Journal of Number Theory 53 (1995), 88-99.

[18] Miyake, K. *Algebraic Investigations of Hilbert's Theorem 94, the Principal Ideal Theorem and Capitulation Problem.* Expos. Math. 7 (1989), 289-346.

[19] Schoof, R. *Infinite class field towers of quadratic fields.* J. reine angew. Math. 372 (1986), 209-220.

[20] Sime, P. J. *On the ideal class group of real biquadratic fields.* Trans. Am. Math. Soc. 347, No. 12 (1995), 4855- 4876.

[21] Sime, P. J., *Hilbert Class Fields of Real Biquadratic Fields.* Journal of number theory 50 (1995), 154-166.

[22] Suzuki, H. *A generalisation of Hilbert Theorem 94.* Nagoya Math. J., vol 121 (1991).

[23] Suzuki, H. *About the capitulation problem.* Proceedings of CFT Conference in class field theory, JAPAN.

[24] Taussky, O. *A Remark on the Class Field Tower.* J.London Math. Soc. 12 (1937), 82-85.

[25] Terada, F. *A principal Ideal Theorem in the Genus Fields.* Tôhoku Math. J. Second Series, vol. 23, 4 (1971), pp. 697-718.

[26] Wada, H. *On the class number and the unit group of certain algebraic number fields.* Tokyo U., Fac. of sc. J., Serie I, 13 (1966), 201-209.

RUNGE-KUTTA NEURAL NETWORKS FOR DELAY DIFFERENTIAL EQUATIONS

P. Balasubramaniam[1] and A. V. A. Kumar[2]
[1]Department of Mathematics,
Gandhigram Rural Institute-Deemed University,
Gandhigram - 624 302, Tamil Nadu, India
email: pbalgri@yahoo.co.in

[2]Department of Mathematics,
PSNA College of Engg., and Tech.,
Dindigul- 624 602, Tamil Nadu, India

Abstract

In this paper Runge-Kutta Neural Networks (RKNN's) for solving delay differential equations(DDE) in high accuracy have been developed. These networks are constructed according to the Runge-Kutta approximation method. The learning algorithm is developed for the RKNN's nonlinear recursive least-squares based algorithm.

KEY WORDS: Runge-Kutta methods, Delay differential equations, Nonlinear recursive least square.

1. INTRODUCTION

The Runge-Kutta method for solving DDE have been studied by several authors [1], [2], [3]. In recent years neural networks have been widely used for identification for dynamical system [4], [5].

Using the neural networks, in this paper a class of feed forward neural networks called Runge-Kutta Neural Networks (RKNN's) for precisely modeling a DDE in the form $x'(t) = f(x_t)$ with an unknown f, have been established.

The neural approximation of f is used in the well known Runge-Kutta integration formulae [6] to obtain an approximation of x. With the designed network structure and learning scheme, the RKNN's perform high order discretization of unknown DDE systems implicitly without the aforementioned complexity and intractability problems. The main attraction of RKNN's that they can precisely estimate the changing rates of system states directly in their subnetwork based on

the space- domain interpolation within one sampling interval such that they can do long-term prediction of systems state trajectories and are good at parallel model prediction. Also, since the RKNN's models the right hand side of DDE in its subnetworks directly, some known continuous relationship of the identified system can be incorporated into the RKNN directly to speed up its learning. Such kind of a priori knowledge is not easy to be used directly in normal neural identifiers. Another important feature of RKNN is that it can predict the system behavior at any time instant, not limited by fixed time step as the case in normal neural modeling. An n-order RKNN consists of n identical subnetworks connected in the way realizing an n-order Runge-Kutta algorithm.

The subnetwork is a normal neural network such as multilayer perceptron network or radial basis function network. Each subnetwork models the right-hand side of DDE directly, and thus the RKNN can approximate an DDE system in high-order accuracy. Here verify theoretically the superior generalization and long-term prediction ability of the RKNN's over the normal neural networks by providing some quantitative measures of the errors involved in the RKNN's modelling. Associated with the RKNN's is a class of learning algorithms derived by the recursive least-square (RLS) method. Especially a class of RLS algorithm, called non linear recursive least-square(NRLS)learning algorithm, is derived to increase the learning rate and prediction accuracy of the RKNN's. The NRLS generalizes the original RLS to non linear cases such that it can tune the parameters in the hidden layers of the RKNN's such as the centers and variances of the radial basis function networks.

2. RUNGE-KUTTA NEURAL NETWORK

Consider a non linear system described by the following DDE

$$x'(t) = f(x_t), \quad t \geq 0, \quad x_0 = \phi. \tag{1}$$

x_t will denote the function with domain $[-r, 0)$ defined by $x_t(\theta) = x(t + \theta), -r \leq \theta \leq 0$, the state vector $x(t) \in \Re^m$. The objective of this paper is to develop a neural network that can model an DDE system precisely whose right-hand side function f is unknown such that it can do long-term prediction of the state trajectory $x(t)$ of the system described in (1). The derived model is to be a parallel-model predictor. It uses the initial system state ϕ, which yield the long-term output $y(t)$, with high accurate prediction of the state $x(t)$ over $t \in [-r, T]$ giving the previous outputs of the identifier back to itself recursively.

To predict the state trajectory of the unknown system described by (1). There

are two methods available, first is known as conventional and the second namely neural-network based approach see [4], [8]. To construct a neural network say $\tilde{N}_f(\cdot)$, this will learn the system state trajectory, that is $x(ih;\phi) \cong \tilde{N}_f\Big(x((i-1)h);\phi\Big)$.

Here a learning algorithm for the RKNN such that the function $f(\cdot)$ in (1) directly approximated from $N_f(\cdot)$ is developed and then the state trajectory $x(t)$ can be predicted by the solution of $y'(t) = N_f(y_t)$ with initial conditions $y(0) = \phi$ where $y(t) \in \Re^m$.

3. THE STRUCTURE OF RKNN

The new neural network $N_f(\cdot)$ such that $y'(t) = N_f(y_t; w)$ can do long-term predication of the state trajectory $x(t)$ to any degree of accuracy. From the universal approximation theory of neural networks [9] and the delay differential equations theory, we prove the following lemma.

Lemma 3.1 *Given any solution $x(t,\phi)$ of the system described in (1)with $\phi \in D$, $t \in [-r, T]$ and a function $f(\cdot)$ satisfying the assumption(see[7]) for any $\epsilon > 0$ there exist a neural network $N_f(\cdot)$ such that the trajectory $y(t;\phi)$ corresponding to the system $y' = N_f(y_t; w)$ with $y(0) = \phi$ satisfies $||y_t(\phi) - x_t(\phi)|| < \epsilon$ for all $-r \le t \le T$, which shows the existence of the neural network that meets the required property.*

Construct a neural network RKNN, that fits Lemma 3.1. which is motivated by the Runge-Kutta algorithm [6] and construct an n- order RKNN to realize the computation flow of an n-order RKNN algorithm.

The input -output relationship of the fourth-order RKNN is described by

$$y(i+1) = y(i) + \frac{h}{6}(k_0 + 2k_1 + 2k_2 + k_3), \tag{2}$$

where

$$\begin{aligned} k_0 &= N_f(y_{t_i}(\cdot); w), \\ k_1 &= N_f(y_{t_i}(\cdot) + \frac{1}{2}hk_0; w), \\ k_2 &= N_f(y_{t_i}(\cdot) + \frac{1}{2}hk_1; w), \\ k_3 &= N_f(y_{t_i}(\cdot) + hk_2; w) \end{aligned}$$

The neural network $N_f(x_t; w)$ with input x and weights w can be the multi layer perceptron network or the radial basis function network. It is noted that four $N_f(x; w)$ are identical, which means, that they have same structure and use the same corresponding weights. It is enough to train only one network for software or hardware implementation.

Using supervised learning algorithms, we can tune the weights, w of $N_f(x_t; w)$ by training the RKNN on training trajectories obtained by the system (1). The total prediction error is sufficiently small or the weight vector w converges to w^*, from this obtain a RKNN $N_f(x_t; w^*)$ which approximates the continuous function $f(x_t)$ of the system in (1) accurately.

4. LEARNING ALGORITHM FOR THE RKNN

Consider an initial state $\phi \in D$ and a trajectory $x(t, \phi)$ which is the solution of system $x'(t) = f(x_t)$ corresponding to the initial state ϕ. At each time step h, the sampling data

$$x(i; \phi) \equiv x(ih; \phi), \quad i = 0, \ldots, (T/h) = L - 1, \text{ where } h \text{ is small.}$$

Collecting $x(i; \phi)$ for several different initial states $\phi \in D$ as a training data of the RKNN. By the learning algorithm developed in this section, the weights w of $N_f(x_t; w)$ in the RKNN are tuned such that the outputs $y(t; \phi)$ of the identified system $y'(t) = N_f(y_t; w), y(0) = \phi$ can approximate the solution $x(t, \phi)$ of $x'(t) = f(x_t), x(0) = \phi$, for all t in some fixed interval $[-r, T]$. We shall develop the non linear recursive least square (NRLS) algorithm, for the RKNN's. The NRLS algorithm, generalize the conventional recursive least square (RLS) algorithms to non linear cases. The number of iteration is increased to improve the accuracy by minimizing the square error function $E(w)$. The RLS method finds the root of the equation $\frac{\partial E(w)}{\partial w} = 0$ to locate the local minimum points directly. . This equation can be transformed into regression model as $\psi^T(\cdot)w = 0$, where w represents network weight and ψ is the regressor. To solve the equation $\psi^T(\cdot)w = 0$ we use a recursive algorithm.

4.1 Zero order NRLS learning algorithm

The radial basis functions $N_f(\cdot)$ in the RKNN is chosen to tune the weights W on the links connecting the radial basis functions nodes $\psi_j(\cdot; t_l, \delta_l)$ to the output

layer of $N_f(\cdot)$. Assume the training trajectories $\{x(i;\phi)|\ i = 0,\ldots,L-1, \phi \in D\}$ from the system $x'(t) = f(x_{t_i})$ with $x(0) = \phi$

Define, $E(W) = \frac{1}{2}\sum(x(i) - y(i))^2, i = 1\ to\ L$, where $y(i)$ is the output of the RKNN. According to the RKNN structure described in (2) the output y(i) of the RKNN with input $x(i-1)$ at time step $i-1$ can be written as

$$y(i) = x(i-1) + \frac{h}{6}\Big[\sum_{l=1}^{N} W_l\psi_l(x(i-1)) + 2\sum_{l=1}^{N} W_l\psi_l(x(i-1) + \frac{h}{2}K_0(i-1))$$

$$+2\sum_{l=1}^{N} W_l\psi_l(x(i-1) + \frac{h}{2}K_1(i-1)) + \sum_{l=1}^{N} W_l\psi_l(x(i-1) + hK_2(i-1))\Big], \qquad (3)$$

where ψ_l is the l^{th}radial basis function, W_l is the connection weight between ψ_l node to the output node of $N_f(\cdot)$ and K_o, K_1, K_2 are the output of $N_f(\cdot)$ subnetwork in the RKNN defined by (2). The regression form $\psi(x(i-1);W)$ of $y(i) - x(i-1)$ in(3) such that the connection weights W_l 's can be obtained using the non linear least-square method in [9].

It observed that minimization of $E(W)$ is equivalent to finding the solution of the following equations in the least square sense:

$$\begin{pmatrix} \psi^T(x(0);W) \\ \vdots \\ \psi^T(x(L-1);W) \end{pmatrix} \begin{pmatrix} W_1 \\ \vdots \\ W_N \end{pmatrix} = \begin{pmatrix} x(1) - x(0) \\ \vdots \\ x(L) - x(L-1) \end{pmatrix} \qquad (4)$$

.

If $$d^T \equiv (x(1) - x(0), x(2) - x(1), \ldots, x(L) - x(L-1)),$$

$$\zeta(x(0), \ldots, x(L-1); W) = [\psi(x(0); W) \ldots, \psi(x(L-1); W)]^T$$

then (4) can be expressed as $\zeta(x(0), \ldots, x(L-1); W)W = d$. The problem of solving (4) is a nonlinear lest-square problem, because the regression matrix $\zeta(x(0), \ldots, x(L-1); W)$ is a function of parameter W. Combining the fixed point method and the RLS algorithm to find the solution W^* of (4) in the least square sense. Consider this method as the zero- order NRLS algorithm for the RKNN's.The algorithm is listed as follows.

step 1. Fix appropriate initial weights $W(0) = W_0$ and fix $i = 0$.

step 2. Replace the $W(i)$ into the regression matrix to get $\zeta(x(0)\dots x(L-1); W(i))$.

step 3. Use RLS algorithm to solve $\zeta(x(0),\dots,x(L-1); W(i))W = d$ to get the solution $W = W^*$.

step 4. Let $W(i+1) = W^*$.

step 5. If the sequence $W(i)$ converges, then stop; otherwise, set $i = i+1$ and go to step 2.

The sufficient condition for the convergence of the nonlinear least square method are given in [9]. To prove the convergence property of the zero-order NRLS learning algorithm which own required sufficient condition in [9].

To simplify the analysis consider the second order RKNN with single state variable and to find the solution of W^* by substituting W and proving Theorem 6.2.2 in [9]. Using this concept it is sure that W will converge to a fixed point W^*. This convergence property for second order RKNN's can be expanded as fourth order RKNN's or even for higher order RKNN's directly.

5. CONCLUSION

In this paper RKNN is constructed for identification of DDE system with unknown right-hand-side function and also derived a zero-order NRLS learning algorithm. The NRLS algorithm to nonlinear cases such that it can tune the parameters in the hidden layers of RKNN's fastly. The convergence property of the proposed NRLS algorithm is studied theoretically. The algorithm derivation and theory focused on the fourth order RKNN's, these can be generalized to any n-order RKNN's directly.

ACKNOWLEDGEMENTS

The work of the first author was supported by the DST,Govt.of India,New Delhi under grant No.SR/FTP/MA-05/2002.

REFERENCES

[1] Harier, E., Norsett, S. and Wanner, G., *Solving ordinary differential equations, nonstiff problems*, Springer, Berlin, 1987.

[2] Hout, K.J. and Spijker, M.N., Stability analysis of numerical methods for delay differential equations, *Numer. Math.*, 59 (1991), 807-814.

[3] Zennaro, M. P-Stability properties of Runge-Kutta methods for delay differential equations, *Numer. Math.*, 49 (1986), 305-318.

[4] Miller, W. T., Sutton, R. S. and Werbos, P. J., *Neural Networks for Control*, Cambridge, MA:MIT press, 1990.

[5] Hunt, K.J., Sabarbaro, D., Zbikowski, R. and Gawthrop,P.J., Neural networks for control systems: A survey, *Automatica*, 28 (1992), 1083-1112.

[6] Lambert, J. D. *Computation Methods in O.D.E.,* New York, Wiley, 1973, ch.4.

[7] Caller, F. and Desoer, C., *Linear System Theory*. New York, Springer-verlag, 1992.

[8] Narendra, K. S. and Parthasarathy, K., Identification and control of dynamical natural system using neural networks, *IEEE Trans. Neural Networks*, 1 (1990), 4-27.

[9] Flechter, R. *Practical Methods of Optimization*, 2nd ed.,Chichester, U.K., Wiley, 1987

ON ROSENBERGER'S CONJECTURE FOR GENERALIZED TRIANGLE GROUPS OF TYPES $(2, 10, 2)$ AND $(2, 20, 2)$

V. Beniash-Kryvets
Department of Algebra
Byelorussian State University, 4, F. Skaryny Ave., 220050, Minsk, Belarus
e-mail: benyash@bsu.by

1. INTRODUCTION

Tits [16] proved that if G is a finitely generated linear group then G contains either a non abelian free subgroup or a solvable subgroup of finite index. Let Γ be an arbitrary finitely generated group. One says that the Tits alternative holds for Γ if Γ contains either a non abelian free subgroup or a solvable subgroup of finite index.

A one-relator free product of a family of groups $\{G_i\}$, $i \in I$, is called the group $G = (*G_i)/N(S)$, where S is a cyclically reduced word in the free product $*G_i$, $N(S)$ is its normal closure. Here S is called the relator. One-relator free products share many properties with one-relator groups (Howie [8]). We consider the case when G_i's are cyclic groups.

Definition 1. *A group Γ having a presentation*

$$\Gamma =< a_1, \ldots, a_n; a_1^{l_1} = \ldots = a_n^{l_n} = R^m(a_1, \ldots, a_n) = 1 >, \qquad (1)$$

where $n \geq 2$, $m \geq 1$, $l_i = 0$ or $l_i \geq 2$ for all i, and $R(a_1, \ldots, a_n)$ is a cyclically reduced word in the free group on $a_1, \ldots, a_n$ which is not a proper power, is called a one-relator product of n cyclic groups.

One relator products of cyclic groups provide a natural algebraic generalization of Fuchsian groups which are one relator products of cyclics relative to the standard Poincare presentation (see Fine and Rosenberger [7])

$$F = \langle a_1, \ldots, a_p, b_1, \ldots, b_t, c_1, d_1, \ldots, c_r, d_r; \\ a_i^{m_i} = a_1 \ldots a_p b_1 \ldots b_t [c_1, d_1] \ldots [c_r, d_r] = 1 \rangle.$$

If $n = 2$ and $m \geq 2$ in (1) then we have so-called *generalized triangle groups*

$$T(k, l, m, R) = \langle a, b; a^k = b^l = R^m(a, b) = 1 \rangle.$$

If $R(a,b) = ab$ then we obtain an ordinary triangle group.

Let Γ be a group of the form (1) and $m \geq 2$. If either $n \geq 4$ or $n = 3$ and $(l_1, l_2, l_3) \neq (2,2,2)$ then Γ contains a free subgroup of rank 2. If $n = 3$ and $(l_1, l_2, l_3) = (2,2,2)$ then Γ contains either a free subgroup of rank 2 or a free abelian subgroup of rank 2 and index 2 (Fine, Levin and Rosenberger[6]).

The case when Γ is a generalized triangle group is much more difficult. Rosenberger [14] stated the following conjecture.

Conjecture 1. *The Tits alternative holds for generalized triangle groups.*

Fine, Levin, and Rosenberger [6] proved this conjecture in the following cases: 1) $l = 0$ or $k = 0$; 2) $m \geq 3$. Now suppose that $k, l, m \geq 2$. Let $s(\Gamma) = 1/k + 1/l + 1/m$. If $s(\Gamma) < 1$ then Baumslag, Morgan and Shalen [1] proved that the group Γ contains a non abelian free subgroup. Using some new methods, Howie [9] proved Conjecture 1 in the case when $s(\Gamma) = 1$ and up to equivalence $R \neq ab$. If $s(\Gamma) = 1$ and $R = ab$ then Γ is an ordinary triangle group. The classical result says that Γ contains $\mathbb{Z}$ as a subgroup of finite index.

Now consider groups of the form

$$\Gamma = T(2, l, 2, R) = \langle a, b; a^2 = b^l = R^2(a,b) = 1 \rangle, \tag{2}$$

where $l > 2$, $R = ab^{v_1} \dots ab^{v_s}$, $0 < v_i < l$. In the following cases Conjecture 1 holds for Γ: 1) $s \leq 4$ (Rosenberger [14], Levin and Rosenberger[10]); 2) $l > 5$ and $l \neq 6, 10, 12, 15, 20, 30, 60$ (Beniash-Kryvets [2], [3]); 3) $l = 6, 12, 30, 60$ (Beniash-Kryvets, Barkovich [4]). In this paper we prove the following theorem.

Theorem 1. Let Γ be a group of the form *(2)* with $s \geq 5$ and $l = 10, 20$. Then Γ contains a free subgroup of rank 2.

Thus, Conjecture 1 is still open for groups $T(k, l, 2, R)$ with $k = 2, 3$ and $l = 3, 4, 5$ and $T(2, 15, 2)$.

2. SOME AUXILIARY RESULTS

In this section we prove several auxiliary results used in the proof of Theorem 1. Throughout we shall denote the ring of algebraic integers in $\mathbb{C}$ by $\mathcal{O}$, the group of units in $\mathcal{O}$ by $\mathcal{O}^*$, the free group of a rank 2 with generators g and h by $F_2 =< g, h >$, the greatest common divisor of integers a and b by (a, b), the image of a matrix

$A \in \mathrm{SL}_2(\mathbb{C})$ in $\mathrm{PSL}_2(\mathbb{C})$ by $[A]$, the trace of a matrix A by $\operatorname{tr} A$, and the identity matrix in $\mathrm{SL}_2(\mathbb{C})$ by E. The following lemma characterizes elements of finite order in $\mathrm{PSL}_2(\mathbb{C})$.

Lemma 1. Let $2 \le m \in \mathbb{Z}$ and $\pm E \ne X \in \mathrm{SL}_2(\mathbb{C})$. Then $[X]^m = 1$ in $\mathrm{PSL}_2(\mathbb{C})$ if and only if $\operatorname{tr} X = 2\cos\frac{r\pi}{m}$ for some $r \in \{1, \dots, m-1\}$.

The proof easily follows from the fact that $\varepsilon, \varepsilon^{-1}$, where ε is a root of unity of degree m, are the eigenvalues of the matrix X.

We shall use standard facts from geometric representation theory (see Culler and Shalen [5], Lubotzky and Magid [11]). Here we recall some notations. Let $F_n = \langle g_1, \dots, g_n \rangle$ be a free group, $R(F_n) = \mathrm{SL}_2(\mathbb{C})^n$ be a representation variety of F_n in $\mathrm{SL}_2(\mathbb{C})$. The group $\mathrm{GL}_2(\mathbb{C})$ acts naturally on $R(F_n)$ (by simultaneous conjugation of components) and its orbits are in one-to-one correspondence with the equivalence classes of representations of F_n. $\mathrm{GL}_2(\mathbb{C})$-orbits are not necessarily closed and so the variety of orbits (the geometric quotient) is not an algebraic variety. However one can consider the categorical quotient $R(F_n)/\mathrm{GL}_2(\mathbb{C})$ (see Mumford [13]), which we shall denote by $X(F_n)$ and call it the variety of characters. By construction, its points parametrize closed $\mathrm{GL}_2(\mathbb{C})$-orbits. It is well known that an orbit of a representation is closed iff the corresponding representation is fully reducible and so the points of the variety $X(F_n)$ are in one-to-one correspondence with the equivalence classes of fully reducible representations of Γ in $\mathrm{SL}_2(\mathbb{C})$.
For an arbitrary element $g \in F_n$ one can consider the regular function

$$\tau_g : R(F_n) \to \mathbb{C}, \qquad \tau_g(\rho) = \operatorname{tr} \rho(g).$$

Usually τ_g is called *a Fricke character* of the element g. It is known that the $\mathbb{C}$-algebra $T(F_n)$ generated by all functions τ_g, $g \in F_n$, is equal to $\mathbb{C}[X(F_n)] = \mathbb{C}[R(F_n)]^{\mathrm{GL}_2(\mathbb{C})}$. Combining results of Culler and Shalen [5], Sibirskij [15], it is easy to see that $T(F_n)$ is generated by Fricke characters $\tau_{g_i} = x_i$, $\tau_{g_i g_j} = y_{ij}$, and $\tau_{g_i g_j g_k} = z_{ijk}$, where $1 \le i < j < k \le n$. Consider a morphism $\pi : R(F_n) \to \mathbb{A}^t$ defined by

$$\pi(\rho) = (x_1(\rho), \dots, x_n(\rho), y_{12}(\rho), \dots, y_{n-1,n}(\rho), z_{123}(\rho), \dots, z_{n-2,n-1,n}(\rho)), \quad (3)$$

where $t = n + n(n-1)/2 + n(n-1)(n-2)/6$. The image $\pi(R(F_n))$ is closed in $\mathbb{A}^t$ (see Culler and Shalen [11]). Since $X(F_n)$ and $\pi(R(F_n))$ are biregularly isomorphic, we shall identify $X(F_n)$ and $\pi(R(F_n))$. Obviously, $\dim R(F_n) = 3n$ and $\dim X(F_n) = 3n - 3$. Set

$$R^s(F_n) = \{\rho \in R(F_n) \mid \rho \text{ is irreducible}\}, \qquad X^s(F_n) = \pi(R^s(F_n)).$$

The sets $R^s(F_n)$ and $X^s(F_n)$ are open in Zariski topology subsets of $R(F_n)$ and $X(F_n)$ respectively (see Culler and Shalen [5]).

Now, consider a free group $F_2 = \langle g, h\rangle$. The ring $T(F_2)$ is generated by the functions τ_g, τ_h, and τ_{gh}.

Lemma 2. For all $\alpha, \beta, \gamma \in \mathbb{C}$ there exist matrices $A, B \in \mathrm{SL}_2(\mathbb{C})$ such that

$$\tau_g(A, B) = \operatorname{tr} A = \alpha, \quad \tau_h(A, B) = \operatorname{tr} B = \beta, \quad \tau_{gh}(A, B) = \operatorname{tr} AB = \gamma.$$

This lemma can be easily proved by straightforward computations.

Lemma 2 implies that $X(F_2) = \pi(R(F_2)) = \mathbb{A}^3$. Moreover, the functions $\tau_g, \tau_h, and \tau_{gh}$ are algebraically independent over $\mathbb{C}$ and for every $u \in F_2$ we have

$$\tau_u = Q_u(\tau_g, \tau_h, \tau_{gh}),$$

where $Q_u \in \mathbb{Z}[x, y, z]$ is a uniquely determined polynomial with integer coefficients (see Culler and Shalen [5]). The polynomial Q_u is usually called the Fricke polynomial of the element u.

Consider polynomials $P_n(\lambda)$ satisfying the initial conditions $P_{-1}(\lambda) = 0$, $P_0(\lambda) = 1$ and the recurrence relation

$$P_n(\lambda) = \lambda P_{n-1}(\lambda) - P_{n-2}(\lambda).$$

If $n < 0$ then we set $P_n(\lambda) = -P_{|n|-2}(\lambda)$. The degree of the polynomial $P_n(\lambda)$ is equal to n if $n > 0$ and to $|n| - 2$ if $n < 0$. It is easy to verify by induction on n that

$$P_n(2\cos\varphi) = \frac{\sin(n+1)\varphi}{\sin\varphi}. \tag{4}$$

It follows from (4) that the polynomial $P_n(\lambda)$, $n \geq 1$, has n zeros described by the formula

$$\lambda_{n,k} = 2\cos\frac{k\pi}{n+1}, \qquad k = 1, 2, \ldots, n. \tag{5}$$

Moreover, it is easy to verify by induction that for $n \geq 0$ we have

$$P_{2n}(\lambda) = \lambda^{2n} + \cdots + (-1)^n,$$
$$P_{2n-1}(\lambda) = \lambda(\lambda^{2n-2} + \cdots + (-1)^{n-1}n). \tag{6}$$

Lemma 3. Let $k, l \in \mathbb{Z}$ and assume that $(k, l) = 1$ and $l \geq 2$ is not a power of a prime. Then $2\sin\frac{k\pi}{l} \in \mathcal{O}^*$.

Proof. Let $l = 2^t u$, where u is odd. If $t = 1$ then k is odd and $2\sin\frac{k\pi}{l} = 2\cos\frac{r\pi}{u}$ with $r = (u-k)/2 \in \mathbb{Z}$. Since $u-1$ is even, it follows from (6) that $2\cos\frac{r\pi}{u} \in \mathcal{O}^*$.

If $t > 1$ then k is odd and $2\sin\frac{k\pi}{l} = 2\cos\frac{r\pi}{2^t u}$ with $r = 2^{t-1}u - k$.

If $t = 0$ then $2\sin\frac{k\pi}{l} = 2\cos\frac{r\pi}{2u}$ with $r = u - 2k$.

Thus, it is sufficient to prove that $2\cos\frac{r\pi}{2^t u} \in \mathcal{O}^*$, where $t \geq 1$, $(r, 2^t u) = 1$, $u > 1$ and u is not a power of a prime in the case $t = 1$. Let $u = p_1^{\alpha_1} \dots p_s^{\alpha_s}$, where p_i is a prime and $0 < \alpha_i \in \mathbb{Z}$ for $i = 1, 2, \dots, s$. By (5) numbers $\lambda_i = 2\cos\frac{i}{2^t u}\pi$, $i = 1, 2, \dots, 2^t u - 1$, are the roots of the polynomial $P_{2^t u-1}(\lambda)$, so that

$$P_{2^t u-1}(\lambda) = \prod_{i=1}^{2^t u-1} (\lambda - \lambda_i)$$

and the constant term of $P_{2^t u-1}$ is equal to $(-1)^{2^{t-1}-1} 2^{t-1} p_1^{\alpha_1} \dots p_s^{\alpha_s}$. On the other hand, the polynomials $P_{2p_i^{\alpha_i}-1}(\lambda)$, $i = 1, 2, \dots, s$, and $P_{2^t-1}(\lambda)$ have the roots $2\cos\frac{j\pi}{2p_i^{\alpha_i}}$, $j = 1, 2, \dots, 2p_i^{\alpha_i} - 1$, and $2\cos\frac{j\pi}{2^t}$, $j = 1, 2, \dots, 2^t - 1$, respectively. Hence, all these polynomials divide $P_{2^t u-1}(\lambda)$ and any two of them have only one common root $\lambda = 0$. Hence,

$$P_{2^t u-1}(\lambda) = F(\lambda) F_1(\lambda),$$

where

$$F(\lambda) = \lambda^{-s} P_{2^t-1}(\lambda) \prod_{i=1}^{s} P_{2p_i^{\alpha_i}-1}(\lambda).$$

By (5) the constant term of $F(\lambda)$ is equal to $(-1)^{2^{t-1}-1} 2^{t-1} p_1^{\alpha_1} \dots p_s^{\alpha_s}$. Consequently, the constant term and the leading coefficient of $F_1(\lambda)$ are equal to 1. Since $2\cos\frac{r\pi}{2^t u}$ is not a root of $F(\lambda)$, it is a root of $F_1(\lambda)$ and we obtain $2\cos\frac{r\pi}{2^t u} \in \mathcal{O}^*$ as required. □

Furthermore, we require more detailed information on the Fricke polynomials. Let $w = g^{\alpha_1} h^{\beta_1} \dots g^{\alpha_s} h^{\beta_s} \in F_2$ and let $x = \tau_g$, $y = \tau_h$, $z = \tau_{gh}$. Let us treat the Fricke polynomial $Q_w(x, y, z)$ as a polynomial in z. Set

$$Q_w(x, y, z) = M_n(x, y) z^n + M_{n-1}(x, y) z^{n-1} + \dots + M_0(x, y).$$

Lemma 4 ([17]). The degree of the Fricke polynomial $Q_w(x, y, z)$ with respect to z is equal to s and its leading coefficient $M_s(x, y)$ has the form

$$M_s(x, y) = \prod_{i=1}^{s} P_{\alpha_i-1}(x) P_{\beta_i-1}(y). \tag{7}$$

A subgroup $H \in \mathrm{PSL}_2(\mathbb{C})$ is called *non-elementary* if H is infinite, irreducible and non-isomorphic to a dihedral group.

Lemma 5 ([12]). Let $H \in \mathrm{PSL}_2(\mathbb{C})$ be a non-elementary subgroup. Then H contains a non-abelian free subgroup.

Lemma 6 ([5]). Let $A, B \in \mathrm{SL}_2(\mathbb{C})$ and $\operatorname{tr} A = x$, $\operatorname{tr} B = y$, and $\operatorname{tr} AB = z$. A subgroup $< A, B >$ is irreducible if and only if

$$\operatorname{tr} ABA^{-1}B^{-1} = x^2 + y^2 + z^2 - xyz - 2 \neq 2.$$

3. Proof of Theorem 1.

Let Γ be a group in Theorem 1, that is,

$$\Gamma = T(2, l, 2, R) = \langle a, b; a^2 = b^l = R^2(a, b) = 1 \rangle, \tag{8}$$

where $l \in \{10, 15, 20\}$, $R = ab^{v_1} \dots ab^{v_s}$, $0 < v_i < l$, $s > 4$. Set $V = \sum_{i=1}^{s} v_i$. If $(V, l) \neq 1$ then Γ contains a non-abelian free subgroup (see Beniash-Kryvets [2]). So we shall assume that $(V, l) = 1$. Without loss of generality we may assume that

$$V \equiv 1 \pmod{l}.$$

If $V \not\equiv 1 \pmod{l}$ then one can apply an automorphism of the free product $\langle a; a^2 = 1 \rangle * \langle b; b^l = 1 \rangle$, $a \mapsto a$ and $b \mapsto b^p$ with $(p, l) = 1$ and $pV \equiv 1 \pmod{l}$ to the word $R(a, b)$. To prove Theorem 1, we construct a representation $\rho : \Gamma \to \mathrm{PSL}_2(\mathbb{C})$ such that $\rho(\Gamma)$ contains a non-abelian free subgroup. Set

$$\beta = 2\cos\frac{\pi}{l}, \qquad f_R(z) = Q_R(0, \beta, z), \tag{9}$$

where Q_R is the Fricke polynomial of R.

Definition 2. *Let z_0 be a root of a polynomial $f_R(z)$ and $A, B \in \mathrm{SL}_2(\mathbb{C})$ be matrices such that $\operatorname{tr} A = 0$, $\operatorname{tr} B = \beta$, and $\operatorname{tr} AB = z_0$. We shall denote a subgroup of $\mathrm{PSL}_2(\mathbb{C})$, generated by $[A], [B]$, by $G(z_0)$.*

The group $G(z_0)$ is an epimorphic image of Γ since by Lemma 1

$$[A]^2 = [B]^l = R^2([A], [B]) = 1.$$

Lemma 7. Numbers $\pm 2\sin\frac{\pi}{l}$ are not roots of the polynomial $f_R(z)$.

Proof. Suppose that $f_R(-2\sin\frac{\pi}{l}) = 0$. Let ε be a primitive root of unity of degree $4l$. Consider a representation $\rho_k : F_2 \to \mathrm{SL}_2(\mathbb{C})$ defined by

$$\rho_k(g) = A = \begin{pmatrix} \varepsilon^l & 0 \\ 1 & \varepsilon^{-l} \end{pmatrix}, \qquad \rho_k(h) = B_k = \begin{pmatrix} \varepsilon^{2k} & x \\ 0 & \varepsilon^{-2k} \end{pmatrix}. \tag{10}$$

Then we have $\operatorname{tr} A = 0$, $\operatorname{tr} B_1 = \beta_k$, and $\operatorname{tr} AB_1 = x - 2\sin\frac{\pi}{l}$. So we obtain

$$f_R(z)(\rho_1) = f_R(x - 2\sin\frac{\pi}{l}) = g(x) = \operatorname{tr} R(A, B_1).$$

Since $-2\sin\frac{\pi}{l}$ is a root of $f_R(z)$, 0 is a root of $g(x)$. This means that a constant term of $g(x)$ is equal to 0. On the other hand, a constant term of $\operatorname{tr} R(A, B_1)$ is equal to

$$\varepsilon^{ls+2V} + \varepsilon^{-ls-2V} = 2\cos(\frac{ls+2V}{2l}\pi) \neq 0,$$

since $(V, l) = 1$ by assumption. This contradiction proves that $-2\sin\frac{\pi}{l}$ is not a root of $f_R(z)$. Analogously, by considering a matrix B_{-1} instead the matrix B_1, we obtain that $2\sin\frac{\pi}{l}$ is not a root of $f_R(z)$. □

Lemma 8. Assume that the polynomial $f_R(z)$ has a root $z_0 \neq 0$. Then Γ contains a non-abelian free subgroup.

Proof. By Lemma 7 we have $z_0 \neq \pm 2\sin\frac{\pi}{l}$. Let us show that $G(z_0)$ is a non-elementary subgroup of $\mathrm{PSL}_2(\mathbb{C})$. First, $G(z_0)$ is irreducible by Lemma 6 since

$$\operatorname{tr} ABA^{-1}B^{-1} - 2 = z_0^2 - 4\sin^2\frac{k\pi}{l} \neq 0.$$

Second, $G(z_0)$ is not a dihedral group since two of three numbers $\operatorname{tr} A$, $\operatorname{tr} B$, $\operatorname{tr} AB$ are not equal to 0 (see Majeed and Mason [12]). Third, it follows from classification of finite subgroups of SLC [12] that $G(z_0)$ is infinite because it is irreducible and contains an element $[B]$ of order greater than 5. Thus, $G(z_0)$ (and consequently Γ) contains a non-abelian free subgroup. □

Bearing in mind lemmas 7 and 8, we shall assume in what follows that

$$f_R(z) = M_R z^s, \tag{11}$$

where by Lemma 4

$$M_R = \prod_{i=1}^{s} P_{v_i - 1}(2\cos\frac{\pi}{l}) = (2\sin\frac{\pi}{l})^{-s} \prod_{i=1}^{s} 2\sin\frac{v_i\pi}{l}. \tag{12}$$

Let A, B_t be matrices defined in (10), $W(A, B_t) = AB_t^{u_1} \dots AB_t^{u_s}$, where $0 < u_i < l$. A set $(u_1, \dots, u_s)$ will be considered as cyclically ordered. Let

$$l_i = |\{j \mid u_j = i\}|, \qquad f_{i,j} = |\{r \mid u_r = i, u_{r+1} = j\}|. \tag{13}$$

We have the following equations:

$$\sum_{i=1}^{l-1} l_i = s, \quad \sum_{i=1}^{l-1} f_{ij} = l_j, \quad \sum_{j=1}^{l-1} f_{ij} = l_i, \quad i, j = 1, \dots, l-1. \tag{14}$$

The following lemma will be heavily used.

Lemma 9. Let $g(x) = \operatorname{tr} W(A, B_t) = a_0 x^s + \dots + a_s$, $h_i = P_{i-1}(\varepsilon^{2t} + \varepsilon^{-2t})$. Then we have $a_0 = \prod_{j=1}^{s} h_{u_j}$ and

$$\begin{aligned} a_2 =& a_0 \sum_{j=1}^{l-1} \frac{f_{ii}}{h_i} \left(\frac{l_i - 2}{h_i} + \sum_{j \neq i} \frac{l_j \varepsilon^{2ti-2tj}}{h_j} \right) + \\ & a_0 \sum_{i \neq j} \frac{f_{ij}}{h_i} \left(\frac{l_i - 1}{h_i} + \frac{(l_j - 1)\varepsilon^{2ti-2tj}}{h_j} + \sum_{k \neq i, k \neq j} \frac{l_k \varepsilon^{2ti-2tk}}{h_k} \right) - \\ & a_0 \left(\sum_{i=1}^{l-1} \frac{l_i(l_i - 1)}{2h_i^2} (\varepsilon^{4ti} + \varepsilon^{-4ti}) + \sum_{i \neq j} \frac{l_i l_j}{h_i h_j} (\varepsilon^{2ti+2tj} + \varepsilon^{-2ti-2tj}) \right). \end{aligned} \tag{15}$$

This lemma can be proved by direct computations.

Lemma 10. Let $f_{R,1}(z) = M_{R,1} z^s$.

1. If $l = 10$ and s is odd then $l_1 + l_9 = l_3 + l_7 + 1$ and $l_2 = l_4 = l_5 = l_6 = l_8 = 0$.

2. If $l = 10$ and $s = 4s_1$ then $l_2 = 1$, $l_1 + l_9 = l_3 + l_7 - 1$, $l_4 = l_5 = l_6 = l_8 = 0$ and $V = 20t + 1$ for some $t \in \mathbb{Z}$.

3. If $l = 10$ and $s = 4s_1 + 2$ then $l_8 = 1$, $l_1 + l_9 = l_3 + l_7 - 1$, $l_2 = l_4 = l_5 = l_6$ and $V = 20t + 11$ for some $t \in \mathbb{Z}$.

Proof. Let ρ_{-1} be a representation defined by (10). Then

$$g(x) = f_{R,1}(x + 2\sin\frac{\pi}{l}) = M_{R,1}(x + 2\sin\frac{\pi}{l}) = \operatorname{tr} R(A, B_{-1}). \tag{16}$$

Comparing constant terms in (16), we obtain

$$\prod_{i=1}^{s} 2\sin\frac{v_i\pi}{l} = 2\cos\frac{ls-2V}{2l}\pi. \tag{17}$$

1. If $l = 10$ and $s = 2s_1 + 1$ then we must have $2\cos\frac{10s_1+5-V}{10}\pi = 2\sin\frac{\pi}{10} \in \mathcal{O}^*$. Hence $2\sin\frac{v_i\pi}{10} \in \mathcal{O}^*$, $i = 1, \dots, s$. By Lemma 3 $(v_i, 10) = 1$ for all v_i's. Since $2\sin\frac{\pi}{10}2\sin\frac{3\pi}{10} = 1$, it follows from (17)

$$(2\sin\frac{\pi}{10})^{l_1+l_9-l_3-l_7-1} = 1.$$

Hence, $l_1 + l_9 - l_3 - l_7 - 1 = 0$ as required.

2. If $l = 10$ and $s = 4s_1$ then we must have

$$2\cos\frac{20s_1 - V}{10}\pi = 2\cos\frac{V\pi}{10} = 2\sin\frac{2\pi}{5}.$$

Hence $V = 20t + 1$ for some $t \in \mathbb{Z}$. Since $2\sin\frac{r\pi}{5} \notin \mathcal{O}^*$ and $2\sin\frac{2\pi}{5}/2\sin\frac{r\pi}{5} \in \mathcal{O}^*$ for any r prime to 5, we obtain from (17) that only one of v_i's is even (let, for example, v_1 is even). If $v_1 = 2$ then

$$(2\sin\frac{\pi}{10})^{l_1+l_9-l_3-l_7+1} = 1,$$

hence $l_1 + l_9 = l_3 + l_7 - 1$ as required. If $v_1 = 4$ or $v_1 = 6$ then it follows from (17) that $l_1 + l_9 = l_3 + l_7$. But in this case $V = v_1 + l_1 + 3l_3 + 7l_7 + 9l_9 = v_1 + 4l_3 + 8l_7 + 8l_9$ is even which is a contradiction. If $v_1 = 8$ then as above $l_1 + l_9 = l_3 + l_7 - 1$ and

$$V = 20t + 1 = 7 + 4l_3 + 8l_7 + 8l_9,$$

which is a contradiction.

3. This case can be proved in the same way as the previous one. □

Let

$$\Gamma_1 = T(2, 5, 2, R) = \langle c, d; c^2 = d^5 = S^2(c, d) = 1\rangle, \tag{18}$$

where $S = cdu_1 \dots cd^{u_s}$, $0 < u_i < 5$, $s > 4$. Set $U = \sum_{i=1}^{s} u_i$. If $(U, 5) \neq 1$ then Γ contains a non-abelian free subgroup (see Beniash-Kryvets [2]). So we shall assume that $(U, 5) = 1$. As above, without loss of generality we may assume that

$$U \equiv 1 \pmod 5.$$

Set

$$h_S(z) = Q_S(0, 2\cos\frac{\pi}{5}, z), \tag{19}$$

where Q_S is the Fricke polynomial of S. Let z_0 be a root of a polynomial $f_R(z)$ and $A, B \in \mathrm{SL}_2(\mathbb{C})$ be matrices such that $\operatorname{tr} A = 0$, $\operatorname{tr} B = \beta$, $\operatorname{tr} AB = z_0$. As above we shall denote a subgroup of $\mathrm{PSL}_2(\mathbb{C})$, generated by $[A], [B]$, by $G(z_0)$. The group $G(z_0)$ is an epimorphic image of Γ_1 since by Lemma 1

$$[A]^2 = [B]^5 = S^2([A], [B]) = 1.$$

Lemma 11. Numbers $\pm 2\sin\frac{\pi}{5}$ are not roots of the polynomial $h_S(z)$.

The proof of Lemma 11 is similar to proof of Lemma 7.

Lemma 12. Assume that the polynomial $h_S(z)$ has a root $z_0 \notin \{0, \pm 1, \pm 2\cos\frac{2\pi}{5}\}$. Then Γ_1 contains a non-abelian free subgroup.

Proof. By Lemma 11 we have $z_0 \neq \pm 2\sin\frac{\pi}{5}$. Let us show that $G(z_0)$ is a non-elementary subgroup of $\mathrm{PSL}_2(\mathbb{C})$. First, $G(z_0)$ is irreducible by Lemma 6 since

$$\operatorname{tr} ABA^{-1}B^{-1} - 2 = z_0^2 - 4\sin^2\frac{\pi}{5} \neq 0.$$

Second, $G(z_0)$ is not a dihedral group since two of the three numbers $\operatorname{tr} A$, $\operatorname{tr} B$, $\operatorname{tr} AB$ are not equal to 0 (see Majeed and Mason [12]). Third, it follows from classification of finite subgroups of $\mathrm{PSL}_2(\mathbb{C})$ (see Vinberg, Kaplinsky [18]) that $G(z_0)$ is infinite. Thus, $G(z_0)$ (and consequently Γ_1) contains a non-abelian free subgroup. □

Bearing in mind lemmas 11 and 12, we shall assume in what follows that

$$h_S(z) = K_S z^{a_1}(z-1)^{a_2}(z+1)^{a_3}(z-2\cos\frac{2\pi}{5})^{a_4}(z+2\cos\frac{2\pi}{5})^{a_5}, \tag{20}$$

where by Lemma 4

$$K_S = \prod_{i=1}^{s} P_{u_i-1}(2\cos\frac{\pi}{5}) = (2\cos\frac{\pi}{5})^{l_2+l_3}. \tag{21}$$

Lemma 13. Let $h_S(z)$ has the form (20). Then

1. If s is even then $a_1 = 0$, $a_2 = a_3$, $a_4 = a_5$ and $l_2 + l_3 = 2a_3 + 1$, where l_2, l_3 are defined by (13).

1. If s is odd then $a_1 = 1$, $a_2 = a_3$, $a_4 = a_5$ and $l_2 + l_3 = 2a_3$.

Proof. Let ρ_{-1} be a representation defined by (10) with $l = 5$. Then

$$p(x) = h_S(x + 2\cos\frac{3\pi}{10}) = \operatorname{tr} R(A, B_{-1}). \tag{22}$$

Comparing constant terms in (22),we obtain

$$(2\cos\frac{\pi}{5})^{l_2+l_3}(2\cos\frac{3\pi}{10})^{a_1}(2\cos\frac{3\pi}{10}-1)^{a_2}(2\cos\frac{3\pi}{10}+1)^{a_3}(2\cos\frac{3\pi}{10}-2\cos\frac{2\pi}{5})^{a_4}$$
$$(2\cos\frac{3\pi}{10}+2\cos\frac{2\pi}{5})^{a_5} = 2\cos\frac{5s+2U}{10}\pi. \tag{23}$$

1. If s is even then we must have $2\cos\frac{5s+2U}{10}\pi = 2\cos\frac{\pi}{5} \in \mathcal{O}^*$ in (23). Hence $a_1 = 0$. Using identities

$$(2\cos\frac{3\pi}{10}-1)(2\cos\frac{3\pi}{10}+1) = (2\cos\frac{\pi}{5})^{-2},$$
$$(2\cos\frac{3\pi}{10}-2\cos\frac{2\pi}{5})(2\cos\frac{3\pi}{10}+2\cos\frac{2\pi}{5}) = 1,$$

we can write (23) in the form

$$(2\cos\frac{\pi}{5})^{l_2+l_3-2a_3-1}(2\cos\frac{3\pi}{10}-1)^{a_2-a_3}(2\cos\frac{3\pi}{10}-2\cos\frac{2\pi}{5})^{a_4-a_5} = 1. \tag{24}$$

It is not difficult to see that (24) implies

$$a_2 = a_3, \quad a_4 = a_5, \quad l_2 + l_3 = 2a_3 + 1,$$

as required.

The case when s is odd can be proved analogously. □

3.1. THE CASE $l = 10$

First, let $s = 4s_1$. Consider a representation $\rho : F_2 \to \mathrm{PSL}_2(\mathbb{C})$, $\rho(g) = A$, $\rho(h) = B_1$, where A, B_1 are defined in (10). Then we have

$$f_1(x) = f_R(z)(\rho) = M_R(x + 2\cos\frac{3\pi}{5})^s = \operatorname{tr} R(A, B_1). \tag{25}$$

Further, $l_2 = 1$, $l_4 = l_5 = l_6 = l_8 = 0$ and equations (14) have the form

$$\begin{aligned}
f11 + f12 + f13 + f17 + f19 &= l1, & f31 + f32 + f33 + f37 + f39 &= l3,\\
f71 + f72 + f73 + f77 + f79 &= l7, & f91 + f92 + f93 + f97 + f99 &= l9,\\
f11 + f21 + f31 + f71 + f91 &= l1, & f13 + f23 + f33 + f73 + f93 &= l3,\\
f17 + f27 + f37 + f77 + f97 &= l7, & f19 + f29 + f39 + f79 + f99 &= l9,\\
f12 + f32 + f72 + f92 &= 1, & f21 + f23 + f27 + f29 &= 1,\\
l1 + l9 &= l3 + l7 - 1, & l3 + 2 * l7 + 2 * l9 &= 5 * t,\\
l1 + l3 + l7 + l9 + 1 &= 4 * s1 & &.
\end{aligned} \tag{26}$$

The coefficient a_2 of the polynomial $f_1(x)$ is equal to

$$a_2 = M_R(2\cos\frac{3\pi}{5})^2 s(s-1)/2$$

by (25). Taking into account Lemma 9, we obtain following equation

$$\begin{aligned}
(2\cos\frac{3\pi}{5})^2 s(s-1)/2 = &\sum_{j=1}^{l-1}\frac{f_{ii}}{h_i}\left(\frac{l_i-2}{h_i} + \sum_{j\neq i}\frac{l_j\varepsilon^{2i-2j}}{h_j}\right) +\\
&\sum_{i\neq j}\frac{f_{ij}}{h_i}\left(\frac{l_i-1}{h_i} + \frac{(l_j-1)\varepsilon^{2i-2j}}{h_j} + \sum_{k\neq i, k\neq j}\frac{l_k\varepsilon^{2i-2k}}{h_k}\right) -\\
&\sum_{i=1}^{l-1}\frac{l_i(l_i-1)}{2h_i^2}(\varepsilon^{4i} + \varepsilon^{-4i}) + \sum_{i\neq j}\frac{l_i l_j}{h_i h_j}(\varepsilon^{2i+2j} + \varepsilon^{-2i-2j}).
\end{aligned} \tag{27}$$

The equation (27) can be written in the form

$$X_0 + X_1\varepsilon^4 + X_2\varepsilon^8 + X_3\varepsilon^{12} = 0, \tag{28}$$

where X_0, X_1, X_2, X_3 are polynomial with integer coefficients of l_i, f_{ij}, t and s_1. Since $1, \varepsilon^4, \varepsilon^8, \varepsilon^{12}$ are linearly independent over $\mathbb{Q}$, one obtains a system

$$X_0 = X_1 = X_2 = X_3 = 0. \tag{29}$$

Now, consider an epimorphic image $\Gamma_1 = \langle c, d; c^2 = d^5 = R^2(c,d) = 1\rangle$ of the group Γ, where $R(c,d) = cd^{v_1}\dots cd^{v_s}$. We can write the word $R(c,d)$ from the free product $\langle c; c^2 = 1\rangle * \langle d; d^5 = 1\rangle$ in the form $S(c,d) = cd^{u_1}\dots cd^{u_s}$, where $u_i = \begin{cases} v_i, & \text{if } v_i < 5,\\ v_i - 5, & \text{if } v_i > 5.\end{cases}$ Let $U = \sum_{i=1}^{s} u_i$. Since $(V, 10) = 1$, we have $(U, 5) = 1$. By Lemma 13

$$h_S(z) = K_S(z^2-1)^{a_2}(z^2 - (2\cos\frac{2\pi}{5})^2)^{a_4}. \tag{30}$$

Let ρ_2 be a representation defined by (10). Then

$$p(x) = h_S(x - 2\cos\frac{3\pi}{10}) = \operatorname{tr} R(A, B_2). \tag{31}$$

The coefficient b_2 of the polynomial $p(x)$ is equal to

$$b_2 = K_S(a_4 + a_2(2\cos\frac{2\pi}{5})^2 + 4(2s_1^2 - s_1)(2\cos\frac{3\pi}{10})^2).$$

On the other hand, we can apply Lemma 9 to compute b_2. Let

$$\begin{array}{llll}
l'_1 = l_1, & l'_2 = l_7 + 1, & l'_3 = l_3, & l'_4 = l_9, \\
f'_{11} = f_{11}, & f'_{12} = f_{17} + f_{12}, & f'_{13} = f_{13}, & f'_{14} = f_{14}, \\
f'_{21} = f_{71} + f_{11}, & f'_{22} = f_{77} + f_{72} + f_{27}, & f'_{23} = f_{23} + f_{73}, & f'_{24} = f_{24} + f_{74}, \\
f'_{31} = f_{31}, & f'_{32} = f_{37} + f_{32}, & f'_{33} = f_{33}, & f'_{34} = f_{34}, \\
f'_{41} = f_{91}, & f'_{42} = f_{97} + f_{92}, & f'_{43} = f_{93}, & f'_{44} = f_{99}.
\end{array} \tag{32}$$

Then we have an equation

$$\begin{aligned}
&a_4 + a_2(2\cos\frac{2\pi}{5})^2 + 4(2s_1^2 - s_1)(2\cos\frac{3\pi}{10})^2 = \\
&\sum_{j=1}^{l-1} \frac{f'_{ii}}{h_i}\left(\frac{l'_i - 2}{h_i} + \sum_{j\neq i}\frac{l'_j\varepsilon^{4i-4j}}{h_j}\right) + \\
&\sum_{i\neq j}\frac{f'_{ij}}{h_i}\left(\frac{l'_i - 1}{h_i} + \frac{(l'_j - 1)\varepsilon^{4i-4j}}{h_j} + \sum_{k\neq i, k\neq j}\frac{l'_k\varepsilon^{4i-4k}}{h_k}\right) - \\
&\sum_{i=1}^{l-1}\frac{l'_i(l'_i - 1)}{2h_i^2}(\varepsilon^{8i} + \varepsilon^{-8i}) + \sum_{i\neq j}\frac{l'_i l'_j}{h_i h_j}(\varepsilon^{4i+4j} + \varepsilon^{-4i-4j}).
\end{aligned} \tag{33}$$

The equation (33) can be written in the form

$$Y_0 + Y_1\varepsilon^4 + Y_2\varepsilon^8 + Y_3\varepsilon^{12} = 0, \tag{34}$$

where Y_0, Y_1, Y_2, Y_3 are polynomial with integer coefficients of l_i, f_{ij}, t and s_1. One obtains a system

$$Y_0 = Y_1 = Y_2 = Y_3 = 0. \tag{35}$$

Thus, we have the equations (26), (29), (35). Solving this system by computer with Maple, one obtains, in particular, that

$$h_{37} = h_{13} - h_{17} - h_{39} + h_{79} + h_{73} + 1/2,$$

which is a contradiction, because h_{ij} is an integer. Therefore, Theorem 1 is proved in the case when $l = 10$ and $s = 4s_1$.

If $s = 4s_1 + 2$ then analogously we obtain $h_{77} < 0$, which is a contradiction. If s is odd then analogously we obtain $h_{77} = -h_{79} - h_{79} + l_9 - h_{99} \pm 1/2$, which is a contradiction. Thus, the case $l = 10$ is proved.

3.2. The case $l = 20$.

Lemma 14. Let $f_{R,20}(z) = M_{R,20}z^s$. Then $v_i \neq 10$ for $i = 1, \ldots, s$.

Proof. Let us assume the contrary. Let, for example, $v_1 = 10$. Consider a representation ρ_{-1} defined by (10). Then

$$g(x) = f_{R,20}(x + 2\sin\frac{\pi}{20}) = M_{R,20}(x + 2\sin\frac{\pi}{20}) = \operatorname{tr} R(A, B_{-1}). \tag{36}$$

Comparing constant terms in (36), we obtain

$$\prod_{i=1}^{s} 2\sin\frac{v_i\pi}{20} = 2\prod_{i=2}^{s} 2\sin\frac{v_i\pi}{20} = 2\cos\frac{ls - 2V}{40}\pi \in \mathcal{O}^* \tag{37}$$

by Lemma 3. Hence $1/2 \in \mathcal{O}^*$ which is a contradiction. □

Now, let

$$\Gamma_1 = \langle c, d; c^2 = d^{10} = R^2(c, d)\rangle$$

be an epimorphic image of Γ, where $R(c, d) = cd^{v_1} \ldots cd^{v_s}$. Since $v_i \neq 10$ for all i, we can write the word $R(c, d)$ from the free product $\langle c; c^2 = 1\rangle * \langle d; d^{10} = 1\rangle$ in the form $S(c, d) = cd^{u_1} \ldots cd^{u_s}$, where $u_i = \begin{cases} v_i, & \text{if } v_i < 10, \\ v_i - 10, & \text{if } v_i > 10. \end{cases}$. It was proved above that Γ_1 contains a non-abelian free subgroup. Hence Γ contains a non-abelian free subgroup as well.

Theorem 1 is proved.

REFERENCES

[1] Baumslag, G., Morgan, J. W. and Shalen, P.B. Generalized triangle groups, *Math. Proc. Cambridge Philos. Soc.*, 102 (1987), 25-31.

[2] Beniash-Kryvets, V. On free subgroups of some generalized triangle groups, *Dokl. Akad. Nauk Belarus*, 47:2 (2003), 29-32.

[3] Beniash-Kryvets, V. On the Tits alternative for some finitely generated groups, *Dokl. Akad. Nauk Belarus*, 47:3 (2003), 14-17.

[4] Beniash-Kryvets, V. and Barkovich, O. On the Tits alternative for some generalized triangle groups, *Algebra and Discrete Mathematics*, 2 (2004), 15-35.

[5] Culler, M. and Shalen, P. Varieties of group representations and splittings of 3 manifolds, *Ann. of Math.*, 117 (1983), 109-147.

[6] Fine, B., Levin, F. and Rosenberger, G. Free subgroups and decompositions of one-relator products of cyclics. Part I: the Tits alternative, *Arch. Math.*, 50 (1988), 97-109.

[7] Fine, B. and Rosenberger, G., *Algebraic generalizations of discrete groups. A path to combinatorial group theory through one-relator products*, Marcel Dekker, New York, 1999.

[8] Howie, J. One-relator products of groups, *Proceedings of groups St. Andrews, Cambridge University Press*, 1985, 216-220.

[9] Howie, J. Free subgroups in groups of small deficiency, *J. of Group Theory*, 1 (1998), 95-112.

[10] Levin, F. and Rosenberger, G. On free subgroups of generalized triangle groups, Part II, *Proceedings of the Ohio State-Denison Conference on Group Theory, (ed. S. Sehgal et al), World Scientific*, 1993, 206-222.

[11] Lubotzky, A. and Magid, A. Varieties of representations of finitely generated groups, *Memoirs AMS*, 58 (1985), 1-116.

[12] Majeed, A. and Mason, A.W. Solvable-by-finite subgroups of GL(2, F), *Glasgow Math. J.*, 19 (1978), 45-48.

[13] Mumford, D. *Geometric invariant theory*, Springer-Verlag, New York, 1965.

[14] Rosenberger, G. On free subgroups of generalized triangle groups, *Algebra i Logika*, 28 (1989), 227-240.

[15] Sibirskij, K.S. Algebraic invariants for a set of matrices, *Sib. Math. J.*, 9:1 (1968), 115-124.

[16] Tits, J. Free subgroups in linear groups, *J. Algebra*, 20 (1972), 250-270.

[17] Traina, C. Trace polynomial for two generated subgroups of $SL_2(\mathbb{C})$, *Proc. Amer. Math. Soc.*, 79 (1980), 369-372.

[18] Vinberg, E. and Kaplinsky, Y. Pseudo-finite generalized triangle groups, *Preprint 00-003, Universität Bielefeld*, 2000.

EXPONENTIAL FITTING: GENERAL APPROACH AND APPLICATIONS FOR ODE-SOLVERS

G. V. Berghe
Department of Applied Mathematics and Computer Science
Ghent University, Krijgslaan 281 (S9), B-9000 Gent, Belgium

Abstract

The idea of exponential fitting will be explained and applied to the construction of ODE-solvers. In particular attention will be given to the extension of the Euler and Numerov methods. A numerical experiment related to the Schrödinger equation will show the merits of the exponential-fitted versions of the Numerov method.

1. SITUATION OF THE PROBLEM

In the field of the numerical solution of ODEs a series of methods have been devised for the case when the solution is known to exhibit a specific oscillatory or exponential behaviour. The idea of using a basis of functions other than polynomials has a long history, going back to papers, where sets of exponential functions were used to derive the coefficients of the methods for the first order ODE

$$y' = f(x, y)\,. \tag{1}$$

Methods using trigonometric polynomials have also been considered; for theoretical aspects see Gautschi [5]. Salzer [18] assumed that the solution is a linear combination of trigonometric functions, of the form

$$y(x) = \sum_{j=0}^{J} [a_j \sin(jx) + b_j \cos(jx)]\,, \tag{2}$$

with arbitrary constant coefficients a_j and b_j; expressions of the coefficients of these methods are given in that paper for small values of J.

Sheffield [19] and Stiefel and Bettis [20] considered the second order ODE of the form

$$y'' = f(x, y)\,, \tag{3}$$

for the orbit problem in celestial mechanics. They constructed multistep methods which are exact if the solution is of the form

$$y(x) = \sum_{j=1}^{J} [f_1^j(x) \sin(\omega_j x) + f_2^j(x) \cos(\omega_j x)]\,, \tag{4}$$

where f_1^j and f_2^j are low degree polynomials. A simple two-step method which is exact for the form (4) with $J = 1$ and constant f_1^1 and f_2^1 has been derived by Denk [3] by means of a principle of coherence.

Coleman [1] considered a special family of methods, hybrid versions included, by means of a technique based on rational approximations of the cosine.

Methods for the solution of

$$y^{(r)} = f(x, y)\,,\ r = 1, 2, \ldots \tag{5}$$

were derived by Vanthournout et al. [28] and Vanden Berghe et al. [21] as an application of the mixed interpolation technique; they are exact if $y(x)$ is of the form

$$y(x) = f_1(x)\sin(\omega\ x) + f_2(x)\cos(\omega\ x) + \phi(x)\ ,$$

where f_1 and f_2 are constants and ϕ is a polynomial of low degree.

The existence of a large variety of techniques, which sometimes seem to have distinct areas of applicability, though this is not always true, is rather discouraging for a user who, without being directly implied in applied mathematics, is interested in getting the pertinent information for his or her own problem. Such a user would be better served if one and the same technique, hopefully transparent and simple enough, would be available for as many cases of interest as possible.

The exponential fitting (EF) technique is the best suited in this context. It is aimed at deriving linear approximation formulae for various operations on functions of the form

$$y(x) = \sum_{i=1}^{I} f_i(x)\exp(\mu_i x)\,, \tag{6}$$

where μ_i, called frequencies, are constants (complex in general) whose exact values, or some reasonable approximations of these, are known; it is also assumed that the functions $f_i(x)$ are smooth enough but only the numerical values of the whole $y(x)$ are available.

The idea of the approach consists in constructing the coefficients of the formula by asking it be exact for each of the following M functions (we choose here two μ_i-values, i.e. $\mu_1 = -\mu_2 = \mu$):

$$x^k\exp(\pm\mu\ x),\ k = 0, 1, 2, \ldots, M-1\ . \tag{7}$$

The value of M depends on the number N of coefficients to be evaluated. As a rule one has $M = N$ but exceptions do also exist.

As for the theoretical background of the procedure, this is inspired from the generalization of Lyche [16] of the approach of Henrici [7] on multistep methods for ODEs. This perhaps explains why for a long period of time the EF procedure was thought to cover only this field. As a matter of fact, the expression EF with the stated meaning seems to have been first used also in the context of solving ODEs, by Liniger and Willoughby [15]. However, as shown by Ixaru and the present author [11, 14], the area of applicability of this procedure is much broader; it covers operations as numerical differentiation, quadrature or interpolation, as well.

2. AN OUTLINE OF THE EF-PROCEDURE

In this section we describe the basic ingredients of the EF approach by treating in detail a simple case. Consider a one-step ODE solver for the first-order ODE $y' = f(x, y)$:

$$y_{n+1} = a_1 y_n + a_2 h f(x_n, y_n) \ , \tag{8}$$

where $y_{n+1} \approx y(x_{n+1}) = y(x_n + h)$. Notice that for the well-known Euler method the occurring parameters $a_1 = a_2$ both have a value 1.

In the EF-procedure one introduces an operator $\mathcal{L}$, acting on $y(x)$ and depending parametrically on h and on the parameters $\mathbf{a} = [a_1, a_2]$:

$$\mathcal{L}[h, \mathbf{a}]y(x) := y(x + h) - a_1 y(x) - a_2 h y'(x) \ ,$$

where $\mathbf{a}$ is the vector of coefficients $a_1, a_2,\ \mathbf{a} = [a_1, a_2]$. One asks for the determination of a_1, a_2 upon the condition that $\mathcal{L}[h, \mathbf{a}]y(x)$ is identically vanishing for some prerequisite forms of $y(x)$.

In first instance one considers as prerequisite form the set of *power functions*, i.e.

$$1, x, x^2, x^3, \ldots$$

The action of the operator $\mathcal{L}$ on these functions results in

$$\begin{aligned} \mathcal{L}[h, \mathbf{a}]1 &= 1 - a_1 \\ \mathcal{L}[h, \mathbf{a}]x &= (1 - a_1)x + h(1 - a_2) \\ \mathcal{L}[h, \mathbf{a}]x^2 &= (1 - a_1)x^2 + 2xh(1 - a_2) + h^2 \\ &\textit{etc.} \end{aligned}$$

Related to these expressions one introduces the so-called *moments*, $L_m(h, \mathbf{a})$, which represent the expressions of $\mathcal{L}[h, \mathbf{a}]x^m, (m = 0, 1, 2, \ldots)$ at $x = 0$, i.e

$$\begin{aligned} L_0(h, \mathbf{a}) &= 1 - a_1 \quad (9) \\ L_1(h, \mathbf{a}) &= h(1 - a_2) \quad (10) \\ L_2(h, \mathbf{a}) &= h^2 \quad (11) \\ &etc. \end{aligned}$$

Since $\mathcal{L}$ is a linear operator it follows that, upon taking $y(x)$ as a linear combination of power functions, $y(x) = y_0 + y_1 x + y_2 x^2 + y_3 x^3 + \ldots$, we have

$$\mathcal{L}[h, \mathbf{a}]y(x) = \sum_{m=0} y_m \mathcal{L}[h, \mathbf{a}]x^m$$
$$= L_0(h, \mathbf{a})(y_0 + y_1 x + y_2 x^2 + y_3 x^3 + \ldots) + L_1(h, \mathbf{a})(y_1 + 2y_2 x + 3y_3 x^2 + \ldots)$$
$$+L_2(h, \mathbf{a})(y_2 + 3y_3 x + 6y_4 x^2 + \ldots) + \ldots = \sum_{m=0}^{\infty} \frac{1}{m!} L_m(h, \mathbf{a}) D^m y(x).$$

We now address the problem of finding out the values of the coefficients a_1 and a_2 such that the function $\mathcal{L}[h, \mathbf{a}]y(x)$ is identically vanishing at any x and at any $h \in (0,\ H]$ for as many successive terms as possible in the classical power set, i.e

$\mathcal{L}[h, \mathbf{a}]1 = 0$ is equivalent to $L_0(h, \mathbf{a}) = 0$
$\mathcal{L}[h, \mathbf{a}]1 = \mathcal{L}[h, \mathbf{a}]x = 0$ is equivalent to $L_0(h, \mathbf{a}) = L_1(h, \mathbf{a}) = 0$, etc...

In general, the set of conditions $\mathcal{L}[h, \mathbf{a}]x^m = 0,\ m = 0, 1, 2, \ldots, M - 1$ is equivalent to

$$L_m(h, \mathbf{a}) = 0,\ m = 0, 1, 2, \ldots, M - 1 \ , \quad (12)$$

which is a set of M linear equations in two unknowns. The stated problem is then equivalent to that of finding the biggest M such that (12) is compatible. On using the expressions under (9-11) we find out that $M = 2$ and that $a_1 = a_2 = 1$ and the method obtained is the classical Euler method.

$$y_{n+1} = y_n + hf(x_x, y_n) \ .$$

On the other hand $L_2(h, \mathbf{a}) = h^2 \neq 0$ showing that finally

$$\begin{aligned}\mathcal{L}[h,\mathbf{a}]y(x) &= \frac{1}{2}L_2(h,\mathbf{a})y^{(2)} + \frac{1}{3!}L_3(h,\mathbf{a})y^{(3)} + \dots \\ &= \frac{1}{2}h^2y^{(2)} + \frac{1}{6}h^3y^{(3)} + \dots\end{aligned}$$

To summarize, we have established the following result: on imposing the stated conditions on the function $\mathcal{L}[h,\mathbf{a}]y(x)$ on the power set, we obtained $a_1 = a_2 = 1$ and these are the classical coefficients. In addition we have found that the leading term of the error of the classical formula is

$$lte_{clas} = \frac{1}{2}h^2y^{(2)} \ ,$$

a well-known expression.

Our derivation has also shown that the maximal M is 2, which at its turn indicates that approximation (8) is exact for two successive power functions 1 and x and for any linear combination of them or, in other words, for any first degree polynomial.

Let us now take some arbitrary real or imaginary μ and introduce the set of pairs of exponentials

$$\exp(\pm\mu x), x\exp(\pm\mu x), x^2\exp(\pm\mu x), \dots \ ,$$

which will be called the exponential fitting set. On applying the operator considered on the members of the set we obtain:

$$\begin{aligned}\mathcal{L}[h,\mathbf{a}]\exp(\mu x) &= \exp(\mu x)(\exp(h\mu) - a_1 - a_2h\mu) \\ &= \exp(\mu x)(\exp(z) - a_1 - a_2z) \\ &= \exp(\mu x)E_0(z,\mathbf{a}) \ , \\ \mathcal{L}[h,\mathbf{a}]\exp(-\mu x) &= \exp(-\mu x)E_0(-z,\mathbf{a}) \ ,\end{aligned}$$

where $z = \mu h$. Notice that in general $E_m := \mathcal{L}[h,\mathbf{a}]x^m\exp(\mu x)|_{x=0}$ and in particular $\mathcal{L}[h,\mathbf{a}]\exp(\mu x) = \exp(\mu x)E_0$. It is easy to verify that $E_m = \dfrac{\partial E_{m-1}}{\partial\mu}, m = 1,2,\dots$. Conditions $\mathcal{L}[h,\mathbf{a}]x^k\exp(\mu x) = 0$ imply for any x and any $h \neq 0$ $E_k(z,\mathbf{a}) = 0$, which means that, upon introducing $Z = \mu^2h^2$ and $G^{\pm}(Z,\mathbf{a}) := \frac{1}{2}[E_0(z,\mathbf{a})] \pm E_0(-z,\mathbf{a})]$, we should equivalently have

$$G^{\pm}(Z,\mathbf{a}) = 0 \ .$$

In general introducing $G^{\pm(p)}(Z,\mathbf{a})$ p-th derivative of $G^{\pm}(Z,\mathbf{a})$ implies that $E_p(\pm z,\mathbf{a}) = 0$ is equivalent with $G^{\pm(p)}(Z,\mathbf{a}) = 0$.

We are now ready to tackle the problem of extending formula (8) for the exponential fitting forms of $y(x)$. As the investigation just performed on the classical set indicated that equation (8) is exact for *two* functions, it is appropriate to consider here the functions $\exp(\pm\mu x)$ and the set of linear equations for the coefficients are:

$$E_0(z,\mathbf{a}) = E_0(-z,\mathbf{a}) = 0 \ .$$

or

$$\begin{cases} \exp(z) - a_1 - a_2 z & = 0 \\ \exp(-z) - a_1 + a_2 z & = 0 \ , \end{cases}$$

which is equivalent with $G^{\pm}(Z,\mathbf{a}) = 0$, i.e.

$$\begin{cases} G^{+}(Z,\mathbf{a}) = \cosh(z) - a_1 & = 0 \\ G^{-}(Z,\mathbf{a}) = \dfrac{\sinh(z)}{z} - a_2 & = 0 \ , \end{cases}$$

resulting in

$$y_{n+1} = \cosh(z) y_n + h \frac{\sinh(z)}{z} f(x_n, y_n) \ .$$

As, for the error, the differential equation $y'' - \mu^2 y = 0$ is the one which has the functions $\exp(\pm\mu x)$ as its linear independent solutions and then the leading term of the error should be of the form

$$A(-\mu^2 y + y'') \ .$$

The factor A is fixed by considering the fact that the coefficient of y should be the same in the classical (polynomial) expansion and in the above equation, i.e.

$$A = -\frac{1}{\mu^2} L_0(h,\mathbf{a}) = -\frac{1}{\mu^2}(1 - \cosh(z)) \ .$$

Notice that in the limit $\mu \to 0$ the new formulae tends to the classical ones.

In general the frequency μ can be

- real: $\mu = \omega$
- pure imaginary: $\mu = i\lambda$

In the latter case all occurring hyperbolic functions transform into trigonometric ones. To generalize the above results we introduce new functions, useful in all cases, i.e. the η-functions. These functions were originally introduced in Section 3.4 of [8]. The η functions are real functions of the real variable Z. The functions $\eta_{-1}(Z)$ and $\eta_0(Z)$ are introduced by the formulae:

$$\eta_{-1}(Z) := \begin{cases} \cos(|Z|^{1/2}) & \text{if } Z < 0 \\ \cosh(Z^{1/2}) & \text{if } Z \geq 0 \end{cases}$$

and

$$\eta_0(Z) := \begin{cases} \dfrac{\sin(|Z|^{1/2})}{|Z|^{1/2}} & \text{if } Z < 0 \\ 1 & \text{if } Z = 0 \\ \dfrac{\sinh(Z^{1/2})}{Z^{1/2}} & \text{if } Z > 0 \end{cases}$$

while $\eta_s(Z)$ with $s > 0$ are subsequently generated by recurrence

$$\eta_s(Z) = \frac{1}{Z}[\eta_{s-2}(Z) - (2s-1)\eta_{s-1}(Z)], \quad s = 1, 2, 3, \ldots$$

These functions satisfy several properties of which the following ones are of interest for the present discussion

- Differentiation:

$$\eta_s'(Z) = \frac{1}{2}\eta_{s+1}(Z), \ s = \ -1, \ 0, \ 1, \ 2, \ \ldots$$

- Generating differential equation: $\eta_s(Z)$ ($s = 0, \ 1, \ \ldots$) is the regular solution of

$$Zw'' + \frac{1}{2}(2s+3)w' - \frac{1}{4}w = 0.$$

- Relation with the spherical Bessel functions:

$$\eta_s(-x^2) = \ x^{-s} j_s(x), \ s = 0, \ 1, \ 2, \ \ldots$$

With the generalized notation the EF Euler method can be written as

$$y_{n+1} = \eta_{-1}(Z) y_n + h\eta_0(Z) f(x_n, y_n) \ ,$$

while the leading term of the corresponding error reads

$$-\frac{1}{\mu^2}(1-\eta_{-1}(Z))(-\mu^2 y+y'')\ .$$

3. SIX-STEP FLOW CHART

For the construction of EF-methods when functions of the forms $f_1(x)\sin(\omega x)+f_2(x)\cos(\omega x)$ or $f_1(x)\sinh(\lambda x)+f_2(x)\cosh(\lambda x)$, are present, the following procedure to get tuned formulae can be followed in general:

Step *i* . Choose the appropriate form of $\mathcal{L}[h,\,\mathbf{a}]$ and find its classical moments:

$$L_m(h,\mathbf{a}),\ m=0,\ 1,\ 2,\ \ldots\ .$$

Step *ii* . Examine the algebraic system

$$L_m(h,\mathbf{a})=0,\ m=0,\ 1,\ 2,\ \ldots,M-1\ ,$$

to find out the maximal M for which it is compatible.

Step *iii* . Denote $z:=\mu h$, construct the formal expression of $E_0(z,\,\mathbf{a})$ and, on this basis, write the expressions of $G^{\pm}(Z,\mathbf{a})$ where $Z:=z^2$. Also write the expressions of their derivatives $G^{\pm(p)}(Z,\mathbf{a}),\ p=1,\ 2,\ \ldots$ with respect to Z.

Step *iv* . Choose the reference set of M functions which is appropriate for the given form of $y(x)$. This is in general a hybrid set:

$$1,\ x,\ x^2\ ,\ldots,\ x^K,$$
$$\exp(\pm\mu x),\ x\exp(\pm\mu x),\ x^2\exp(\pm\mu x),\ \ldots,x^P\exp(\pm\mu x),$$

with $K+2P=M-3$.

The reference set is thus characterized by two integer parameters, K and P. The set in which there is no classical component is identified by $K=-1$ while the set in which there is no exponential fitting component is identified by $P=-1$. Parameter P is the level of tuning.

Step *v* . Solve the algebraic system

$$\begin{cases} L_k(h,\mathbf{a})=0, & 0\le k\le K \\ G^{\pm(p)}(Z,\mathbf{a})=0, & 0\le p\le P \end{cases}$$

for the Z dependent coefficients $\mathbf{a}(Z)$.

Step *vi* . The leading term of the error of the formula is proportional with

$$\frac{L_{K+1}(h,\mathbf{a})}{(K+1)!Z^{P+1}}\ D^{K+1}(D^2-\mu^2)^{P+1}y(X),$$

where $\mu = i\ \omega$ or $\mu = \lambda$, according to the case.

4. THE EF-NUMEROV METHODS

4.1 The construction of the methods

As explained in the previous sections and in [11, 14] the six-step procedure is followed for every construction of EF-methods; here we shall illustrate the technique for the construction of EF versions of the Numerov method. This method is often used for the derivation of the numerical solution of the initial value problem for second order differential equations of the special form:

$$y'' = f(x,y),\ x \in [a,\ b],\ y(a) = y_0,\ y'(a) = y'_0\,. \tag{13}$$

The form of these algorithms is:

$$y_{n+1} + a_1 y_n + y_{n-1} = h^2[b_0(f_{n+1} + f_{n-1}) + b_1 f_n], \tag{14}$$

where $x_{n\pm1} = x_n \pm h$, y_n is an approximation to $y(x_n)$ and $f_n = f(x_n, y_n)$. The coefficients of the classical version are

$$a_1 = -2,\ b_0 = \frac{1}{12},\ b_1 = \frac{5}{6}. \tag{15}$$

Following steps of the six-steps procedure [11, 14] can be applied:
Step *i*. Choose the appropriate form of $\mathcal{L}[h,\ \mathbf{a}]$ and find the expressions of its classical moments $L_m(h,\mathbf{a}),\ m = 0,\ 1,\ 2,\ \ldots$. With $\mathbf{a} := [a_1,\ b_0,\ b_1]$ we define $\mathcal{L}[h,\mathbf{a}]$ by

$$\begin{aligned}\mathcal{L}[h,\mathbf{a}]y(x) &:= y(x+h) + a_1 y(x) + y(x-h) \\ &- h^2[b_0(y''(x+h) + y''(x-h)) + b_1 y''(x)].\end{aligned} \tag{16}$$

The expressions of the classical moments (powers of h omitted) are:

$$L_0(\mathbf{a}) = 2 + a_1,\ L_2(\mathbf{a}) = 2(1 - 2b_0 - b_1),$$

$$L_{2k}(\mathbf{a}) = 2 - 4k(2k-1)b_0,\ k = 2,\ 3,\ \ldots$$

$$L_{2k+1}(\mathbf{a}) = 0,\ k = 0,\ 1,\ \ldots$$

Step *ii*. Examine the algebraic system

$$L_m(h, \mathbf{a}) = 0, \quad m = 0,\ 1,\ 2,\ \ldots, M-1 \tag{17}$$

to find out the maximal M for which it is compatible. System $L_k(\mathbf{a}) = 0,\ 0 \le k \le 5$ is compatible by which $M = 6$ and it has solution (15), further denoted as S_0.
Step *iii*. Let $z = \mu h$. We construct $E_0(z, \mathbf{a})$ by applying $\mathcal{L}$ on $y(x) = \exp(\mu x)$ to obtain

$$E_0(z, \mathbf{a}) = \exp(z) + \exp(-z) + a_1 - z^2[b_0(\exp(z) + \exp(-z)) + b_1],$$

and then

$$G^+(Z, \mathbf{a}) = 2\eta_{-1}(Z) + a_1 - Z[2b_0\eta_{-1}(Z) + b_1], \quad G^-(Z, \mathbf{a}) = 0,$$

where $Z = z^2$. It follows that

$$G^{+(1)}(Z, \mathbf{a}) = \eta_0(Z) - (2\eta_{-1}(Z) + Z\eta_0(Z))b_0 - b_1,$$

$$G^{+(m)}(Z, \mathbf{a}) = 2^{-m+1}[\eta_{m-1}(Z) - (3\eta_{m-2}(Z) + \eta_{m-3}(Z))b_0],\ m = 2,\ 3,\ \ldots$$

$$G^{-(m)}(Z, \mathbf{a}) = 0,\ m = 1,\ 2,\ \ldots$$

Step *iv*. Choose the reference set of M functions which is appropriate for the given form of $y(x)$. This is in general a hybrid set:

$$\begin{aligned} y = {} & 1,\ x,\ x^2\ , \ldots,\ x^K, \\ & \exp(\pm\mu x),\ x\exp(\pm\mu x),\ x^2\exp(\pm\mu x),\ \ldots,\ x^P\exp(\pm\mu x), \end{aligned} \tag{18}$$

with

$$K + 2P = M - 3. \tag{19}$$

Step *v*. Solve the algebraic system

$$L_k(\mathbf{a}) = 0, \quad 0 \le k \le K, \quad G^{\pm(p)}(Z, \mathbf{a}) = 0, \quad 0 \le p \le P \tag{20}$$

for the Z dependent coefficients and let $\mathbf{a}(Z) = [a_0(Z),\ b_0(Z),\ b_1(Z)]$ be its solution. Three options for tuning are then available: (*i*) $P = 0,\ K = 3$, (*ii*) $P = 1,\ K = 1$, and (*iii*) $P = 2,\ K = -1$ and these lead to the schemes S_1, S_2 and S_3, respectively.

S_1. The six functions to be integrated exactly by the algorithm are 1, x, x^2, x^3, and the pair $\exp(\pm\mu x)$ and therefore the system to be solved is $L_k(\mathbf{a}) = 0$, $0 \le k \le 3$ and $G^{\pm}(Z, \mathbf{a}) = 0$. The system is compatible and its solution is

$$a_1(Z) = -2, \quad b_0(Z) = \frac{1}{Z} - \frac{1}{2(\eta_{-1}(Z) - 1)}, \quad b_1(Z) = 1 - 2b_0(Z). \tag{21}$$

S_2. The functions to be integrated exactly are now 1, x, and the pairs $\exp(\pm\mu x)$ and $x\exp(\pm\mu x)$. Since all classical moments with odd indices and all G functions with a minus sign for the upper index are identically vanishing, the system to be solved is simply $L_0(\mathbf{a}) = G^+(Z,\mathbf{a}) = G^{+(1)}(Z,\mathbf{a}) = 0$, with the solution

$$\begin{aligned} &a_1(Z) = -2,\ b_0(Z) = \frac{1}{Z} - \frac{2(\eta_{-1}(Z)-1)}{Z^2\eta_0(Z)} = \frac{\eta_1(Z/4)}{4\eta_{-1}(Z/4)}, \\ &b_1(Z) = 2[\frac{\eta_{-1}(Z)-1}{Z} - b_0(Z)\eta_{-1}(Z)] = \eta_0^2(Z/4) - 2b_0(Z)\eta_{-1}(Z). \end{aligned} \tag{22}$$

S_3. The reference set of six functions is $\exp(\pm\mu x)$, $x\exp(\pm\mu x)$, $x^2\exp(\pm\mu x)$ and thus the system to be solved is

$$G^+(Z,\mathbf{a}) = G^{+(1)}(Z,\mathbf{a}) = G^{+(2)}(Z,\mathbf{a}) = 0,$$

with the solution

$$\begin{aligned} &a_1(Z) = -(6\eta_{-1}(Z)\eta_0(Z) - 2\eta_{-1}^2(Z) + 4)/D(Z), \\ &b_0(Z) = \eta_1(Z)/D(Z),\ b_1(Z) = (4\eta_0^2(Z) - 2\eta_1(Z)\eta_{-1}(Z))/D(Z), \end{aligned} \tag{23}$$

where $D(Z) = 3\eta_0(Z) + \eta_{-1}(Z)$.

Step *vi*. The leading term of the error of the formula can be obtained.

S_0.

$$lte_{clas} = -\frac{h^6}{240}y^{(6)}(x_n).$$

S_1.

$$lte_{ef} = -h^6\frac{1-12b_0(Z)}{12Z}(-\mu^2 y^{(4)}(x_n) + y^{(6)}(x_n)).$$

S_2.

$$lte_{ef} = h^6\frac{Z^2\eta_0(Z) - 4(\eta_{-1}(Z)-1)^2}{Z^4\eta_0(Z)}[\mu^4 y''(x_n) - 2\mu^2 y^{(4)}(x_n) + y^{(6)}(x_n)].$$

S_3.

$$lte_{ef} = h^6\ \frac{N(Z)}{F(Z)} \times [-\mu^6 y(x_n) + 3\mu^4 y^{(2))}(x_n) - 3\mu^2 y^{(4)}(x_n) + y^{(6)}(x_n)]$$

where

$$N(Z) = 6\eta_0(Z) + 2\eta_{-1}(Z) - 6\eta_{-1}(Z)\eta_0(Z) + 2\eta_{-1}^2(Z) - 4$$

and

$$F(Z) = Z^3(3\eta_0(Z) + \eta_{-1}(Z))$$

. **4.2 A typical application of the Numerov Method**

The relative merits of each of the four versions $S_i, i = 0, 1, 2, 3$ can be evaluated by comparing the expressions of the *lte*. Notice that each *lte* has one and the same h^6-dependence.

- This means that the order of each version is 4.
- The factors in the middle are close to $-1/240$ when Z is not too large.
- The difference in accuracy is contained in the third factors.

To illustrate the differences in the third factors we consider the case of the Schrödinger equation. The problem to be solved is

$$y'' + (E - V(x))y = 0, \quad x > 0$$

where $\lim_{x\to\infty} V(x) = 0$, with the conditions

$$y(0) = 0$$

and $y(x)$ is finite for any $x > 0$. For each of the S_i methods the expressions of the *lte* for this particular problem can be evaluated.

For such an equation the knowledge of the potential function $V(x)$ and of the energy E is sufficient to get reasonable approximations for frequencies: the integration domain $[a, b]$ is divided in subintervals and on each of them the function $V(x)$ is approximated by a constant $\bar{V}$. On all steps in such a subinterval one and the same μ^2 is used, $\mu^2 = \bar{V} - E$. If this is done the order of each version remains four but the errors will be very different when big values of the energy are involved. To see this let us denote $\Delta V(x) = V(x) - \bar{V}$, express the higher order derivatives of y in terms of y, y', μ^2, $\Delta V(x)$ and the derivatives of $V(x)$, for example $y''(x) = (V(x)-E)y(x) = (\mu^2+\Delta V(x))y(x)$, $y^{(4)}(x) = V'(x)y(x)+(\mu^2+\Delta V(x))y'(x)$ etc., and finally introduce them in the expressions of the last factors in the *lte*, which will be denoted $\Delta_i(x_n)$, $i = 0, 1, 2, 3$. The expression of each $\Delta_i(x_n)$ resulting from such a treatment will consist in a sum of y and y' with coefficients which depend on μ^2, $\Delta V(x)$ and on the derivatives of $V(x)$.

If $E >> \bar{V}$, $|\mu^2|$ has a big value and then the μ^2 dependence of $\Delta_i(x_n)$ will become dominating; the approximation $\mu^2 \approx -E$ will hold as well. To compare the errors it is then sufficient to organize the coefficients of y and of y' as polynomials in E and to retain only the terms with the highest power. This gives:

$$\Delta_0(x_n) \approx -E^3 y - 6EV'y' \,,$$

$$\Delta_1(x_n) \approx E^2 \Delta V y - 4EV'y' \,,$$

$$\Delta_2(x_n) \approx -E[5V^{(2)} + (\Delta V)^2]y - 2EV'y' \,,$$

$$\Delta_3(x_n) \approx -4EV^{(2)}y + (4V^{(3)} + 6V'\Delta V)y' \,.$$

Since in the discussed range of energies the solution is of oscillatory type with almost constant coefficients, the amplitude of the first derivative is bigger by a factor $E^{1/2}$ than that of the solution itself and then the error from the four schemes increases with E as E^3, E^2, $E^{3/2}$ and E, respectively.

For illustration we take as potential function the sum of the Woods–Saxon potential and its first derivative, that is

$$V(x) = v_0/(1+t) + v_1 t/(1+t)^2, \; t = \exp[(x - x_0)/a],$$

where $v_0 = -50$, $x_0 = 7$, $a = 0.6$ and $v_1 = -v_0/a$. Its shape is such that only two values for $\bar{V}$ are sufficient: $\bar{V} = -50$ for $0 \leq x \leq 6.5$ and $\bar{V} = 0$ for $x \geq 6.5$.

We solve the resonance problem which consists in the determination of the positive eigenvalues corresponding to the boundary conditions

$$y(0) = 0, \quad y(x) = \cos(E^{1/2}x) \text{ for big } x.$$

The physical interval $x \geq 0$ is cut at $b = 20$ and the eigenvalues are obtained by shooting at $x_c = 6.5$. The error in the eigenvalues will then reflect directly the quality of the solvers for the initial value problem used for the determination of the solution $y(x)$.

In table 1 we list the absolute errors for one particular eigenvalue for all four schemes of the Numerov method for several step length h; reference values, which are exact in the written figures, have been generated in a separate run with the method CPM(2) from [8] at $h = 1/16$. It is seen that, as expected, all these versions are of order four. In table 2 the absolute errors of four such eigenvalue for all schemes are listed. From these data it is obvious that the way in which the error increases with

h	S_0	S_1	S_2	S_3
1/16		79579	9093	721
1/32	595230	4734	525	46
1/64	36661	292	32	2
1/128	2287	18	1	0

Table 1: $E = 163.215298$

h	$S_0(E^3)$	$S_1(E^2)$	$S_2(E^{(3/2)})$	$S_3(E)$
$E = 53.58$				
1/64	989	22	5	1
$E = 163.21$				
1/64	36661	292	32	2
$E = 341.49$				
1/64	560909	2215	126	7
$E = 989.70$				
1/64		46269	975	28

Table 2: The behaviour of the absolute errors of the eigenvalues as a function of E

the energy differs from one version to another. The theoretical predictions of the behaviour of these errors with respect to E is confirmed.

5. SOME GENERAL COMMENTS

In this review paper we have demonstrated some applications of the EF-technique applied to ODE-solvers. Quite a lot of other research in that field have be done. We give here a short survey of these results and some references where some of the results have been presented:

- EF versions of the two-step bdf algorithm

$$a_0 y_n + a_1 y_{n+1} + y_{n+2} = h b_2 f(x_{n+2}, y_{n+2}) \ ,$$

  including stability theory, variable step form (see [12, 13]).

- Numerov method for second-order ODEs of the type $y'' = f(x, y)$

$$y_{n+1} + a_1 y_n + y_{n-1} = h^2 [b_0 (f_{n+1} + f_{n-1}) + b_1 f_n] \ ,$$

(see previous section).

- Symmetric four-step methods

$$y_{n+2} + a_1 y_{n+1} + a_2 y_n + a_1 y_{n-1} + y_{n-2} \\ = h^2(b_0 f_{n+2} + b_1 f_{n+1} + b_2 f_n + b_1 f_{n-1} + b_0 f_{n-2}),$$

(see [9, 10])

- Explicit Runge-Kutta EF methods including 2-,3- and 4-step methods, embedded pairs, etc.. (see [4, 23, 24, 26]).
- Implicit 2-step Runge-Kutta EF methods of order 2, 3 and 4 (see [25, 27]).
- Runge-Kutta-Nyström EF methods (see [17]).

More general applications of these EF technique can be found in [14].

REFERENCES

[1] Coleman, J. P. Numerical methods for $y'' = f(x, y)$ via rational approximations for the cosine. *IMA J. Numer. Anal.*, 9 (1989), 145–165.

[2] De Meyer, H., Vanthournout, J. and Vanden Berghe, G., On a new type of mixed interpolation. *J. Comput. Appl. Math.*, 30 (1990), 55–69.

[3] Denk, G. A new numerical method for the integration of highly oscillatory second-order ordinary differential equations. *Appl. Numer. Math.*, 13 (1993), 57–67.

[4] Franco, J. M., An embedded pair of exponentially fitted explicit Runge-Kutta methods. *J. Comp. Appl. Math.*, 149 (2002), 407–414.

[5] Gautschi, W. Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math.*, 3 (1961), 381–397.

[6] Greenwood, R. E. Numerical integration of linear sums of exponential functions. *Ann. Math. Stat.*, 20 (1949), 608–611.

[7] Henrici, P. *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York, 1962.

[8] Ixaru, L. Gr. *Numerical Methods for Differential Equations and Applications*, Reidel, Dortrecht - Boston - Lancaster, 1984.

[9] Ixaru, L. Gr. , Vanden Berghe, G. , De Meyer, H. and Van Daele, M., Four step exponential fitted methods for nonlinear physical problems. *Comput. Phys. Commun.*, 100 (1997), 56–70.

[10] Ixaru, L. Gr. , Vanden Berghe, G. , De Meyer, H. and Van Daele, M., `EXPFIT4` –A FORTRAN program for the numerical solution of systems of nonlinear second order initial value problems. *Comput. Phys. Commun.*, 100 (1997), 71–80.

[11] Ixaru, L. Gr. Operations on oscillatory functions. *Comput. Phys. Commun.*, 105 (1997), 1–19.

[12] Ixaru, L. Gr. , Vanden Berghe, G. and De Meyer, H., Frequency evaluation in exponential mutistep algorithms for ODEs. *J. Comput. Appl. Math.*, 140 (2002), 423–434.

[13] Ixaru, L. Gr. , Vanden Berghe, G. and De Meyer, H., Exponentially fitted variable two-step BDF algorithm for first order ODE. *Comput. Phys. Commun.*, 150 (2003), 116–128.

[14] Ixaru, L. Gr. and Vanden Berghe G., *Exponential fitting*. Kluwer Academic Publishers, Dortrecht (ISBN 1-4040-2099-6, eBook ISBN 1-4020-2100-3), (2004).

[15] Liniger, W. and Willoughby, R. A., Efficient integration methods for stiff systems of ordinary differential equations. *SIAM J. Numer. Anal.*, 7 (1970), 47–66.

[16] Lyche, T. Chebyshevian multistep methods for ordinary differential equations. *Numer. Math.*, 19 (1974), 65–75.

[17] Paternoster, B. Runge-Kutta(-Nyström) methods for ODEs with periodic solutions based in trigonometric polynomials, *Appl. Num. Math.*, 28 (1998), 401–412.

[18] Salzer, H. E. Trigonometric interpolation and predictor–corrector formulas for numerical integration. *ZAMM*, 9 (1962), 403–412.

[19] Sheffield, C. Generalized multi-step methods with an application to orbit computation. *Celestial Mech.*, 1 (1969), 46–58.

[20] Stiefel, E. and Bettis, D. G., Stabilization of Cowell's method. *Numer. Math.*, 13 (1969) 154–175.

[21] Vanden Berghe, G., De Meyer, H. and Vanthournout, J., A modified Numerov integration method for second order periodic initial–value problems. *Intern. J. Computer Math.*, 32 (1990), 233–242.

[22] Vanden Berghe, G. and De Meyer, H., Accurate computation of higher Sturm–Liouville eigenvalues. *Numer. Math.*, 59 (1991), 243–254.

[23] Vanden Berghe,G. , De Meyer, H. , Van Daele, M. and Van Hecke, T., Exponentially-fitted explicit Runge-Kutta methods. *Computer Phys. Comm.*, 123 (1999), 7–15.

[24] Vanden Berghe, G. , De Meyer, H. , Van Daele, M. and Van Hecke, T., Exponentially-fitted Runge-Kutta methods. *J. Comp. Appl. Math.*, 125 (2000), 107–115.

[25] Vanden Berghe, G. , Ixaru, L. Gr. and Van Daele, M., Optimal implicit exponentially-fitted Runge-Kutta methods. *Comp. Phys. Commun.*, 140 (2001), 346–357.

[26] Vanden Berghe, G. , Ixaru, L. Gr. and De Meyer, H., Frequency determination and step–length control for exponentially-fitted Runge–Kutta methods. *J. Comp. Appl. Math.*, 132 (2001), 95–105.

[27] Vanden Berghe, G. , Van Daele, M. and Vande Vyver, H., Exponential-fitted Runge–Kutta methods of collocation type: fixed or variable knot points? *J. Comp. Appl. Math.*, 159 (2003), 217–239.

[28] Vanthournout, J., Vanden Berghe, G. and De Meyer, H., Families of backward differentiation methods based on a new type of mixed interpolation. *Computers Math. Applic.*, 11 (1990), 19–30.

MATHEMATICS IN MEDICINE: BIOACOUSTIC MODELING AND COMPUTATIONS FOR AN ULTRASONIC IMAGING TECHNIQUE

A. Bounaïm[1,2], S. Holm[1,2], W. Chen[2], A. Ødegard[2]
[1] Department of Informatics,
University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway
email:(aichab,wenc,aasmund)@simula.no

[2] Simula Research Laboratory, PO. Box 134
NO-1325 Lysaker, Norway
email: sverre@ifi.uio.no.

Abstract:

Biomathematics use mathematics to quantitatively represent the dynamics of biological or biomedical systems and thereby analyze and predict system behavior. As an example, this work addresses an application of bioacoustic modeling and computations to a clinical imaging technique for breast cancer detection. The mathematical model consists in a damped wave equation incorporating a frequency-dependent attenuation, which describes ultrasound propagating in the human breast tissue. 3D numerical simulations are presented to investigate the dectectability of breast tumors. An extension to a more general model for the acoustic attenuation is also discussed. For this, 2D numerical experiments are presented to illustrate the issue in the case of the CARI technique.

1. INTRODUCTION - CLINICAL DESCRIPTION

Breast cancer became the most widespread female disease, in particular in western countries. Lives can be saved and treatment can be more effective if the diagnosis is made early. Ultrasonography is a common technique used in breast screening due to its low cost and large availability. Moreover, it is a good adjunct to mammography in differentiating cancerous from non-cancerous breast tumors. In this study, we are interested in the CARI (clinical amplitude-velocity reconstruction imaging) ultrasonic technique that was developed by Dr. K. Richter [1, 2].

The breast, in the CARI device, is fixed between two plates as schematically illustrated in Figure 1. The stainless steel plate, opposite to the transducer, plays the role of a reference structure producing a reflecting line (RL). The CARI modality operates in such that the RL is straight if the sound velocity in the intervening tissues is roughly homogeneous while it is elevated if the tissue contains a suspicious

tumor as shown from the CARI-ultrasonic image in Figure 2. The CARI technique is characterized by two important acoustic components of breast evaluation, namely the sound speed and the attenuation. Moreover, the CARI method is more sensitive than the conventional ultrasound, especially in assessing cancer surrounded by the breast fatty tissue.

In general, experimental study in living tissues is not practical, and acoustic phantoms are useful but limited. Therefore, mathematical computer modeling of ultrasound propagation and scattering complement to both approaches, although it has its own limitations. Moreover, recent advances in high-performance computing enable large-scale simulations such those occurring in high frequency acoustic wave propagation.

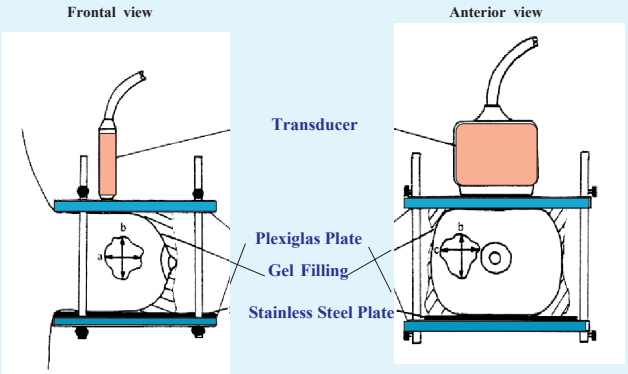


Figure 1: Two frontal views of the ultrasonic CARI technique for breast tumor detection taken from [2]

2. MATHEMATICAL AND GEOMETRICAL MODELING

To simulate ultrasound in breast models taking into account the two tissue parameters of the CARI technique, we solve the damped linear wave equation in an inhomogeneous lossy acoustic medium representing the breast fatty tissue:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} + \gamma \frac{\partial p}{\partial t} = \nabla^2 p, \tag{1}$$

where p is the pressure, c is the sound velocity, and γ is the damping or attenuation parameter. Note that wave attenuation is an essential tissue characteristic. There are various attenuation mechanisms where few of them can be isolated, and commonly the attenuation follows a power law in frequency f expressed as

$$\gamma = \alpha f^{\mathrm{y}}, \tag{2}$$

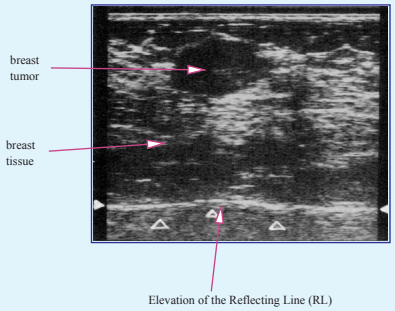


Figure 2: A CARI-ultrasound image showing the elevation of the reflecting line due to the presence of a tumor in the breast tissue

where coefficients α and y depend on the tissue. For example, in water $\alpha \approx 0.0022\text{dB/cm/MHz}^{\text{y}}$, y= 2.0, and in muscle tissue $\alpha \approx 0.7\text{dB/cm/MHz}^{\text{y}}$, y= 1.1. In our simulations, $\alpha = 2\alpha_0/c$ and the values of c, α_0 and y are deduced from clinical experiments and will be specified later. In the last section we introduce a more general attenuation model using Laplacian fractional derivative.

The equation (1) is supplemented with initial and boundary conditions according to the 3D configuration in Figure 3. The transducer is represented by a Dirichlet condition

$$p(x_{\texttt{trsd}}, t) = p_{\texttt{trsd}}(x, t). \tag{3}$$

The RL in the CARI setup is modeled by reflecting boundary (RB) conditions

$$\frac{\partial p}{\partial n}(x_{\texttt{RB}}, t) = 0, \tag{4}$$

while the remaining boundaries are represented by first-order absorbing or non-reflecting boundary (NRB) conditions:

$$\frac{\partial p}{\partial n}(x_{\texttt{NRB}}, t) = -\frac{1}{c}\frac{\partial p}{\partial t}. \tag{5}$$

The system is initialized with the conditions:

$$p(x, t_0) = p_{\texttt{atm}} \quad \texttt{and} \quad \frac{\partial p}{\partial t}(x, t_0) = 0. \tag{6}$$

The FETD (finite element time domain) approach used to discretize the equation (1) and the corresponding boundary conditions consists of a finite element method

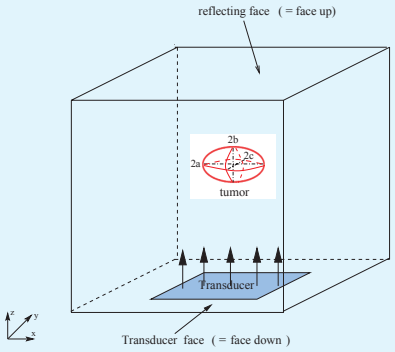


Figure 3: 3D configuration of the CARI technique for ultrasound breast tumor detection

in the spatial domain and a second order finite difference representation to evaluate the time derivatives. A semi-discrete time scheme of (1) writes

$$\frac{p^{n+1} - 2p^n + p^{n-1}}{\Delta t^2} + \gamma c^2 \frac{p^{n+1} - p^{n-1}}{2\Delta t} = c^2 \nabla^2 p^n, \tag{7}$$

where Δt is the time step size and the superscript n denotes the nth time iterate of the pressure field. Then, by decomposing p^n in a finite element basis and incorporating the boundary conditions, (7) leads to

$$A_1 p^{n+1} = A_2 p^n + A_3 p^{n-1}, \tag{8}$$

where the matrices $A_i, (i = 1, 2, 3)$ result from the finite element matrices and depend on the parameters c, γ, and Δt. In summary, the problem is reduced to the solution of a linear system at each time step. The numerical implementation is carried out using Diffpack, a finite element library based on C++ and object-oriented programming [3]. We refer to [7] for a detailed description of the FETD discretization method as well as a review on the stability of the numerical scheme.

3. NUMERICAL RESULTS AND DISCUSSIONS

Geometrically, the breast tissue is assimilated to a 3D box of size 22mmx24mmx20mm containing an ellipsoid-shaped tumor of axes 2a, 2b and 2c as shown in Figure 3. The transducer is a 12mmx8mm-rectangle from which a 3.5MHz signal is transmitted into the breast tissue. The sound speed in the breast tissue and the tumor are extracted from clinical experiments [4] together with the attenuation parameters, and are summarized in Table 1. Note that the transducer signal has a wavelength of $\lambda = \frac{f}{c} \approx 0.4$mm. Thus, for a better resolution of the spatial features, a grid

Table 1: Sound velocity and coefficients of frequency-dependent power law attenuation of the breast tissue and tumor.

	breast fatty tissue	breast cancer
c(m/s)	1475	1527
α_0(dB/MHz$^{\mathrm{y}}$)	15.8	57.0
y(s/m^2)	1.7	1.3

is chosen able to resolve 2 finite elements per wavelength which requires a grid of approximately 1.5x10^6 nodes. The numerical scheme is then stable for a time step Δt =26x10^{-9}s, and two ways of the wave traveling along the 3D breast model is achieved over 1125 time steps.

As shown in Figure 5, the waves are perturbed due to the presence of the tumor in the abnormal tissue compared to the healthy (or homogeneous) one. Here, the tumor is an ellipsoid of axes (12mm,8mm,8mm), and the views represent cross-sections normal to the wave propagation direction.
Clinically, the ultrasound imaging techniques have some limitations and lesions as small as 1cm-diameter can hardly be detected. The numerical experiments show instead that smaller lesions can be readily recognized in the tissue, an observation confirmed by 2D simulations in [7], which mimic cross-sections in 3D breast model. Besides the disturbance of the echo pattern around the lesion, snapshots from Figure 6 show that the ultrasound pressure is attenuated as the wave propagates along the tissue towards the RL and back to the transducer. Moreover, ultrasound pressure of a layer traversing the tumor (z = 5mm) displayed at successive time steps on Figure 7 gives a quantative evaluation in detecting the tumor and recognizing its shape.

4. ON A FRACTIONAL DERIVATIVE ATTENUATION MODEL

Acoustic waves propagating in media exhibiting arbitrary frequency power law attenuation can be modeled by time-domain partial differential equations given by (1). However, for non-integer power exponent y of the attenuation parameter γ, these models may not accurately describe more realistic media such as soft biological tissues. Therefore, we introduce in this section a new model for the dissipative term using a Laplacian fractional derivative developed by Chen and Holm [9, 10]:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} + 2\frac{\alpha_0}{c^{1-\mathrm{y}}\frac{\partial}{\partial t}((-\Delta)^{\mathrm{y}}/2}) = \nabla^2 p, \tag{9}$$

where the coefficients are similar to those introduced earlier. Chen and Holm note

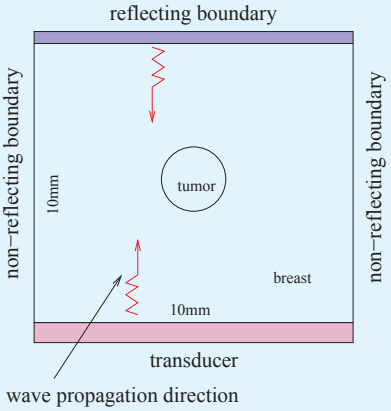


Figure 4: 2D configuration used for the FEM simulations of the wave propagating in attenuated medium

that the spatial fractional Laplacian models reflect the fractal microstructures of the media and describe quite well the frequency power law attenuation.
We aim in this section to develop a finite element approach to the new wave equation.

Using a finite element approach to the equation (9) and the Green's formula for the right hand side term, we assume we obtain the *pseudo-spatial approximation*:

$$\frac{1}{c^2}\frac{\partial^2(Mp)}{\partial t^2} + 2\frac{\alpha_0}{c^{1-\mathrm{y}}}\frac{\partial}{\partial t}(K^{\mathrm{y}/2}p) = -Kp + Bp. \tag{10}$$

The matrices M, K, and B are given, respectively, by

$$[M_{ij}] = \int_\Omega N_i N_j dx, \quad [K_{ij}] = \int_\Omega \nabla N_i \nabla N_j dx, \quad [B_{ij}] = \int_{\partial\Omega_{NR}} N_i N_j d\sigma, \tag{11}$$

where N_i are the finite element basis functions and ∇N_i their gradients. Ω_{NR} refers to as the non-reflecting boundaries of the computational domain as illustrated on Figure 4.
Then, using a second-order finite difference approximation in time, we get the discrete matrix form

$$\begin{aligned}\left[2M + 2\alpha_0 c^{1+\mathrm{y}}\Delta t K^{\mathrm{y}/2} + c\Delta t B\right] p^{n+1} &= -2c^2\Delta t^2 K p^n + 4Mp^n - 2Mp^{n-1} \\ &+ 2\alpha_0 c^{1+\mathrm{y}}\Delta t K^{\mathrm{y}/2} p^{n-1} + c\Delta t B p^{n-1}.\end{aligned}$$

The pressure is then calculated by a process solving a linear system at each time step.

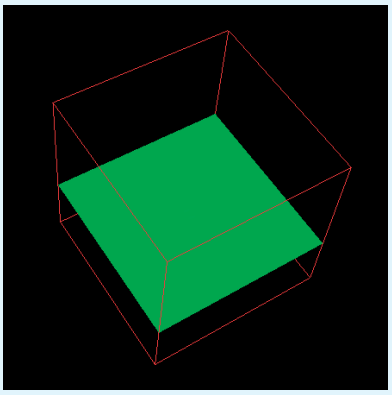
Some 2D numerical results

To our knowledge, numerical simulations involving fractional derivative are not widely treated in the literature. In this section, we investigate the feasibility of the finite element approach to give quantitative results on the behavior of the wave propagating in media presenting a fractional derivative attenuation model. We especially study the effect of the power exponent y on the presence or not of the oscillations within or outside the tumor region.
The domain is a 10mmx10mm square meshed with a 26x26nodes-grid, the sound speeds in the breast and the tumor respectively are $c_{\texttt{breast}} = 1475ms^{-1}$, $c_{\texttt{tumor}} = 1527ms^{-1}$, and $\alpha_0 = 1$ is the first attenuation coefficient. The matrix $K^{\texttt{y}}/2$ is computed using one of the matrix functionalities of matlab. The process is time consuming and results, in particular, in a full matrix.
Two series of numerical experiments are carried out for different values of y when the wave travels in the 2D breast model outside the tumor region: (1) for 5 values of y close to 0, Figure 8 shows that the oscillations are very similar; (2) for 5 values of y between 0.2 and 2, the results from Figure 9 show that the oscillations are present for y=0.2 and y=0.5, but then disappear when $\texttt{y} \geq 1$. It is also observed that the amplitude is smaller than the above case.
In Figures 10 and 11, the same observations are made from the results when the wave propagates across the tumor region. Other results are presented in [8].

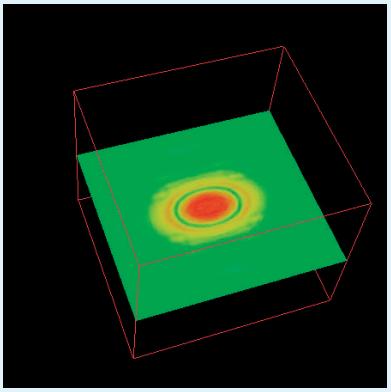
5. CONCLUSIONS AND PERSPECTIVES

This paper addresses bioacoustic numerical modeling for the CARI ultrasonic breast imaging technique. A finite element approach is presented and numerical experiments for a 3D breast model illustrate the detectability of lesions in the breast fatty tissue.

The attenuation model is extended by introducing a Laplacian fractional derivative. The discretization of the wave equation incorporating the new attenuation model is achieved by a finite element method. The numerical results, although limited to bi-dimensional case and simple boundary conditions, give insights in the feasibility of the attenuation modeling in human soft tissues. However, further analysis can be done to achieve a more accurate numerical approximation of the presented attenuation model.

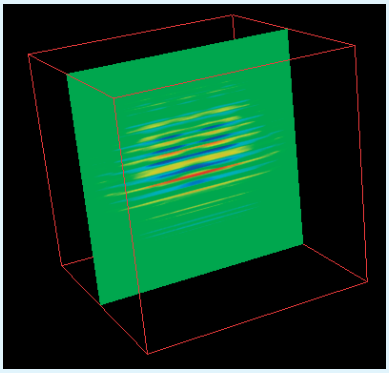


(a)

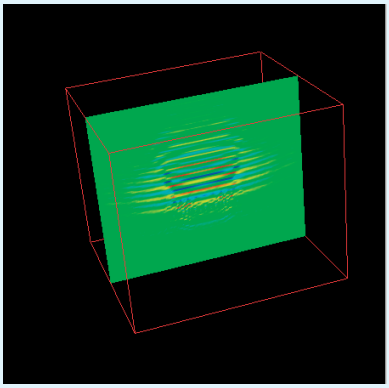


(b)

Figure 5: Ultrasound pressure in a cross-section (z=7mm) normal to the z-axis and traversing the tumor for two breast fatty tissues at t=23.9μs: (a) homogeneous tissue; (b) containing a (12mm,8mm,8mm)-ellipsoid tumor. The shape of the section is readily recongnized in the background medium.



(a)



(b)

Figure 6: Ultrasound pressure in a cross-section normal to the y-axis and traversing the tumor for two breast fatty tissues at t=23.9μs: (a) homogeneous tissue; (b) containing a (12mm,8mm,8mm)-ellipsoid tumor. The wave travels back to the transducer, and it is noted that it is disturbed around the tumor compared to the tissue without tumor.

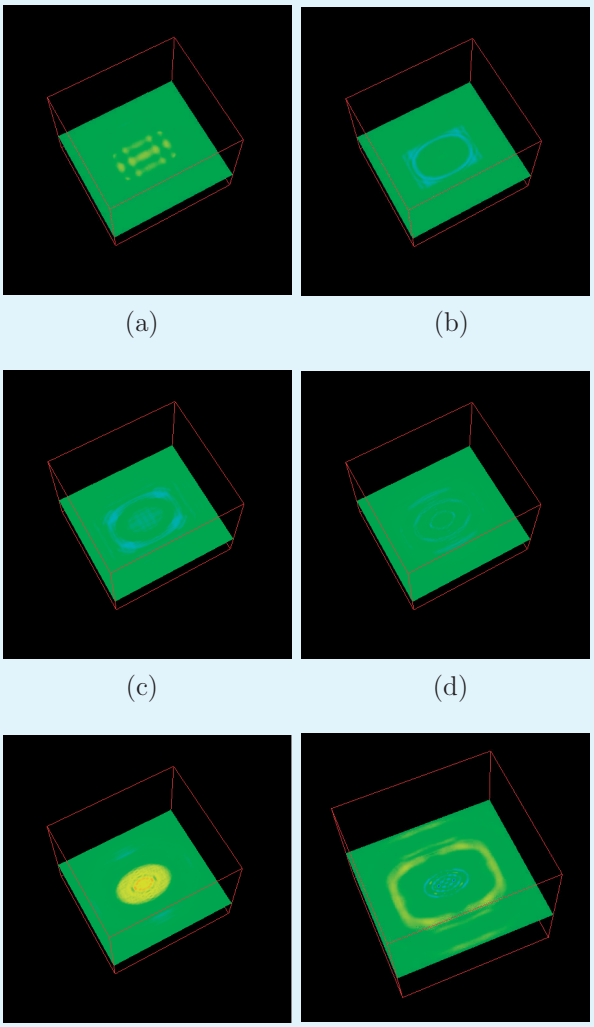


Figure 7: History of the ultrasound pressure in the breast tissue containing an ellipsoid tumor at 8 successive time steps:(a) t=26.6ns; (b) t=2.6μs; (c) t=7.9μs ; (d) t=15.9μs; (e) t=18.6μs; (f) t=29.3μs. The color scale shows also the attenuation of the pressure during the two-way travel of the wave along the tissue.

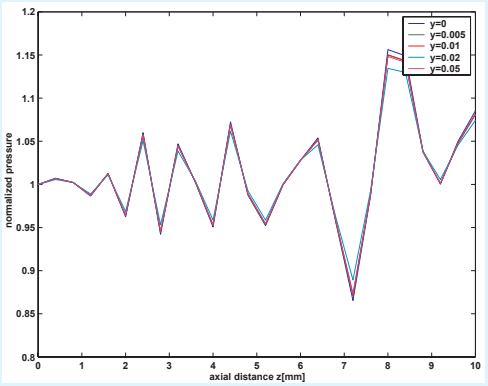


Figure 8: Normalized ultrasound pressure in 5 different media presenting a fractional Laplacian derivative attenuation model. The media are varying according to 5 values (close to 0) of the power exponent parameter y. The plots represent the pressure field as a function of the axial distance (z) when the lateral distance is fixed to x=-3mm, i.e., the wave travels outside the tumor region.

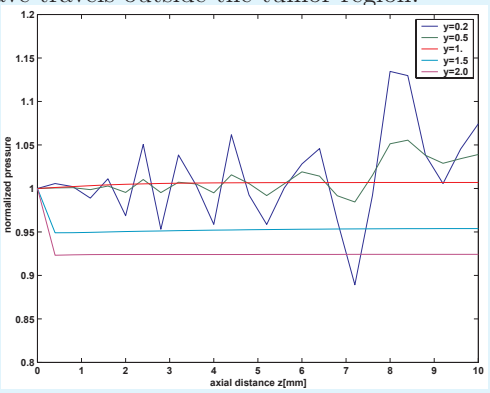


Figure 9: Normalized ultrasound pressure in 5 different media presenting a fractional Laplacian derivative attenuation model. The media are varying according to 5 values (between 0.2 and 2) of the power exponent parameter y. The plots represent the pressure field as a function of the axial distance (z) when the lateral distance is fixed to x=-3mm, i.e., the wave travels outside the tumor region.

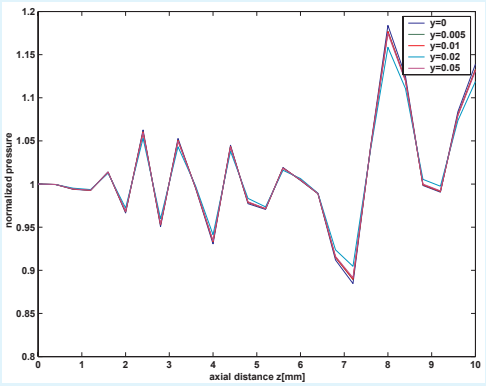


Figure 10: Normalized ultrasound pressure in 5 different media presenting a fractional Laplacian derivative attenuation model. The media are varying according to 5 values (close to 0) of the power exponent parameter y. The plots represent the pressure field as a function of the axial distance (z) when the lateral distance is fixed to x=1mm, i.e., the wave travels through the tumor region.

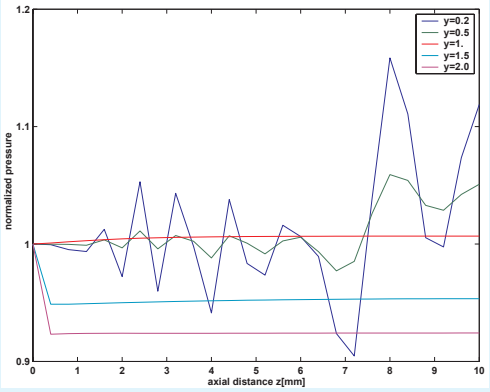


Figure 11: Normalized ultrasound pressure in 5 different media presenting a fractional Laplacian derivative attenuation model. The media are varying according to 5 values (between 0.2 and 2) of the power exponent parameter y. The plots represent the pressure field as a function of the axial distance (z) when the lateral distance is fixed to x=1mm, i.e., the wave travels through the tumor region.

REFERENCES

[1] Richter K. and SH. Heywang-Köbrunner, Quantitative parameters measured by a new sonographic method for detecting breast lesions, Invest. Radiol., Vol 30, pp. 401-411, 1995.

[2] Richter, K. Clinical amplitude/velocity reconstructive imaging (CARI)-a new sonographic method for detecting breast lesions, The British Journal of Radiology, Vol 68, pp. 375-384, 1995.

[3] Langtangen, H. P. Computational partial differential equations - Numerical methods and Diffpack programming, Springer-Verlag, 1999.

[4] Weiwad, W., Heinig, A., Goetz, L., Hartmann, H., Lampe, D., Buchmann, J., Millner, R., Spielmann, R.P. and Heywang-Koebrunner, S. H., Direct measurement of sound velocity in various specimens of breast tissue, Investigative Radiology, 35(12), pp. 721-726, 2000.

[5] Bounaïm, A., Holm, S., Chen, W., Ødegard, A., Tveito, A. and Thomenius, K., FETD simulation of wave propagation modeling the CARI breast sonography, Proc. ICCSA'2003, LNCS 2668, pp. 705-714, Springer-Verlag, 2003.

[6] Bounaïm, A., Holm, S., Chen, W., Ødegard, A., Sensitivity of the ultrasonic CARI technique for breast tumor detection using a FETD scheme, Journal Ultrasonics, Vol 42, Issues 1-9, pp. 919-925, 2004.

[7] Bounaïm, A., Holm, S., Chen, W., Ødegard, A., Quantification of the CARI breast imaging sensitivity by 2D/3D numerical time-domain ultrasound wave propagation, J. Mathematics and Computers in Simulation, Vol 65, Issues 4-5, pp. 521-534, 2004.

[8] Bounaïm, A., Holm, S., Chen, W., Ødegard, A., Modified fractional derivative model for the attenuation of acoustic waves propagating in human soft tissue: Finite element simulations using Diffpack, Simula Report 2004-02, Simula Research Laboratory, 2004.

[9] Chen, W. and Holm, S., Modified Szabo's wave equation models for lossy media obeying frequency power law, J. Acoust. Soc. Amer., 114(5), pp. 2570-2584, 2003.

[10] Chen, W. and Holm, S., Fractional Laplacian time-space models for linear and nonlinear lossy media exhibiting arbitrary frequency dependency, J. Acoust. Soc. Amer., 114(5), pp. 1424-1430, 2004.

ON THE FRACTIONAL HEAT EQUATION

L. Boyadjiev[1] and R. Scherer[2]
[1]Applied Mathematics and Informatics Department,
Technical University of Sofia, P. O. Box 384, 1000 Sofia, Bulgaria
e-mail: boyadjievl@yahoo.com

[2]Institute of Practical Mathematics,
University of Karlsruhe (TH), D-76128 Karlsruhe, Germany
e-mail: scherer@math.uni-karlsruhe.de

1. INTRODUCTION

The method of integral transforms is used to solve the temperature field problem in oil stratum described by a fractional heat equation. The case of the incomplete lumped formulation of radial fluid injection in the stratum is considered. By using the Riemann-Liouville differintegration operator of arbitrary order and the Laplace transform, the solution containing special functions of Wright's type in the integrand is obtained.

A porous medium (sandstone) which is saturated with oil is called an oil stratum. It is possible to consider the stratum depth as equal to infinity since the depth of oil stratum usually varies from one to several kilometers. The rock surrounding a stratum (cap and base rock) is considered impermeable to the fluid. A standard method of oil extraction is to pump the oil out from a series of production wells which are drilled in the center of the oil deposit. At particular time of the exploitation period a water or steam is injected into injection wells drilled along the boundary of the oil reservoir. There are two cases of fluid injection to be considered: linear and radial injection. The problem of describing the temperature field $u = u(x, y, z, t)$ in a single or multiple layer oil stratum arises when a hot water (or steam) whose temperature differs from that of the stratum, is injected into the injection wells. In the so–called radial case, a hot fluid is forced into the stratum through an infinitely thin well, which is considered as a linear source of incompressible fluid whose volume rate is positive.

Beside the so–called exact formulation of the problem, the following three approximate formulations are treated (cf. Antimirov, Kolyshkin and Vaillancourt [1]):

[1]This paper has been partially supported by NSF, Bulgarian Ministry of Education and Science, under Grant MM 1305. – It was prepared when the first author was a Visiting Professor at the University of Karlsruhe.

- *the lumped formulation,* where the thermal conductivity of the stratum is infinitely large in the vertical direction,
- *the incomplete lumped formulation,* where the horizontal heat transfer in the cap and base rock is neglected,
- *the formulation of Lauwerier,* where the horizontal heat transfer is neglected also in the stratum.

In the radial case of the incomplete lumped formulation, the temperature field $u = u(r, z, t)$ in cylindrical coordinates satisfies the equation

$$\frac{\partial u}{\partial t} = a^2 \, \frac{\partial^2 u}{\partial z^2}, \quad a > 0, \quad 0 < r, z, t < \infty \tag{1}$$

subject to the boundary condition

$$z = 0 : \quad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1 - 2\nu}{r} \, \frac{\partial u}{\partial r} + \alpha \, \frac{\partial u}{\partial z}, \quad 0 < r, t < \infty, \tag{2}$$

and the conditions

$$\begin{array}{ll} (a) & r = 0, \ z = 0 : \ u = 1, \\ (b) & u \to 0 \quad \text{as} \quad r^2 + z^2 \to \infty, \\ (c) & t = 0 : \ u = 0. \end{array} \tag{3}$$

We should make clear that: The constant $a > 0$ depends on the *coefficient of thermal diffusivity* of the cap rock and the stratum; the constant $\alpha > 0$ is a ratio of the *coefficients of thermal conductivity* of the cap rock and the stratum; the constant $\nu > 0$ depends on the volume rate and the *volumic heat capacity* of the fluid as well as the coefficient of the thermal conductivity of the stratum. For later consideration we set $b = \frac{\alpha}{a}$.

By using the Laplace transform

$$\bar{f}(p) = L[f(t)] = \int_0^\infty e^{-pt} f(t) dt, \tag{4}$$

the solution of the problem (1), (2), (3) is given in the form (cf. [1], 8.2.54)

$$u(r, z, t) = \frac{1}{\Gamma(\nu)} \int_0^t \frac{1}{\tau} \left(\frac{r^2}{4\tau} \right)^\nu e^{-\frac{r^2}{4\tau}} \, \operatorname{erfc} \left(\frac{b\tau + \frac{z}{a}}{2\sqrt{t - \tau}} \right) d\tau, \tag{5}$$

where $\Gamma(\nu)$ is Euler's gamma function,

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt = \frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}, x^2)$$

is the complementary error function and

$$\Gamma(a,x)=\int_x^\infty t^{a-1}e^{-t}dt,\quad a>0.$$

Formula (5) is the well–known formula (cf. Avdonin [2]) which was used in numerous computations of oil stratum temperature fields in the radial case. Similar boundary value problems are also studied by using the Laplace and the general Hankel transforms (cf. Ben Nakhi and Kalla [3]).

The concept of a non–integer differentiation and integration of a function is almost as old as calculus itself. Many famous mathematicians including Leibniz, Euler, Lagrange, Laplace, Fourier and Abel made some contribution to it. But it was before Liouville and Riemann made the idea more precisely formulated. Nowadays there exist specialized treaties where mathematical aspects and applications of the fractional calculus are extensively discussed (cf. Bride [4], Kiryakova [10], Miller and Ross [13], Oldham and Spanier [14], Samko, Kilbas and Marichev [16]).

Fractional calculus became a significant topic in mathematical analysis as a result of its increasing range of potential applications. Operators for fractional differentiation and integration *(differintegration operators)* have been used in various fields as: hydraulics of dams, potential fields, diffusion problems and waves in liquids and gases (cf. Schneider and Wyss [17]). The use of half–order derivatives and integrals leads to a formulation of certain electro–chemical problems which is more economical and useful than the classical approach in terms of Fick's law of diffusion (cf. Crank [5]). Maybe the main advantage of the fractional calculus is that the fractional derivatives provide an excellent instrument for the description of memory and hereditary properties of various materials and processes.

Thus motivated, we extend the problem (1), (2), (3) by replacing the partial time derivative of the field temperature by the fractional time derivative

$$\frac{\partial^{2\beta}u}{\partial t^{2\beta}}\quad\text{of real order}\quad\alpha=2\beta,\ 0<\beta\le 1/2.$$

Then the radial case of the fractional incomplete lumped formulation reads as

$$\frac{\partial^{2\beta}u}{\partial t^{2\beta}}=a^2\frac{\partial^2u}{\partial z^2},\quad a>0,\quad 0<r,z,t<\infty,\quad 0<\beta\le 1/2,\tag{6}$$

subject to the boundary condition

$$z=0:\quad \frac{\partial^{2\beta}u}{\partial t^{2\beta}}=\frac{\partial^2u}{\partial r^2}+\frac{1-2\nu}{r}\,\frac{\partial u}{\partial r}+\alpha\,\frac{\partial u}{\partial z},\quad 0<r,t<\infty,\tag{7}$$

and the conditions

$$\begin{array}{ll} (a) & r=0,\ z=0:\ u=1, \\ (b) & u\to 0 \quad \text{as} \quad r^2+z^2\to\infty, \\ (c) & t=0:\ u=0. \end{array} \tag{8}$$

Using the Riemann–Liouville fractional derivative operator and the Laplace transform, we get an integral form of the solution involving into the integrand special functions of Wright's type. The solution obtained contains (5) as particular cases as $\beta=\frac{1}{2}$.

Publications of fractional calculus based on Laplace transform (cf. Gorenflo and Rutman [9], Mainardi [12], Podlubny [15] and others) confirm the present interest in using the potentialities of fractional calculus in mathematical physics.

2. THE RIEMANN–LIOUVILLE DIFFERINTEGRATION OPERATOR

To present some essentials of the Riemann–Liouville fractional calculus we follow in this section the Handbook of function and generalized function transformations (Zayed [18]).

Definition 1: *If $\alpha>0$, the Riemann–Liouville fractional integral of order α of a function $f(t)$ is defined by*

$$I_\alpha[f(t)](x)=\frac{d^{-\alpha}}{dx^{-\alpha}}f(x)=\frac{1}{\Gamma(\alpha)}\int_a^x (x-t)^{\alpha-1}f(t)dt\ .$$

By using Definition 1, it is possible to define the fractional derivative as follows.

Definition 2: *If $\alpha\geq 0$, the Riemann–Liouville fractional derivative of order α of a function $f(t)$ is defined by*

$$D_\alpha f(x)=\frac{d^\alpha}{dx^\alpha}f(x)=\frac{d^m}{dx^m}\ \frac{d^{-(m-\alpha)}}{dx^{-(m-\alpha)}}f(x)=\frac{d^m}{dx^m}\ \frac{1}{\Gamma(m-\alpha)}\int_a^x (x-t)^{m-\alpha-1}f(t)dt,$$

where m is nonnegative integer such that $m-1\leq\alpha<m$.

It is convenient to introduce a notation that unifies fractional differentiation and integration.

Definition 3: *The Riemann–Liouville differintegration operator of order α is defined by*

$$R^{\alpha} f(x) = \begin{cases} I_{\alpha}\left[f(t)\right](x), & \text{if } \alpha < 0, \\ D_{\alpha} f(x), & \text{if } \alpha \geq 0. \end{cases}$$

In particular, if $0 < \alpha < 1$, then

$$R^{\alpha} f(x) = \frac{d}{dx}\, I_{1-\alpha}\left[f(t)\right](x).$$

For our further discussion in this paper, it is important to mention that using Definition 3 it is possible to be established that the relationship of the Riemann–Liouville differintegration operator of order α with the Laplace transform (4) is given by the equation

$$L\left[R^{\alpha} f(t)\right](p) = p^{\alpha} L[f(t)](p) - \sum_{k=0}^{n-1} p^{k}\, \frac{d^{\alpha-1-k} f}{dt^{\alpha-1-k}}\,(0), \qquad \text{for all } \alpha, \tag{9}$$

where n is an integer such that $n-1 < \alpha \leq n$.

3. EFROS' THEOREM

A key role in obtaining the solution (5) is given to the following generalized multiplication theorem proved by A.M. Efros (cf. Antimirov and Vaillancourt [1], p. 12, Ditkin and Prudnikov [6], pp. 35–36). Because this result is very important and the proof not popular, we will add it.

Theorem 1 [Efros' Theorem] Let be given analytic functions $G(p)$ and $q(p)$ and the relations

$$F(p) = L[f(t)], \quad G(p)\; e^{-\tau q(p)} = L[g(t,\tau)],$$

then it holds

$$G(p)\; F\left(q(p)\right) = L\left[\int_0^{\infty} f(\tau) g(t,\tau) d\tau\right]. \tag{10}$$

Proof: The right–hand side of (10) is

$$\begin{aligned} L\left[\int_0^{\infty} f(\tau) g(t,\tau) d\tau\right] &= \int_0^{\infty} e^{-pt} \int_0^{\infty} f(\tau) g(t,\tau) d\tau dt \\ &= \int_0^{\infty} f(\tau) \int_0^{\infty} g(t,\tau) e^{-pt} dt d\tau, \end{aligned}$$

provided we can reverse the order of integration. But the inner integral in the last one is the Laplace transform of $g(t,\tau)$ and hence by (9), we can write

$$L\left[\int_0^\infty f(\tau)g(t,\tau)d\tau\right] = G(p)\int_0^\infty f(\tau)e^{-q(p)\tau}d\tau = G(p)F[q(p)],$$

that completes the proof. □

If in particular, we take $q(p) = p$, then

$$L[g(t,\tau)] = e^{-p\tau}G(p)$$

and by the p–shift theorem, $g(t,\tau) = g(t-\tau)$. Hence, formula (10) becomes

$$F(p)G(p) = L\left[\int_0^\infty f(\tau)g(t-\tau)d\tau\right] = L\left[\int_0^t f(\tau)g(t-\tau)d\tau\right],$$

since for original functions we have $g(t-\tau) = 0$ for $\tau > t$. The last formula shows that Efros' theorem is a generalization of the convolutional theorem for the Laplace transform.

4. A SPECIAL FUNCTION OF WRIGHT'S TYPE

In studying the time fractional diffusion equation (6), the fundamental solution of the basic Cauchy problem can be expressed in term of an auxiliary function, defined as (cf. Mainardi [12])

$$M(z;\beta) = \frac{1}{2\pi i}\int_{Ha} e^{\sigma - z\sigma^\beta}\frac{d\sigma}{\sigma^{1-\beta}}, \quad 0 < \beta < 1,$$

where Ha denotes the Hankel path of integration that begins at $\sigma = -\infty - ib_1$ ($b_1 > 0$), encircles the branch cut that lies along the negative real axis, and ends up at $\sigma = -\infty + ib_2$ ($b_2 > 0$). It is proved that

$$M(z;\beta) = W(-z;-\beta,1-\beta),$$

where

$$W(z;\lambda,\mu) = \sum_{n=0}^\infty \frac{z^n}{n!\Gamma(\lambda n+\mu)} = \frac{1}{2\pi i}\int_{Ha} e^{\sigma + z\sigma^{-\lambda}}\frac{d\sigma}{\sigma^\mu}, \quad \lambda > -1, \mu > 0,$$

is an entire function of z referred to as the Wright's function (cf. Erdélyi [8], vol. III, Chapter 18). In the particular case $\beta = \frac{1}{2}$, it holds

$$M(z;\frac{1}{2}) = \frac{1}{\sqrt{\pi}}\sum_{m=0}^\infty (-1)^m\left(\frac{1}{2}\right)^m\frac{z^{2m}}{(2m)!} = \frac{1}{\sqrt{\pi}}e^{-\frac{z^2}{4}}. \tag{11}$$

Further, we use the function

$$N(\xi;\beta)=\frac{1}{2\pi i}\int_{Ha}e^{\sigma-\xi\sigma^{2\beta}}\frac{d\sigma}{\sigma}.$$

By adopting Mainardi's approach [12], we obtain the following auxiliary result.

Lemma 2 If $0<\beta\leq\frac{1}{2}$, and $0<t,\tau,z<\infty$, then we obtain

$$(i)\qquad e^{-(b\tau+\frac{z}{a})p^{\beta}}=L[g_1(t,\tau;\beta)],$$

$$(ii)\qquad \tfrac{1}{p}e^{-\tau p^{2\beta}}=L\left[g_2(t,\tau;\beta)\right],$$

where

$$g_1(t,\tau;\beta)=\frac{(b\tau+\frac{z}{a})\beta}{t^{\beta+1}}M\left(\frac{b\tau+\frac{z}{a}}{t^{\beta}};\beta\right), \tag{12}$$

and

$$g_2(t,\tau;\beta)=N\left(\frac{\tau}{t^{2\beta}};\beta\right). \tag{13}$$

Proof: Part (i) is a direct consequence of [12] (formula (3.4) and (3.6)).
To prove (ii) consider the Laplace transform

$$\bar{g}_2(\tau,p;\beta)=\tfrac{1}{p}\ e^{-\tau p^{2\beta}}.$$

According to the inversion formula of the Laplace transform,

$$g_2(\tau,t;\beta)=\frac{1}{2\pi i}\int_{Ha}e^{pt-\tau p^{2\beta}}\frac{dp}{p},$$

putting $\sigma=pt$ and introducing the variable $\xi=\frac{\tau}{t^{2\beta}}$, we obtain

$$g_2(\tau,t;\beta)=N(\xi;\beta)=\frac{1}{2\pi i}\int_{Ha}e^{\sigma-\xi\sigma^{2\beta}}\frac{d\sigma}{\sigma},$$

where $N(\xi;\beta)$ is the auxiliary function. Using Taylor's representation of the exponential function and Hankel's representation of the reciprocal of the Euler gamma function, we arrive at the following series representation

$$N(\xi;\beta)=\sum_{n=0}^{\infty}\frac{(-1)^n\ \xi^n}{n!\ \Gamma(-2\beta n+1)},\quad 0<\beta<\frac{1}{2}.$$

Hence for $0 < \beta < \frac{1}{2}$, the auxiliary function is Wright's function, since

$$N(\xi;\beta) = W(-\xi; -2\beta, 1).$$

Therefore we can state that

$$\tfrac{1}{p}\, e^{-\tau p^{2\beta}} = L\left[g_2(t,\tau;\beta)\right],$$

where

$$g_2(t,\tau;\beta) = N\left(\frac{\tau}{t^{2\beta}};\beta\right),$$

that completes the proof of part (ii). □

5. FRACTIONAL INCOMPLETE LUMPED FORMULATION

In this section, we apply the Laplace transform to solve the fractional problem stated in Section 2. Recall that $\nu > 0$ and $\alpha > 0$ are parameters as specified in Section 1. By using essentially the relationship (9) of the Riemann–Liouville differintegration operator of arbitrary order α with the Laplace transform, we prove the main statement given by the following theorem.

Theorem 3 If $0 < \beta \leq \frac{1}{2}$, the solution of the radial case of the fractional incomplete formulation (6), (7), (8) is given by the integral

$$u(r,z,t) = \frac{2}{\Gamma(\nu)} \int_0^\infty \frac{1}{\tau} \left(\frac{r^2}{4\tau}\right)^\nu e^{-\frac{r^2}{4\tau}}\; g(t,\tau;\beta)\; d\tau, \tag{14}$$

where

$$g(t,\tau;\beta) = g_1(t,\tau;\beta) * g_2(t,\tau;\beta), \tag{15}$$

and $g_1(t,\tau;\beta)$ and $g_2(t,\tau;\beta)$ are the functions defined by (12) and (13), respectively.

Proof: Let $\bar{u}(r,z,p) = L[u(r,z,t)]$, where L again denotes the Laplace transform operator. Applying the Laplace transform to (6), (7), (8), we obtain according (8)(c) and (9):

$$p^{2\beta}\bar{u} = a^2 \frac{\partial^2 \bar{u}}{\partial z^2}, \qquad 0 < r,\ z < \infty, \tag{16}$$

$$z = 0: \quad p^{2\beta}\bar{u} = \frac{\partial^2 \bar{u}}{\partial r^2} + \frac{1-2\nu}{r}\frac{\partial \bar{u}}{\partial r} + \alpha \frac{\partial \bar{u}}{\partial z}, \quad 0 < r < \infty, \tag{17}$$

$$\begin{array}{ll}(a) & r=0, z=0: \ \bar{u}=\frac{1}{p}, \\ (b) & \bar{u}\to 0 \ \text{ as } \ r^2+z^2\to\infty. \end{array} . \tag{18}$$

The solution of (16), which remains bounded as $z\to\infty$, reads

$$\bar{u}(r,z,p)=c(r,p)e^{-\frac{p^\beta}{a}z}, \tag{19}$$

where the function $c(r,p)$ has to be determined. Substituting (19) into (17) leads to the following ordinary differential equation for $c(r,p)$,

$$\frac{d^2c}{dr^2}+\frac{1-2\nu}{r}\frac{dc}{dr}-\left[p^{2\beta}+bp^\beta\right]c=0. \tag{20}$$

It follows from (18) and (19) that

$$r=0: \ c(r,p)=\tfrac{1}{p}, \quad \lim_{r\to\infty}c(r,p)=0. \tag{21}$$

Let us abbreviate $\mu=\sqrt{p^{2\beta}+bp^\beta}$. Then the solution of (20), which remains bounded as $r\to\infty$ (cf. Lebedev [11], p. 106, formula (5.4.11)), is given by

$$c(r,p)=c_1(p)\left(\frac{\mu r}{2}\right)^\nu K_\nu(\mu r), \tag{22}$$

where $K_\nu(z)$ is the modified Bessel function of the second kind of order ν, and $c_1(p)$ is a constant that must conform with (21). To apply conditions (21), we consider $\lim_{r\to 0}c(r,p)$ by refering to the following formula (cf. [16], p. 111)

$$K_\nu(z)\sim\tfrac{1}{2}\ \Gamma(\nu)\left(\frac{2}{z}\right)^\nu \ \text{ as } \ z\to 0.$$

Then it follows from (21) and (22) that

$$c_1(p)=\frac{2}{p\ \Gamma(\nu)}. \tag{23}$$

Substituting (23) and (22) into (19), gives the Laplace transform of the solution

$$\bar{u}(r,z,p)=\frac{2}{p\Gamma(\nu)}\left(\frac{\mu r}{2}\right)^\nu K_\nu(\mu r)e^{-\frac{p^\beta}{a}z}. \tag{24}$$

To make use of Theorem 1, let us represent (24) in the form

$$\bar{u}(r,z,p)=G(p;\beta)F[q(p;\beta)], \tag{25}$$

where

$$G(p;\beta)=\frac{2}{p\Gamma(\nu)}e^{-\frac{p^\beta}{a}z} \quad \text{and} \quad q(p;\beta)=p^{2\beta}+bp^\beta.$$

It is well-known (cf. Ditkin and Prudnikov [7], p. 345, formula 2.10.128) that

$$p^{\frac{\nu}{2}} K_\nu(\alpha\sqrt{p}) = L\left[\frac{\alpha^\nu}{(2t)^{\nu+1}} e^{-\frac{\alpha^2}{4t}}\right].$$

It then follows from (24) and (25) that

$$F(p) = \left(\frac{r\sqrt{p}}{2}\right)^\nu K_\nu(r\sqrt{p}) = L[f(t;\beta)],$$

where

$$f(t;\beta) = \left(\frac{r}{2}\right)^\nu \frac{r^\nu}{(2t)^{\nu+1}} e^{-\frac{r^2}{4t}}. \tag{26}$$

Furthermore, we obviously have

$$G(p;\beta)e^{-\tau q(p;\beta)} = \frac{2}{\Gamma(\nu)}\ \frac{1}{p} e^{-\tau p^{2\beta}} e^{-(b\tau+\frac{z}{a})p^\beta}.$$

Then Lemma 2 and the convolution theorem yield

$$G(p;\beta)\ e^{-\tau q(p;\beta)} = \frac{2}{\Gamma(\nu)}\ L[g(t,\tau;\beta)],$$

where

$$g(t,\tau;\beta) = g_1(t,\tau;\beta) * g_2(t,\tau;\beta), \tag{27}$$

and $g_1(t,\tau;\beta)$ and $g_2(t,\tau;\beta)$ are defined by (12) and (13), respectively. Taking into account (26) and (27), by Theorem 1 we obtain the solution of the radial case of the fractional incomplete lumped formulation in the form (14) that proves the theorem. □

Corollary 4 For the particular case $\beta = \frac{1}{2}$, the result (14) in Theorem 3 yields the representation (5) of Avdonin.

Proof: In the particular case $\beta = \frac{1}{2}$, we have

$$N(\xi;\frac{1}{2}) = \frac{1}{2\pi i}\int_{Ha} e^{pt}\left(\frac{1}{p}\ e^{-\tau p}\right) dp = H(t-\tau), \tag{28}$$

where $H(t-\tau)$ is the Heaviside function. To make sure that the solution (5) occurs as a particular case of the solution obtained, let us consider the convolution (15)

$$\begin{aligned} g_1(t,\tau;\beta) * g_2(t,\tau;\beta) &= \int_0^t g_1(t-s,\tau;\beta)\ g_2(s,\tau;\beta)\ ds \\ &= \int_0^t \frac{(b\tau+\frac{z}{a})\beta}{(t-s)^{\beta+1}}\ M\left(\frac{b\tau+\frac{z}{a}}{(t-s)^\beta};\beta\right)\ N\left(\frac{\tau}{s^{2\beta}};\beta\right)\ ds. \end{aligned}$$

For the case $\beta = \frac{1}{2}$, according to (11) and (28) the convolution becomes

$$\begin{aligned} g_1(t,\tau;\frac{1}{2}) * g_2(t,\tau;\frac{1}{2}) &= \frac{1}{2\sqrt{\pi}} H(t-\tau) \int_\tau^t \frac{b\tau + \frac{z}{a}}{(t-s)^{\frac{3}{2}}} e^{-\left(\frac{b\tau+\frac{z}{a}}{2\sqrt{t-s}}\right)^2} ds = \\ &= \frac{2}{\sqrt{\pi}} H(t-\tau) \int_{\frac{b\tau+\frac{z}{a}}{2\sqrt{t-\tau}}}^\infty e^{-w^2}\, dw = \frac{2}{\sqrt{\pi}} H(t-\tau)\ \mathrm{erfc}\ \left(\frac{b\tau + \frac{z}{a}}{2\sqrt{t-\tau}}\right). \end{aligned}$$

Hence if $\beta = \frac{1}{2}$, the solution (14) yields the formula (5). □

ACKNOWLEDGEMENT

The authors are indebted to Professor Shyam L. Kalla (Kuwait University) for many valuable comments.

REFERENCES

[1] Antimirov, M.Ya., Kolyshkin, A.A. and Vaillancourt, R., *Applied Integral Transforms*, American Mathematical Society, 1993.

[2] Avdonin, N.A., Some formulas for calculating the temperature field of a stratum subject to thermal injection, *Izvestia VUZ, Neft'i Gaz*, 7 (1964), 37–41 (Russian).

[3] Ben Nakhi, Y. and Kalla, S.L., Some boundary value problems of temperature fields in oil strata, *Applied Mathematics and Computation*, 146 (2003), 105–119.

[4] Bride, A.C., *Fractional Calculus and Integral Transforms of Generalized Functions*, Pitman Research Notes in Mathematics, No. 31, Pitman, London, 1979.

[5] Crank, J., *The Mathematics of Diffusion*, Oxford University Press, 1975.

[6] Ditkin, V.A. and Prudnikov, A.P., *Integral Transforms and Operational Calculus*, Pergamon Press, Oxford, New York, 1965.

[7] Ditkin, V.A. and Prudnikov, A.P., *Formulaire pour le calcul opérationnel*, Masson, Paris, 1967.

[8] Erdélyi, A., *Higher Transcendental Functions*, Vol. I, II, III, McGraw–Hill, New York, 1955.

[9] Gorenflo, R. and Rutmann, R., On ultraslow and intermediate processes, In: Rusev, P. (ed.) et al., *Transform methods and special functions*, Proc. Intern. Workshop, Bankya, Bulgaria, August 12–17, 1994, Sofia: SCT Publishing, (1995), 61-81.

[10] Kiryakova, V. *Generalized Fractional Calculus and Applications*, Pitman Research Notes in Mathematics Series, 301, Longmann, Harlow, 1994.

[11] Lebedev, N.N. *Special Functions and their Applications*, Prentice-Hall, Englewood Cliffs, N.J., 1965.

[12] Mainardi, F. The Fundamental Solutions for the fractional diffusion–wave equation. *Appl. Math. Lett.*, 9 (1996) 23–28.

[13] Miller, K.S. and Ross, B., *An Introduction to the Fractional Calculus and Fractional Differential Equations*, Wiley, New York, 1993.

[14] Oldham, K.B. and Spanier, J., *The Fractional Calculus*, Academic Press, New York, 1974.

[15] Podlubny, I. The Laplace transform method for linear differential equations of fractional order, *Preprint UEF–02–94*, Inst. Exp. Phys., Slovak Acad. Sci., Kosice, 1994.

[16] Samko, S.G., Kilbas, A.A. and Marichev, O.I., *The Fractional Integrals and Derivatives*, Theory and Applications, Gordon and Breach, Amsterdam, 1993.

[17] Schneider, W.R. and Wyss, W., Fractional diffusion and wave equations, *J. Math. Phys.* 30 (1989), 134-144.

[18] Zayed, Ahmed I., *Handbook of Function and Generalized Function Transformations*, CRC Press, 1996.

GALOIS CORINGS APPLIED TO PARTIAL GALOIS THEORY

S. Caenepeel and E. De Groot
Faculty of Engineering Sciences,
Vrije Universiteit Brussel, VUB, B-1050 Brussels, Belgium
email:scaenepe@vub.ac.be
email: edegroot@vub.ac.be

Abstract

Partial Galois extensions were recently introduced by Dokuchaev, Ferrero and Paques. We introduce partial Galois extensions for noncommutative rings, using the theory of Galois corings. We associate a Morita context to a partial action on a ring.

INTRODUCTION

Partial actions of groups originate from the theory of operator algebras, see for example [16]. Partial representations of groups on Hilbert spaces were introduced independently in [17] and [19]. Several applications are given in the literature, we refer to [14] for a more extensive bibliography. More recently, partial actions were studied from a purely algebraic point of view, in [12, 13, 15].
In [14], the authors consider partial actions on commutative rings, with the additional assumption that the associated ideals are generated by idempotents. Then they generalize Galois theory for commutative rings, as introduced in [10] for usual group actions, to partial actions.
Corings were introduced by Sweedler in 1975 in [23]. There has been a revived interest in corings since the beginning of the century, based on an observation made by Takeuchi that various types of modules, such as Hopf modules, relative Hopf modules, graded modules, entwined modules and Yetter-Drinfeld modules may be viewed as comodules over a coring. Brzeziński [1] noticed the importance of this observation: the language of corings can be applied successfully to give a unified and more elegant treatment to properties related to all these kinds of modules. An overview can be found in [4].
One of the nice applications is descent and Galois theory: Galois corings were introduced in [1], and studied in [6] and [24].

1991 Mathematics Subject Classification. 16W30.

The corings approach provides a unified theory for various types of Galois theories, including the classical Chase-Harrison-Rosenberg theory [10], Hopf-Galois theory (see [11, 18, 20]), coalgebra Galois theory (see [3]) and weak Hopf-Galois theory (see the forthcoming [7]).
The aim of this note is to develop partial Galois theory starting from Galois corings. The strategy is basically the following: given a set of idempotents e_σ indexed by a finite group G in a ring A, we investigate when the direct sum of the Ae_σ is a coring; it turns out that this is the case if a partial action of G on A is given. Then we investigate when this coring is a Galois coring, and apply the results in [6]. This procedure still works in the case where the ring A is not commutative. In the case where A is commutative, we recover some of the results in [14]. This is done in Section 2. In Section 3, we associate a Morita context to a partial action on a ring A, and show that the context is strict if A is a faithfully flat partial Galois extension of the invariants ring A^G.

1. PRELIMINARY RESULTS

1.1 Galois corings

Let A be a ring. An A-coring $\mathcal{C}$ is a coalgebra in the category ${}_A\mathcal{M}_A$ of A-bimodules. Thus an A-coring is a triple $\mathcal{C} = (\mathcal{C}, \Delta_\mathcal{C}, \varepsilon_\mathcal{C})$, where $\mathcal{C}$ is an A-bimodule, and $\Delta_\mathcal{C} : \mathcal{C} \to \mathcal{C} \otimes_A \mathcal{C}$ and $\varepsilon_\mathcal{C} : \mathcal{C} \to A$ are A-bimodule maps such that

$$(\Delta_\mathcal{C} \otimes_A \mathcal{C}) \circ \Delta_\mathcal{C} = (\mathcal{C} \otimes_A \Delta_\mathcal{C}) \circ \Delta_\mathcal{C}, \tag{1}$$

and

$$(\mathcal{C} \otimes_A \varepsilon_\mathcal{C}) \circ \Delta_\mathcal{C} = (\varepsilon_\mathcal{C} \otimes_A \mathcal{C}) \circ \Delta_\mathcal{C} = \mathcal{C}. \tag{2}$$

We use the Sweedler-Heyneman notation for the comultiplication:

$$\Delta_\mathcal{C}(c) = c_{(1)} \otimes_A c_{(2)}.$$

A right $\mathcal{C}$-comodule $M = (M, \rho)$ consists of a right A-module M together with a right A-linear map $\rho : M \to M \otimes_A \mathcal{C}$ such that:

$$(\rho \otimes_A \mathcal{C}) \circ \rho = (M \otimes_A \Delta_\mathcal{C}) \circ \rho, \tag{3}$$

and

$$(M \otimes_A \varepsilon_\mathcal{C}) \circ \rho = M. \tag{4}$$

We then say that $\mathcal{C}$ coacts from the right on M, and we denote

$$\rho(m) = m_{[0]} \otimes_A m_{[1]}.$$

A right A-linear map $f:\ M \to N$ between two right $\mathcal{C}$-comodules M and N is called right $\mathcal{C}$-colinear if $\rho(f(m)) = f(m_{[0]}) \otimes m_{[1]}$, for all $m \in M$. The category of right $\mathcal{C}$-comodules and $\mathcal{C}$-colinear maps is denoted by $\mathcal{M}^{\mathcal{C}}$.
$x \in \mathcal{C}$ is called grouplike if $\Delta_{\mathcal{C}}(x) = x \otimes x$ and $\varepsilon_{\mathcal{C}}(x) = 1$. Grouplike elements of $\mathcal{C}$ correspond bijectively to right $\mathcal{C}$-coactions on A: if A is grouplike, then we have the following right $\mathcal{C}$-coaction ρ on A: $\rho(a) = xa$.
Let $(\mathcal{C}, x)$ be a coring with a fixed grouplike element. For $M \in \mathcal{M}^{\mathcal{C}}$, we call

$$M^{\mathrm{co}\mathcal{C}} = \{m \in M \mid \rho(m) = m \otimes_A x\}$$

the submodule of coinvariants of M. Observe that

$$A^{\mathrm{co}\mathcal{C}} = \{b \in A \mid bx = xb\}$$

is a subring of A. Let $i:\ B \to A$ be a ring morphism. i factorizes through $A^{\mathrm{co}\mathcal{C}}$ if and only if

$$x \in G(\mathcal{C})^B = \{x \in G(\mathcal{C}) \mid xb = bx,\ \text{for all } b \in B\}.$$

We then have a pair of adjoint functors (F, G), respectively between the categories $\mathcal{M}_B$ and $\mathcal{M}^{\mathcal{C}}$ and the categories ${}_B\mathcal{M}$ and ${}^{\mathcal{C}}\mathcal{M}$. For $N \in \mathcal{M}_B$ and $M \in \mathcal{M}^{\mathcal{C}}$,

$$F(N) = N \otimes_B A \ \text{ and } \ G(M) = M^{\mathrm{co}\mathcal{C}}.$$

The unit and counit of the adjunction are

$$\nu_N:\ N \to (N \otimes_B A)^{\mathrm{co}\mathcal{C}},\ \ \nu_N(n) = n \otimes_B 1;$$

$$\zeta_M:\ M^{\mathrm{co}\mathcal{C}} \otimes_B A \to M,\ \ \zeta_M(m \otimes_B a) = ma.$$

Let $i:\ B \to A$ be a morphism of rings. The associated canonical coring is $\mathcal{D} = A \otimes_B A$, with comultiplication and counit given by the formulas

$$\Delta_{\mathcal{D}}:\ \mathcal{D} \to \mathcal{D} \otimes_A \mathcal{D} \cong A \otimes_B A \otimes_B A,\ \ \Delta_{\mathcal{D}}(a \otimes_B a') = a \otimes_B 1 \otimes_B a'$$

and

$$\varepsilon_{\mathcal{D}}:\ \mathcal{D} = A \otimes_B A \to A,\ \ \varepsilon_{\mathcal{D}}(a \otimes_B a') = aa'.$$

If $i:\ B \to A$ is pure as a morphism of left and right B-modules, then the categories $\mathcal{M}_B$ and $\mathcal{M}^{\mathcal{D}}$ are equivalent.
Let $(\mathcal{C}, x)$ be a coring with a fixed grouplike element, and $i:\ B \to A^{\mathrm{co}\mathcal{C}}$ a ring morphism. We then have a morphism of corings

$$\mathrm{can}:\ \mathcal{D} = A \otimes_B A \to \mathcal{C},\ \ \mathrm{can}(a \otimes_B a') = axa'.$$

If F is fully faithful, then $B \cong A^{\mathrm{co}\mathcal{C}}$; if G is fully faithful, then can is an isomorphism. $(\mathcal{C}, x)$ is called a Galois coring if can : $A \to_{A^{\mathrm{co}\mathcal{C}}} A \to \mathcal{C}$ is bijective. From [6], we recall the following results.

Theorem 1.1

Let $(\mathcal{C}, x)$ be an A-coring with fixed grouplike element, and $B = A^{\mathrm{co}\mathcal{C}}$. Then the following statements are equivalent.

1. $(\mathcal{C}, x)$ is Galois and A is faithfully flat as a left B-module;
2. (F, G) is an equivalence and A is flat as a left B-module.

Let $(\mathcal{C}, x)$ be a coring with a fixed grouplike element, and take $T = A^{\mathrm{co}\mathcal{C}}$. Then ${}^*\mathcal{C} = {}_A\mathrm{Hom}(\mathcal{C}, A)$ is a ring, with multiplication given by

$$(f\#g)(c) = g(c_{(1)}f(c_{(2)})). \tag{5}$$

We have a morphism of rings $j : A \to {}^*\mathcal{C}$, given by

$$j(a)(c) = \varepsilon_{\mathcal{C}}(c)a.$$

This makes ${}^*\mathcal{C}$ into an A-bimodule, via the formula

$$(afb)(c) = f(ca)b.$$

Consider the left dual of the canonical map:

$${}^*\mathrm{can} : {}^*\mathcal{C} \to {}^*\mathcal{D} \cong {}_T\mathrm{End}(A)^{\mathrm{op}}, \quad {}^*\mathrm{can}(f)(a) = f(xa).$$

We then have the following result.

Proposition 1.2

If $(\mathcal{C}, x)$ is Galois, then *can is an isomorphism. The converse property holds if $\mathcal{C}$ and A are finitely generated projective, respectively as a left A-module, and a left T-module.

Let $Q = \{q \in {}^*\mathcal{C} \mid c_{(1)}q(c_{(2)}) = q(c)x, \text{ for all } c \in \mathcal{C}\}$. A straightforward computation shows that Q is a $({}^*\mathcal{C}, T)$-bimodule. Also A is a left $(T, {}^*\mathcal{C})$-bimodule; the

right ${}^*\mathcal{C}$-action is induced by the right $\mathcal{C}$-coaction: $a \cdot f = f(xa)$. Now consider the maps

$$\tau : \quad A \otimes_{*\mathcal{C}} Q \to T, \quad \tau(a \otimes_{*\mathcal{C}} q) = q(xa); \tag{6}$$

$$\mu : \quad Q \otimes_T A \to {}^*\mathcal{C}, \quad \mu(q \otimes_T a) = q\#i(a). \tag{7}$$

With this notation, we have the following property (see [6]).

Proposition 1.3

$(T, {}^*\mathcal{C}, A, Q, \tau, \mu)$ is a Morita context.

We also have (see [6]):

Theorem 1.4

Let $(\mathcal{C}, x)$ be a coring with fixed grouplike element, and assume that $\mathcal{C}$ is a left A-progenerator. We take a subring B of $T = A^{\mathrm{co}\mathcal{C}}$, and consider the map

$$\mathrm{can} : \ \mathcal{D} = A \otimes_B A \to \mathcal{C}, \quad \mathrm{can}(a \otimes_T a') = axa'$$

Then the following statements are equivalent:

1.
   - can is an isomorphism;
   - A is faithfully flat as a left B-module.
2.
   - *can is an isomorphism;
   - A is a left B-progenerator.
3.
   - $B = T$;
   - the Morita context $(B, {}^*\mathcal{C}, A, Q, \tau, \mu)$ is strict.
4.
   - $B = T$;
   - (F, G) is an equivalence of categories.

1.2 Partial group actions

Let G be a finite group, and $R \to S$ a commutative ring extension. From [13], we recall that a partial action α of G on S is a collection of ideals S_σ and isomorphisms of ideals $\alpha_\sigma : \ S_{\sigma^{-1}} \to S_\sigma$ such that

1. $S_1 = S$, and $\alpha_1 = S$, the identity on S;

2. $S_{(\sigma\tau)^{-1}} \supset \alpha_\tau^{-1}(S_\tau \cap S_{\sigma^{-1}})$;

3. $(\alpha_\sigma \circ \alpha_\tau)(x) = \alpha_{\sigma\tau}(x)$, for all $x \in \alpha_\tau^{-1}(S_\tau \cap S_{\sigma^{-1}})$.

In [14], the following particular situation is considered: every S_σ is of the form $S_\sigma = Se_\sigma$, where e_σ is an idempotent of S. In this case, we can show that

$$\alpha_\sigma(\alpha_\tau(xe_{\tau^{-1}})e_{\sigma^{-1}}) = \alpha_{\sigma\tau}(xe_{\tau^{-1}\sigma^{-1}})e_\sigma, \tag{8}$$

for all $\sigma, \tau \in G$ and $x \in S$. We then have an associative ring with unit

$$A \star_\alpha G = \bigoplus_{\sigma\in G} Ae_\sigma u\sigma,$$

with multiplication

$$(a_\sigma u_\sigma)(a_\tau u_\tau) = \alpha_\sigma(\alpha_{\sigma^{-1}}(a_\sigma)b_\tau)u_{\sigma\tau}. \tag{9}$$

2. PARTIAL GALOIS THEORY FOR NONCOMMUTATIVE RINGS

Let A be a (noncommutative) ring, and G a finite group. For every $\sigma \in G$, we assume that there is a central idempotent $e_\sigma \in A$, and a ring automorphism

$$\alpha_\sigma : \ Ae_{\sigma^{-1}} \to Ae_\sigma.$$

In particular, it follows that $\alpha_\sigma(e_{\sigma^{-1}}) = e_\sigma$. We can extend α_σ to A, by putting $\alpha_\sigma(a) = \alpha_\sigma(ae_\sigma)$, for all $a \in A$.
Then we consider the direct sum $\mathcal{C}$ of all the Ae_σ. Let v_σ be the element of $\mathcal{C}$ with e_σ in the Ae_σ-component, and 0 elsewhere. We then have

$$\mathcal{C} = \bigoplus_{\sigma\in G} Ae_\sigma v_\sigma = \bigoplus_{\sigma\in G} Av_\sigma.$$

Obviously $\mathcal{C}$ is a left A-module.

Lemma 2.1

$\mathcal{C}$ is an A-bimodule. The right A-action is given by the formula

$$(a'v_\sigma)a = a'\alpha_\sigma(ae_{\sigma^{-1}})v_\sigma \tag{10}$$

Proof. Let us show that (10) is an associative action: for all $a, a' \in A$, we have

$$\begin{aligned} v_\sigma(aa') &= \alpha_\sigma(aa'e_{\sigma^{-1}})v_\sigma = \alpha_\sigma(ae_{\sigma^{-1}}a'e_{\sigma^{-1}})v_\sigma \\ &= \alpha_\sigma(ae_{\sigma^{-1}})\alpha_\sigma(a'e_{\sigma^{-1}})v_\sigma = \alpha_\sigma(ae_{\sigma^{-1}})v_\sigma a' = (v_\sigma a)a'. \end{aligned}$$

□

We now consider the left A-linear maps

$$\Delta_{\mathcal{C}} :\ \mathcal{C} \to \mathcal{C} \otimes_A \mathcal{C}, \quad \Delta_{\mathcal{C}}(av_\sigma) = \sum_{\tau \in G} av_\tau \otimes_A v_{\tau^{-1}\sigma};$$

$$\varepsilon_{\mathcal{C}} :\ \mathcal{C} \to A, \quad \varepsilon_{\mathcal{C}}(\sum_{\sigma \in G} a_\sigma v_\sigma) = a_1.$$

Proposition 2.2

With notation as above, $(\mathcal{C}, \Delta_{\mathcal{C}}, \varepsilon_{\mathcal{C}})$ is an A-coring if and only if $e_1 = A$, $\alpha_1 = A$ and

$$\alpha_\sigma(\alpha_\tau(ae_{\tau^{-1}})e_{\sigma^{-1}}) = \alpha_{\sigma\tau}(ae_{\tau^{-1}\sigma^{-1}})e_\sigma, \tag{11}$$

for all $a \in A$ and $\sigma, \tau \in G$.

Proof. We compute

$$\begin{aligned} \Delta_{\mathcal{C}}(v_\sigma a) &= \Delta_{\mathcal{C}}(\alpha_\sigma(ae_{\sigma^{-1}})v_\sigma \\ &= \sum_{\tau \in G} \alpha_\sigma(ae_{\sigma^{-1}})v_\tau \otimes_A v_{\tau^{-1}\sigma}; \\ \Delta_{\mathcal{C}}(v_\sigma)a &= \sum_{\tau \in G} v_\tau \otimes_A v_{\tau^{-1}\sigma}a \\ &= \sum_{\tau \in G} v_\tau \otimes_A \alpha_{\tau^{-1}\sigma}(ae_{\sigma^{-1}\tau})v_{\tau^{-1}\sigma} \\ &= \sum_{\tau \in G} \alpha_\tau(\alpha_{\tau^{-1}\sigma}(ae_{\sigma^{-1}\tau})e_{\tau^{-1}})v_\tau \otimes_A v_{\tau^{-1}\sigma}. \end{aligned}$$

Hence $\Delta_{\mathcal{C}}$ is right A-linear if and only if

$$\alpha_\sigma(ae_{\sigma^{-1}})e_\tau = \alpha_\tau(\alpha_{\tau^{-1}\sigma}(ae_{\sigma^{-1}\tau})e_{\tau^{-1}}),$$

for all $\sigma, \tau \in G$ and $a \in A$. Substituting $\lambda = \tau^{-1}\sigma$, we find that this is equivalent to (11).
Let us now investigate when $\varepsilon_{\mathcal{C}}$ is right A-linear. We have

$$\varepsilon_{\mathcal{C}}(\sum_{\sigma \in G} a_\sigma v_\sigma)a = a_1 a)$$

and

$$\varepsilon_{\mathcal{C}}(\sum_{\sigma\in G} a_\sigma v_\sigma a) = \varepsilon_{\mathcal{C}}(\sum_{\sigma\in G} a_\sigma \alpha_\sigma(ae_{\sigma^{-1}})v_\sigma) = a_1\alpha_1(ae_1)$$

If $\varepsilon_{\mathcal{C}}$ is right A-linear, then we find that $\alpha_1(ae_1) = a$, for all $a \in A$. In particular,

$$e_1 = \alpha_1(e_1e_1) = \alpha_1(1e_1) = 1,$$

and then it follows that $\alpha_1(a) = a$, for all $a \in A$. Conversely, if $e_1 = 1$ and $\alpha_1 = A$, then it follows that $\varepsilon_{\mathcal{C}}$ is right A-linear.
Now assume that (11) holds, and that $e_1 = 1$ and $\alpha_1 = A$. The coassociativity and counit property then follow in a straightforward way. □

From now on, we will assume that $\mathcal{C} = \bigoplus_{\sigma\in G} Av_\sigma$ is an A-coring. The set of data $(e_\sigma, \alpha_\sigma)_{\sigma\in G}$ will be called an idempotent partial action of G on A. This is the case for the partial actions discussed in [14], that we recalled in Section 1.2, in view of (8).

Lemma 2.3

$x = \sum_{\sigma\in G} v_\sigma$ is a grouplike element of $\mathcal{C}$.

Proof. $\varepsilon_{\mathcal{C}}(x) = 1$, and

$$\Delta_{\mathcal{C}}(x) = \sum_{\sigma,\tau\in G} v_\tau \otimes_A v_{\tau^{-1}\sigma} = \sum_{\rho,\tau\in G} v_\tau \otimes_A v_\rho = x \otimes_A x.$$

□

Consider the left A-linear maps

$$u_\sigma :\ \mathcal{C} \to Ae_\sigma,\quad u_\sigma(\sum_{\tau\in G} a_\tau v_\tau) = a_\sigma e_\sigma. \tag{12}$$

Then for all $c \in \mathcal{C}$, we have

$$c = \sum_{\sigma\in G} u_\sigma(c)v_\sigma,$$

hence $\{(u_\sigma, v_\sigma) \mid \sigma \in G\}$ is a dual basis of $\mathcal{C}$ as a left A-module.
Now let (M, ρ) be a right $\mathcal{C}$-comodule. We have a right A-linear map

$$\rho :\ M \to \bigoplus_{\sigma\in G} M \otimes_A Av_\sigma.$$

Consider the maps

$$\rho_\sigma = (M \otimes_A u_\sigma) \circ \rho : \ M \to Me_\sigma.$$

We then have

$$\rho(m) = m_{[0]} \otimes_A m_{[1]} = \sum_{\sigma \in G} \rho_\sigma(m) \otimes_A v_\sigma,$$

for all $m \in M$. From the fact ρ is right A-linear, it follows that

$$\rho(ma) = \sum_{\sigma \in G} \rho_\sigma(ma) \otimes_A v_\sigma = \rho(m)a = \sum_{\sigma \in G} \rho_\sigma(m) \otimes_A \alpha_\sigma(ae_{\sigma^{-1}})v_\sigma,$$

hence

$$\rho_\sigma(ma) = \rho_\sigma(m)\alpha_\sigma(ae_{\sigma^{-1}}), \tag{13}$$

for all $m \in M$ and $\sigma \in G$. It follows from (13) that

$$\rho_\sigma(me_{\sigma^{-1}}) = \rho_\sigma(m)\alpha_\sigma(e_{\sigma^{-1}}) = \rho_\sigma(m)e_\sigma.$$

This means that $\rho_\sigma : \ M \to Me_\sigma$ factors through the projection $M \to Me_{\sigma^{-1}}$, so we obtain a map

$$\rho_\sigma : \ Me_{\sigma^{-1}} \to Me_\sigma.$$

Since $(M \otimes_A \varepsilon_{\mathcal{C}}) \circ \rho = M$, we have, for all $m \in M$:

$$m = \sum_{\sigma \in G} \rho_\sigma(m) \otimes_A \varepsilon_{\mathcal{C}}(v_\sigma) = \rho_1(m)e_1 = \rho_1(m).$$

Hence $\rho_1 : \ Me_1 = M \to Me_1 = M$ is the identity. From the coassociativity of ρ, we deduce that

$$\begin{aligned} \sum_{\sigma,\tau \in G} \rho_\tau(\rho_\sigma(m)) \otimes_A v_\tau \otimes_A v_\sigma &= \sum_{\sigma,\rho \in G} \rho_\sigma(m) \otimes_A v_\mu \otimes_A v_{\mu^{-1}\sigma} \\ = \sum_{\kappa,\mu \in G} v_{\mu\circ\kappa} \otimes_A v_\mu \otimes_A v_\kappa &= \sum_{\sigma,\tau \in G} v_{\tau\circ\sigma} \otimes_A v_\tau \otimes_A v_\sigma, \end{aligned}$$

hence

$$\rho_\tau(\rho_\sigma(m)) = \rho_{\tau\circ\sigma}(m), \tag{14}$$

for all $m \in M$ and $\sigma, \tau \in G$. In particular,

$$\rho_\tau(\rho_\sigma(me_{\sigma^{-1}}e_{\tau^{-1}})) = \rho_{\tau\circ\sigma}(me_{\sigma^{-1}\tau^{-1}})e_\tau. \tag{15}$$

It follows from (15) that $\rho_{\tau^{-1}} : \ Me_\tau \to Me_{\tau^{-1}}$ is the inverse of $\rho_\tau : \ Me_{\tau^{-1}} \to Me_\tau$.

Definition 2.4

Let $(e_\sigma, \alpha_\sigma)_{\sigma\in G}$ be an idempotent partial action of G on A, and M a right A-module. A partial Galois descent datum consists of a set of maps

$$\rho_\sigma :\ M \to Me_\tau$$

such that $\rho_1 = M$, the identity on M, the restriction of ρ_σ to $Me_{\tau^{-1}}$ is an isomorphism, and (13) and (15) hold for all $m \in M$, $a \in A$ and $\sigma, \tau \in G$.

Proposition 2.5

Let $(e_\sigma, \alpha_\sigma)_{\sigma\in G}$ be an idempotent partial action of G on A, and $\mathcal{C}$ the corresponding A-coring. Then right $\mathcal{C}$-coactions on M correspond bijectively to partial Galois descent data.

Proof. We have already explained above how a right $\mathcal{C}$-coaction ρ on M can be transformed into a partial Galois descent datum. Conversely, let $(\rho_\sigma)_{\sigma\in G}$ be a partial Galois descent datum, and define $\rho :\ M \to M \otimes_A \mathcal{C}$ by

$$\rho(m) = \sum_{\sigma\in G} \rho_\sigma(m) \otimes_A v_\sigma.$$

Straightforward computations show that ρ is a coaction, and that the two constructions are inverse to each other. □

Let M be a right $\mathcal{C}$-comodule. Then $m \in M^{\mathrm{co}\mathcal{C}}$ if and only if

$$\rho(m) = \sum_{\sigma\in G} \rho_\sigma(m) \otimes_A v_\sigma = \sum_{\sigma\in G} m \otimes_A v_\sigma = \sum_{\sigma\in G} me_\sigma \otimes_A v_\sigma$$

if and only if

$$\rho_\sigma(m) = \rho_\sigma(me_{\sigma^{-1}}) = me_\sigma,$$

for all $\sigma \in G$. We define

$$M^G = \{m \in M \mid \rho_\sigma(me_{\sigma^{-1}}) = me_{\sigma^{-1}},\ \text{for all } \sigma \in G\} = M^{\mathcal{C}}.$$

The grouplike element $x = \sum_{\sigma\in G} v_\sigma$ makes A into a right $\mathcal{C}$-comodule:

$$\rho(a) = 1 \otimes_A xa = \sum_{\sigma\in G} \alpha_\sigma(ae_{\sigma^{-1}}) \otimes_A v_\sigma,$$

and we have

$$T = A^G = \{a \in A \mid \alpha_\sigma(ae_{\sigma^{-1}}) = ae_\sigma, \text{ for all } \sigma \in G\}.$$

Let $i: B \to T$ be a ring morphism. We have seen in Section 1.1 that we have a pair of adjoint functors (F, G):

$$F: \mathcal{M}_B \to \mathcal{M}^{\mathcal{C}}, \quad F(N) = N \otimes_B A;$$

$$G: \mathcal{M}^{\mathcal{C}} \to \mathcal{M}_B, \quad G(N) = N^G.$$

$F(N) = N \otimes_B A$ is a right $\mathcal{C}$-comodule in the following way:

$$\rho_\sigma(n \otimes_A a) = n \otimes \alpha_\sigma(a).$$

The canonical map is the following:

$$\text{can}: A \otimes_B A \to \bigoplus_{\sigma \in G} A_e v_\sigma, \quad \text{can}(a \otimes b) = \sum_{\sigma \in G} a\alpha_\sigma(be_{\sigma^{-1}})v_\sigma.$$

$\bigoplus_{\sigma \in G} A_e v_\sigma$ is a Galois coring if can : $A \otimes_{A^G} A \to \bigoplus_{\sigma \in G} A_e v_\sigma$ is an isomorphism. We will then say that A is a partial G-Galois extension of A^G. From Theorem 1.1, we immediately obtain the following result.

Theorem 2.6

Let $(e_\sigma, \alpha_\sigma)_{\sigma \in G}$ be an idempotent partial action of G on A, and $T = A^G$. Then the following assertions are equivalent.

1. A is a partial G-Galois extension of T and T is faithfully flat as a left T-module;
2. (F, G) is a category equivalence and A is flat as a left T-module.

3. PARTIAL ACTIONS AND MORITA THEORY

Let us now compute the multiplication on

$${}^*\mathcal{C} = {}_A\text{Hom}(\mathcal{C}, A) = \bigoplus_{\sigma \in G} {}_A\text{Hom}(Ae_\sigma, A).$$

We will use the maps u_σ defined in (12). Also recall that ${}^*\mathcal{C}$ is an A-bimodule, with left and right A-action

$$(afb)(c) = f(ca)b.$$

Take $f \in {}_A\mathrm{Hom}(\mathcal{C}, A)$. For all $c \in \mathcal{C}$, we have

$$f(c) = \sum_{\sigma \in \mathcal{C}} u_\sigma(c) f(v_\sigma).$$

Now $f(v_\sigma) = f(e_\sigma v_\sigma) = e_\sigma f(v_\sigma) \in Ae_\sigma$, so we can conclude that

$${}^*\mathcal{C} = \bigoplus_{\sigma \in \mathcal{C}} u_\sigma A e_\sigma.$$

For $b \in A$, we compute

$$(bu_\tau)(\sum_{\sigma \in G} a_\sigma v_\sigma) = u_\tau(\sum_{\sigma \in G} a_\sigma v_\sigma b) = u_\tau(\sum_{\sigma \in G} a_\sigma \alpha_\sigma(be_{\sigma^{-1}})v_\sigma) = a_\tau \alpha_\tau(be_{\tau^{-1}}),$$

and we conclude that

$$bu_\tau = u_\tau \alpha_\tau(be_{\tau^{-1}}). \tag{16}$$

We next compute, using (11):

$$\begin{aligned}(u_\rho \# u_\nu)(\sum_{\sigma \in G} a_\sigma v_\sigma) &= u_\nu(\sum_{\sigma,\tau} a_\sigma v_\tau u_\rho(v_{\tau^{-1}\sigma})) = u_\nu(\sum_{\sigma,\tau} a_\sigma v_\tau \delta_{\tau\rho,\sigma}) \\ &= u_\nu(\sum_{\tau} a_{\tau\rho} v_\tau) = a_{\nu\rho} = u_{\nu\rho}(\sum_{\sigma \in G} a_\sigma v_\sigma),\end{aligned}$$

and we conclude that

$$u_\sigma \# u_\tau = u_{\sigma\tau}. \tag{17}$$

We can summarize this as follows:

Proposition 3.1

Let $\mathcal{C} = \bigoplus_{\sigma \in G} Av_\sigma$. The left dual ring is

$${}^*\mathcal{C} = \bigoplus_{\sigma \in G} u_\sigma A e_\sigma,$$

with multiplication rule

$$u_\tau b_\tau \# u_\sigma a_\sigma = u_{\sigma\tau} \alpha_\sigma(b_\tau e_{\sigma^{-1}}) a_\sigma. \tag{18}$$

If A is commutative, then ${}^*\mathcal{C}$ is isomorphic to $(A \star_\alpha G)^{\mathrm{op}}$, as introduced in [14], see (9). Indeed, for $a \in Ae_\sigma$ and $b \in Ae_\tau$, we compute that

$$\begin{aligned}\alpha_\sigma(\alpha_{\sigma^{-1}}(a_\sigma)b_\tau) &= \alpha_\sigma(\alpha_{\sigma^{-1}}(a_\sigma)b_\tau e_{\sigma^{-1}}) \\ &= a_\sigma \alpha_\sigma(b_\tau e_{\sigma^{-1}}) \alpha_\sigma(b_\tau e_{\sigma^{-1}}) a_\sigma\end{aligned}$$

Recall that a ring morphism $A \to R$ is called *Frobenius* if there exists an A-bimodule map $\overline{\nu}: R \to A$ and $e = e^1 \otimes_A e^2 \in R \otimes_A R$ (summation implicitly understood) such that

$$re^1 \otimes_A e^2 = e^1 \otimes_A e^2 r \tag{19}$$

for all $r \in R$, and

$$\overline{\nu}(e^1)e^2 = e^1\overline{\nu}(e^2) = 1. \tag{20}$$

This is equivalent to the restrictions of scalars $\mathcal{M}_R \to \mathcal{M}_A$ being a Frobenius functor, which means that its left and right adjoints are isomorphic (see [8, Sec. 3.1 and 3.2]). $(e, \overline{\nu})$ is then called a Frobenius system.

Proposition 3.2

Suppose that we have an idempotent partial action of G on A. Then the ring morphism $A \to {}^*\mathcal{C}$ is Frobenius.

Proof. The Frobenius system is $(e = \sum_{\sigma\in G} u_{\sigma^{-1}} \otimes_A u_\sigma, \overline{\nu})$, with

$$\overline{\nu}(\sum_{\sigma\in G} u_\sigma a_\sigma) = a_1.$$

We compute that, for all $a \in A$,

$$\begin{aligned} a\sum_{\sigma\in G} u_{\sigma^{-1}} \otimes_A u_\sigma &= \sum_{\sigma\in G} u_{\sigma^{-1}}\alpha_{\sigma^{-1}}(ae_\sigma) \otimes_A u_\sigma \\ &= \sum_{\sigma\in G} u_{\sigma^{-1}} \otimes_A u_\sigma\alpha_\sigma(\alpha_{\sigma^{-1}}(ae_\sigma)) = \sum_{\sigma\in G} u_{\sigma^{-1}} \otimes_A u_\sigma ae_\sigma \\ &= \sum_{\sigma\in G} u_{\sigma^{-1}} \otimes_A u_\sigma a. \end{aligned}$$

The rest is obvious. □

Let $i: B \to T = A^{\mathrm{co}\mathcal{C}}$ be a ring morphism. We have the canonical morphism

$$\mathrm{can}: \mathcal{D} = A \otimes_B A \to \mathcal{C} = \bigoplus_{\sigma\in G} Av_\sigma,$$

given by

$$\mathrm{can}(a \otimes b) = \sum_{\sigma\in G} av_\sigma b = \sum_{\sigma\in G} a\alpha_\sigma(be_{\sigma^{-1}})v_\sigma$$

We can also compute that

$${}^*\mathrm{can}: {}^*\mathcal{C} = \bigoplus_{\sigma\in G} u_\sigma A \to {}^*\mathcal{D} \cong {}_B\mathrm{End}(A)^{\mathrm{op}}$$

is given by

$$^*\mathrm{can}(u_\tau b_\tau)(a) = \alpha_\tau(ae_{\tau^{-1}})b_\tau.$$

Let us now compute the module $Q \subset {}^*\mathcal{C}$ introduced in Section 1.1 Recall that $q \in Q$ if and only if

$$c_{(1)}q(c_{(2)}) = q(c)x, \tag{21}$$

for all $c \in \mathcal{C}$.

Proposition 3.3

$$Q = \{\textstyle\sum_{\sigma\in G} u_\sigma\alpha_\sigma(ae_{\sigma^{-1}}) \mid a \in A\}.$$

Proof. Take $q = \sum_{\sigma\in G} u_\sigma a_\sigma \in Q$, with $a_\sigma \in Ae_\sigma$, and put $c = v_\tau$ in (21). Recall that $\Delta_{\mathcal{C}}(v_\sigma) = \sum_{\rho\in G} v_\rho \otimes_A v_{\rho^{-1}\tau}$. Then we calculate that

$$c_{(1)}q(c_{(2)}) = \sum_{\rho,\sigma\in G} v_\rho\delta_{\sigma,\rho^{-1}\tau}a_\sigma = \sum_{\rho\in G} v_\rho a_{\rho^{-1}\tau} = \sum_{\rho\in G} \alpha_\rho(a_{\rho^{-1}\tau}e_{\rho^{-1}})v_\rho,$$

$$q(c) = (\sum_{\sigma\in G} u_\sigma a_\sigma)(v_\tau) = a_\tau$$

and

$$q(c)x = \sum_{\rho\in G} a_\tau v_\rho$$

Hence it follows that

$$a_\tau e_\rho = \alpha_\rho(a_{\rho^{-1}\tau}e_{\rho^{-1}}), \tag{22}$$

for all $\tau, \rho \in G$. Taking $\tau = \rho$, we find that

$$a_\tau = a_\tau e_\tau = \alpha_\tau(a_1 e_{\tau^{-1}}), \tag{23}$$

and we find that

$$q = \sum_{\sigma\in G} u_\sigma a_\sigma = \sum_{\sigma\in G} u_\sigma\alpha_\sigma(a_1 e_{\sigma^{-1}}) \tag{24}$$

is of the desired form. Conversely, take q of the form (24). Then (23) holds. Using (11), we compute

$$\alpha_\rho(a_{\rho^{-1}\tau}e_{\rho^{-1}}) = \alpha_\rho(\alpha_{\rho^{-1}\tau}(a_1 e_{\tau^{-1}\rho})e_{\rho^{-1}}) = \alpha_\tau(a_1 e_{\tau^{-1}})e_\rho = a_\tau e_\rho,$$

and (22) follows, which means that (21) holds for $c = v_\tau$. Using the left A-linearity of q and $\Delta_{\mathcal{C}}$, it follows that (21) holds for arbitrary $c \in \mathcal{C}$. □

It follows from proposition 3.3 that we have an isomorphism of abelian groups

$$A \to Q, \quad a \mapsto \sum_{\sigma \in G} u_\sigma \alpha_\sigma(ae_{\sigma^{-1}}).$$

This can also be seen using Proposition 3.2 and [6, Theorem 2.7]. The $({}^*\mathcal{C}, T)$-bimodule structure on Q (see Proposition 1.2) can be transported to A. The right T-action on A is then given by right multiplication, and the left ${}^*\mathcal{C}$-action is the following:

$$(u_\tau a_\tau) \cdot a = \alpha_{\tau^{-1}}(a_\tau a e_\tau).$$

Recall also from Proposition 1.2 that $A \in {}_T\mathcal{M}_{{}^*\mathcal{C}}$. The left T-action is given by left multiplication. The right ${}^*\mathcal{C}$-action is the following:

$$\begin{aligned} a \cdot (u_\tau a_\tau) &= (a u_\tau a_\tau)(x) \\ &= \sum_{\sigma \in G} u_\tau \alpha_\tau(a e_{\tau^{-1}}) a_\tau)(v_\sigma) = \alpha_\tau(a e_{\tau^{-1}}) b_\tau. \end{aligned}$$

We have seen in Proposition 1.2 that we have a Morita context $(T, {}^*\mathcal{C}, A, Q, \tau, \mu)$. Using the isomorphism between A and Q, we find a Morita context $(T, {}^*\mathcal{C}, A, A, \tau, \mu)$. Let us compute the connecting maps $\tau : A \otimes_{{}^*\mathcal{C}} A \to T$ and $\mu : A \otimes_T A \to {}^*\mathcal{C}$, using (6-7).

$$\begin{aligned} \tau(b \otimes a) &= (\sum_{\sigma \in G} u_\sigma \alpha_\sigma(a e_{\sigma^{-1}}))(\sum_{\tau \in G} (v_\tau b)) \\ &= \sum_{\sigma, \tau \in G} u_\sigma(\alpha_\tau(b e_{\tau^{-1}}) v_\tau) \alpha_\sigma(a e_{\sigma^{-1}}) \\ &= \sum_{\sigma \in G} \alpha_\sigma(b e_{\sigma^{-1}}) \alpha_\sigma(a e_{\sigma^{-1}}) \\ &= \sum_{\sigma \in G} \alpha_\sigma(b a e_{\sigma^{-1}}); \\ \mu(a \otimes b) &= \sum_{\sigma \in G} u_\sigma \alpha_\sigma(a e_{\sigma^{-1}}) b. \end{aligned}$$

We summarize our results as follows.

Proposition 3.4

We have a Morita context $(T, {}^*\mathcal{C}, A, Q, \tau, \mu)$. The connecting maps are given by the formulas

$$\tau(b \otimes a) = \sum_{\sigma \in G} \alpha_\sigma(b a e_{\sigma^{-1}}); \tag{25}$$

$$\mu(a \otimes b) = \sum_{\sigma \in G} u_\sigma \alpha_\sigma(a e_{\sigma^{-1}}) b. \tag{26}$$

Proposition 3.5

The map τ in the Morita context from Proposition 3.4 is surjective if and only if there exists $a \in A$ such that

$$\sum_{\sigma \in G} \alpha_\sigma(ae_{\sigma^{-1}}) = 1.$$

Proof. According to [9, Theorem 3.3] τ is surjective if and only if there exists $q \in Q$ such that $q(x) = 1$. Let $a \in A$ correspond to $q \in Q$. Then we compute that

$$q(x) = (\sum_{\sigma \in G} u_\sigma \alpha_\sigma(ae_{\sigma^{-1}}))(\sum_{\tau \in G} v_\tau) = \sum_{\sigma \in G} \alpha_\sigma(ae_{\sigma^{-1}}),$$

and the result follows. □

From Theorem 1.4, we obtain:

Theorem 3.6

Let G be a finite group, and $(e_\sigma, \alpha_\sigma)_{\sigma \in G}$ an idempotent partial action of G on A. Let $i: B \to T = A^{\mathrm{co}\mathcal{C}}$ a ring morphism, and consider can : $A \otimes_B A \to \mathcal{C}$. Then the following assertions are equivalent.

1. - can is an isomorphism;
   - A is faithfully flat as a left B-module.
2. - *can is an isomorphism;
   - A is a left B-progenerator.
3. - $B = T$;
   - the Morita context $(B, {}^*\mathcal{C}, A, A, \tau, \mu)$ is strict.
4. - $B = T$;
   - (F, G) is an equivalence of categories.

If we take A and B commutative, then Theorem 3.6 implies part of [14, Theorem 3.1], namely the equivalence of the conditions (i), (ii) and (iii).

REFERENCES

[1] Brzeziński, T. The structure of corings. Induction functors, Maschke-type theorem, and Frobenius and Galois properties, *Algebr. Representat. Theory* 5 (2002), 389–410.

[2] Brzeziński, T. The structure of corings with a grouplike element, preprint math. RA/0108117.

[3] Brzeziński, T. and Hajac, P.M., *Coalgebra extensions and algebra coextensions of Galois type*, Comm. Algebra 27 (1999), 1347-1367.

[4] Brzeziński, T. and Wisbauer, R., "Corings and comodules", *London Math. Soc. Lect. Note Ser.* 309, Cambridge University Press, Cambridge, 2003.

[5] Caenepeel, S. Brauer groups, Hopf algebras and Galois theory, *K-Monographs Math.* 4, Kluwer Academic Publishers, Dordrecht, 1998.

[6] Caenepeel, S. *Galois corings from the descent theory point of view*, *Fields Inst. Comm*, to appear.

[7] Caenepeel, S., Groot, E. De., Galois theory for weak Hopf algebras, in preparation.

[8] Caenepeel, S., Militaru, G. and Shenglin Zhu, "Frobenius and separable functors for generalized module categories and nonlinear equations", *Lect. Notes in Math.* 1787, Springer Verlag, Berlin, 2002.

[9] Caenepeel, S., Vercruysse, J. and Shuanhong Wang, *Morita Theory for corings and cleft entwining structures*, *J. Algebra* 276 (2004), 210–235.

[10] Chase, S., Harrison, D. and Rosenberg, A., Galois theory and Galois cohomology of commutative rings, *Mem. Amer. Math. Soc.* 52 (1965), 1–19.

[11] Chase, S. and Sweedler, M. E., "Hopf algebras and Galois theory", *Lect. Notes in Math.* 97, Springer Verlag, Berlin, 1969.

[12] Dokuchaev, M., Exel, R. and Piccione, P., Partial representations and partial group algebras, *J. Algebra* 226 (2000), 251–268.

[13] Dokuchaev, M. and Exel, R., Associativity of crossed products by partial actions, enveloping actions and partial representations, *Trans. Amer. Math. Soc.*, to appear.

[14] Dokuchaev, M., Ferrero, M. and Pacques, A., Partial Galois theory of commutative rings, preprint 2004.

[15] Dokuchaev, M., and Zhukavets, N., On finite degree partial representations of groups, *J. Algebra*, to appear.

[16] Exel, R. Twisted partial actions: a classification of regular C^*-algebraic bundels, *Proc. London Math. Soc.* 74 (1997), 417-443.

[17] Exel, R. Partial actions of groups and actions of semigroups, *Proc. Amer. Math. Soc.* 126 (1998), 3481–3494.

[18] Kreimer, H., Takeuchi M.

[19] Quigg, J., and Raeburn, I. Characterizations of crossed products by partial actions, *J. Operator Theory* 37 (1997), 311–340.

[20] Schneider, H.J., Principal homogeneous spaces for arbitrary Hopf algebras, *Israel J. Math.* 70 (1990), 167–195.

[21] Sweedler, M. E. Cohomology of algebras over Hopf algebras, *Trans. Amer. Math. Soc.* 133 (1968), 205-239.

[22] Sweedler, M. E., "Hopf algebras", Benjamin, New York, 1969.

[23] Sweedler, M. E. The predual Theorem to the Jacobson-Bourbaki Theorem, *Trans. Amer. Math. Soc.* 213 (1975), 391–406.

[24] Wisbauer, R. On Galois corings, in "Hopf algebras in non-commutative geometry and physics", S. Caenepeel and F. Van Oystaeyen (eds.), *Lecture Notes Pure Appl. Math.* 239, Dekker, New York, to appear.

COMODULES OVER SEMIPERFECT CORINGS

S. Caenepeel[1] and M. Iovanov[2]
[1]Faculty of Applied Sciences, Vrije Universiteit Brussel,
VUB, B-1050 Brussels, Belgium
e-mail: scaenepe@vub.ac.be

[2]Faculty of Mathematics, University of Bucharest,
Str. Academiei 14, RO-70109 Bucharest, Romania
e-mail: myo30@lycos.com

Abstract

We discuss when the Rat functor associated to a coring satisfying the left α-condition is exact. We study the category of comodules over a semiperfect coring. We characterize semiperfect corings over artinian rings and over qF-rings.

INTRODUCTION

The aim of this note is to generalize properties of semiperfect coalgebras over fields, as discussed in [13], see also [8], to semiperfect corings. We also extend some results given in [4].

Corings were introduced by Sweedler [14]. A coring over a (possibly noncommutative) ring R is a coalgebra (or comonoid) in the category of R-bimodules. Since the beginning of the 21st century, there has been a renewed interest in corings and comodules over a coring, iniated by Brzeziński's paper [3]. The key point is that Hopf modules and most of their generalizations (relative Hopf modules, graded modules, Yetter-Drinfeld modules and many more) are comodules over a certain coring. This observation appeared in MR 2000c 16047 written by Masuoka, who tributed it to Takeuchi, but apparently it was already known by Sweedler, at least in the case of Hopf modules. It has lead to a unified and simplified treatment of the above mentioned modules, and new viewpoints on subjects like descent theory and Galois theory. For an extensive treatment, we refer to [4].

1991 Mathematics Subject Classification. 16W30.
Keywords and phrases. semiperfect coring, qF-ring, rational module.
Research supported by the bilateral project "Hopf Algebras in Algebra, Topology, Geometry and Physics" of the Flemish and Romanian governments.

In this paper, we study semiperfect corings. A coring is called right semiperfect if it satisfies the left α-condition, and the (abelian) category of right $\mathcal{C}$-comodules is semiperfect, which means that every simple object has a projective cover. It turns out that this notion is closely related to rationality properties of modules over the dual of the coring (which is a ring). Rationality properties have been studied in [1] and [6]. The Rat functor sends a module over the dual of the coring to its largest rational submodule. It can be described using the category $\sigma[M]$. The category $\sigma[M]$ is discussed briefly in Section 1, and the Rat functor is introduced in Section 2. General facts on the category $\sigma[M]$ show that the exactness of the Rat functor is connected to some topological properties of the base ring R, more precisely the M-adic topology on M. In the case of corings, the $\mathcal{C}$-adic topology on ${}^*\mathcal{C}$ coincides with the finite topology, motivating a general study of the properties of the finite topology. We then give some connections between density properties, direct sum decompositions and the exactness of Rat. We show (see Corollary 2.7) that the Rat functor is exact if the coring $\mathcal{C}$ can be decomposed as a direct sum of finitely generated left $\mathcal{C}$-comodules. Under certain conditions, which hold if R is a qF-ring, we can prove the converse, namely if Rat is exact, then there is a direct sum decomposition of $\mathcal{C}$ into finitely generated comodules. This is in fact an application of the duality between left and right finitely generated modules over qF-rings.
In Section 3, we characterize semiperfect corings over artinian rings. The main result is Theorem 3.1, stating that a coring over an artinian ring is right semiperfect if and only if the category of right comodules has enough projectives, if and only if it has a projective generator, if and only if every finitely generated comodule has a finitely generated projective cover.
In Section 4, we discuss some applications and examples. First, we apply our results to the case where R is a qF-ring. We recover a result of [10] telling that a left and right (locally) projective coring over a qF-ring is right semiperfect if and only if the Rat functor is exact. Also two-sided prefectness is equivalent to two-sided semiperfectness for corings over qF-rings.
finally, we give some examples, focussing on the Sweedler coring associated to a ring morphism. In particular, we can describe the Rat functor in this situation, and we can discuss when the assumptions of the results in Section 3 and 4.1 are satisfied.

1. PRELIMINARY RESULTS

1.1 The category $\sigma[M]$

Let R be a ring, and $M \in {}_R\mathcal{M}$. Recall from [15, Sec. 15] that $\sigma[M]$ is the full subcategory of ${}_R\mathcal{M}$ consisting of R-modules that are subgenerated by M, that is,

submodules of an epimorphic image of $M^{(I)}$, for some index set I. $\sigma[M]$ is the smallest closed subcategory of ${}_R\mathcal{M}$ containing M. Since epimorphic images of objects of $\sigma[M]$ belong to $\sigma[M]$ (see [15, Prop. 15.1]), we have for any $N \in {}_R\mathcal{M}$ that

$$\mathcal{T}^M(N) = \sum\{f(X) \mid X \in \sigma[M],\ f \in {}_R\mathrm{Hom}(N, X)\} \in \sigma[M].$$

$\mathcal{T}^M :\ {}_R\mathcal{M} \to \sigma[M]$ is called the trace functor, and it is straightforward to show that $\mathcal{T}^M$ is the right adjoint of the inclusion functor $i :\ \sigma[M] \to {}_R\mathcal{M}$. Therefore $\mathcal{T}^M$ is left exact; it is also not difficult to see that

$$\mathcal{T}^M(N) = \sum\{X \mid X \subset \sigma[M],\ X \subset M\}.$$

For $X, Y \in {}_R\mathrm{Hom}(X, Y)$, we consider the finite topology on ${}_R\mathrm{Hom}(X, Y)$. A basis of open sets consists of

$$\mathcal{O}(f, x_1, \cdots, x_n) = \{g \in {}_A\mathrm{Hom}(X, Y) \mid g(x_i) = f(x_i),\ \text{for all } i = 1, \cdots, n\}$$

We have a natural map $r :\ R \to {}_{\mathbb{Z}}\mathrm{Hom}(M, M)$, $r_a(m) = am$. The finite topology on ${}_Z Z\mathrm{Hom}(M, M)$ induces a topology on R, called the M-adic topology.
An ideal T of R is called M-dense in R if it is dense in the M-adic topology. This means that for all $a \in R$ and $m_1, \cdots, m_n \in M$, there exists a $b \in T$ such that $am_i = bm_i$, for all i. A left T-module N is called unital if for every $n \in N$, there exists $t \in T$ such that $tn = n$, or, equivalently, for every finite $\{n_1, \cdots, n_k\} \subset N$, there exists $t \in N$ such that $tn_i = n_i$, for all i.
The proof of Proposition 1.1 is straightforward; we also refer to [4, Sec. 41].

Proposition 1.1 Let R be a ring, and $M \in {}_R\mathcal{M}$.
(a) For an ideal T of R, and a faithful R-module M, the following assertions are equivalent.

(i) T is M-dense in R;

(ii) M is a unital T-module (with the induced structure from R);

(iii) $TN = N$ for all $N \in \sigma[M]$;

(iv) the multiplication map $T \otimes_R N \to N$ is an isomorphism.

(b) $T = \mathcal{T}^M(A)$ is an ideal of A, and the following assertions are equivalent.

(i) T is M-dense in A;

(ii) M is a T-unital module;

(iii) $\mathcal{T}^M$ is exact;

(iv) $T^2 = T$ and T is a generator in $\sigma[M]$.

Let K be an A-submodule of M. Recall (see e.g. [15, 19.1]) that K is called superfluous or small, written $K \ll M$, if for every submodule $L \subset M$, $K + L = M$ implies that $L = M$. An epimorphism $f : M \to N$ is called superfluous if $\mathrm{Ker}\, f \ll M$. Note that this definition can be extended to abelian categories.

Proposition 1.2

Assume that $\mathcal{T}^M$ is exact.

(i) The class $\sigma[M]$ is closed under small epimorphisms in ${}_A\mathcal{M}$;

(ii) the inclusion functor $\sigma[M] \to {}_A\mathcal{M}$ preserves projectives.

Proof (i) Take $N \in \sigma[M]$, and let

$$0 \to K \to X \to N$$

be an exact sequence in ${}_A\mathcal{M}$ such that K is small in X. then $Y = X/(K + \mathcal{T}^M(X))$ is a quotient of $X/K = N \in \sigma[M]$, so $Y \in \sigma[M]$, by [15, 15.1], and $\mathcal{T}^M(Y) = Y$. Consider the exact sequence

$$0 \to \mathcal{T}^M(X) \to X \to X/\mathcal{T}^M(X) \to 0.$$

Since $\mathcal{T}^M$ is exact and idempotent, it follows that $\mathcal{T}^M(X/\mathcal{T}^M(X)) = 0$. Now Y is a quotient of $X/\mathcal{T}^M(X)$, and it follows from the exactness of $\mathcal{T}^M$ that $\mathcal{T}^M(Y) = 0$. Thus $Y = 0$, and $K + \mathcal{T}^M(X) = X$. Since $K \ll X$, we have that $\mathcal{T}^M(X) = X$, so $X \in \sigma[M]$, as needed.

1.2. Properties of the finite topology

Proposition 1.3 Let R be a ring, and fix a right R-module T. Density will mean density in the finite topology.

(i) Let $M = M_1 \oplus M_2$ in $\mathcal{M}_R$, and $X_1 \subset \mathrm{Hom}_R(M_1, T)$, $X_2 \subset \mathrm{Hom}_R(M_2, T)$ If $X_1 \oplus X_2$ is dense in $\mathrm{Hom}_R(M, T) = \mathrm{Hom}_R(M_1, T) \oplus \mathrm{Hom}_R(M_1, T)$, then each X_i is dense in $\mathrm{Hom}_R(M_i, T)$.

(ii) Let $(M_i)_{i\in I}$ be a family of R-modules, and $X_i \subset \mathrm{Hom}_R(M_i, T)$ such that each X_i is dense in $\mathrm{Hom}_R(M_i, T)$. Let $M = \bigoplus_{i\in I} M_i$. Then $\bigoplus_{i\in I} X_i$ is dense in $\mathrm{Hom}_R(M, T) = \prod_{i\in I} \mathrm{Hom}_R(M_i, T)$.

Proof (i) Take $f \in \mathrm{Hom}_R(M_1, T)$ and F is a finite subset of M_1. Viewing f as the pair $(f, 0) \in \mathrm{Hom}_R(M_1, T) \oplus \mathrm{Hom}_R(M_2, T)$ and $F \subset M_1 \subset M_1 \oplus M_2$, we find a pair $(g, h) \in X_1 \oplus X_2 \subset \mathrm{Hom}_R(M, T) = \mathrm{Hom}_R(M_1, T) \oplus \mathrm{Hom}_R(M_2, T)$ such that $(g, h) = (f, 0)$ on F, so $g = f$ on all $m \in F$, with $g \in X_1 \subset \mathrm{Hom}_R(M_1, T)$.
(ii) Take $(f_i)_{i\in I} \in \mathrm{Hom}_R(M, T) = \prod_{i\in I} \mathrm{Hom}_R(M_i, T)$ and a finite subset $F \subset \bigoplus_{i\in I} M_i$. Then there is a finite subset $J \subset I$ such that $F \subset \bigoplus_{i\in J} M_i$. $F_i = \{m_i \mid m \in F\}$ is finite, and, using the density of X_i in $\mathrm{Hom}_R(M_i, T)$, we find $g_i \in X_i$ such that $g_i = f_i$ on F_i. Now let $g \in \prod_{i\in I} \mathrm{Hom}_R(M_i, T) = \mathrm{Hom}_R(M, T)$ be defined as follows: the i-th component of g is g_i if $i \in J$, and it is zero otherwise. Then $g \in \bigoplus_{i\in I} X_i$ and $g = f$ on all F_i, and a fortiori on F, by linearity.

Corollary 1.4

If $(M_i)_{i\in I}$ is a family of R-modules and $X_i \subset \mathrm{Hom}_R(M_i, T)$ then $\bigoplus_{i\in I} X_i$ is dense in $\prod_{i\in I} \mathrm{Hom}_R(M_i, T) = \mathrm{Hom}_R(\bigoplus_{i\in I} M_i, T)$ if and only if all X_i are dense in $\mathrm{Hom}_R(M_i, T)$. Consequently, the direct sum $\bigoplus_{i\in I} \mathrm{Hom}_R(M_i, T)$ is dense in the direct product $\prod_{i\in I} \mathrm{Hom}_R(M_i, T)$.

Proposition 1.5

Let $T \in \mathcal{M}_R$ be an injective module, and $u : X \to Y$ a monomorphism in $\mathcal{M}_R$. If V is dense in $\mathrm{Hom}_R(Y, T)$, then $\mathrm{Hom}_R(u, T)(V)$ is dense in $\mathrm{Hom}_R(X, T)$.

Proof

Take $f \in \mathrm{Hom}_R(X, T)$, and a finite subset $F \subset X$. As T is an injective module, we can find $g \in \mathrm{Hom}_R(Y, T)$ such that $g \circ u = f$. As $u(F)$ is a finite subset of Y we can find $h \in V$ such that h equals g on $u(F)$. Now we obviously have that $\mathrm{Hom}_R(u, T)(h) = h \circ u$ equals $g \circ u = f$ on F, hence $\mathrm{Hom}_R(u, T)(V)$ is dense in $\mathrm{Hom}_R(X, T)$.

2. CORINGS AND THE RAT FUNCTOR

2.1 Corings Let R be a ring. An R-coring is a coalgebra in the monoidal category ${}_R\mathcal{M}_R$. It consists of a triple $\mathcal{C} = (\mathcal{C}, \Delta, \varepsilon)$, where $\mathcal{C}$ is an R-bimodule, and $\Delta : \mathcal{C} \to \mathcal{C} \otimes_R \mathcal{C}$ and $\varepsilon : Cc \to R$ are R-bimodule maps satisfying appropriate coassociativity

and counit properties. We refer to [3] and [4] for more detail about corings. We use the Sweedler-Heyneman notation

$$\Delta(c) = c_{(1)} \otimes_R c_{(2)},$$

where the summation is implicitely understood. If $\mathcal{C}$ is an R-coring, then ${}^*\mathcal{C} = {}_R\mathrm{Hom}(\mathcal{C}, R)$ is a ring with multiplication given by the formula

$$(f\#g)(c) = g(c_{(1)}f(c_{(2)})).$$

The unit of the multiplication is ε. We have a ring morphism

$$\iota : \ R \to {}^*\mathcal{C}, \ \iota(r)(c) = \varepsilon(c)r.$$

A right $\mathcal{C}$-comodule consists of a pair (M, ρ^r), where $M \in \mathcal{M}_R$ and $\rho^r : \ M \to M \otimes_R \mathcal{C}$ is a right A-linear map satisfying the conditions

$$(\rho^r \otimes_R \mathcal{C}) \circ \rho^r = (M \otimes_R \Delta) \circ \rho^r \quad \text{and} \quad (M \otimes_R \varepsilon) \circ \rho^r = M.$$

Left $\mathcal{C}$-comodules are defined in a similar way, and the categories of left and right $\mathcal{C}$-comodules are respectively denoted by $\mathcal{M}^{\mathcal{C}}$ and ${}^{\mathcal{C}}\mathcal{M}$. We use the Sweedler-Heyneman notation

$$\rho^r(m) = m_{[0]} \otimes_R m_{[1]} \quad \text{and} \quad \rho^l(m) = m_{[-1]} \otimes_R m_{[0]}$$

for right and left $\mathcal{C}$-coactions. We have a functor $F : \ \mathcal{M}^{\mathcal{C}} \to \mathcal{M}_{*\mathcal{C}}$, with $F(M) = M$ as an R-module, equipped with the right ${}^*\mathcal{C}$-action $m \cdot f = m_{[0]}f(m_{[1]})$. In particular, $\mathcal{C}$ is a right and left ${}^*\mathcal{C}$-module. If M and N are right $\mathcal{C}$-comodules, then the set of R-linear maps preserving the $\mathcal{C}$-coaction is denoted by $\mathrm{Hom}^{\mathcal{C}}(M, N)$.

2.2 The α-condition $M \in {}_R\mathcal{M}$ satisfies the (left) α-condition if the canonical map

$$\alpha_{N,M} : \ N \otimes_R M \to \mathrm{Hom}_R({}^*M, N), \quad \alpha(n \otimes_R m)(f) = nf(m)$$

is injective, for all $N \in \mathcal{M}_R$. Otherwise stated: if $n \otimes_R m \in N \otimes_R M$ is such that $nf(m) = 0$ for all $f \in {}^*M$, then $n \otimes m = 0$. M satisfies the α-condition if and only if M is locally projective in ${}_R\mathcal{M}$. An R-coring $\mathcal{C}$ satisfies the left α-condition if and only if $\mathcal{M}^{\mathcal{C}}$ is a full subcategory of $\mathcal{M}_{*\mathcal{C}}$, and the natural functor $\mathcal{M}^{\mathcal{C}} \to \sigma[C_{*\mathcal{C}}]$ is an isomorphism. In this case, $\mathcal{C}$ is flat as a left R-module, hence $\mathcal{M}^{\mathcal{C}}$ is a Grothendieck category in such a way that the forgetful functor $\mathcal{M}^{\mathcal{C}} \to \mathcal{M}_A$ is exact (see [4, Sec. 19]).

If $\mathcal{C} \in {}_R\mathcal{M}$ is locally projective, then for all $M \in \mathcal{M}^{\mathcal{C}}$, the lattices consisting respectively of all $\mathcal{C}$-subcomodules and of all ${}^*\mathcal{C}$-submodules of M coincide, so it

makes sense to talk about the subcomodule generated by a subset of M. ¿From the proof of [4, 19.12], we deduce the following result.

Theorem 2.1

(Finiteness Theorem) If $\mathcal{C} \in {}_R\mathcal{M}$ is locally projective, then a right $\mathcal{C}$-comodule M is finitely generated as a right $\mathcal{C}$-comodule if and only if it is finitely generated as a right R-module.

Let $\mathcal{C}$ be locally projective as a left R-module, and M a right ${}^*\mathcal{C}$-module. $\mathrm{Rat}^{\mathcal{C}}(M)$ is by definition the largest ${}^*\mathcal{C}$-submodule N of M, on which there exists a right $\mathcal{C}$-coaction ρ such that $F(N,\rho) = N$. Otherwise stated, $\mathrm{Rat}^{\mathcal{C}}$ is the preradical functor $\mathcal{T}^{\mathcal{C}}$, with $\mathcal{C}$ considered as a right ${}^*\mathcal{C}$-module. We also have that $\mathrm{Rat}^{\mathcal{C}}(M)$ consists of the elements $m \in M$ such that there exists $m_{[0]} \otimes_R m_{[1]} \in M \otimes_R \mathcal{C}$ with $m \cdot f = m_{[0]} f(m_{[1]})$, for all $f \in {}^*\mathcal{C}$. In a similar way, we define the left Rat functor ${}^{\mathcal{C}}\mathrm{Rat}$. The proof of Proposition 2.2 is straightforward, and left to the reader.

Proposition 2.2

Let $\mathcal{C}$ be an R-coring, and $M \in {}^{\mathcal{C}}\mathcal{M}$.

(i) The R-modules ${}^{\mathcal{C}}\mathrm{Hom}(M,\mathcal{C})$ and ${}^*M = {}_R\mathrm{Hom}(M,R)$ are isomorphic;

(ii) ${}^{\mathcal{C}}\mathrm{Hom}(M,\mathcal{C})$ is a right ${}^*\mathcal{C}$-module, via

$$(\varphi \cdot f)(m) = f(\varphi(m));$$

(iii) we have isomorphic functors ${}^{\mathcal{C}}\mathrm{Hom}(-,\mathcal{C})$ and ${}_R\mathrm{Hom}(-,R)$ from ${}^{\mathcal{C}}\mathcal{M}$ to $\mathcal{M}_{*\mathcal{C}}$; these functors are left exact if $\mathcal{C}$ is locally projective in $\mathcal{M}_R$, and exact if R is injective as a left R-module;

(iv) the isomorphism from (i) defines a ring isomorphism ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C}) \cong {}^*\mathcal{C}$, where the multiplication on ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C})$ is the oppositie composition;

(v) ${}^{\mathcal{C}}\mathrm{Hom}(M,\mathcal{C})$ is a right ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C})$-module, via

$$(\varphi \cdot f)(m) = f(\varphi(m)).$$

Observe that the right coactions defined in (ii) and (v) are the same after we identify ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C})$ and ${}^*\mathcal{C}$ using (iv).

Let ${}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ be the category of finitely generated left $\mathcal{C}$-comodules. If R is left noetherian, then the kernel of a morphism in ${}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ is still finitely generated, hence ${}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ has kernels (and cokernels), and is an abelian category.

Proposition 2.3

Let R be a left noetherian ring, and $\mathcal{C}$ a locally projective R-coring.

(i) For any finitely generated $M \in {}_R\mathcal{M}$, the evaluation map

$$\psi_M : {}_R\mathrm{Hom}(M,R) \otimes \mathcal{C} \to {}_R\mathrm{Hom}(M,\mathcal{C}), \quad \psi_M(f \otimes c)(m) = f(m)c$$

is an isomorphism.

(ii) Let $(M, \rho_M) \in {}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ and consider the map

$$\phi_M : {}^*M \to {}^*M \otimes_R \mathcal{C}, \quad \phi_M(f) = \psi_M^{-1}((\mathcal{C} \otimes f) \circ \rho_M)$$

Then $({}^*M, \phi_M) \in \mathcal{M}^{\mathcal{C}}$, and the associated ${}^*\mathcal{C}$-module structure is as defined in Proposition 2.2

Proof

(i) It is straightforward to prove the statement for free modules. Then we can easily show it for finitely presented modules, using the flatness of $\mathcal{C}$ over R. Since R is noetherian, every finitely presented module is finitely generated.
(ii) Take $f \in {}^*M$, and write

$$\phi_M(f) = f_{[0]} \otimes f_{[1]} \in {}^*M \otimes_R \mathcal{C}.$$

Then $m_{[-1]}f(m_{[0]}) = f_{[0]}(m)f_{[1]}$, and for very ${}^*c \in {}^*\mathcal{C}$, we find that

$$\begin{aligned}(f \cdot {}^*c)(m) &= {}^*c(m_{[-1]}f(m_{[0]})) = {}^*c(f_{[0]}(m)f_{[1]}) \\ &= f_{[0]}(m){}^*c(f_{[1]}) = (f_{[0]} \cdot {}^*c(f_{[1]}))(m)\end{aligned}$$

This shows that ${}^*M$ is a rational ${}^*\mathcal{C}$-module, and that ϕ_M is a right $\mathcal{C}$-coaction.

2.3 The Rat functor

Assume that $\mathcal{C}$ is a coring satisfying the left α-condition. Then the functor $\mathrm{Rat}^{\mathcal{C}}$ is additive and left exact.

Proposition 2.4

The following assertions are equivalent.

(i) $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ is dense in ${}^*\mathcal{C}$ in the $\mathcal{C}$-adic topology;

(ii) $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ is dense in ${}^*\mathcal{C}$ in the finite topology;

(iii) $\mathrm{Rat}^{\mathcal{C}}$ is an exact functor.

Proof

The equivalence of (i) and (iii) follows from Proposition 1.1, invoking the fact that $\mathcal{C}$ is faithful as a right ${}^*\mathcal{C}$-module.
Note that the sets

$$\mathcal{O}_a(F) = \{{}^*c \mid c \cdot {}^*c = 0, \text{ for all } c \in F\},$$

with $F \subset \mathcal{C}$ finite, form a basis of open neighborhoods of $0 \in \mathcal{C}$ in the $\mathcal{C}$-adic topology, which is a linear topology. Also

$$\mathcal{O}_f(F) = \{{}^*c \mid {}^*c(c) = 0, \text{ for all } c \in F\},$$

with $F \subset \mathcal{C}$ finite, form a basis of open neighborhoods of 0 for the finite topology, which is also linear.
Let $F \subset \mathcal{C}$ be finite. For each $c \in F$, we fix a tensor representation of $\Delta(c)$, and then consider the finite set F' of all second tensor components. Then we easily see that

$$\mathcal{O}_f(F') \subseteq \mathcal{O}_a(F) \subseteq \mathcal{O}_f(F)$$

and it follows that the two linear topologies on ${}^*\mathcal{C}$ coincide, so it follows that (i) is equivalent to (ii).

Proposition 2.5

Suppose we have a decomposition $\mathcal{C} = \bigoplus_{i\in I} C_i$ as left $\mathcal{C}$-comodules. Then $\mathrm{Rat}^{\mathcal{C}}({}^*C_i)$ is dense in ${}^*C_i$ for all $i \in I$ if and only if $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ is dense in ${}^*\mathcal{C}$.

Proof

Assume that each $\mathrm{Rat}^{\mathcal{C}}({}^*C_i)$ is dense in ${}^*C_i$. It follows from Proposition 1.3 that $\bigoplus_{i\in I} \mathrm{Rat}^{\mathcal{C}}({}^*C_i)$ is dense in ${}^*C$, and then $\mathrm{Rat}^{\mathcal{C}}({}^*C) \supset \bigoplus_{i\in I} \mathrm{Rat}^{\mathcal{C}}({}^*C_i)$ is also dense.

Conversely, let $M = \bigoplus_{j\in I, j\neq i} C_j$, for each $i \in I$. Then $C = C_i \oplus M$ and ${}^*C = {}^*C_i \oplus {}^*M$, hence $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C}) = \mathrm{Rat}^{\mathcal{C}}({}^*C_i) \oplus Rat^{\mathcal{C}}({}^*M)$ is dense in ${}^*C = {}^*C_i \oplus {}^*M$ ($\mathrm{Rat}^{\mathcal{C}}$ is an aditive functor). The result then follows from Proposition 1.3 (ii).

Lemma 2.6

(i) Assume that $M \in {}^{\mathcal{C}}\mathcal{M}$ is finitely generated and projective as a left R-module. Then ${}^*M$ is a rational right ${}^*\mathcal{C}$-module.

(ii) Suppose that $\mathcal{C} = M \oplus N$ in ${}^{\mathcal{C}}\mathcal{M}$. Then ${}^*M$ is rational if and only if M is finitely generated as a left R-module.

Proof

(i) We take a finite dual basis $\{(x^i, f_i) \mid i = 1, \cdots, n\}$ of $M \in {}_R\mathcal{M}$. For all $h \in {}^*M$ and $\alpha \in {}^*\mathcal{C}$, we have

$$h \cdot \alpha = \sum_i f_i \cdot (h \cdot \alpha)(x^i) = \sum_i f_i \alpha(x^i_{[-1]} h(x^i_{[0]}))$$

This shows that $h_{[0]} \otimes h_{[1]} = f_i \otimes x^i_{[-1]} h(x^i_{[0]}) \in {}^*M \otimes \mathcal{C}$ is such that $h \cdot \alpha = h_{[0]} \alpha(h_{[1]})$, and this proves that ${}^*M$ is rational.
(ii) One direction follows from (i). Conversely, assume that ${}^*M$ is rational. Take $e = \varepsilon_{|M} \in {}^*M$. We can identify ${}^*\mathcal{C} = {}^*M \oplus {}^*N$ as right ${}^*\mathcal{C}$ modules. For $h \in {}^*M$ and $c \in \mathcal{C}$, $(e \cdot h)(c) = h(c_{(1)} e(c_{(2)})) = h(c_{(1)} \varepsilon(c_{(2)})) = h(c)$ if $c \in M$ ($c_{(1)} \otimes c_{(2)} \in C \otimes M$) and $(e \cdot h)(c) = h(c_{(1)} e(c_{(2)})) = 0$ if $c \in N$ ($c_{(1)} \otimes c_{(2)} \in C \otimes N$) showing that $e \cdot h = h$ (the h in the $e \cdot h$ is regarded as belonging to ${}^*\mathcal{C}$). As ${}^*M$ is rational there is $\sum_i f_i \otimes x^i \in {}^*M \otimes \mathcal{C}$ such that $e \cdot \alpha = \sum_i f_i \alpha(x^i)$, for all $\alpha \in {}^*\mathcal{C}$. Then for any $h \in {}^*M$, $h = e \cdot h = \sum_i f_i h(x^i)$, and, for all $m \in M$, we have $h(m) = \sum_i f_i(m) h(x^i) = h(\sum_i f_i(m) x^i) = h(\sum_i f_i(m) m^i)$, where $x^i = m^i + n^i \in M \oplus N$ is the unique representation of x^i in the direct sum $\mathcal{C} = M \oplus N$ and the last equality holds as $h_{|N} = 0$. As this last equality holds for all $h \in {}^*M$, we can easily see that it actually holds for all $\alpha = (h, g) \in {}^*\mathcal{C} = {}^*M \oplus {}^*N$ because $m \in M$, and so we now obtain, using the left α-condition on ${}^*\mathcal{C}$, that $m = f_i(m) m^i$, where $m \in M$ is arbitrary and $m^i \in M$ are fixed. Thus M is finitely generated.

Corollary 2.7

Assume that $\mathcal{C} = \bigoplus_{i\in I} C_i$ as left $\mathcal{C}$-comodules, and that each C_i is finitely generated. Then $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ is dense in ${}^*\mathcal{C}$, and, equivalently, $\mathrm{Rat}^{\mathcal{C}}$ is an exact functor.

Proof

This is a direct consequence of Proposition 1.3 (ii) and Lemma 2.6.

Example 2.8

We now present an example of a coring for which we can explicitly construct the Rat functor. Let G be a group, k a commutative ring, and R a G-graded k-algebra. It is well-known that $\mathcal{C} = R \otimes kG$ is an R-coring. The structure maps are given by the formulas

$$r(s \otimes \sigma)t = \sum_{\rho \in G} rst_\rho \otimes \sigma\rho;$$

$$\Delta_{\mathcal{C}}(s \otimes \sigma) = (s \otimes \sigma) \otimes_R (1 \otimes \sigma) \quad ; \quad \varepsilon(s \otimes \sigma) = s.$$

Here t_ρ is the homogeneous part of degree ρ of t. Clearly $\mathcal{C} = \bigoplus_{\sigma \in G} R \otimes \sigma$ decomposes as the direct sum of finitely generated (free of rank one) left $\mathcal{C}$-comodules, hence it follows from Corollary 2.7 that Rat is exact. We will illustrate this, computing Rat. First observe that

$${}^*\mathcal{C} = {}_R\mathrm{Hom}(R \otimes kG, R) \cong \mathrm{Hom}(kG, R) \cong \mathrm{Map}(G, R).$$

The multiplication on ${}^*\mathcal{C}$ can be transported into a multiplication on $\mathrm{Map}(G, R)$. This multiplication is the following. For $f, g : G \to R$ and $\tau \in G$:

$$(f\#g)(\tau) = \sum_\rho f(\tau)_\rho g(\tau\rho) \tag{1}$$

Let $(kG)^*$ be the dual of the group algebra kG, with free basis $\{v_\sigma \mid \sigma \in G\}$, such that $v_\sigma(\tau) = \delta_{\sigma,\tau}$. then v_σ can also be viewed as a map $G \to R$, and this gives us an algebra embedding $(kG)^* \subset \mathrm{Map}(G, R)$. Indeed, using (1), we easily compute that $v_\sigma \# v_\tau = \delta_{\sigma,\tau} v_\sigma$.
We also have an algebra embedding

$$\iota : R \to \mathrm{Map}(G, R), \quad \iota_r(\sigma) = r.$$

Indeed, using (1), we find

$$(\iota_r \# \iota_s)(\tau) = \sum_\rho \iota_r(\tau)_\rho \iota_s(\tau\rho) = \sum_\rho r_\rho s = rs = \iota_{rs}(\tau).$$

Let $r \in R$ be homogeneous of degree ρ, and $f : G \to R$. Using (1), we compute

$$v_\sigma \# \iota_r = \iota_r \# v_{\sigma\rho} \quad \text{and} \quad v_\sigma \# f = v_\sigma \# \iota_{f(\sigma)}. \tag{2}$$

Now take $M \in \mathcal{M}_{^*\mathcal{C}} \cong \mathcal{M}_{\mathrm{Map}(G,R)}$. By restriction of scalars, M is also a right R-module and a right $(kG)^*$-module. Now put $M_\sigma = M \cdot v_\sigma$.
1) If $\sigma \neq \tau$, then $M_\sigma \cap M_\tau = 0$. Indeed, if $m \cdot v_\sigma = n \cdot v_\tau$, then

$$m \cdot v_\sigma = m \cdot (v_\sigma \# v_\sigma) = (m \cdot v_\sigma) \cdot v_\sigma = (n \cdot v_\tau) \cdot v_\sigma = n \cdot (v_\tau \# v_\sigma) = 0.$$

2) $M_\sigma R_\rho \subset M_{\sigma\rho}$. Take $m \cdot v_\sigma \in M_\sigma$ and $r \in R_\rho$. Using (2), we find

$$(m \cdot v_\sigma)r = m \cdot (v_\sigma \# \iota_r) = m \cdot (\iota_r \# v_{\sigma\rho}) = (mr) \cdot v_{\sigma\rho} \in M_{\sigma\rho}.$$

This shows that $\bigoplus_{\sigma\in G} M_\sigma$ is a G-graded R-module; we will show that it is the rational part of M.
3) $M_\sigma \subset \mathrm{Rat}(M)$. Take $m \cdot v_\sigma \in M_\sigma$ and $f \in \mathrm{Map}(G, R)$. Using (2), we find

$$(m \cdot v_\sigma) \cdot f = m \cdot (v_\sigma \# f) = m \cdot (v_\sigma \# \iota_{f(\sigma)}) = (m \cdot v_\sigma) f(\sigma),$$

so $m \cdot v_\sigma$ is rational.
4) It follows from 3) that $\bigoplus_{\sigma\in G} M_\sigma \subseteq \mathrm{Rat}(M)$.
5) Let $m \in \mathrm{Rat}(M)$. Then there exist $m_1, \cdots, m_n \in M$, $r_1, \cdots, r_n \in R$ and $\sigma_1, \cdots, \sigma_n \in G$ such that, for all $\varphi \in {}^*\mathcal{C}$:

$$m \cdot \varphi = \sum_i m_i \varphi(r_i \otimes \sigma_i).$$

Making the identification ${}^*\mathcal{C} \cong \mathrm{Map}(G, R)$, we find for all $f : G \to R$:

$$m \cdot f = \sum_i m_i r_i f(\sigma_i).$$

Replacing m_i by $m_i r_i$, it is no restriction to take $r_i = 1$. We can also take the σ_i pairwise different. Taking $f = v_\sigma$, we find that

$$m_\sigma = \sum_i m_i \delta_{\sigma,\sigma_i}$$

so $m_\sigma \neq 0$ for only a finite number of σ, and $m_{\sigma_i} = m_i$. Finally

$$m = m \cdot \iota_1 = \sum_i m_i \iota_1(\sigma_i) = \sum_i m_i = \sum_i m_{\sigma_i} \in \bigoplus_{\sigma\in G} M_\sigma.$$

We conclude that

$$\mathrm{Rat}(M) = \bigoplus_{\sigma\in G} M \cdot v_\sigma,$$

and it is clear that Rat is exact.

In some situations, the converse of Corollary 2.7 also holds. If R is left artinian, then any left comodule contains a simple comodule. The same holds for comodules that are locally artinian, in the sense that any finitely generated submodule is artinian. If this is the case for $\mathcal{C}$, then the left socle of C is essential in $\mathcal{C}$. If moreover $\mathcal{C}$ is injective in ${}^{\mathcal{C}}\mathcal{M}$, then a decomposition $\mathcal{C} = \bigoplus_{i\in I} E(S_i)$ holds with usual arguments, where $\bigoplus_{i\in I} S_i = {}^{\mathcal{C}}s(\mathcal{C})$ is a decomposition of the left socle ${}^{\mathcal{C}}s(\mathcal{C})$ of $\mathcal{C}$ and $E(S_i)$ is the injective hull of S_i contained in $\mathcal{C}$. We will assume that $\mathcal{C}$ is locally projective as a right R-module, which implies that ${}^{\mathcal{C}}\mathcal{M}$ is abelian, so that we have a categorical definition of injective hulls.

Proposition 2.9

Assume that $\mathcal{C}$ also satisfies the right α-condition, and that the two following conditions hold:

1. $\mathcal{C}$ is an injective object of ${}^{\mathcal{C}}\mathcal{M}$;
2. R is left artinian or $\mathcal{C}$ is locally artinian in ${}_R\mathcal{M}$ (equivalently in ${}^{\mathcal{C}}\mathcal{M}$).

Let $\bigoplus_{i\in I} S_i$ be the decomposition of the left socle of $\mathcal{C} \in {}^{\mathcal{C}}\mathcal{M}$ into simple left $\mathcal{C}$-comodules, and $E(S_i)$ an injective envelope of S_i contained in $\mathcal{C}$. Then $\mathrm{Rat}^{\mathcal{C}}$ is exact if and only if each $E(S_i)$ is finitely generated.

Proof

We have that $\mathcal{C} = \bigoplus_{i\in I} E(S_i)$, so one direction follows from Corollary 2.7. Conversely, assume that $\mathrm{Rat}^{\mathcal{C}}$ is exact, and let S be a simple subcomodule of C, and $E(S)$ an injective envelope of S contained in C. Then there is a left subcomodule X of $\mathcal{C}$ such that $E(S) \oplus X = \mathcal{C}$ in ${}^{\mathcal{C}}\mathcal{M}$. The functor ${}^{\mathcal{C}}\mathrm{Hom}(-,\mathcal{C})$ is exact since $\mathcal{C} \in \mathcal{M}^{\mathcal{C}}$ is injective, and the composition of ${}^{\mathcal{C}}\mathrm{Hom}(-,\mathcal{C})$ with the natural functor $\mathcal{M}^{\mathcal{C}} \to \mathcal{M}_{*\mathcal{C}}$ is also exact. Thus we obtain an epimorphism $\pi : {}^*E(S) \to {}^*S$, with kernel ${}^{\perp}S = \{f \in {}^*E(S) \mid f_{|S} = 0\}$.
We will first show that ${}^{\perp}S \ll {}^*E(S)$. Using the isomorphisms in Proposition 2.2, we can regard π as a left ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C})$-module morphism ${}^{\mathcal{C}}\mathrm{Hom}(E(S),\mathcal{C}) \to {}^{\mathcal{C}}\mathrm{Hom}(S,\mathcal{C})$.
Take $f \in {}^{\mathcal{C}}\mathrm{Hom}(E(S),\mathcal{C}) \setminus {}^{\perp}S$, i.e. $f : E(S) \to \mathcal{C}$ such that $f_{|S} \neq 0$. Then $\mathrm{Ker}\, f \cap S = 0$ since S is simple, and therefore $\mathrm{Ker}\, f = 0$, since S is essential in $E(S)$. So $E(S) \cong f(E(S))$, and there exists a left $\mathcal{C}$-subcomodule M of $\mathcal{C}$ such that $\mathcal{C} \cong f(E(S)) \oplus M$. We can extend f to a left $\mathcal{C}$-comodule isomorphism $\overline{f} : \mathcal{C} \to \mathcal{C}$, since $X \cong M$. Let h be the inverse of $\overline{f}$. Take an arbitrary $g \in {}^{\mathcal{C}}\mathrm{Hom}(E(S),\mathcal{C})$,and extend

g to $\overline{g}:\ \mathcal{C} = E(S) \oplus X \to \mathcal{C}$ by putting $\overline{g}_{|X} = 0$. Then $\overline{g} = \overline{g} \circ h \circ \overline{f}$, which means that ${}^{\mathcal{C}}\mathrm{Hom}(E(S), \mathcal{C})$ is generated by $\overline{f}$ as a left ${}^{\mathcal{C}}\mathrm{End}(\mathcal{C})$-module. Consequently ${}^{\perp}S \ll {}^{*}E(S)$.
The Finiteness Theorem 2.1 shows that S is finitely generated and then it follows from Proposition 2.3 (ii) that ${}^{*}S$ is a rational ${}^{*}C$-comodule, so $\mathrm{Rat}^{\mathcal{C}}({}^{*}S) = {}^{*}S$. $\mathrm{Rat}^{\mathcal{C}}$ is exact, so we have an exact sequence

$$0 \longrightarrow \mathrm{Rat}^{\mathcal{C}}({}^{\perp}S) \longrightarrow \mathrm{Rat}^{\mathcal{C}}({}^{*}E(S)) \xrightarrow{\pi} \mathrm{Rat}^{\mathcal{C}}({}^{*}S) = {}^{*}S \longrightarrow 0.$$

We obtain $\pi(\mathrm{Rat}^{\mathcal{C}}({}^{*}E(S))) = {}^{*}S$, so ${}^{\perp}S + \mathrm{Rat}^{\mathcal{C}}({}^{*}E(S)) = {}^{*}E(S)$. It then follows that ${}^{*}E(S)$ is rational. This last part can also be seen as follows. We have an exact sequence

$$0 \longrightarrow {}^{\perp}S \longrightarrow {}^{*}E(S) \longrightarrow {}^{*}S \longrightarrow 0,$$

with ${}^{\perp}S \ll {}^{*}E(S)$ and ${}^{*}S$ rational, so ${}^{*}E(S)$ is rational by Proposition 1.2 (i). Using Lemma 2.6, we find that ${}_{R}E(S)$ is finitely generated.

3. SEMIPERFECT CORINGS

Let $\mathcal{C}$ be an abelian category. A projective object $P \in \mathcal{C}$ together with a superfluous epimorphism $P \to M$ is called a projective cover of M. $\mathcal{C}$ is called semiperfect if every simple object has a projective cover. If a coring $\mathcal{C}$ satisfies the left α-condition, then $\mathcal{M}^{\mathcal{C}}$ is an abelian category, and $\mathcal{C}$ is called right semiperfect if $\mathcal{M}^{\mathcal{C}}$ is semiperfect. Semiperfect corings were introduced first in [10].

Theorem 3.1

Let R be a right artinian ring, and $\mathcal{C}$ an R-coring satisfying the left α-condition. The following statements are equivalent.

(i) $\mathcal{C}$ is right semiperfect;

(ii) Every finitely generated right comodule has a projective cover;

(iii) every finitely generated right comodule has a finitely generated projective cover;

(iv) the category $\mathcal{M}^{\mathcal{C}}$ has enough projectives;

(v) every simple right comodule has a finitely generated projective cover;

(vi) the category $\mathcal{M}^{\mathcal{C}}$ has a progenerator (=projective generator).

Proof

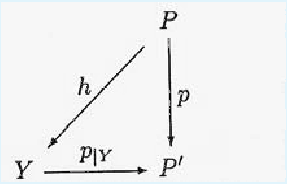
(i)⇒(ii). First notice that an R-module is finitely generated if and only if it has finite length. Every finitely generated comodule M has a maximal subcomodule, so its Jacobson radical $J(M)$ in $\mathcal{M}^{\mathcal{C}}$ is different from the comodule itself. $J(M) \ll M$, and $M/J(M)$ is a semisimple finitely generated comodule. Every simple component of $M/J(M)$ has a projective cover, and the direct sum of all these projective covers is a projective cover $f : P \to M/J(M)$ of $M/J(M)$. Since P is projective, there exists $g : P \to M$ such that $u \circ g = f$, with $u : M \to M/J(M)$ the canonical projection. Then a usual argument shows that $g : P \to M$ is a projective cover: $u(g(P)) = f(P) = M/J(M)$, hence $u(J(M) + g(P)) = M/J(M)$ and it follows that $J(M) + g(P) = M$. From the fact that $J(M)$ is small in M, it follows that $g(P) = M$ and g is surjective. Finally $\mathrm{Ker}\, g \subset \mathrm{Ker}\, f \ll P$, so $\mathrm{Ker}\, g \ll P$, and $g : P \to M$ is a projective cover of M.

(iv)⇒(iii). Let M be a finitely generated comodule. We know that there exists a projective object $P \in \mathcal{M}^{\mathcal{C}}$ and a $\mathcal{C}$-colinear epimorphism $f : P \to M$. Let $(M_i)_{i\in I}$ be a family of finitely generated comodules such that we have a $\mathcal{C}$-colinear epimorphism $f : \bigoplus_{i\in I} M_i \to P$. As P is projective, we have that $\bigoplus_{i\in I} M_i \cong P \oplus X$ as comodules. Since R is artinian, we can assume that the M_i are indecomposable. As they have finite length in $\mathcal{M}_R$, they also have finite length in $\mathcal{M}^{\mathcal{C}}$ and $\mathcal{M}_{^*\mathcal{C}}$, so their $^*\mathcal{C}$-endomorphism rings are local, by the Krull-Schmidt Theorem (see [2, 12.8]). It then follows from the Crawley-Jønsson-Warfield Theorem (see [2, 26.5]) that $P \cong \bigoplus_{i\in J} M_i$, with $J \subset I$. The M_i are finitely generated (rational) $^*\mathcal{C}$-modules, and are projective objects of $\mathcal{M}^{\mathcal{C}}$, since they are direct summands of P. Since M is finitely generated, we can find a finite $F \subset J$ and a projection $\bigoplus_{i\in F} M_i \to M$, induced by f. Thus we have found a finitely generated projective object $P \in {}^{\mathcal{C}}\mathcal{M}$ and a $\mathcal{C}$-colinear epimorphism $f : P \to M$. Dualizing the proof of the Eckmann-Schopf Theorem on the existence of the injective envelope of a module, see e.g. [2, 18.10], we can show that M has a projective cover. This works as follows.
• Let $K = \mathrm{Ker}\, f$, and consider the set V consisting of subcomodules $H \subset K$ such that $K/H \ll P/H$, which is equivalent to

$$H \subset T \subset P, K + T = P \implies T = P$$

$V \neq \emptyset$ since $K \in V$. V contains a minimal element K' since R is artinian.
• Then consider the set W consisting of subcomodules $Y \subset P$ such that $K' + Y = P$. This set is nonempty, since P belongs to it. Then take an element in this set such that $K' \cap Y$ is minimal. Let $p : P \to P' = P/K'$ be the projection. Since P is

projective, there exists a comodule morphism $h: P \to Y$ such that $p_{|Y} \circ h = p$, that is, the following diagram commutes:



We will now show that $p_{|Y}$ is an isomorphism.

• h is surjective. Take $y \in Y$. Then

$$p(y - h(y)) = p(y) - p(h(y)) = p(y) - p(y) = 0$$

so $y - h(y) \in K'$ and

$$y = (y - h(y)) + h(y) \in (Y \cap K') + \operatorname{Im} h.$$

It follows that $Y \subset (Y \cap K') + \operatorname{Im} h$. The converse implication is obvious, so

$$Y = (Y \cap K') + \operatorname{Im} h$$

It then follows that

$$P = Y + K' = (Y \cap K') + \operatorname{Im} h + K' = \operatorname{Im} h + K'$$

The minimality condition on Y then yields that $Y = \operatorname{Im} h$, so h is surjective.

• $Y \cap K' \ll Y$. If $H \subset Y$ and $H + (Y \cap K') = Y$, then $H + K' = H + (Y \cap K') + K' = Y + K' = P$. This means that $H \in W$, and the minimality condition on Y gives us that $H \cap K' \supset Y \cap K'$, and $H \cap K' \subset Y \cap K'$ since $H \subset Y$. Then we find that $Y = H + (Y \cap K') = H + (H \cap K') = H$, as needed.

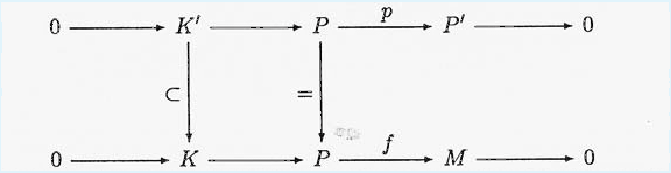
• From the fact that $0 = p(K') = (p \circ h)(K')$, it follows that $h(K') \subset \operatorname{Ker}(p_{|Y}) = Y \cap K'$.

• $\operatorname{Ker} h = K'$. It is clear that $\operatorname{Ker} h \subset K'$. It follows that $K' \subset \operatorname{Ker} h$ if we can show that $\operatorname{Ker} h \in V$, or

$$\operatorname{Ker} h \subset T \subset P, K + T = P \quad \Rightarrow \quad T = P$$

Assume $\operatorname{Ker} h \subset T \subset P$ and $K + T = P$. Since $K' \subset P$, we find that $K + K' + T = P$. Also $K' \subset T + K' \subset P$, so it follows from the fact that $K' \in V$ that $K' + T = P$.Then $h(K') + h(T) = h(P) = Y$, since K is surjective. Since $h(K') \subset Y \cap K'$, this implies

that $Y \cap K' + h(T) = Y$, hence $h(T) = Y$, since $Y \cap K' \ll Y$, and finally $T = P$ because $T \subset \operatorname{Ker} h$.
• $p_{|Y}$ is surjective, as $p = p_{|Y} \circ h$ and p is an epimorphism.
• $p_{|Y}$ is injective. Take $y \in Y$ such that $p(y) = 0$. h is surjective, so $y = h(z)$. Then $0 = p(y) = p(h(z)) = p(z)$, so $z \in K' = \ker h$, and $y = h(z) = 0$.
• It now follows that $Y \cap K' = 0$. We know from the definition of Y that $Y + K' = P$. Hence $Y \oplus K' = P$, and $P' \cong Y$ is finitely generated projective, being a direct factor of P. Now look at the commutative diagram
It follows that we have an epimorphism $P' \to M$ in $\mathcal{M}^{\mathcal{C}}$, with kernel K/K'. This



is a projective cover, since $K/K' \ll P' = P/K'$. Moreover, P' is finitely generated as a quotient of P.

(ii)⇒(vi). Take a family $(M_i)_{i\in I}$ consisting of finitely generated comodules that generate $\mathcal{M}^{\mathcal{C}}$. Let $P_i \to M_i$ be a projective cover of M_i. Then $\bigoplus_{i\in I} P_i$ is a projective generator of $\mathcal{M}^{\mathcal{C}}$.

(vi)⇒(iv), (iii)⇒(ii)⇒(i) and (iii)⇒(v)⇒(i) are obvious.

Proposition 3.2 Let R be a right artinian ring, and $\mathcal{C}$ an R-coring satisfying the left α-condition.

(i) $\mathcal{M}^{\mathrm{fg}\mathcal{C}}$ is an abelian category;

(ii) $Q \in \mathcal{M}^{\mathrm{fg}\mathcal{C}}$ is injective if and only if Q is an injective object in $\mathcal{M}^{\mathcal{C}}$;

(iii) $P \in \mathcal{M}^{\mathrm{fg}\mathcal{C}}$ is projective if and only if P is a projective object in $\mathcal{M}^{\mathcal{C}}$.

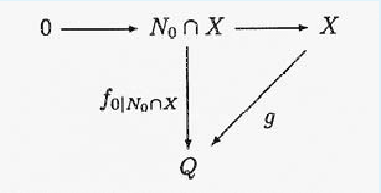
Proof (i) The fact that $\mathcal{M}^{\mathrm{fg}\mathcal{C}}$ has kernels follows from the assumption that R is right artinian and the Finiteness Theorem.

(ii) This is a straightforward adaptation of the corresponding result on comodules over a coalgebra. Let $u: N \to M$ be a monomorphism in $\mathcal{M}^{\mathcal{C}}$ and $f: N \to Q$.

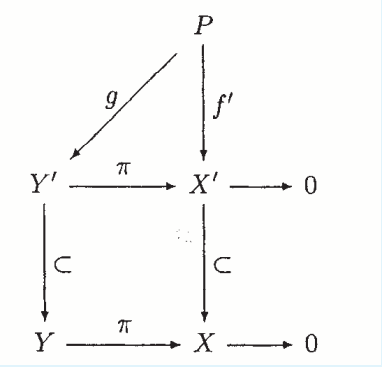
Consider the set

$$X = \{(N', f') \mid N \subset N' \subset M,\ f' :\ N' \to Q,\ f'_{|N} = f\}$$

ordered by the relation $(N', f') < (N'', f'')$ if $N' \subset N''$ and $f''_{N'} = f'$. Take a maximal element (N_0, f_0) in X, and assume that $N_0 \neq M$. Take $m \in M \setminus N_0$ and X the subcomodule of M generated by M. By the Finiteness Theorem for comodules, X is finitely generated, so there exists $g :\ X \to Q$ such that the following diagram commutes:



Then consider the map $f' :\ N' = N_0 + X \to Q$, defined by $f'(n_0 + x) = f_0(n_0) + g(x)$. The usual computation shows that f' is well-defined, and (N', f') is an element in X that is strictly greater than (N_0, f_0), a contradiction.



(3)

(iii) Let $\pi :\ Y \to X$ and $f :\ P \to X$ be morphisms in $\mathcal{M}^{\mathcal{C}}$, with π surjective. Let $\{p_1, \cdots, p_n\}$ be a set of generators of P as an R-module (and a fortiori as a $\mathcal{C}$-comodule). Then $X' = \mathrm{Im}\, f$ is generated by $\{x_1, \cdots, x_n\}$, with $x_i = f(p_i)$. Take $y_i \in Y_i$ such that $\pi(y_i) = x_i$, and let Y' be the $\mathcal{C}$-submodule (or ${}^*\mathcal{C}$-submodule) of Y generated by $\{y_1, \cdots, y_n\}$. Let $f' :\ P \to X'$ be the corestriction of f. Since X' and Y' are finitely generated and $\pi_{|Y'}$ is still an epimorphism, there exists $g :\ P \to X'$ such that $f' = \pi \circ g$, and the projectivity of P in $\mathcal{M}^{\mathcal{C}}$ follows from the commutativity of the diagram (3).

4. APPLICATIONS AND EXAMPLES

4.1 Application to qF-rings In Theorem 3.1, we gave equivalent conditions for the semiperfectness of a left locally projective coring $\mathcal{C}$ over a right artinian ring R. In the case where R is a qF-ring, more characterizations are possible. This has been studied recently by El Kaoutit and Gómex-Torrecillas (see [10, Theorems 3.5, 3.8, 4.2]. Using the results of the previous Sections, we find a different proof of these results.

First recall that a qF ring, or quasi-Frobenius ring, is a ring which is right artinian and injective as a right R-module, or, equivalently, left artinian and injective as a left R-module (in [15], these rings are called noetherian QF rings). In this situation, R is a cogenerator of $\mathcal{M}_R$ and ${}_R\mathcal{M}$, see [15, 48.15]. Since a qF-ring is a left and right perfect ring, local projectivity is equivalent to projectivity. Also recall that flat modules over qF-rings are projective. Let R be a qF-ring, and assume that $\mathcal{C} \in {}_R\mathcal{M}$ is flat (or, equivalently, (locally) projective). Then ${}^{\mathcal{C}}\mathcal{M}$ is a Grothendieck category, and the forgetful functor ${}^{\mathcal{C}}\mathcal{M} \to {}_R\mathcal{M}$ is exact and has a right adjoint $\mathcal{C} \otimes_R -$. Since ${}_R\mathcal{M}$ has enough injectives and the forgetful functor is exact, $\mathcal{C} \otimes_R -$ preserves injectives. Now $R \in {}_R\mathcal{M}$ is injective because R is a qF-ring, so $\mathcal{C} = \mathcal{C} \otimes_R R$ is an injective object of ${}^{\mathcal{C}}\mathcal{M}$, and we can apply Proposition 2.9. We find that $\mathcal{C} = \bigoplus_{i\in I} E(S_i)$, with $\bigoplus_{i\in I} S_i$ the decomposition of the left socle of $\mathcal{C} \in {}^{\mathcal{C}}\mathcal{M}$.

If R is a qF-ring, then the contravariant functors

$$(-)^* = \mathrm{Hom}_R(-, R) : \ \mathcal{M}_R \to {}_R\mathcal{M}, \ \ {}^*(-) = {}_R\mathrm{Hom}(-, R) : \ {}_R\mathcal{M} \to \mathcal{M}_R,$$

define an equivalence duality between the categories of finitely generated left R-modules and finitely generated right R-modules. More explicitely, every finitely generated left R-module M is reflexive, that is, the map

$$\Phi_M : \ M \to ({}^*M)^*, \ \ \Phi_M(m)(f) = f(m)$$

is an isomorphism. This result follows, for example, after we take $U = M = R$ in [15, 47.13(2)].

If M is not finitely generated, then we still have the following result.

Lemma 4.1

Let R be a qF ring and $M \in {}_R\mathcal{M}$-module. Then $\mathrm{Im}\,(\Phi_M)$ is dense in $({}^*M)^*$ with respect to the finite topology on $\mathrm{Hom}_R({}^*M, R)$.

Proof Take $T \in ({}^*M)^*$ and $F = \{f_1, \ldots, f_n\} \subset {}^*M$. We have to prove that there exists an $m \in M$ such that $T(f_i) = \Phi_M(f_i) = f_i(m)$. Let ${}^{\perp}F = \bigcap_{i=1}^n \mathrm{Ker}\, f_i \subset M$

and $N = M/^{\perp}F$. Then we have a natural inclusion

$$\frac{M}{\bigcap_{i=\overline{1,n}} \mathrm{Ker}\ f_i} \hookrightarrow \bigoplus_{i=1}^{n} \frac{M}{\mathrm{Ker}\ f_i} \simeq \bigoplus_{i=1}^{n} \mathrm{Im}\ f_i \hookrightarrow R^n$$

and this shows that $N = M/^{\perp}F$ has finite length. Let $\pi : M \longrightarrow M/^{\perp}F = N$ be the canonical projection and consider its dual $\pi^* : {}^*N \longrightarrow {}^*M$. By the construction of N as a factor module, there are left R-linear maps $\overline{f_i} : N \longrightarrow R$ such that $\overline{f_i} \circ \pi = f_i$. Consider $t = T \circ \pi^* \in ({}^*N)^*$. As N is finitely generated, Φ_N is an isomorphism (it gives the above stated duality between ${}_R\mathcal{M}$ and $\mathcal{M}_R$), so there is $n = \hat{m} = \pi(m) \in N$ such that $t = \Phi_N(n)$. Then $T(f_i) = T(\overline{f_i} \circ \pi) = (T \circ \pi^*)(\overline{f_i}) = t(\overline{f_i}) = \Phi_N(n)(\overline{f_i}) = \overline{f_i}(\pi(m)) = f_i(m)$, as needed.

If $\mathcal{C}$ is a left and right projective R-coring, then the duality is kept after we pass to the categories of finitely generated $\mathcal{C}$-comodules: he functors ${}^*(-) = {}_R\mathrm{Hom}(-, R)$ and $(-)^* = \mathrm{Hom}_R(-, R)$ define an equivalence between the categories ${}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ and $\mathcal{M}^{\mathrm{fg}\mathcal{C}}$. To prove this, it suffices to show that Φ_M is left $\mathcal{C}$-colinear, or, equivalently, right $\mathcal{C}^*$-linear, for every finitely generated left $\mathcal{C}$-comodule M, and this is a standard computation. From this duality and Proposition 3.2, we obtain the following result.

Corollary 4.2

Let R be a qF-ring, and $\mathcal{C}$ an R-coring that is projective as a left and right R-module. A finitely generated right $\mathcal{C}$-comodule M is injective (resp. projective) in $\mathcal{M}^{\mathcal{C}}$ if and only if M^* is projective (resp. injective) in ${}^{\mathcal{C}}\mathcal{M}$.

Theorem 4.3

Let R be a qF-ring, and $\mathcal{C}$ an R-coring that is (locally) projective as a left and right R-module. The following assertions are equivalent.

(i) $\mathrm{Rat}^{\mathcal{C}}$ is exact;

(ii) $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ is dense in ${}^*\mathcal{C}$;

(iii) $\mathrm{Rat}^{\mathcal{C}}({}^*M)$ is dense in ${}^*M$ for every left $\mathcal{C}$-comodule M;

(iv) $\mathrm{Rat}^{\mathcal{C}}({}^*Q)$ is dense in ${}^*Q$ for every left injective $\mathcal{C}$-comodule Q;

(v) $\mathrm{Rat}^{\mathcal{C}}({}^*Q)$ is dense in ${}^*Q$ for every left injective indecomposable $\mathcal{C}$-comodule Q;

(vi) ${}^*Q$ is ${}^*\mathcal{C}$-rational for every left injective indecomposable $\mathcal{C}$-comodule Q;

(vii) $E(S)$ is finitely generated for every simple left comodule S;

(viii) every simple right $\mathcal{C}$-comodule has a finitely generated projective cover;

(ix) $\mathcal{C}$ is right semiperfect.

Proof (i)$\Longleftrightarrow$(ii) follows from Proposition 2.4.
(ii)$\Longleftrightarrow$(v). As we have seen, $\mathcal{C} = \bigoplus_{i\in I} E(S_i)$, and each injective indecomposable left $\mathcal{C}$-comodule is isomorphic to one of the $E(S_i)$'s, because every comodule contains a simple comodule. The equivalence of (ii) and (v) then follows from Proposition 2.5.
(v)$\Longrightarrow$(iv). Every left injective comodule Q is a direct sum of injective indecomposable left $\mathcal{C}$-comodules (because its socle is essential), $Q = \bigoplus_{i\in I} Q_i$. Then we have ${}^*Q = \prod_{i\in I} {}^*Q_i$ in $\mathcal{M}_{{}^*C}$ and $\bigoplus_{i\in I} \mathrm{Rat}^{\mathcal{C}}({}^*Q_i) \subseteq \mathrm{Rat}^{\mathcal{C}}({}^*Q) \subset \prod_{i\in I} {}^*Q_i$ and then it all follows from Proposition 1.3.
(iv)$\Longrightarrow$(iii). Take $M \in {}^{\mathcal{C}}\mathcal{M}$ and an injective envelope $f : M \to Q$ in ${}^{\mathcal{C}}\mathcal{M}$. We know that $\mathrm{Rat}^{\mathcal{C}}({}^*Q)$ is dense in ${}^*Q = {}_R\mathrm{Hom}(Q, R)$. Proposition 1.5 then yields that ${}^*f(\mathrm{Rat}^{\mathcal{C}}({}^*Q))$ is dense in ${}_R\mathrm{Hom}(M, R) = {}^*M$. But ${}^*f(\mathrm{Rat}^{\mathcal{C}}({}^*Q)) \subset \mathrm{Rat}^{\mathcal{C}}({}^*M)$, so $\mathrm{Rat}^{\mathcal{C}}({}^*M)$ is dense in ${}^*M$.
(iii)$\Longrightarrow$(iv)$\Longrightarrow$(v): trivial.
(i)$\Longleftrightarrow$(vii) follows from Proposition 2.9.
(vi)$\Longleftrightarrow$(vii) follows from Lemma 2.6 and the fact that every injective indecomposable is isomorphic to one of the $E(S_i)$'s.
(vii)$\Longleftrightarrow$(viii). Let T be a simple right $\mathcal{C}$-comodule. Then T is finitely generated, and therefore a simple object in $\mathcal{M}^{\mathrm{fg}\mathcal{C}}$. By the duality between ${}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ and $\mathcal{M}^{\mathrm{fg}\mathcal{C}}$, $T^* \in {}^{\mathrm{fg}\mathcal{C}}\mathcal{M}$ is simple, and $E(T^*)$ is finitely generated by assumption. The monomorphism $T^* \to E(T^*)$ is essential, so, using the duality, the dual map is a superfluous epimorphism ${}^*E(T^*) \to {}^*(T^*) \simeq T$. It follows from Corollory 4.2 that ${}^*E(T^*)$ is projective, and, using again the duality, that it is finitely generated. Hence ${}^*E(T^*)$ is a finitely generated projective cover of T.

A coring $\mathcal{C}$ is called left (resp. right) perfect if every object in ${}^{\mathcal{C}}\mathcal{M}$ (resp. $\mathcal{M}^{\mathcal{C}}$) has a projective cover. We will now see that, over a qF-ring, perfectness on both sides is equivalent to semiperfectness on both sides. First we need a Lemma.

Lemma 4.4

Let R be a qF-ring, and $\mathcal{C}$ a right semiperfect coring that is both left and right profective over R. Then every $0 \neq M \in {}^{\mathcal{C}}\mathcal{M}$ contains a maximal subcomodule. Consequently the Jacobson radical $J(M)$ is small in M.

Proof

${}^*M \in \mathcal{M}_{{}^*\mathcal{C}}$, and $\mathrm{Rat}^{\mathcal{C}}({}^*M)$ is dense in ${}^*M$, by Theorem 4.3. Thus, if $\mathrm{Rat}^{\mathcal{C}}({}^*M) = 0$, then ${}^*M = 0$, which is impossible since R is a cogenerator in ${}_R\mathcal{M}$. So $\mathrm{Rat}^{\mathcal{C}}({}^*M) \neq 0$, and we can take a nonzero simple right subcomodule S of $\mathrm{Rat}^{\mathcal{C}}({}^*M)$. Let $u: S \to \mathrm{Rat}^{\mathcal{C}}({}^*M)$ and $v: \mathrm{Rat}^{\mathcal{C}}({}^*M) \to {}^*M$ be the inclusion maps. Then u is right $\mathcal{C}$-colinear, and v is right ${}^*\mathcal{C}$-linear. Now consider the composition $f = u^* \circ v^* \circ \phi$.

$$M \xrightarrow{\phi} ({}^*M)^* \xrightarrow{v^*} (Rat^{\downarrow}({}^*M))^* \xrightarrow{u^*} S^*.$$

A straightforward computation shows that $v^* \circ \phi$ is left $\mathcal{C}^*$-linear, and therefore $f = u^* \circ v^* \circ \phi$ is also left $\mathcal{C}^*$-linear. Now $u^* \circ v^*$ is surjective, $\mathrm{Im}\,\phi$ is dense in $({}^*M)^*$, by Lemma 4.1, so $\mathrm{Im}\, f = (u^* \circ v^*)(\mathrm{Im}\,\phi)$ is dense in S^*, by Proposition 1.5. Since S is simple, and therefore finitely generated, the only dense submodule of S^* is S^* itself. So $f: M \to S^*$ is a surjective $\mathcal{C}^*$-linear morphism between the left $\mathcal{C}$-comodules M and S^*, hence it is a left $\mathcal{C}$-colinear surjection. Since S^* is simple in ${}^*\mathcal{M}$, $\mathrm{Ker}\, f$ is a maximal subcomodule of M.

Proposition 4.5

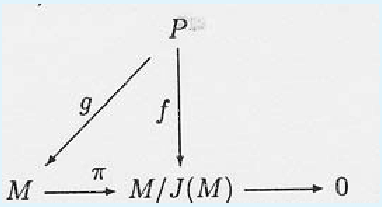
Let R be a qF-ring, and $\mathcal{C}$ an R-coring which is left and right (locally) projective over R. Then the following assertions are equivalent.

(i) $\mathcal{C}$ is left and right perfect;

(ii) $\mathcal{C}$ is left and right semiperfect.

Proof

The implication (i)⇒(ii) is trivial. Conversely, we will first show that $M/J(M)$ is a semisimple object in $\mathcal{M}^{\mathcal{C}}$, for any $M \in \mathcal{M}^{\mathcal{C}}$. Take $\overline{x} \in M/J(M)$, and let N be the subcomodule of $M/J(M)$ generated by $\overline{x}$. Then $N \subset M/J(M)$, hence $J(N) \subset J(M/J(M)) = 0$. N is finitely generated, and therefore artinian. Let $N_1, \cdots, N_n$ be maximal subcomodules of N such that $\bigcap_{i=1}^n N_i = 0$. Then $N = \bigoplus_{i=1}^n N/N_i$ is semisimple. This shows that every $\overline{x} \in M/J(M)$ belongs to a semisimple subcomodule, so $M/J(M)$ is semisimple.
Since $\mathcal{C}$ is right semiperfect, there exists a projective cover $f: P \to M/J(M)$. Since P is projective, there exists $g \in \mathcal{M}^{\mathcal{C}}$ making the following diagram commutative (π



is the canonical projection):

Now $\pi(J(M)+g(P)) = \pi(g(P)) = f(P) = M/J(M)$, so $J(M)+g(P) = M$, since π is surjective. $\mathcal{C}$ is left semiperfect, hence, by Lemma 4.4, $J(M) \ll M$, and we conclude that $g(P) = M$. So g is surjective. $\mathrm{Ker}\, f \ll P$ and $\mathrm{Ker}\, g \subset \mathrm{Ker}\, f$, hence $\mathrm{Ker}\, g \ll P$, and we conclude that $g:\ P \to M$ is a projective cover of M.

4.2. Examples

Example 4.6 Let $\mathcal{C}$ be a coring, and assume that $\mathcal{C}$ is finitely generated and projective as a left R-module. Then $\mathcal{M}^{\mathcal{C}}$ is isomorphic to $\mathcal{M}_{*\mathcal{C}}$, and $\mathrm{Rat}^{\mathcal{C}}$ is an isomorphism of categories. Hence $\mathrm{Rat}^{\mathcal{C}}$ is exact. $\mathcal{M}^{\mathcal{C}}$ has enough projectives, but not necessarily projective covers. As an example, let R be a non-semiperfect ring, and $\mathcal{C} = R$, the trivial R-coring. Then $\mathcal{M}^{\mathcal{C}} = \mathcal{M}_R$ is not semiperfect.

Example 4.7 Let $\mathcal{C}$ be a cosemisimple coring. Then $\mathcal{C}$ is left and right semiperfect, since the categories of left and right $\mathcal{C}$-comodules are semisimple, see [4, 19.14], [9] and [11]. In this case, $\mathcal{C}$ is projective in ${}_R\mathcal{M}$ and $\mathcal{M}_R$, so $\mathcal{C}$ satisfies the left and right α-condition. $\mathcal{C}$ can then be written as a direct sum of finitely generated left (or right) $\mathcal{C}$-comodules, and the functors $\mathrm{Rat}^{\mathcal{C}}$ and ${}^{\mathcal{C}}\mathrm{Rat}$ are exact. So all the equivalent statements of Theorem 4.3 hold, without the assumption that the base ring R is a qF-ring.

Example 4.8 To a ring morphism $\iota:\ R \to S$, we can associate the Sweedler coring $\mathcal{C}$. As an S-bimodule, $\mathcal{C} = S \otimes_R S$, and the comultiplication and counit are given by the formulas

$$\Delta(s \otimes_R s') = (s \otimes_R 1) \otimes_S (1 \otimes_R s') \quad ; \quad \varepsilon(s \otimes_R s') = ss'$$

The Sweedler coring is important in descent theory: the comodules over $\mathcal{C}$ are exactly the descent data from [12] (in the commutative case) and [7] (in the noncommutative

case). If $M \in \mathcal{M}^{\mathcal{C}}$, then M descends to an R-module

$$M^{\mathrm{co}\mathcal{C}} = \{m \in M \mid \rho(m) = m \otimes_R 1\}$$

For a detailed discussion, we refer to [5]. It is also easy to see that we have an isomorphism of R-algebras

$${}^*\mathcal{C} = {}_S\mathrm{Hom}(S \otimes_R S, S) \cong {}_R\mathrm{End}(S)$$

(again, ${}_R\mathrm{End}(S)$ is a ring with the opposite composition as multiplication). Also notice that $S \subset {}_R\mathrm{End}(S)$ as algebras, by right multiplication.
If we assume that $S \in {}_R\mathcal{M}$ is locally projective, then $\mathcal{C} \in {}_S\mathcal{M}$ is locally projective, and we can consider the functor

$$\mathrm{Rat}^{\mathcal{C}} : \mathcal{M}_{{}_R\mathrm{End}(S)} \to \mathcal{M}^{\mathcal{C}}$$

Let M be a right ${}_R\mathrm{End}(S)$-module, and take $m \in M$. Then $m \in \mathrm{Rat}^{\mathcal{C}}(M)$ if and only if there exists $m_{[0]} \otimes_R m_{[1]} \in M \otimes_R S$ such that $m \cdot f = m_{[0]} f(m_{[1]})$, for all $f \in {}_R\mathrm{End}(S)$. In particular,

$$\begin{aligned} \mathrm{Rat}^{\mathcal{C}}(M)^{\mathrm{co}\mathcal{C}} &= \{m \in \mathrm{Rat}^{\mathcal{C}}(M) \mid \rho(m) = m \otimes_R 1\} \\ &= \{m \in M \mid m \cdot f = mf(1), \text{ for all } f \in {}_R\mathrm{End}(S)\} \end{aligned}$$

$\mathrm{Rat}^{\mathcal{C}}(M)$ is a right $\mathcal{C}$-comodule, and therefore a right ${}_R\mathrm{End}(S)$-module, and, by restriction of scalars, a right S-module. Therefore

$$\mathrm{Rat}^{\mathcal{C}}(M)^{\mathrm{co}\mathcal{C}} \cdot S \subset \mathrm{Rat}^{\mathcal{C}}(M) \tag{4}$$

If we take $M = {}_R\mathrm{End}(S)$, then we see that

$$\mathrm{Rat}^{\mathcal{C}}(M)^{\mathrm{co}\mathcal{C}} = \{g \in {}_R\mathrm{End}(S) \mid (f \circ g)(s) = g(s)f(1), \text{ for all } f \in {}_R\mathrm{End}(S)\}$$

Take $h \in {}^*S = {}_R\mathrm{Hom}(S, R)$. Then $\overline{h} = h \circ \iota \in {}_R\mathrm{Hom}(S, S)$, and it follows easily that $\overline{h} \in \mathrm{Rat}^{\mathcal{C}}({}_R\mathrm{End}(S))^{\mathrm{co}\mathcal{C}}$. We will use this to show that $\mathrm{Rat}^{\mathcal{C}}(\mathrm{End}(S))$ is dense in ${}_R\mathrm{End}(S)$.
Take $f \in {}_R\mathrm{End}(S)$, and $F \subset S$ finite. Since S is locally projective, there are $h_1, \cdots, h_n \in {}^*S$ and $x_1, \cdots, x_n \in S$ such that

$$x = \sum_{k=1}^{n} h_k(x) x_k$$

for $x \in F$ and then a simple computation shows that

$$f(x) = \sum_{k=1}^{n} h_k(x) f(x_k) = \Big(\sum_{k=1}^{n} \overline{h}_k \cdot f(x_k)\Big)(x)$$

By (4) and the the fact the above argument, $\sum_{k=1}^{n} \overline{h}_k \cdot f(x_k) \in \mathrm{Rat}^{\mathcal{C}}(\mathrm{End}(S))$. So we have shown that f coincides on F to an element in $\mathrm{Rat}^{\mathcal{C}}(\mathrm{End}(S))$. We conclude that $\mathrm{Rat}^{\mathcal{C}}({}^*\mathcal{C})$ lies dense in ${}^*\mathcal{C}$, and, by Proposition 2.4, $\mathrm{Rat}^{\mathcal{C}}$ is exact.
If S is pure as a left and right R-module, in particular if $S \in {}_R\mathcal{M}$ is faithfully flat, then the categories $\mathcal{M}_R$ and $\mathcal{M}^{\mathcal{C}}$ are equivalent (see [5, 7, 12]). In this case, $\mathcal{M}^{\mathcal{C}}$ has enough projectives.
If $S \in {}_R\mathcal{M}$ is faithfully flat and locally projective, then we have an explicit description of $\mathrm{Rat}^{\mathcal{C}}(M)$, namely

$$\mathrm{Rat}^{\mathcal{C}}(M) = \mathrm{Rat}^{\mathcal{C}}(M)^{\mathrm{co}\mathcal{C}} \otimes_R S,$$

with $\mathrm{Rat}^{\mathcal{C}}(M)^{\mathrm{co}\mathcal{C}}$ given by (4).

REFERENCES

[1] Abuhlail, J. Y. Rational modules for corings, *Comm. Algebra* 31 (2003), 5793–5840.

[2] Anderson, F. and Fuller, K., "Rings and categories of modules", *Grad. Texts Math.* 13, Springer Verlag, Berlin, 1992.

[3] Brzeziński, T. The structure of corings. Induction functors, Maschke-type theorem, and Frobenius and Galois-type properties, *Algebras and Representation Theory* 5 (2002), 389–410.

[4] Brzeziński, T. and Wisbauer, R., "Corings and comodules", *London Math. Soc. Lect. Notes Ser.* 309, Cambridge University Press, Cambridge, 2003.

[5] Caenepeel, S. *Galois corings from the descent theory point of view*, *Fields Inst. Comm* 43 (2004).

[6] Caenepeel, S., Vercruysse, J. and Shuanhong Wang, Rationality properties for Morita contexts associated to corings, in "Hopf algebras in non-commutative geometry and physics", S. Caenepeel and F. Van Oystaeyen, eds., *Lecture Notes Pure Appl. Math.* 239, Dekker, New York, 2004.

[7] Cipolla, M. Discesa fedelmente piatta dei moduli, *Rendiconti del Circolo Matematico di Palermo, Serie II* 25 (1976), 43–46.

[8] Dăscălescu, S., Năstăsescu, C. and Ş.aianu, S., "Hopf algebras: an Introduction", *Monographs Textbooks in Pure Appl. Math.* 235 Marcel Dekker, New York, 2001.

[9] El Kaoutit, L. and Gómez, Torrecillas J., Comatrix corings: Galois corings, descent theory, and a structure Theorem for cosemisimple corings, *Math. Z.*, 244 (2003), 887–906.

[10] El Kaoutit, L. and Gómez, Torrecillas J., Morita duality for corings over quasi-Frobenius rings, in "Hopf algebras in non-commutative geometry and physics", S. Caenepeel and F. Van Oystaeyen, eds., *Lecture Notes Pure Appl. Math.* 239, Dekker, New York, 2004.

[11] El Kaoutit, L., Gómez, Torrecillas J. and Lobillo, F. J., Semisimple corings, *Algebra Coll.*, to appear.

[12] Knus, M. and Ojanguren, M., "Théorie de la Descente et Algèbres d'Azumaya", *Lect. Notes in Math.* 389, Springer Verlag, Berlin, 1974.

[13] Năstăsescu, C. and Gómez Torrecillas, J., Quasi-coFrobenius coalgebras, *J. Algebra* 174 (1995), 909–923.

[14] Sweedler, M. E. The predual Theorem to the Jacobson-Bourbaki Theorem, *Trans. Amer. Math. Soc.* 213 (1975), 391–406.

[15] Wisbauer, R. "Foundations of module and ring theory", Gordon and Breach, Philadelphia, 1991

NONLINEAR FOURIER ANALYSIS OF SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS WITH COMPUTER ALGEBRA: SURVEY AND NEW RESULTS

D. K. Callebaut[1], G. K. Karugila[1] and A. H. Khater[1,2]
[1]Physics Department, Campus Drie Eiken UA
University of Antwerp, B-2610 Antwerp, Belgium
e-mail: Dirk.Callebaut@ua.ac.be
Geoffrey.Karugila@ua.ac.be

[2]Department of Mathematics, Faculty of Science
Cairo University, Beni-Suef, Egypt
e-mail: Khater_ah@hotmail.com

Abstract

We explain the essence of our method to solve systems of nonlinear partial differential equations using a kind of continuation of the Fourier analysis for the linearized system. An illustration shows how to determine the convergence limit. A brief survey of previous results is given. We study the equilibrium and stability of an inhomogeneous universe in a Newtonian approximation supplemented with a cosmological constant in view to explain the formation of galaxies, clusters of galaxies and the tessellation of the universe. Some qualitative results are obtained.

Key words: Waves and wave propagation; Perturbation theory; Fourier analysis; Cosmology; Gravitation.

1. INTRODUCTION

The theoretical, experimental and computational investigations of oscillations and instabilities of physical systems in plasma- and astrophysics, in hydro- and magnetodynamics, in transport systems, etc. are very intriguing and complicated. They flourished in the second half of the 20th century. Several causes contributed to this blossoming: the goal of providing a quasi unlimited amount of energy by nuclear fusion; the radio- and radarwaves and the investigation of the ionosphere; the magnetic phenomena related to the sun and to astrophysics in general; a lot of physical, chemical and engineering problems with all sort of boundary conditions and practical applications. The linearized perturbation methods (Chandrasekhar [1]) yield the dispersion relation: very useful but often not sufficient. Since around 1970 various nonlinear methods were conceived (Malfliet and Hereman [2]; Verheest [3]; Pillay, Rao and Bharuthram [4]). One of those is a nonlinear Fourier analysis

due to Callebaut [5]. It will be sketched below (section 2).

The problems are usually expressed in mathematical form by a system of partial differential equations (PDEs) (and boundary and/or initial conditions) including in general the continuity equation, the equation of motion, the Poisson equation, some "closing" assumption (an equation of state, e.g., the polytropic equation for the pressure or some equation(s) concerning the heat transport) and, if electromagnetism is involved, completed by Maxwell equations (in full or with simplifications). Clearly such a system of PDEs was and still is usually impossible or very difficult to solve straightforwardly. Plateau [6] had done a lot of work experimentally and theoretically concerning the investigation of the stability of liquid cylinders and tori (oil and mercury). Lord Rayleigh [7], extending the work of Plateau, developed his method and completed the treatment of Plateau and developed his theory of sound waves. Following Lord Rayleigh one (a) applies an "infinitely small" perturbation to an equilibrium (or steady state), (b) linearizes the set of equations and applies a Fourier analysis, allowing to work with a single term, and (c) obtains the (linear) dispersion relation, i.e., a relation between the frequency of oscillation or the growth rate of the instability and the wavenumber and the equilibrium quantities. This is a most valuable information, but often it is not enough and one needs a nonlinear approach. However, in strong contrast to the linear cases, there is no unique systematic method for nonlinear problems but a variety of approaches each more or less suited for specific situations. Roughly speaking there are two methods: (a) methods yielding exact solutions and (b) methods yielding approximations. In the class (a) one has e.g. the Bäcklund transformations (Khater et al.[8]-[12]) to obtain new solutions from a known one. There is the Ablowitz, Kaup, Newell and Segur (AKNS) systems (Khater et al. [13]) which may lead to a solution in some cases. For solitary waves the tanh method of Malfliet (Malfliet and Hereman [2]; Malfliet[14], Malfliet and Wieërs [15]) and some extensions (Hereman et al. [16]-[18]; Verheest [19]) are very adequate. For approximate solutions, class (b), the time scale method working with a hierarchy of orders is often used. Callebaut [5] developed a kind of nonlinear Fourier analysis, and the present work is addressed to this method.

2. THE "NONLINEAR FOURIER ANALYSIS"

In the footsteps of Lord Rayleigh and his followers - as stated above - the way to obtain the (linear) dispersion relation is to linearize the system of equations and next to apply a Fourier analysis. This has the result that at least part of the PDEs (partial differential equations) may become algebraic and that one may work with one single Fourier term, say $Ae^{i(\omega t+\boldsymbol{k}\cdot\boldsymbol{r})}$, in which A is the (arbitrary) amplitude,

ω the angular frequency, $\boldsymbol{k}$ the wave vector, $\boldsymbol{r}$ the space vector and t the time. In view of the (linearized) boundary conditions it is often the case that one or two dimensions do not allow a Fourier analysis and that a differential equation has still to be solved, e.g., in the cylindrical case this leads to the inclusion of Bessel functions combined with the Fourier terms for the other directions. Of course the full linearized solution is then a sum/integral over such terms, still with arbitrary (but very small) amplitudes. Let us call this Fourier analysis the *horizontal* Fourier integral/sum. However, having derived the dispersion relation and sticking to one specific Fourier term with a specific amplitude A, specific ω and specific k, one may look at the higher order terms generated by this specific term by iteration on the nonlinear system which leads to a series of terms of the type $a_n A^n e^{ni(\omega t+\boldsymbol{k}\cdot\boldsymbol{r})}$ as solution. Here a_n are the coefficients fixed by substitution in the nonlinear basic system of equations. We may call this the *vertical* Fourier series. In fact it is the normal Fourier series of the function defined by the nonlinear system and by its first term $Ae^{i(\omega t+\boldsymbol{k}\cdot\boldsymbol{r})}$. Actually, this function (supposing that it exists, satisfies the very broad conditions required for a Fourier series which means essentially that it is periodic) is fully fixed as a solution of the system of equations and its first Fourier term which in a sense acts as a kind of initial condition, or rather as the "linear approximation condition".

Clearly the horizontal Fourier integral and the vertical Fourier series have distinct characters.

The horizontal Fourier integral/sum has the following features:

(a) The value of ω is fixed by $\boldsymbol{k}$ according to the linear dispersion relation.

(b) The components of $\boldsymbol{k}$ are arbitrary except for the (linearized) boundary conditions which may discretize them or limit their domain. Consequently the full horizontal solution is a trifold integral or sum or mixture of both.

(c) All amplitudes are arbitrary (supposed "small").

(d) All terms are approximate solutions, satisfying only the linearized equations (plus boundary conditions).

(e) There is no nonlinearity involved at all.

On the other hand the *vertical Fourier series* has the following features:

(a) As above the value of ω is fixed by $\boldsymbol{k}$ according to the linear dispersion relation.

(b) It consists of terms in which all arguments are multiples of $\omega t + \boldsymbol{k} \cdot \boldsymbol{r}$.

(c) All the amplitudes stand in a well defined relationship to each other. The choice of one of their amplitudes (usually the one corresponding to $\omega t + \boldsymbol{k} \cdot \boldsymbol{r}$ itself) fixes them all through the basic equations and the boundary conditions.

(d) All terms together constitute a special but exact solution of the system of partial differential equations and boundary conditions.

(e) The nonlinearity is in the equations, not in the Fourier series itself, which is still a "traditional Fourier series" although the way to determine it is not. The real nonlinearity appears if several "initial terms" are used.

Involving more "initial terms"

Suppose now that we consider two "initial terms", taken from the horizontal solutions $A_1 e^{i(\omega_1 t + \boldsymbol{k}_1 \cdot \boldsymbol{r})}$ and $A_2 e^{i(\omega_2 t + \boldsymbol{k}_2 \cdot \boldsymbol{r})}$, where the pairs $(\omega_1, \boldsymbol{k},)$ and $(\omega_2, \boldsymbol{k}_2)$ are incommensurable (meaning that $\chi_1 \equiv \omega_1 t + \boldsymbol{k}_1 \cdot \boldsymbol{r}$ and $\chi_2 \equiv \omega_2 t + \boldsymbol{k}_2 \cdot \boldsymbol{r}$ do not satisfy a relation of the type $n\chi_1 = m\chi_2$ with n and m integers, otherwise the approach is somewhat different). Substituting their sum in the system of equations leads to three parts: a vertical Fourier series corresponding to $A_1 e^{i\chi_1}$, a similar vertical Fourier series corresponding to $A_2 e^{i\chi_2}$ and a mixed series. In particular for the mixed series one has just to use combinatorial coefficients for each order.

The same is true when considering several "initial terms". Clearly once we consider mixing (interference) of initial terms we are working with an authentic nonlinear solution and this method may justly be called a kind of *nonlinear Fourier analysis*. We shall not deal here with the more general situation in which the solution is not quite periodic, so that the frequency in higher order terms depends on the amplitude as well as on $\boldsymbol{k}$.

3. ILLUSTRATION BY THE ELECTRON PLASMA

As an illustration we consider the fairly simple case called the electron plasma: a uniform plasma consisting of electrons and ions, infinite in all directions. We neglect gravity, viscosity, resistivity and the magnetic contributions. The basic equations are then respectively the equation of continuity, motion, Poisson, and polytropics:

$$\partial_t n + div(n\boldsymbol{v}) = 0, \tag{1}$$

$$nm\left(\partial_t \boldsymbol{v} + \boldsymbol{v}\cdot\boldsymbol{\nabla}\boldsymbol{v}\right) = -\boldsymbol{\nabla}p + e\,n\boldsymbol{\nabla}\varphi, \tag{2}$$

$$\Delta\varphi = e\left(n - n_0\right)/\varepsilon, \tag{3}$$

$$p = K\left(mn\right)^{\Gamma} + p_i, \tag{4}$$

where n is the number density of the electrons, n_0 their equilibrium density, $\boldsymbol{v}$ their velocity, φ is the electrical potential, p the total pressure with pressure p_i of ions which is supposed to be constant, e and m are the electron charge and mass, ε is the permittivity (in vacuum 8.85×10^{-12} C/Vm), K and Γ (polytropic exponent) are constants.

The linear perturbation is expressed as a Fourier series. We fix one term say $A\exp[i\chi]$ as mentioned above. We thus develop only one family of higher order terms corresponding to a single Fourier term of the linearized analysis. The nonlinear terms then generate $a_2A^2\exp[2i\chi]$, $a_3A^3\exp[3i\chi]$, etc., with coefficients a_2, a_3, $\cdots$ to be determined.

Using $\chi = \omega t + \boldsymbol{k}\cdot\boldsymbol{r}$ as a single variable reduces the system to ordinary differential equations:

$$\omega n_-' + \boldsymbol{k}\cdot\left(n_-\boldsymbol{v}_-\right)' = 0, \tag{5}$$

$$m_-\left(\omega + \boldsymbol{v}_-\cdot\boldsymbol{k}\right)\boldsymbol{v}_-' = e\boldsymbol{k}\varphi' - \boldsymbol{k}p', \tag{6}$$

$$k^2\varphi'' = \frac{e}{\varepsilon}\left(n_- - n_0\right), \tag{7}$$

$$p = K\left(mn\right)^{\Gamma} + p_i \Rightarrow p' = \Gamma K_-\, n^{\Gamma-1}n', \tag{8}$$

where the accent means the derivative with respect to χ, $K_- = Km^{\Gamma}$ and $p_i = \textit{constant}$ (ions are assumed to be massive and hence immobile). Integrating the continuity equation (5) we obtain, with $\in_-$ the constant of integration,

$$\left(\omega + \boldsymbol{k}\cdot\boldsymbol{v}_-\right)n_- = \in_- = \omega n_0, \tag{9}$$

which is used to reduce the system of equations (5) - (8) to a differential equation of second order

$$\left[\frac{\left(\Gamma-2\right)k^2\,v_{s-}^2\,n^{\Gamma+1}}{n_0^{\Gamma+1}} + 3\,\omega^2\right]n'^2 + \left(\frac{k^2\,v_{s-}^2\,n^{\Gamma+1}}{n_0^{\Gamma+1}} - \omega^2\right)n\,n'' = \frac{\omega_e^2 n^4(n-n_0)}{n_0^3}, \tag{10}$$

where n_0 is the equilibrium density, $v_{s-}^2 = K_-\,\Gamma\,n_0^{\Gamma-1}/m$ is the sound velocity of electrons and $\omega_e^2 = (e^2n_0)/(m\varepsilon)$ is the square of the electron plasma frequency. We

calculated a number of coefficients numerically using *mathematica* and then inferred the analytic expression. E.g., in case of the cold plasma we obtained

$$n = n_0 \left(1 + \sum_{j=1}^{N} \frac{j^j}{j!} A^j e^{ij\chi}\right),\ \boldsymbol{v} = \frac{\boldsymbol{k}\,\omega}{k^2} \sum_{j=1}^{N} \frac{j^{j-1}}{j!} A^j e^{ij\chi},\ \varphi = \frac{e n_0}{k^2 \varepsilon} \sum_{j=1}^{N} \frac{j^{j-2}}{j!} A^j e^{ij\chi}. \tag{11}$$

We integrated as well equation (10) to obtain a differential equation of first order and even to obtain a fully integrated equation. Insertion of the Fourier series led to the same coefficients as given in (11), but the time needed for the three procedures to calculate the coefficients was different and strongly dependent on any simplification or additional effect taken into account (e.g., when the ions are mobile as well). The series (11) is convergent provided $A < e^{-1}$ ($e \approx 2.71828\cdots$); but if the linearized perturbed density has an amplitude larger than 37% of the equilibrium density n_0, the series is no more convergent.

A graphical method confirmed the convergence obtained analytically. Summing up N terms of density n and plotting the result for χ in the interval 0, 2π (or even 0, π), yields an oscillating graph. If there was any value in this interval for n less than zero the series has to be rejected since the electron number density may not become negative. It turns out that this graphical method confirms the radius of convergence found analytically rather well; we performed the summation in some cases up to $N = 7000$ to verify the result accurately. Actually when we exceeded the limit of convergence for A slightly the number of terms might be rather limited (say ten or twenty), except very close to the limit of convergence. This numerical/graphical method turned out a powerful tool in all cases where we did not have a systematic analytical expression for the coefficients.

We developed as well the theory using cosines instead of exponentials. It turned out that the radius of convergence was doubled: now the series converges for $A < 2e^{-1}$, i.e. the linearized perturbed density amplitude may reach nearly 74% of the equilibrium density and still not lead to breakdown. In fact the result had to be expected: an exponential $e^{i\chi}$ corresponds to a sine and a cosine, thus two waves instead of one, thus leading to halving the convergence limit.

Even if only the second order term is calculated it yields valuable information above the linearized theory. We suggested to verify the results experimentally e.g. by applying a (strong) external perturbation (electric field) in a Q-machine (Quiescent Plasma Machine) having a magnetized alkaline plasma or unmagnetized argon plasma of a DP-Machine (Double Plasma Machine).

4. BRIEF SURVEY OF RESULTS

The preceding results have been generalized to a multiple species plasma, by including the motion of ions as well, by including their pressure (even applying different approaches) (Callebaut and Karugila [20]; [21]) and by including the magnetic effects involved in motion, too. The method can be generalized to instabilities when a growth rate occurs instead of ω. This was elaborated for a liquid jet involving surface tension and the result agreed very well with experiments (Callebaut [5]; [22]; [23]). Similarly it was applied to various plasma columns and extended to involve magnetic fields internally (i.e. in the plasma column) and externally (around the column).

The method was applied to gravitational problems, too. E.g., an infinitely long gravitational cylinder of homogeneous density was studied to higher order (Callebaut [5]). The linear theory was performed by Chandrasekhar and Fermi [24] and Chandrasekhar [1]; they considered it as a model to investigate the stability of arms of spiral galaxies. The nonlinear theory was further extended to include a magnetic field in the gravitational cylinder and outside it. Similarly the case of a plane-parallel homogeneous gravitating medium (cf. a flat galaxy) was studied (Callebaut [5]), including magnetic fields inside and outside. The results of the linear theory confirmed Jeans' criterion, which was derived on an inappropriate basis but was made plausible on various grounds and explaining some observational data (Callebaut [25]). Again convergence and influence of various effects are calculated and discussed.

5. INHOMOGENEOUS UNIVERSE

We consider the medium infinite in all space dimensions and obeying the Newtonian law of gravitation to which a cosmological term is added. Here we analyze the equilibria involving an inhomogeneous density and a varying gravitational potential. The basic equations are the continuity equation, the equation of motion, the Poisson equation (i.e. the field equation for Newtonian gravitation with a cosmological constant and the polytropic equation). Thus, the basic system of equations is written as (Callebaut [5])

$$\partial_t \rho + div(\rho \boldsymbol{v}) = 0, \tag{12}$$

$$\rho \frac{d\boldsymbol{v}}{dt} = -\boldsymbol{\nabla} p - \rho \boldsymbol{\nabla} \varphi, \tag{13}$$

$$\Delta \varphi + \Lambda \varphi = 4\pi G \rho, \tag{14}$$

$$p = K\rho^{\Gamma}, \tag{15}$$

where $\rho(\boldsymbol{r}, t)$ is the mass density, $\boldsymbol{v}(\boldsymbol{r}, t)$ is the velocity, $\varphi(\boldsymbol{r}, t)$ is the gravitational potential, Λ is the cosmological constant, G is the gravitational constant ($\approx (6.6726 \pm 0.0005) \times 10^{-11}\mathrm{m}^3/\mathrm{kg\ s}^2$), $p(\boldsymbol{r}, t)$ is the pressure, K and Γ are constants, $\boldsymbol{r}$ is the space and t is the time.

Equation (14) is an approximation like the one leading to the Newtonian theory from Einstein's gravitational field equation of 1917, where he introduced the cosmological term. It may be noted that the interest in the cosmological term has been revived in recent years in particular due to the recent findings of the supernovae of type 1a acting as standard candles to look back into the distant past of the universe.

Equilibrium

The basic system of equations to be considered here for equilibrium are to be inferred from the equations (12) - (15). Putting zero order quantities, with $\boldsymbol{v}_0 = \boldsymbol{0}$, into these equations we have

$$\partial_t \rho_0 = 0, \tag{16}$$

$$\boldsymbol{\nabla} p_0 = -\rho_0 \boldsymbol{\nabla} \varphi_0, \tag{17}$$

$$\Delta\varphi_0 + \Lambda\varphi_0 = 4\pi G\rho_0, \tag{18}$$

$$p_0 = K\rho_0^{\Gamma}. \tag{19}$$

Here we note that ρ_0 and hence p_0 and φ_0 are independent of time t but dependent on space, $\boldsymbol{r}$. In order to differentiate them from the previously seen quantities, we denote them by $\rho_0(\boldsymbol{r})$, $p_0(\boldsymbol{r})$ and $\varphi_0(\boldsymbol{r})$, respectively.

Substituting (19) into (17) we get

$$\boldsymbol{\nabla} K\rho_0^{\Gamma}(\boldsymbol{r}) = K\Gamma\rho_0^{\Gamma-1}(\boldsymbol{r})\boldsymbol{\nabla}\rho_0(\boldsymbol{r}) = -\rho_0(\boldsymbol{r})\boldsymbol{\nabla}\varphi_0(\boldsymbol{r})$$

and further

$$\frac{K\Gamma}{\Gamma-1}\boldsymbol{\nabla}\rho_0^{\Gamma-1}(\boldsymbol{r}) = -\boldsymbol{\nabla}\varphi_0(\boldsymbol{r}) \quad \text{and} \quad \varphi_0(\boldsymbol{r}) = -\frac{\mathrm{K\Gamma}}{\Gamma-1}\rho_0^{\Gamma-1}(\boldsymbol{r}) + \varphi_{00},$$

where φ_{00} is a kind of cosmological background potential. (This is arbitrary in the Newtonian physics, but no more when the cosmological term is added.) Substituting this into the field equation (18), we get

$$-\frac{K\Gamma}{\Gamma-1}\Delta\rho_0^{\Gamma-1}(\boldsymbol{r}) - \frac{K\Gamma\Lambda}{\Gamma-1}\rho_0^{\Gamma-1}(\boldsymbol{r}) + \Lambda\varphi_{00} = 4\pi G\rho_0(\boldsymbol{r}). \tag{20}$$

This constitutes to a generalization of the Lane-Emden equation, famous in the early days of stellar structure (cf. Chandrasekhar [26]), for which $\Lambda = 0$.

It is a nonlinear partial differential equation. In the case of spherical symmetry with $\Lambda = 0$ (stellar case), one has simple closed solutions for the following values of Γ: ∞ (incompressible), 2 and 6/5. Moreover, for $\Lambda = 0$, there is, for any Γ ($\neq 1$) and for spherical, cylindrical or Cartesian coordinates, the so-called singular solution of the type Ar^p, with A and p fixed in terms of Γ (if $p < 0$ the solution is rejected because it is singular at the center of the star, if $p > 0$ the solution is rejected as it yields zero density in its center).

6. AN IMPORTANT EXAMPLE

If $\Gamma = 2$ then equation (20) becomes linear

$$\Delta\rho_0(\boldsymbol{r}) + \left(\Lambda + \frac{2\pi G}{K}\right)\rho_0(\boldsymbol{r}) = \frac{\Lambda\varphi_{00}}{2K}. \tag{21}$$

Replacing $\rho_0(\boldsymbol{r})$ by

$$\rho_0(\boldsymbol{r}) = a_0 + \sum_{j=1}^{N}(a_j \cos j\boldsymbol{k}_g \cdot \boldsymbol{r} + b_j \sin j\boldsymbol{k}_g \cdot \boldsymbol{r}), \tag{22}$$

and comparing the coefficients implies

$$a_0 = \frac{\Lambda\varphi_{00}}{2(K\Lambda + 2\pi G)}, \tag{23}$$

$$[2K(k_g^2 j^2 - \Lambda) - 4\pi G]a_j = 0 \tag{24}$$

and

$$[2K(k_g^2 j^2 - \Lambda) - 4\pi G]b_j = 0. \tag{25}$$

Hence it follows

$$(k_g j)^2 = (2\pi G + K\Lambda)/K \quad \text{or } a_j = 0$$

and similarly the same for b_j in (25). With a suitable choice of the origin

$$\rho_0(\boldsymbol{r}) = B + a\cos k_0(\boldsymbol{r}), \tag{26}$$

where

$$B = \frac{\Lambda\varphi_{00}}{2(K\Lambda + 2\pi G)}, \quad k_0 = \sqrt{\frac{2\pi G}{K} + \Lambda}, \tag{27}$$

and as $\rho_0(\boldsymbol{r}) \geq 0$, it follows

$$\varphi_{00}\,\Lambda \geq 2\,(K\Lambda + 2\pi G)\,|a|. \tag{28}$$

Note that this implies $\varphi_{00}\,\Lambda > 0$: if Λ is zero then $B = 0$ and $a = 0$, which reduces the situation to an empty universe. It is remarkable that even in an inhomogeneous universe the cosmological constant is required. Similarly φ_{00}, although arbitrary, may not simply be put equal to zero if we want a varying density. Notice that (26) implies as well a more general form with a sum of cosines

$$\rho_0(\boldsymbol{r}) = B + a\cos k_0 x + b\cos k_0 y + c\cos k_0 z, \tag{29}$$

where a, b and c are arbitrary but restricted by the condition $\rho_0(\boldsymbol{r}) \geq 0$ or

$$B \geq |a| + |b| + |c|. \tag{30}$$

The other equilibrium quantities are given by

$$p_0(\boldsymbol{r}) = K\rho_0^2(\boldsymbol{r}), \quad \varphi_0(\boldsymbol{r}) = -2K\rho_0(\boldsymbol{r}) + \varphi_{00}. \tag{31}$$

Remarks

If $K\Lambda + 2\pi G = 0$ (or $B = \infty$), which seems physically unrealistic in view of the supposed extreme smallness of Λ, then equation (21) has a solution of the type

$$\rho_0(\boldsymbol{r}) = ax^2 + by^2 + cz^2 + dx + ey + fz + g, \tag{32}$$

where a, b, c, $\cdots$, g are arbitrary constants, except for the condition $\rho_0(\boldsymbol{r}) \geq 0$. As x, y and z may have arbitrary positive and negative values in an infinite universe this condition requires $d = e = f = 0$ and a, b, c, ≥ 0. However, this would still yield infinite density for $|x|$, $|y|$ or $|z| \to \infty$ requiring $a = b = c = 0$ which leads back to a homogeneous density.

Discussion of equation (29)

It is remarkable that the amplitudes in (29) are not fixed. An equilibrium is allowed whatever the coefficients B, a, b or c are, provided that the condition (30) is satisfied. This suggests the idea that a certain equilibrium may evolve smoothly to another equilibrium through a continuous series of equilibria, not really by instability, but

rather like a ball which is in a horizontal gutter or on a horizontal plane at any place. However, in the gutter there are still transversal oscillations possible, and if the gutter is inversed to be a ridge there are transversal instabilities. For the ball on the horizontal plane we speak of an indifferent equilibrium in all directions, for all perturbations.

The kind of equilibrium of the form (29) is quite interesting in view of the so-called tessellation of the universe. Indeed during the last decennium it was observed that there were a kind of accumulating "walls" in the universe, where the density of the galaxies is higher than inside the regions surrounded by those 'walls' forming irregular polyhedra (of several hundred million light-years across).

The physical interpretation of the tessellation is based on the interpretation of Jeans' instability. In the customary view of it an accumulation of matter in a gravitating medium will increase provided it is sufficiently large (depending on pressure, etc.). However, if the medium has a depletion of matter at a certain place and of sufficient magnitude, this depletion will be enhanced by the same mechanism of Jeans' instability and its surroundings, its sides, (its "walls"), will increase their density. Hence the tessellation of the universe seems a natural phenomenon and the strengthening of this tessellation a quite natural evolution. Actually equation (29) indicates such a kind of tessellation and the fact that the amplitudes are arbitrary suggests the possibility of a (not too difficult) enhancement of the tessellation structure. This makes the investigation of the stability of the inhomogeneous universe corresponding to (29) very interesting. However, the numerical value of the wavelength corresponding to equation (27) is about 30,000 lightyears or $3 \cdot 10^{20}$km which rather corresponds to the dimensions of galaxies. Using the radiation pressure may raise this with a factor 10^4 (i.e. the tessellation scale) but then the analysis has to be redone including the radiation, which is appropriate when the medium is optically thick (i.e., before matter and radiation became disentangled, that is when the universe had a temperature of about 5000 K).

7. STABILITY ANALYSIS

Perturbing and linearizing the relevant equations (12) - (15) yields

$$\partial_t \rho_1 + div(\rho_0 \boldsymbol{v}_1) = 0, \tag{33}$$

$$\rho_0 \partial_t \boldsymbol{v}_1 = -\boldsymbol{\nabla} p_1 - \rho_0 \boldsymbol{\nabla} \varphi_1 - \rho_1 \boldsymbol{\nabla} \varphi_0(\boldsymbol{r}), \tag{34}$$

$$\Delta \varphi_1 + \Lambda \varphi_1 = 4\pi G \rho_1, \tag{35}$$

$$p_1 = 2K \rho_0 \rho_1, \tag{36}$$

where we recall that ρ_0, p_0 and φ_0 are not constants but given by the equations (27), (29) - (31). Eliminating v_1 and p_1 yields

$$\partial_{tt}^2\rho_1 = 2K\Delta(\rho_0\rho_1) + \rho_0\Delta\varphi_1 + \rho_1\Delta\varphi_0(\boldsymbol{r}) + \boldsymbol{\nabla}\rho_0\cdot\boldsymbol{\nabla}\varphi_1 + \boldsymbol{\nabla}\rho_1\cdot\boldsymbol{\nabla}\varphi_0(\boldsymbol{r})$$

$$= 2K\,(\rho_0\Delta\rho_1 + \rho_1\Delta\rho_0 + 2\boldsymbol{\nabla}\rho_0\cdot\boldsymbol{\nabla}\rho_1) + \rho_0\Delta\varphi_1 + \rho_1\Delta\varphi_0(\boldsymbol{r}) + \boldsymbol{\nabla}\rho_0\cdot\boldsymbol{\nabla}\varphi_1 + \boldsymbol{\nabla}\rho_1\cdot\boldsymbol{\nabla}\varphi_0(\boldsymbol{r}).$$

Using the expressions for ρ_0 and φ_0 we may simplify this to

$$\partial_{tt}^2\rho_1 = 2K\rho_0\Delta\rho_1 + 2K\boldsymbol{\nabla}\rho_0\cdot\boldsymbol{\nabla}\rho_1 + \rho_0\Delta\varphi_1 + \boldsymbol{\nabla}\rho_0\cdot\boldsymbol{\nabla}\varphi_1. \tag{37}$$

The elimination of φ_1 using (35) is not simple especially in view of the cosmological term. It is simple to eliminate ρ_1, yielding a linear partial differential equation of fourth order in φ, however with non-constant coefficients. Hence, we rather try first to handle (35) by taking a Fourier series or integral for ρ_1 and restricting ourselves to the x-dependence only for the sake of convenience. Actually the problem is linear in the perturbed quantities, however, bilinear in equilibrium and perturbed quantities. Thus we consider the sum or integral

$$\rho_1 = \sum_{j=0}^{\infty} a_j e^{\sigma_j t}\cos(jx + \psi_j), \tag{38}$$

where we have included the time dependence explicitly; we have taken it as exponential in view of equation (37). Equations (35) and (38) yield,

$$\varphi_1 = 4\pi G\sum_{j=0}^{\infty}\frac{a_j\, e^{\sigma_j t}}{\Lambda - j^2}\cos(jx + \psi_j) + b\, e^{\sigma t}\cos(\sqrt{\Lambda}x + \psi).$$

Here a_j, ψ_j, σ_j, b, ψ and σ are still arbitrary constants, σ_j and σ may still be complex. Note that if $\sqrt{\Lambda}$ coincides with a particular j we have 'absorption' in equation (35) and the corresponding solution is $bxe^{\sigma t}\cos(\sqrt{\Lambda}x + \psi)$. However, we do not expect it as an appropriate perturbation in an infinite universe. As equation (37) is linear in ρ_1 and φ_1 we may work with a single term of ρ_1:

$$\sigma_j^2\, a_j\cos(jx + \psi_j) = -\left(2K + \frac{4\pi G}{\Lambda - j^2}\right) j^2\, a_j\, \rho_0(\boldsymbol{r})\cos(jx + \psi_j) + \left(2K + \frac{4\pi G}{\Lambda - j^2}\right)$$

$$\times k_0\, j\, a\, a_j\, \sin k_0 x\sin(jx + \psi_j) - b\Lambda\rho_0\cos(\sqrt{\Lambda}x + \psi) + a\, b\, k_0\sqrt{\Lambda}\sin k_0 x\sin(\sqrt{\Lambda}x + \psi),$$

where we have dropped the time factor. Replacing $2K$ by $4\pi G/(k_0^2 - \Lambda)$ and eliminating ρ_0 yields

$$\sigma_j^2\, a_j\cos(jx + \psi_j) = -4\pi G\, j\, a_j\, a\left(\frac{1}{k_0^2 - \Lambda} + \frac{1}{\Lambda - j^2}\right)\Bigg[j\cos k_0\, x\cos(jx + \psi_j)$$

$$-k_0 \sin k_0\, x \sin(jx+\psi_j) + jB\cos(jx+\psi_j)\Big] - a\,b\,\sqrt{\Lambda}\Big[\sqrt{\Lambda}\,\cos k_0\, x\cos(\sqrt{\Lambda}\,x+\psi)$$

$$-k_0 \sin k_0\, x \sin(\sqrt{\Lambda}\,x+\psi) + \sqrt{\Lambda}\,B\cos(\sqrt{\Lambda}x+\psi)\Big]. \tag{39}$$

As $j \neq \pm\sqrt{\Lambda}$ there is no match between terms on the right-hand side. The terms with products of (co)sines can not match the left-hand side and have to vanish.

1. Take $b \neq 0$. For $k_0 = \pm\sqrt{\Lambda}$ and a suitable choice of ψ the term

   $$\sqrt{\Lambda}\,\cos k_0\, x\cos(\sqrt{\Lambda}\,x+\psi) - k_0 \sin k_0\, x \sin(\sqrt{\Lambda}\,x+\psi)$$

   becomes a constant which cannot cancel with another term ($k_0 = \pm\sqrt{\Lambda} \neq j$). However, with another choice of ψ the expression vanishes. However, the last term in (39) cannot match another term as $j \neq \pm\sqrt{\Lambda}$. Hence $B = 0$, which kills the similar term on the right-hand side, too, and there is no match left for the left-hand side.

2. Take $b = 0$. Then the term

   $$j\cos k_0\, x\cos(jx+\psi_j) - k_0 \sin k_0\, x \sin(jx+\psi_j)$$

   still has to disappear. This expression may vanish if $k_0 = \pm j$ but then its coefficient vanishes, too, reducing the right hand side to zero. If the previous expression does not vanish, its coefficient has to vanish (i.e., $a_j = 0$ or $j^2 = k^2$) and this requires the left-hand side to vanish, i.e., $\sigma_j = 0$ or $a_j = 0$.

Finally, only $a_j = a_{k_0}$ with $\sigma_j = \sigma_{k_0} = 0$ remains and we obtain with adapted notation

$$\rho_1 = a_{k_0}\cos(k_0 x+\psi_{k_0}), \quad \varphi_1 = \frac{4\pi G\, a_{k_0}}{\Lambda - k_0^2}\cos(k_0 x+\psi_{k_0}). \tag{40}$$

Hence we have marginal stability and moreover, in view of the arbitrariness of a_{k_0} and ψ_{k_0} the inhomogeneity of ρ may be enhanced or diminished by this perturbation. All this suits very well with the conjecture of indifferent equilibrium above.

8. THE NONLINEAR FOURIER ANALYSIS METHOD USING THE COMBINED ARGUMENT

Now we analyze the case $\Gamma = 2$ using χ as the combined argument. From equations (12) - (15) it follows

$$\omega\rho' + \boldsymbol{k}\cdot(\rho\boldsymbol{v})' = 0, \tag{41}$$

$$\rho\,(\omega + \boldsymbol{v}\cdot\boldsymbol{k})\,\boldsymbol{v}' = -\rho\boldsymbol{k}\varphi' - \boldsymbol{k}p', \tag{42}$$

$$k^2\varphi'' + \Lambda\varphi = 4\pi G\rho, \tag{43}$$

$$p = K\rho^2. \tag{44}$$

Integrating equation (41) with respect to χ we obtain

$$(\omega + \boldsymbol{k}\cdot\boldsymbol{v})\,\rho = \in_g = \omega\rho_0. \tag{45}$$

However, putting now $\boldsymbol{v} = \boldsymbol{0}$ (more general $\boldsymbol{k}\cdot\boldsymbol{v} = 0$) to fix the constant $\in_g$, leads to a surprise as ρ_0 is not a constant in a nonuniform medium. Thus $\in_g$ and ω have to be zero. It follows that the left hand side of (42) vanishes too and the system is reduced at once to its equilibrium equations. This confirms our previous analysis. Starting from an equilibrium (or more generally from a situation with $\boldsymbol{k}\cdot\boldsymbol{v} = 0$) leads to neither an oscillation nor an instability. The equilibrium (or motion) is indifferent to motions based on a Fourier analysis. This result of the preceding section is now generalized to cases with motion for which $\boldsymbol{k}\cdot\boldsymbol{v} = 0$. Further generalization considers $\omega + \boldsymbol{k}\cdot\boldsymbol{v} = 0$. But this requires the velocity parallel to $\boldsymbol{k}$, $v_{\parallel} = -\omega/k$, to be constant which is just an irrelevant parallel displacement and then the situation is the same for $\omega = 0$.

The only motion which is allowed is then with $\boldsymbol{k}\cdot\boldsymbol{v} = 0$, restricting $\boldsymbol{v}$ to be one-dimensional (but of varying magnitude and sign) or possibly two dimensional. If the medium has a density varying like $\cos kx$ the amplitude of this may increase or decrease as this involves only motions perpendicular to x as exemplified in the previous section.

This cuts short at once the further analysis when starting from an inhomogeneous equilibrium. Further investigations should not use the hypothesis concerning χ, but rather a nonperiodic consideration, maybe just using a series development in t or an analysis in which ω depends on the amplitude in the higher order terms.

Note:

(a) The above situation applies to most inhomogeneous media as the continuity equation is of general validity and similarly are the inertia terms in the equation of motion.

(b) One may further generalize the above by considering $(\omega + \boldsymbol{k}\cdot\boldsymbol{v})\rho = \in$. Then the nonlinear Fourier analysis using χ is still fully applicable, e.g., in studying a dynamic inhomogeneous universe (Callebaut and Karugila [27]).

9. CONCLUSION

Usually one (including us) considers a uniform equilibrium as a model of the universe. Here we obtained a *non-uniform* solution: a specific cosinusoidal equilibrium for the polytropic exponent $\Gamma = 2$. Moreover this has another remarkable feature: its amplitude is arbitrary. This suggests that along the equilibria with varying amplitudes the stability is neutral or indifferent. A detailed linearized perturbation analysis confirmed this view. Using the method of nonlinear Fourier analysis with a combined variable confirmed this easily.

This non-uniform equilibrium may be a first step towards the so-called tessellation of the universe: observations indicate that the galaxies accumulate at certain (irregular) polyhedric walls and desert the interior of those polyhedra (cf. Jeans' criterion). However, the numerical values rather indicate dimensions of the order of magnitude of galaxies, which is a good result in itself, but far too small for the tessellation. Extending the calculation to include the radiation pressure (for the situation when radiation and matter were still entangled, i.e. when the temperature of the universe was still above 5000 K) allows to reach the scale of the tessellation: however in that case the analysis has to be repeated including the radiation.

REFERENCES

[1] Chandrasekhar, S. *Hydrodynamic and hydromagnetic stability*, The Clarendon Press, Oxford, 1961.

[2] Malfliet, W., and Hereman, W. The tanh method: II. Perturbation technique for conservative systems, *Physica Scripta*, 54 (1996), 569-675.

[3] Verheest, F. *Waves in Dusty Space Plasmas*, Kluwer Academic Publishers, Dordrecht, Boston and London, 2000.

[4] Pillay, S. R., Rao, N. N. and Bharuthram, R. Linear and non-linear dust acoustic waves in non-ideal dusty plasmas with grain charge fluctuations, In: *Waves in Dusty, Solar, and Space Plasmas* (Edited by F. Verheest, M. Goossens, M. A. Hellberg and R. Bharuthram), Melville, New York, 2000.

[5] Callebaut, D. K. Lineaire en niet-lineaire perturbaties in hydro-, magneto- en gravitodynamica, *Simon Stevin*, University of Ghent, Ghent, (1972), 1-315.

[6] Plateau, J. *Statique expérimentale et théorique des liquides soumis aux seules forces moléculaires* (Volume 1 and 2), Gauthier-Villars, 1873.

[7] Lord Rayleigh. *The theory of sound*, Volume I (1877, 1894), Volume II (1878, 1896). (Reprinted, Dover Publications, New York, 1945).

[8] Khater, A. K., Ibrahim, R. S., Shamardan, A. B. and Callebaut, D. K. Bäcklund transformations and Painlevé analysis: exact solutions for a Grad-Shafranov-type magnetohydrodynamic equilibrium, *IMA Journal of Applied Mathematics* 58 (1), (1997), 51-69.

[9] Khater, A. K., El-Kalaawy, O. H. and Callebaut, D. K. Bäcklund Transformations and Exact Solutions for Alfvén Solitons in a Relativistic Electron-Positron Plasma, *Physica Scripta* 58 (6), (1998), 545-548.

[10] Khater, A.H., Callebaut, D.K. and El-Kalaawy, O.H. Bäcklund transformations and exact solutions for a nonlinear elliptic equation modelling isothermal magnetostatic atmosphere, *IMA Journal of Applied Mathematics* 65 (1), (2000), 97-108.

[11] Khater, A.H., Callebaut, D.K., Ibrahim, R.S. Bäcklund transformations and Painlevé analysis: Exact solutions for the nonlinear isothermal magnetostatic atmospheres, *Physics of Plasmas* 4 (8), (1997), 2853-2863.

[12] Khater, A.H., Callebaut, D.K., Shamardan, A.B., Ibrahim, R.S. Bäcklund transformations and Painlevé analysis: Exact soliton solutions for strongly rarefied relativistic cold plasma, *Physics of Plasmas* 4 (11), (1997), 3910-3922.

[13] Khater, A.H., Ibrahim, R.S., El-Kalaawy, O.H., Callebaut, D. K. Bäcklund transformations and exact soliton solutions for some nonlinear evolution equations of the ZS/AKNS system, *Chaos, Solitons & Fractals* 9 (11), (1998), 1847-1855.

[14] Malfliet, W. Solitary wave solutions of nonlinear wave equations, American Journal of Physics 60 (7), (1992), 650-654.

[15] Malfliet, W., and Wieërs, E. The theory of nonlinear ion-acoustic waves revisited, *Journal of Plasma Physics* 56, (1996), 441-450.

[16] Hereman, W., Banerjee, P.P., Korpel, A., Assanto, G., Van Immerzeele, A. and Meerpoel, A. Exact solitary wave solutions of non-linear evolution and wave equations using a direct algebraic method, *Journal of Physics A - Mathematical & General* 19 (5), (1986), 607-628.

[17] Hereman, W., and Nuseir, A. Symbolic methods to construct exact solutions of nonlinear partial differential equations, Mathematics and Computers in Simulation 43 (1), (1997), 13-27.

[18] Baldwin, D., Göktas, Ü., Hereman, W., Hong, L., Martino, R.S., and Miller, J.C. Symbolic computation of exact solutions expressible in hyperbolic and elliptic functions for nonlinear PDEs, *Journal of Symbolic Computation*, (in press).

[19] Verheest, F. Ion-Acoustic Solitons in Multi-component Plasmas including Negative-Ions at Critical Densities, *Journal of Plasma Physics* 39, Part 1, (1988), 71-79;

[20] Callebaut, D. K. and Karugila, G. K., Nonlinear Fourier Analysis for Unmagnetized Plasma Waves, *Physica Scripta*, 68 (2003), 7-21.

[21] Callebaut, D. K. and Karugila, G. K., Nonlinear Fourier Analysis for Unmagnetized Multi-species Plasma Waves, *Journal of Mathematical Physics. Submitted.*

[22] Callebaut, D.K., Second order magnetodynamic stability of an infinite cylinder, *Plasma Physics*, 10 (1968), 440.

[23] Callebaut, D. K., A Kind of Nonlinear Analysis, In: *Proceedings of the 10th European School on Plasma Physics* (Edited by N. L. Tsintsadze), World Scientific, Singapore, New Jesey, London and Hong Kong, (1991), 286-302.

[24] Chandrasekhar, S. and Fermi, E. Problems of gravitational stability in the presence of magnetic fields, *Astrophysical Journal* 118, (1953), 116-141.

[25] Callebaut, D. K. Introduction to Stability Problems in Fluid Mechanics and Plasma Physics, *Lecture Notes*, UIA Press, Antwerp, (1986), 1-190. (Revised 2004).

[26] Chandrasekhar, S. *An introduction to the study of stellar structure*, Chicago University Press, Chicago, 1939. (Dover edition, 1957).

[27] Callebaut, D. K. and Karugila, G. K., "Evolving Non-homogeneous Universe, Jeans' Criterion and Tessellation", in *The Book of Abstracts for the 59th Dutch Astronomical Conference, 26-28 May 2004*, Vlieland, The Netherlands, (2004), 38.

ON QUASI EINSTEIN AND SPECIAL QUASI EINSTEIN MANIFOLDS

U.C.De and G. C. Ghosh
Department of Mathematics, University of Kalyani
Kalyani-741235, W. B., India
e-mail: ucde@klyuniv.ernet.in

1. INTRODUCTION

The notion of quasi Einstein manifold was introduced by Chaki and Maity [1]. A non-flat Riemannian manifold $(M^n, g)(n > 2)$ is defined to be a quasi Einstein manifold if its Ricci tensor S of type $(0, 2)$ is not identically zero and satisfies the condition

$$S(X, Y) = a\, g(X, Y) + b\, A(X)A(Y) \tag{1}$$

where a, b are scalars of which $b \neq 0$ and A is a non-zero 1-form such that

$$g(X, U) = A(X) \tag{2}$$

for all vector fields X; U being a unit vector field. If $b = 0$,then the manifold reduces to an Einstein manifold. In such a case a, b are called associated scalars. A is called the associated 1-form and U is called the generator of the manifold. An n-dimensional manifold of this kind is denoted by the symbol $(QE)_n$.

A Riemannian manifold of quasi-constant curvature was given by Chen and Yano [2] as a conformally flat manifold with the curvature tensor $'R$ of type $(0, 4)$ satisfies the condition

$$\begin{aligned} {}'R(X, Y, Z, W) = a[g(Y, Z)g(X, W) - g(X, Z)g(Y, W)] \\ +b[g(X, W)T(Y)T(Z) - g(X, Z)T(Y)T(W) \\ +g(Y, Z)T(X)T(W) - g(Y, W)T(X)T(Z)] \end{aligned} \tag{3}$$

where ${}'R(X, Y, Z, W) = g(R(X, Y)Z, W)$, R is the curvature tensor of type $(1, 3)$, a, b are scalar functions and ρ is a unit vector field defined by

$$g(X, \rho) = T(X). \tag{4}$$

It can be easily seen that if the curvature tensor $'R$ is of the form (3), then the manifold is conformally flat. On the otherhand, Vranceanu [3] defined the notion of

almost constant curvature by the same expression as (3) without assuming conformally flat manifold. Later Mocanu [4] shows that the manifold introduced by Chen and Yano [2] are manifolds of the same type introduced by Vranceanu [3]. Hence a Riemannian manifold is said to be of quasi-constant curvature if the curvature tensor $'R$ satisfies the relation (3). If $b = 0$, then the manifold reduces to a manifold of constant curvature.

A quasi Einstein manifold is said to be a special quasi Einstein manifold if the associated scalar a is constant and such a manifold is denoted by $S(QE)_n$.

In the present paper at first we state some examples of a quasi Einstein manifold $(QE)_n$. Next we prove a theorem for the existence of a $(QE)_n$. Section 4 deals with the hypersurfaces of a Euclidean space. In section 5 we give the physical interpretation of a $(QE)_n$. In the next section we construct a metric of special quasi Einstein manifold $S(QE)_n$. Then we study conformally flat $S(QE)_n$. We prove that a conformally flat $S(QE)_n$ can be expressed as a Warped Product $I \times_{e^q} M^*$ where M^* is an Einstein manifold. As an application we prove that a conformally flat special quasi Einstein manifold is the Robertson-Walker space time.

2. EXAMPLES OF A QUASI EINSTEIN MANIFOLD

Example 2.1. A manifold of quasi-constant curvature defined by (3) is a quasi Einstein manifold.

Putting $X = W = e_i$ in (3) where $\{e_i\}$ is an orthonormal basis of the tangent space at each point of the manifold and taking summation over i, $1 \leq i \leq n$, we get

$$S(Y, Z) = [a(n-1) + b]g(Y, Z) + b(n-2)T(Y)T(Z) \tag{5}$$

Hence the manifold is a quasi Einstein manifold.

Example 2.2. De and Ghosh [5] studied conformally flat weakly Ricci symmetric manifold and prove that such a manifold is a manifold of quasi-constant curvature. Hence a conformally flat weakly Ricci symmetric manifold is a quasi Einstein manifold.

Example 2.3. A special para-sasakian manifold with vanishing D-concircular curvature tensor V in the sense of Chuman [6] is a quasi Einstein manifold.

Example 2.4. A semi-Riemannian manifold (M, g) is said to be Pseudo symmetric in the sence of Deszcz [7] if the Riemannian curvature tensor R satisfies the equality

$R.R = fQ(g, R)$
where f is some function and the tensors $R.R$ and $Q(g, R)$ are defined by

$$\begin{aligned}(R(X,Y).R)(U,V)W = R(X,Y)R(U,V)W - R(R(X,Y)U,V)W \\ -R(U,R(X,Y)V)W - R(U,V)R(X,Y)W\end{aligned} \tag{6}$$

and

$$\begin{aligned}Q(g,R) = (X \wedge Y)R(U,V)W - R((X \wedge Y)U,V)W \\ -R(U,(X \wedge Y)V)W - R(U,V)(X \wedge Y)W\end{aligned} \tag{7}$$

for all $X, Y, U, V, W \in \chi(M)$. Here the endomorphism is defined by

$$(X \wedge Y)Z = g(Y,Z)X - g(X,Z)Y.$$

It is known Deszcz [7] that the Ricci tensor S of a 3-dimensional pseudosymmtric semi-Riemanian manifold satisfies at every point $x \in M$ the relation

$$S(X,Y) = \alpha g(X,Y) + \beta u(X)u(Y), \quad \alpha, \ \ \beta \in R$$

and u is a non-zero 1-form, i.e., (M, g) is a quasi Einstein manifold. Hence a 3-dimensional pseudo symmetric semi-Riemannian manifold in the sense of Deszcz [7] is a quasi Einstein manifold.

3. EXISTENCE THEOREM OF $(QE)_n$

In this section we prove the following:
Theorem 3.1. If the Ricci tensor S of a Riemannian manifold satisfies the relation

$$S(Y,Z)S(X,W) \ - \ S(X,Z)S(Y,W) = \rho[g(Y,Z)g(X,W) - g(X,Z)g(Y,W)], \tag{8}$$

then the manifold is a quasi Einstein manifold.
Proof : If the rank of the Ricci tensor is equal to 1, then the relation (8) satisfies trivially. Hence we assume that the rank of the Ricci tensor is > 1, so ρ is non-zero.

Let U be a vector field defined by

$$g(X,U) = A(X) \qquad \forall\, X.$$

Putting $X = W = U$ in the above relation we have

$$S(U,U)S(Y,Z) - S(U,Z)S(Y,U) = \rho(g(U,U)g(Y,Z) - g(U,Z)g(Y,U))$$

$$\begin{aligned}\text{or,} \quad \overline{\gamma}S(Y,Z) - A(QY)A(QZ) = \rho(|U|^2 g(Y,Z) - A(Z)A(Y)) \\ \text{where} \ \ S(U,U) = \overline{\gamma} \text{and} \ \ A(QY) = g(QY,U) = S(Y,U)\end{aligned}$$

$$\text{or,} \quad S(Y,Z) = \gamma A(QY)A(QZ) + \rho\gamma(|U|^2 g(Y,Z) - A(Y)A(Z))$$
$$\text{where } \gamma = \tfrac{1}{\overline{\gamma}}$$
$$\text{or,} \qquad S(Y,Z) = \gamma B(Y)B(Z) + \rho\gamma(|U|^2 g(Y,Z) - A(Y)A(Z)) \tag{9}$$
$$\text{where } \; B(Y) = A(QY)$$

Again putting $X = U$ in (8) we have

$$S(U,W)S(Y,Z) - S(U,Z)S(Y,W) = \rho(A(W)g(Y,Z) - A(Z)g(Y,W))$$

$$\text{i.e.,} \qquad B(W)S(Y,Z) - B(Z)S(Y,W) = \rho(A(W)g(Y,Z) - A(Z)g(Y,W)). \tag{10}$$

Similarly we can write

$$B(X)S(Y,Z) - B(Z)S(Y,X) = \rho(A(X)g(Y,Z) - A(Z)g(Y,X)). \tag{11}$$

Using (9) in (11) we get

$$\begin{gathered}\rho\gamma\left(|U|^2 B(X)g(Y,Z) - |U|^2 B(Z)g(Y,X) - B(X)A(Y)A(Z)\right. \\ \left. +B(Z)A(Y)A(X)\right) = \rho\left(A(X)g(Y,Z) - A(Z)g(Y,X)\right).\end{gathered} \tag{12}$$

Putting $Y = Z = e_i$ in (12) and taking $g(U,U) = |U|^2 = 1$ we obtain

$$B(X) = \overline{\gamma}A(X). \tag{13}$$

Putting the value of B(X) in (9) we have

$$S(Y,Z) = \rho\gamma g(Y,Z) + (\overline{\gamma} - \rho\gamma)A(Y)A(Z)$$
$$\text{i.e.,} \qquad S(Y,Z) = ag(Y,Z) + bA(Y)A(Z) \qquad \text{where } a = \rho\gamma \text{ and } b = \overline{\gamma} - \rho\gamma$$

which shows that the manifold is a quasi Einstein manifold.

4. HYPERSURFACES OF A EUCLIDEAN SPACE

LetM^n be a hypersurface of a Euclidean space E^{n+1}, such that the tensor induced by the metric of E^{n+1} is the metric tensor ofM^n. The Gauss equation ofM^n in E^{n+1} can be written as follows

$$\tilde{g}(\tilde{R}(X,Y)Z,W) = \tilde{g}(H(X,W),H(Y,Z)) - \tilde{g}(H(Y,W),H(X,Z)) \tag{14}$$

where $\tilde{R}$ is the Riemannian curvature tensor corresponding to the induced metric $\tilde{g}$, H is the second fundamental tensor ofM^n and X, Y, Z, W are vector fields tangent to M^n.

If A is the (1,1) tensor corresponding to the normal valued second fundamental tensor H, then we have Chen [8].

$$\tilde{g}(A_\xi(X),Y) = g(H(X,Y),\xi) \tag{15}$$

where ξ is the unit normal vector field and X, Y are tangent vector field.

Let H_ξ be the symmetric (0,2) tensor associated with A_ξ in the hypersurface defined by

$$\tilde{g}(A_\xi(X), Y) = H_\xi(X, Y). \tag{16}$$

A hypersurface of a Riemannian manifold (M^n, g) is called quasi-umbilical (Chen [8]) if its second fundamental tensor has the form

$$H_\xi(X, Y) = \alpha\tilde{g}(X, Y) + \beta\omega(X)\omega(Y) \tag{17}$$

where ω is a 1-form, the vector field corresponding to the 1-form ω is a unit vector field, and α, β are scalars. If$\alpha = 0$ (res. $\beta = 0$ or $\alpha = \beta = 0$) holds then it is called cylindrical (res. umbilical or geodesic).

Now from (15), (16) and(17) we obtain

$$g(H(X, Y), \xi) = \alpha g(X, Y)g(\xi, \xi) + \beta\omega(X)\omega(Y)g(\xi, \xi)$$

which implies that

$$H(X, Y) = \alpha g(X, Y)\xi + \beta\omega(X)\omega(Y)\xi, \tag{18}$$

since ξ is the only unit normal vector field.

Let us suppose that the hypersurface is a quasi-umbilical. Then from (18) and (14) it follows that

$$\begin{aligned}\tilde{g}(\tilde{R}(X, Y)Z, W) = \alpha^2 &(g(X, W)g(Y, Z) - g(Y, W)g(X, Z)) \\ +\alpha\beta &(g(X, W)\omega(Y)\omega(Z) + g(Y, Z)\omega(X)\omega(W) \\ -g(Y, W)\omega(X)\omega(Z) &- g(X, Z)\omega(Y)\omega(W)).\end{aligned} \tag{19}$$

From (19) we get on contraction

$$\tilde{S}(Y, Z) = \big(\alpha^2(n-1) + \alpha\beta)g(Y, Z) + \alpha\beta(n-2)\omega(Y)\omega(Z)\big)$$

Thus we can state the following :
Theorem 4.1. A quasi-umbilical hypersurface of a Euclidean space is a quasi Einstein manifold.

Let us consider a conformally flat hypersurface of a Euclidean space. It is known Schouten [9] that if a hypersurface of a conformally flat space is conformally flat,then

the hypersurface is quasi-umbilical.
Hence from the above theorem we get the following:

Corollary 4.1. A conformally flat hypersurface of a Euclidean space E^n is quasi Einstein.

5. PHYSICAL INTERPRETATION OF $(QE)_n$

It is known O'Neill [10] that for a perfect fluid spacetime of general relativity, the Einstein field equation without cosmological constant is of the form

$$S(X,Y) - \frac{r}{2}g(X,Y) = (\rho + p)U(X)U(Y) + pg(X,Y) \tag{20}$$

where U is a non-zero 1-form, ρ and p are the energy density and the isotropic pressure of the fluid respectively.
The above equation (20) can be written in the form

$$S(X,Y) = \alpha g(X,Y) + \beta U(X)U(Y)$$

where $\alpha = (\frac{r}{2} + p)$ and $\beta = \rho + p$, $\alpha,\ \beta$ are scalars and $\beta = \rho + p \neq 0$.

So we conclude that a perfect fluid spacetime of general relativity is a four dimensional semi-Riemannian quasi Einstein manifold of Lorentz signature $(-\ +\ +\ +)$ and whose associated scalars are $\frac{r}{2} + p$ and $\rho + p$ respectively.

6. METRIC OF A SPECIAL QUASI EINSTEIN MANIFOLD

A quasi Einstein manifold $(QE)_n$ is said to be a special quasi Einstien manifold $S(QE)_n$ if the associated scalar a is constant. In this section we construct a metric of the special quasi Einstein manifold.

Let E^5 be a Euclidean space with cartesian coordinates $(x^1,\ x^2,\ y^1,\ y^2,\ z)$ or $(x^\alpha,\ y^\alpha,\ z)$ $(\alpha = 1,\ 2)$.

Let us consider

$$A = \sqrt{\frac{3}{2b}}(dz - \sum_{\alpha=1}^{2} y^\alpha dx^\alpha), \text{(where b is a scalar)} \tag{21}$$

If we put

$$x^{\alpha^*} \equiv x^{2+\alpha} = y^\alpha, \quad x^\triangle = z, \quad \triangle = 5 \tag{22}$$

we have from (21) that

$$A_i = \left(-\sqrt{\frac{3}{2b}}y^\alpha, 0, \sqrt{\frac{3}{2b}}\right) \tag{23}$$

Now, we consider a symmetric tensor field in E^5 defined by

$$g_{ij} = \begin{pmatrix} \frac{1}{4}(1+(y^1)^2) & \frac{y^1y^2}{4} & 0 & 0 & -\frac{1}{4}y^1 \\ \frac{1}{4}y^1y^2 & \frac{1}{4}(1+(y^2)^2) & 0 & 0 & -\frac{1}{4}y^2 \\ 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 0 \\ -\frac{1}{4}y^1 & -\frac{1}{4}y^2 & 0 & 0 & \frac{1}{4} \end{pmatrix}. \tag{24}$$

Then (g_{ij}) defines a positive definite Riemannian metric. The contravariant components of the tensor (g^{ij}) are given by

$$g^{ij} = \begin{pmatrix} 4 & 0 & 0 & 0 & 4y^1 \\ 0 & 4 & 0 & 0 & 4y^2 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 4y^1 & 4y^2 & 0 & 0 & 4(1+(y^1)+(y^2)) \end{pmatrix}. \tag{25}$$

We find out the Christoffel symbols, by means of (24) and

$$[ij, r] = \frac{1}{2}[\frac{\partial g_{jr}}{\partial x^i} + \frac{\partial g_{ir}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^r}].$$

We can varify that

$$\left.\begin{array}{l} [\alpha\ \beta^*, \gamma] = \frac{1}{8}(\delta_{\gamma\beta}y^\alpha + \delta_{\alpha\beta}y^\gamma) \\ [\alpha\beta, \gamma^*] = -\frac{1}{8}(\delta_{\gamma\alpha}y^\beta + \delta_{\beta\gamma}y^\alpha) \\ [\alpha\ \beta^*,\ \triangle] = -\frac{\delta_{\alpha\gamma}}{8}, \quad [\alpha\ \triangle,\ \gamma^*] = \frac{\delta_{\alpha\gamma}}{8}, \\ [\alpha^*\ \triangle,\ \gamma] = -\frac{\delta_{\alpha\gamma}}{8}, \\ \\ \text{the other components are zero.} \end{array}\right\} \tag{26}$$

Equations (25) and (26) with $\left\{ \begin{array}{cc} h \\ j & i \end{array} \right\} = g^{hr}[j\ i, r]$ implies that

$$\left.\begin{array}{l} \left\{ \begin{array}{cc} \mu \\ \alpha & \beta^* \end{array} \right\} = \frac{1}{2}\delta_{\mu\beta}y^{\alpha}, \quad \left\{ \begin{array}{cc} \mu^* \\ \alpha & \beta \end{array} \right\} = -\frac{1}{2}\left(\delta_{\alpha\mu}y^{\beta} + \delta_{\mu\beta}y^{\alpha}\right), \\ \left\{ \begin{array}{cc} \mu^* \\ \alpha & \triangle \end{array} \right\} = \frac{1}{2}\delta_{\alpha\mu}, \quad \left\{ \begin{array}{cc} \triangle \\ \alpha & \beta^* \end{array} \right\} = \frac{1}{2}\left(y^{\alpha}y^{\beta} - \delta_{\alpha\beta}\right) \\ \left\{ \begin{array}{cc} \triangle \\ \triangle & \beta^* \end{array} \right\} = -\frac{1}{2}y^{\beta}, \quad \left\{ \begin{array}{cc} \mu \\ \alpha^* & \triangle \end{array} \right\} = -\frac{1}{2}\delta_{\alpha\mu}, \\ \text{the other components are zero.} \end{array}\right\} \tag{27}$$

After straightforward calculations, we obtain the independent components of the curvature tensor R_{kjih} as follows

$$\left.\begin{array}{l} R_{\delta\gamma\beta\alpha} = \frac{1}{16}\left(\delta_{\alpha\beta}y^{\beta}y^{\gamma} + \delta_{\beta\gamma}y^{\alpha}y^{\delta} - \delta_{\alpha\gamma}y^{\beta}\delta - \delta_{\beta\delta}y^{\alpha}y^{\gamma}\right), \\ R_{\delta^*\gamma^*\beta\alpha} = \frac{1}{16}\left(\delta_{\alpha\gamma}\delta_{\beta\delta} - \delta_{\alpha\delta}\delta_{\beta\gamma}\right), \\ R_{\delta\gamma\triangle\alpha} = \left(\delta_{\alpha\gamma}y^{\delta} - \delta_{\alpha\delta}y^{\gamma}\right) \\ R_{\delta\gamma^*\beta^*\alpha} = \frac{1}{16}\left(\delta_{\beta\gamma}y^{\alpha}y^{\delta} - 2\delta_{\alpha\beta}\delta_{\gamma\delta} - \delta_{\alpha\gamma}\delta_{\beta\gamma}\right) \\ R_{\delta\triangle\triangle\alpha} = \frac{1}{16}\delta_{\alpha\delta}, \\ R_{\triangle\gamma^*\triangle\alpha^*} = -\frac{1}{16}\delta_{\alpha\gamma}, \\ R_{\delta^*\triangle\beta^*\alpha} = \frac{1}{16}\delta_{\beta\delta}y^{\alpha}, \\ \text{the other independent components are zero.} \end{array}\right\} \tag{28}$$

According to (25) and (28), the Ricci tensor has the following components.

$$\left.\begin{array}{l} R_{\delta\alpha} = -\frac{1}{2}\left(\delta_{\alpha\delta} - 2y^{\alpha}y^{\delta}\right), \quad R_{\delta^*\alpha} = 0, \\ R_{\delta^*\alpha^*} = -\frac{\delta_{\alpha\delta}}{2}, \quad R_{\triangle\triangle} = 1, \quad R_{\gamma^*\triangle} = 0, \\ R_{\gamma\triangle} = -y^{\gamma}. \end{array}\right\} \tag{29}$$

Substituting (24) and (23) into (29), we have

$$R_{ij} = -2g_{ij} + bA_iA_j. \tag{30}$$

Hence E^5 with the metric (24) is a special quasi Einstein manifold.

7. CONFORMALLY FLAT S(QE)$_n$

In this section we assume that the manifold $S(QE)_n$ is conformally flat. Then $div\ C = 0$ where C denotes the Weyl's conformal curvature tensor and $'div'$ denotes divergence.
Hence we have

$$(\nabla_X S)(Y,Z) - (\nabla_Z S)(Y,X) = \frac{1}{2(n-1)}[g(Y,Z)dr(X) - g(X,Y)dr(Z)]. \tag{31}$$

Contracting (1) we get

$$r = an + b. \tag{32}$$

From (32) it follows that

$$dr(X) = db(X), \quad since\ a\ is\ a\ constant. \tag{33}$$

(1) implies that

$$(\nabla_Z S)(X,Y) = db(Z)A(X)A(Y) + b[(\nabla_Z A)(X)A(Y) + A(X)(\nabla_Z A)(Y)]. \tag{34}$$

Substituting (34) in (31) and using (33) we get

$$\begin{aligned} &dr(X)A(Z)A(Y) + b[(\nabla_X A)(Z)A(Y) + A(Z)(\nabla_X A)(Y)] \\ &-dr(Z)A(Y)A(X) - b[(\nabla_Z A)(Y)A(X) + A(Y)(\nabla_Z A)(X)] \\ &= \tfrac{1}{2(n-1)}[g(Y,Z)dr(X) - g(X,Y)dr(Z)]. \end{aligned} \tag{35}$$

Puttint $Y = Z = e_i$ in the above expression where $\{e_i\}$ is an orthonormal basis of the tangent space at each point of the manifold and taking summation over i, $1 \leq i \leq n$, we get

$$\frac{1}{2}dr(X) = dr(\rho)A(X) + b(\nabla_\rho A)(X) + b(\nabla_{e_i} A)(e_i)A(X). \tag{36}$$

Again putting $Y = Z = \rho$ in (35) yields

$$b(\nabla_\rho A)(X) = \frac{2n-3}{2(n-1)}[dr(X) - dr(\rho)A(X)]. \tag{37}$$

Substituting (37) in (36) we get

$$\frac{(n-2)}{2(n-1)}\, dr(X) + \frac{1}{2(n-1)}dr(\rho)A(X) + b(\nabla_{e_i} A)(e_i)A(X) = 0. \tag{38}$$

Now putting$X = \rho$ in (36) yields

$$b(\nabla_{e_i} A)(e_i) = -\frac{1}{2}dr(\rho). \tag{39}$$

From (38) and (39) it follows that

$$dr(X) = dr(\rho)A(X). \tag{40}$$

Putting $Y = \rho$ in (35) and using (40) we obtain

$$(\nabla_Z A)(X) - (\nabla_X A)(Z) = 0 \tag{41}$$

which implies that the 1-form A is closed. From (37) we get by virtue of (40)

$$(\nabla_\rho A)(Z) = 0, \quad since\ b{\neq}0. \tag{42}$$

Now we consider the scalar function

$$f = \frac{1}{2(n-1)} \frac{dr(\rho)}{b}.$$

We have

$$\nabla_X f = \frac{1}{2(n-1)} \frac{dr(\rho)}{b^2} dr(X) + \frac{1}{2(n-1)b} d^2r(\rho, X). \tag{43}$$

On the otherhand, (40) implies

$$d^2r(Y, X) = d^2r(\rho, Y)A(X) + dr(\rho)(\nabla_X A)(Y)$$

from which we get

$$d^2r(\rho, Y)A(X) = d^2r(\rho, X)A(Y). \tag{44}$$

Putting $X = \rho$ in (44) it follows that

$$\begin{aligned} &d^2(\rho, Y) = d^2r(\rho, \rho)A(Y) \\ &= hA(Y), \quad where\ h\ is\ a\ scalar\ function. \end{aligned}$$

Thus

$$\nabla_X f = \mu A(X) \tag{45}$$

where $\mu = \frac{1}{2(n-1)b}[h + \frac{dr(\rho)}{b} dr(\rho)]$, using (40).
Using (45) it is easy to show that $\omega(X) = \frac{1}{2(n-1)} \frac{dr(\rho)}{b} A(X) = fA(X)$ *is closed.*
In fact,

$$d\omega(X, Y) = 0.$$

Using (40) and (41) in (37) we get

$$\begin{aligned} &b[A(Z)(\nabla_X A)(Y) - A(X)(\nabla_Z A)(Y)] \\ &= \tfrac{dr(\rho)}{2(n-1)}[g(Y, Z)A(X) - g(X, Y)A(Z)] \end{aligned}$$

Now putting $Z = \rho$ in the above expression yields

$$(\nabla_X A)(Y) = \frac{1}{2(n-1)} \frac{dr(\rho)}{b}[A(X)A(Y) - g(X,Y)]. \tag{46}$$

Thus (46) can be written as follows:

$$(\nabla_X A)(Y) = -fg(X,Y) + \omega(X)A(Y) \tag{47}$$

where ω is closed. But this means that the vector field ρ corresponding to the 1-form A defined by $g(X,\rho) = A(X)$ is a proper concircular vector field (Schouten [11], Yano [12]). Hence we can state the following :

Theorem 7.1. In a conformally flat $S(QE)_n$ $(n > 3)$, the vector field ρ defined by $g(X,\rho) = A(X)$ is a proper concircular vector field.

It is known Adati [13] that if a conformally flat manifold (M^n,g) $(n > 3)$ admits a proper concircular vector field, then the manifold is a subprojective manifold in the sense of Kagan. Since a conformally flat $S(QE)_n$ admits a proper concircular vector field, namely the vector field ρ, we can state as follows:

Theorem 7.2. A conformally flat $S(QE)_n$ is a subprojective manifold in the sense of Kagan.

Yano [14] proved that in order that a Riemannian space admits a concircular vector field, it is necessary and sufficient that there exists a coordinate system with respect to which the fundamental quadratic differential form may be written in the form

$$ds^2 = (dx^1)^2 + e^q g^*_{\alpha\beta} dx^\alpha dx^\beta$$

where $g^*_{\alpha\beta} = g^*_{\alpha\beta}(x^\gamma)$ are the functions of x^γ only $(\alpha,\ \beta.\ \gamma,\ \delta = 2,\ 3,\ ...,\ n)$ and $q = q(x^1) \neq$constant is a function of x^1 only. Thus if a $S(QE)_n$ is conformally flat i.e., if is satisfies (31), it is a warped product $IX_{e^q}M^*$, where (M^*, g^*) is an (n-1)-dimensional Riemannian manifold. Gebarowski [15] proved that warped product $IX_{e^q}M^*$ satisfies (2.1) if and only ifM^* is an Einstein manifold. Thus if $S(QE)_n$ satisfies (2.1), it must be a warped product $IX_{e^q}M^*$ where M^* is an Einstein manifold. Thus we can state the following result:

Theorem 7.3. A conformally flat $S(QE)_n$ $(n > 3)$ can be expressed as a warped product $IX_{e^q}M^*$ where M^* is an Einstein manifold.

8. SPECIAL CONFORMALLY FLAT S(QE)$_n$ $(n > 3)$

The notion of a special conformally flat manifold which generalizes the notion of subprojective manifold was introduced by Chen and Yano [16]. According to them a conformally flat manifold is said to be a special conformally flat manifold if the tensor H of type (0,2) defined by

$$H(X,Y) = -\frac{1}{(n-2)}S(X,Y) + \frac{r}{2(n-1)(n-2)}g(X,Y) \tag{48}$$

is expressible in the form

$$H(X,Y) = -\frac{\alpha^2}{2}g(X,Y) + \beta(X\alpha)(Y\alpha) \tag{49}$$

where α and β are two scalars such that α is positive. In virtue of (1) we can express (48) as

$$H(X,Y) = [\,-\frac{a}{n-2} + \frac{r}{2(n-1)(n-2)}]g(X,Y) - \frac{b}{n-2}A(X)A(Y). \tag{50}$$

We now put

$$\begin{aligned}\alpha^2 &= \tfrac{2a}{n-2} - \tfrac{r}{(n-1)(n-2)} \\ &= \tfrac{a(n-2)-b}{(n-1)(n-2)}\end{aligned} \tag{51}$$

Then

$$2\alpha(X\alpha) = -\frac{dr(\rho)}{(n-1)(n-2)}A(X), \qquad \textit{using } (40). \tag{52}$$

Hence (50) can be expressed as

$$H(X,Y) = -\frac{\alpha^2}{2}g(X,Y) + \beta A(X)A(Y) \tag{53}$$

where $\beta = \frac{4b(r-2an+2a)(n-1)}{\lambda^2}$, $\lambda = dr(\rho)$.
Suppose that $a(n-2) - b > 0$, then α is not zero. Hence from (51) it follows that α may be taken as positive. From (53) we conclude that the manifold under consideration is a special conformally flat manifold.

It is known from a theorem of Chen and Yano [5] that every simply connected special conformally flat manifold can be isometrically immersed in a Euclidean space E^{n+1} as a hypersurface.
We can therefore state the following :
Theorem 8.1. Every simply connected conformally flat $S(QE)_n$ $(n > 3)$ satisfying $a(n-2) - b > 0$ can be isometrically immersed in a Euclidean space E^{n+1} as a hypersurface.

9. PHYSICAL INTERPRETATION OF S(QE)$_n$

In this section we consider conformally flat $S(QE)_n$ $(n > 3)$ spacetime. By a spacetime, we will mean a 4-dimensional semi-Riemannian manifold endowed with Lorentz metric of signature $(-\ +\ +\ +)$. By the similar proof as in Theorem 7.3 we get a conformally flat $S(QE)_n$ $(n > 3)$ spacetime can be expressed as a warped product $IX_{e^q}M^*$ where M^* is an Einstein manifold. Since we consider a 4-dimensional manifold, M^* is a 3-dimensional Einstein manifold. It is known that a 3-dimensional Einstein manifold is a manifold of constant curvature. Hence a conformally flat $S(QE)_n$ spacetime is the warped product $IX_{e^q}M^*$, where M^* is a manifold of constant curvature. But such a warped product is the Robertson-Walker spacetime O'Neill [10]
Thus we have the following

Theorem 9.1. A conformally flat special quasi Einstein manifold is the Robertson-Walker spacetime.

ACKNOWLEDGEMENT

The first author would like to express his gratitude to the organizing committee and Prof. Adnan Al-Aqeel for financial support to attend the conference and local hospitality.

REFERENCES

[1] Chaki, M. C. and Maity, R. K. On quasi Einstein manifold, Publ.Math.Debrecen 57(2000),297-306

[2] Chen, B. Y. and Yano, K. Hypersurfaces of a conformally flat space, Tensor, N.S.,26(1972), 318-322.

[3] Vranceanu, Gh. Lecons des Geometrie Differential, Vol.4, Ed.de l'Academie, Bucharest, 1968.

[4] Mocanu, A. L. Les Variétés a curbure quasi-constant de type vrânceanu, Lucr. conf. Nat. de. Geom.Si Top., Trigoviste, 1987.

[5] De, U. C. and Ghosh, S. K. On weakly Ricci-symmetric spaces, Publ. Math. Debrecen 60(2002), 201-208.

[6] Chuman, G. D-conformal changes in para-Sasakian manifold, Tensor, N. S., 39(1982), 117-123.

[7] Deszcz, R. On pseudosymmetric spaces, Bull.Belg.Math.Soc.SerA, 44(1992), 1-34.

[8] Chen, B. Y. Geometry of submanifolds, Marcel Dekker.Inc.New York, 1973.

[9] Schouten, J. A. $\ddot{U}$ber die konforme Abbildung n-dimensionaler Mannigfaltigkeiten mit quadratischer Massbestimmung auf eine Mannigfaltigkeit mit euklidischer Massbestimmung, Math. Z., 11(1921), 58-88.

[10] O'Neill, B. Semi-Riemannian Geometry, Academic Press, Inc. 1983.

[11] Schouten, J. A. Ricci-Calculus, Springer, Berlin, 1954.

[12] Yano, K. Concircular geometry, I, Proc. Imp. Acad. Tokyo, 16(1940), 195-200.

[13] Adati, T. On subprojective spaces, III, Tohoku Math. J., 3(1951), 343-358.

[14] Yano, K. On the torseforming direction in Riemannian spaces, Proc. Imp. Acad. Tokyo, 20(1944), 340-345.

[15] Gebarowski, A. Nearly conformally symmetric warped product manifolds, Bulletin of the Institute of Mathematics Academia Sinica, 20: 4(1992), 359-371.

[16] Chen, B. Y. and Yano, K. Special conformally flat spaces and canal hypersurfaces, Tohoku Math. J., 25(1973), 177-184.

FOUR MAJOR DISCOVERIES IN APPLIED MATHEMATICS DURING THE SECOND HALF OF THE TWENTIETH CENTURY

L. Debnath
Department of Mathematics
University of Texas–Pan American, Edinburg, Texas 78539, USA
email: debnathl@panam.edu

"... the progress of physics will to a large extent depend on the progress of nonlinear mathematics, of method to solve nonlinear equations ... and therefore we can learn by comparing different nonlinear problems."

WERNER HEISENBERG

"... as Sir Cyril Hinshelwood has observed ... fluid dynamicists were divided into hydraulic engineers who observed things that could not be explained and mathematicians who explained things that could not be observed."

JAMES LIGHTHILL

Abstract

In the 1960s, the soliton, (Zabusky and Kruskal [1]), interaction of solitons and the Inverse Scattering Transform (Gardner et al. [2]) were discovered. These discoveries have led to an extensive study of solitons and the mathematical theory of nonlinear waves and their applications. In the 1970s, fractals and fractal dimensions were first introduced by Mandelbrot [3, 4] in order study the geometry of irregular curves and surfaces. His book [5] on *The Fractal Geometry of Nature* contains both the elementary ideas and an unusually wide range of new and advanced topics including multifractals, dynamical systems and chaotic attractors. Many beautiful fractals have been drawn with the aid of a computer or graphical analysis. Mandelbrot also constructed self similar fractals by an iteractive process using an initial and standard polygon. Henri Poincaré (1856-1912) discovered the theory of what is now called *dynamical systems*. As far as fractals are concerned, iterative mappings in the plane – so called the *Poincaré mappings* – are of special importance. Recent developments in complex dynamics produced many totally unexpected results. It turns out that

computational experiments on the quadratic complex transformation generate new geometric structures that are very complex and extremely beautiful. In the 1980s, wavelets and wavelet transforms were discovered by Morlet et al. [6, 7] in order to provide a new mathematical tool for seismic wave analysis. Following this discovery, considerable attention has been given to the mathematical theory of wavelets and wavelet transforms with applications to signal processing, image processing and biomedical engineering. A new and remarkable idea of multiresolution analysis (MRA) was first formulated by Mallat [8] in the context of wavelet analysis. This idea is related to the study of signals or images at different levels of resolution – almost like a pyramid, and deals with a general formalism for construction of an orthogonal basis of wavelets. Mallat's remarkable work has been the major source of many new developments in wavelet analysis and its wide variety of applications. In the 1990s, the compacton (the soliton with compact support) and interaction of compactons were discovered by Rosenau and Hyman [9]. As an application, new intrinsic localized modes in anharmonic crystals were discovered by Sievers and Takeno [10] and Page [11]. It was shown that anharmonicity is fully responsible for the existence of the new intrinsic localized modes in an harmonic quantum crystals at finite temperature. The general compacton solution can describe these new intrinsic localized modes in crystals.

The major objectives of this article is to present the recent developments of the above discoveries and their applications. Special attention is given to open questions and unsolved problems in these areas. A new class of strongly dispersive and nonlinear equations $K(m, n)$ will be discussed with applications to science and engineering. All major discoveries in applied mathematics during the second half of the twentieth century are essentially based on the mathematical theories, physical experiments and mathematical computations. An updated list of references is provided to stimulate new interest in future study and research.

1. THE SOLITON AND THE INVERSE SCATTERING TRANSFORM

Historically, John Scott Russell (1808-1882), a Scotish engineer and naval architect, first experimentally observed the solitary wave, a long water wave without change in shape, on the Edinburgh-Glasgow Canal in 1834. He called it the "great wave of translation" and then reported his experimental observations at the British Association in his 1844 paper "Report on Waves". Thus, the solitary wave represents, not a periodic wave, but the propagation of a single isolated symmetrical hump of unchanged form. His discovery of this remarkable phenomenon inspired him further to conduct a series of extensive experiments on the generation and propa-

gation of waves in natural environments – on canals, lakes and rivers – as well as in his laboratory which was a specially designed small reservoir in his own garden. Based on his numerous experimental findings, Russell discovered empirically several major properties of the solitary wave.

(i) An isolated solitary wave propagates without change of shape and with a constant velocity.

(ii) The velocity U of the solitary wave and its maximum amplitude a above the free surface of water of depth h are related by the formula

$$U^2 \;=\; g\,(a+h)\,, \quad (a<h)\,, \tag{1}$$

where g is the acceleration due to gravity.

(iii) A solitary wave of very high amplitude breaks into two or more smaller solitary waves.

Russell once made a comment that "the great primary waves of translation cross each other without change of any kind in the same manner as the small oscillations produced on the surface of a pool by a falling stone."

Russell's Report on Waves was received considerable attention by two distinguished British scientists – G.B. Airy (1801-1892) and G.G. Stokes (1819-1903). Both Airy and Stokes raised serious questions on the existence of the solitary wave and predicted that such waves cannot propagate in a liquid medium without change of form. In fact, Airy strongly criticized the existence of the solitary wave and published a paper on "Tides and Waves" in 1845. In the paper, Airy stated that Russell's formula for the velocity of the solitary wave was in contradiction with his theory of long waves on shallow water. Furthermore, he argued vigorously against Russell's observations and stated that "We are not disposed to recognize this wave as deserving of the epithets "great" or "primary"...." At the same time, Stokes investigated Russell's observations and emphirical results more carefully than Airy in his 1847 paper "On the Theory of Oscillating Waves" and concluded that the solitary wave cannot exist even in liquids with vanishing viscosity.

It was not until 1870s that Russell's prediction was finally and independently confirmed by both J. Boussinesq (1842-1929) and Lord Rayleigh (1842-1919). Based on the Euler equation of motion and the continuity equation in an invisid and incompressible liquid, they derived the formula and showed the solitary wave profile

$z = \eta(x,t)$ (see Figure 1) is given by

$$\eta(x,t) \;=\; a\,\mathrm{sech}^2\left[\beta(x-Ut)\right], \tag{2}$$

where $\beta^2 = 3a \div \{4h^2(a+h)\}$ for any $a > 0$.

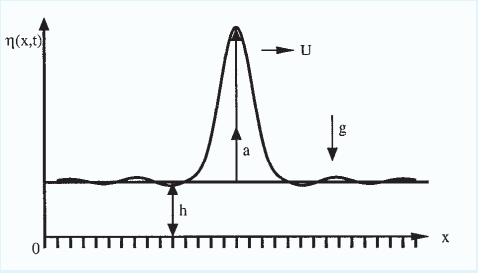


Figure 1: A solitary wave.

More than 60 years later, in 1895, two Dutch mathematicians, D.J. Korteweg and G. de Vries [12] derived a nonlinear mathematical model equation which is known as the *Korteweg-de Vries (KdV)* equation, to provide an explanation of the remarkable observation by Scott Russell. Thus, the KdV equation has the form

$$\eta_x + c\left(1 + \frac{3}{2h}\,\eta\right)\eta_x + \frac{ch^2}{6}\eta_{xxx} \;=\; 0, \tag{3}$$

where $\eta(x,t)$ is the free surface displacement of water of depth h and $c = \sqrt{gh}$ is the shallow water speed and g is the acceleration due to gravity. So the 1895 paper by Korteweg and de Vries finally resolved the famous controversy on the existence of the solitary wave and its various aspects. On the other hand, modern research on soliton dynamics during the last 40 years reveals that the KdV equations played a new and significant role on modern pure and applied mathematics. Its importance for physical sciences lies in its ability to describe not only nonlinear shallow water waves, but also many other nonlinear waves. In pure mathematics, the KdV equation was a starting point for developing a deep and beautiful mathematical theory.

Modern developments in the theory and applications of the KdV solitary waves began with the seminal work published as a Los Alamos Scientific Laboratory Report in 1955 by Fermi, Pasta and Ulam on a numerical model of a discrete nonlinear mass-spring system. In 1914, Debye suggested that the finite thermal conductivity of an anharmonic lattice is due to the nonlinear forces in the springs. This suggestion led Fermi, Pasta and Ulam to believe that a smooth initial state would eventually relax to an equipartition of energy among all modes because of nonlinearity. But their

study led to the striking conclusion that there is no equipartition of energy among the modes. Although all the energy was initially in the lowest modes, after flowing back and forth among various low-order modes, it eventually returns to the lowest mode, and the end state is a series of recurring states. This remarkable fact has become known as the *Fermi-Pasta-Ulam (FPU) recurrence phenomenon.*

This curious result of the FPU experiment inspired Norman Zabusky and Martin Kruskal [1] to formulate a continuum model for the nonlinear mass-spring system to understand why recurrence occurred. In fact, they considered the initial-value problem for the KdV equation,

$$u_t + uu_x + \delta\, u_{xxx} \;=\; 0, \tag{4}$$

where $\delta = \left(\frac{h}{\ell}\right)^2$, ℓ is a typical horizontal length scale, with the initial condition

$$u\,(x,0) \;=\; \cos \pi x, \qquad 0 \le x \le 2, \tag{5}$$

and the periodic boundary conditions with period 2, so that $u\,(x,t) = u\,(x+2,t)$ for all t. Their numerical study with $\sqrt{\delta} = 0.022$ produced a lot of new interesting results, which are shown in Figure 2.

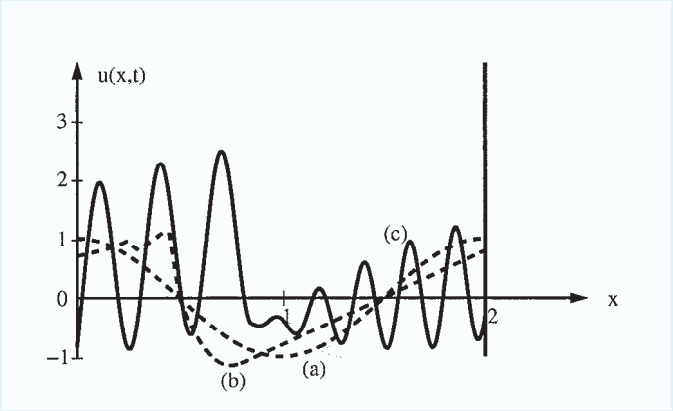


Figure 2: Development of solitary waves: (a) initial profile at $t = 0$, (b) profile at $t = \pi^{-1}$, and (c) wave profile at $t = (3.6)\,\pi^{-1}$ (From Zabusky and Kruskal [1]).

They observed that, initially, the wave steepened in regions where it had a negative slope, a consequence of the dominant effects of nonlinearity over the dispersive term, δu_{xxx}. As the wave steepens, the dispersive effect, then, becomes significant and balances the nonlinearity. At later times, the solution develops a series of *eight* well-defined waves, each like sech^2 functions with the taller (faster) waves

ever catching up and overtaking the shorter (slower) waves. These waves undergo nonlinear interaction according to the KdV equation and, then, emerge from the interaction without change of form and amplitude, but with only a small change in their phases. So the most remarkable feature is that these waves retain their identities after the nonlinear interaction. And another surprising fact is that the initial profile reappears, that is, very close to the FPU recurrence phenomenon. In view of their preservation of shape and their resemblance to the particlelike character of these waves, Zabusky and Kruskal called these solitary waves, *solitons* like photon, proton, electron, and other elementary particles.

Historically, the famous 1965 paper of Zabusky and Kruskal marked the birth of the new concept of the *soliton*, a name intended to signify particlelike quantities. Subsequently, Zabusky [13] confirmed, numerically, the actual physical interaction of two solitons, and Lax [14] gave a rigorous analytical proof that the identities of two distinct solitons are preserved through the nonlinear interaction governed by the KdV equation. Physically, when two solitons of different amplitudes (and hence, of different speeds) are placed far apart on the real line, the taller (faster) wave to the left of the shorter (slower), the taller one eventually catches up to the shorter one and, then, overtakes it. When this happens, they undergo a nonlinear interaction according to the KdV equation and emerge from the interaction completely preserved in form and speed with only a small phase shift. Thus, these two remarkable features, (i) steady progressive pulselike solutions and (ii) the preservation of their shapes and speeds, confirmed the particlelike property of the waves and, hence, the definition of the soliton. Experimental confirmation of solitons and their interactions has been provided successfully by many authors. Thus, these discoveries have led, in turn, to extensive theoretical, experimental, and computational studies over the last 40 years. Many nonlinear model equations have now been found that possess similar properties, and diverse branches of pure and applied mathematics have been required to explain many of the novel features that have appeared.

The computational work of Zabusky and Kruskal [1] led to the discovery of the remarkable stability of the soliton. Their computer experiment can be described as follows. When two or more solitons travel in a dispersive medium, the taller (faster) ones will overtake the shorter (slower) ones, and after nonlinear interaction, these solitons separate from each other without change of their shapes and amplitude, but only a small change in their phases. The end result is that the taller soliton reappears in front and shorter one behind as they move to the right, except for a slight delay. Indeed, the computational analysis reveals another remarkable result that every solution of the KdV equation (4) with any prescribed initial condition

$\eta(x,0) = f(x)$, decomposes as $t \to \infty$ into a finite number of solitons of different velocities and different amplitudes and a dispersive tails which gradually decay as shown as in Figure 3.

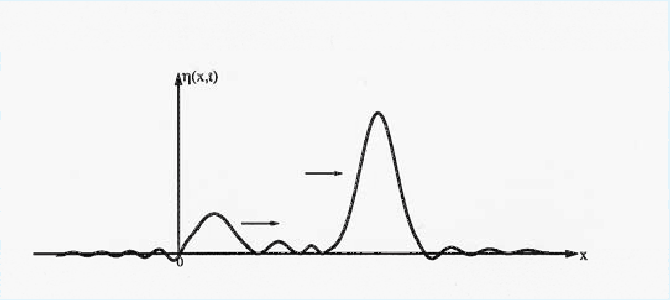


Figure 3: Solitons with dispersive tails.

We seek a traveling wave solution of (1), that is stationary in the frame x so that $\eta = \eta(X)$, $X = x - Ut$ with $\eta \to 0$ as $|X| \to \infty$. Substituting this solution into (1) gives

$$(c-U)\,\eta' + \left(\frac{3c}{2h}\right)\eta\eta' + \frac{ch^2}{6}\eta''' \;=\; 0, \tag{6}$$

where $\eta' = \frac{d\eta}{dx}$. Integrating this equation twice with respect to X yields a *nonlinear ordinary differential equation*

$$(c-U)\,\eta^2 + \left(\frac{c}{2h}\right)\eta^3 + \left(\frac{ch^2}{6}\right)\eta'^2 \;=\; 2A\eta + B, \tag{7}$$

where A and B are integrating constants.

Two special cases are of interest: (i) $A = B = 0$ and (ii) $A \neq 0$ and $B \neq 0$. For the case (i) with η and η' tend to zero at infinity, equation (7) becomes

$$\left(\frac{d\eta}{dx}\right)^2 \;=\; \frac{3}{h^3}\,\eta^2\,(a-\eta)\,, \tag{8}$$

where $a = 2h\left(\frac{U}{c} - 1\right)$.

Substituting $\eta = a\,\mathrm{sech}^2\theta$ in (8) gives the exact solution

$$\eta(X) \;=\; a\,\mathrm{sech}^2\,bX, \quad b = \sqrt{\frac{3a}{4h^3}}. \tag{9}$$

Thus

$$\eta(x,t) \;=\; a\,\mathrm{sech}^2\left[\sqrt{\frac{3a}{4h^3}}\,(x-Ut)\right], \tag{10}$$

where the velocity of the traveling wave is

$$U \;=\; c\left(1+\frac{a}{2h}\right) > c. \tag{11}$$

The solution (10) is called a *solitary wave* (see Figure 1) describing a single symmetric hump of amplitude a above the undisturbed water depth h and decaying to zero exponentially as $|x| \to \infty$. The solitary wave solution is an excellent agreement with the observational result of Scott Russell. Thus, the solitary wave travels to the right in the medium without change of shape. The velocity of the wave is $U\,(>c)$ which is directly proportional to the amplitude a. The width of the wave is $b^{-1} = \left(\frac{3a}{4h^3}\right)^{-\frac{1}{2}}$ that is inversely proportional to the square root of the amplitude a. In fact, the taller and thiner solitary wave travel faster, whereas shorter and fatter one propagate slowly. This kind of behaivor is usually expected for linear differential problems, since the solution can be described by the linear superpositionn of eigenfunctions (corresponding to eigenvalues) as each eigenfunction evolves separately. The existence of the soliton for the nonlinear KdV equation was a total surprise at the time of its discovery. The soliton solution is one of the good examples that can be used to refute the 1948 interesting quotation of Sir James Lighthill as stated at the beginning of this article.

For the case (ii) where A and B are non-zero, equation (7) can be rewritten as

$$\frac{h^3}{3}\,\eta_X^2 \;=\; -\eta^3 + 2h\left(\frac{U}{c}-1\right)\eta^2 + \frac{2h}{c}\,(2A\eta+B) \equiv F(\eta), \tag{12}$$

where $F(\eta)$ is a cubic with simple zeros. Following the detailed calculation (see Debnath [15, 16]) the solution of (12) can be expressed in terms of Jacobi's elliptic functions

$$\eta(X) \;=\; a\left[1 - sn^2\left\{\left(\frac{3b}{4h^3}\right)^{\frac{1}{2}} X\right\}\right] = a\;cn^2\left[\left(\frac{3b}{4h^3}\right)^{\frac{1}{2}} X\right], \tag{13}$$

where three zeros of $F(\eta)$ are 0, a, $-(b-a)$, $(b>a>0)$, $sn(z,m)$ and $cn(z,m)$ are Jacobi's elliptic functions with modulus $m = \sqrt{\left(\frac{a}{b}\right)}$. The solution $\eta(X)$ represents a train of periodic waves in shallow water. These waves are called *cnoidal waves* with wavelength $\lambda = 2\sqrt{\left(\frac{4h^3}{3b}\right)}\,K(m)$ where $K(m)$ is the complete elliptic integral of the first kind. A typical cnoidal wave is shown in Figure 4.

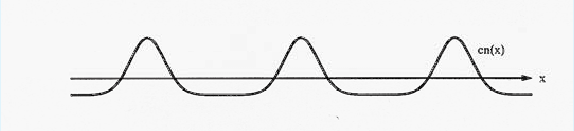


Figure 4: A Cnoidal Wave.

In the limit as $m \to 1\,(a \to b)\, cnz \to \mathrm{sech} z$, the cnoidal waves are in perfect agreement with the classical KdV solitary wave where the wavelength $\lambda \to \infty$ because $K(a) = \infty$ and $K(0) = \frac{\pi}{2}$. On the other hand, in the limit as $m \to 0\,(a \to 0)$, $snz \to \sin z$, $cnz \to \cos z$. The corresponding solution represents the small-amplitude waves associated with the linearized KdV equation. The solution is

$$\eta(x,t) \;=\; \frac{1}{2}a\left[1+\cos\left(kx-\omega t\right)\right], \quad k=\left(\frac{3b}{h^3}\right)^{\frac{1}{2}}, \tag{14}$$

where the corresponding dispersion relation is given by

$$\omega \;=\; Uk = ck\left(1-\frac{1}{6}k^2h^2\right). \tag{15}$$

This corresponds to the first two terms of the series expansion of $\sqrt{gk\ \tanh\left(kh\right)} = \omega$ which is the famous *dispersion relation* of water waves in water of depth h. Thus, these results are in perfect agreement with the linearized water wave theory.

The experimental discovery of the solitary wave by Scott Russell, the mathematical theory of KdV equation as well as the computer experiment of Zabusky and Kruskal provided the conclusive evidence for the existence of the soliton and its remarkable stability property. Thus, the computational work would play a major role on many unexpected future discoveries in mathematics and physics. In 1946, John Von Neumann raised the question. "What phases of the pure and applied mathematics can be furthered by use of large-scale automatic computing instruments?" His detailed and precise answer is given below:

"Our present analytical methods seem unsuitable for the solution of the important problems arising in connection with nonlinear partial differential equations and, in fact, with virtually all types of nonlinear problems in pure mathematics. The truth of this statement is particularly, striking in the field of fluid dynamics. Only the most elementary problems have been solved analytically in this field...

The advance of analysis is, at this moment, stagnant along the entire front of nonlinear problems.... Really efficient high-speed computing device may... provide us with those heuristic hints which are needed in all parts of mathematics for genuine progress...."

In connection with the discovery of the soliton, the usefulness of computer experiments was quite clear from Norman Zabusky's statement: "Almost everyone using computers has experienced instances where computational results have sparked new insights."

1.1 The Inverse Scattering Transform (IST)

Historically, Gardner et al. [2] formulated an ingenious method for finding the exact solution of the KdV equation. The main problem is to find the exact solution of the general initial value problem for the canonical form of the KdV equation

$$u_t - 6uu_x + u_{xxx} = 0, \quad x \in R, \quad t > 0, \tag{16}$$

with the initial condition

$$u(x,0) = u_0(x), \quad x \in R, \tag{17}$$

where $u_0(x)$ satisfies certain fairly weak conditions so that $u(x,t)$ exists for all x and t.

To solve the initial value problem for the KdV equation means finding $u(x,t)$ for any given initial condition $u(x,0) = u_0(x)$. For nonlinear problems, this is an extremely difficult problem because the linear superposition principle does not hold.

In fact, a systematic development of the mathematical theory of solitons began in 1967 when Gardner et al. [2] proposed the method of solving (16)–(17) based on the so-called the *Inverse Scattering Transform (IST)*. The discovery of the Inverse Scattering Transform (often called the nonlinear Fourier transform method) is one of the most remarkable achievements of mathematics in the 20th century.

Making reference to Debnath [15, 16], we briefly outline the major steps involved in the inverse scattering transform. The exact solution of the KdV equation is obtained by associating its solution with the potential of a time-dependent Schrödinger equation. The next step is to find out the solution of the quantum mechanical problem with the initial value for the KdV equation taken as the potential. This involves calculations of the discrete (bound) eigenfunctions, their normalization constants and eigenvalues, and the reflection and transmission coefficients of the

continuous (unbounded) states. These results are collectively called the *scattering data* of the Schrödinger equation. The next step is to determine the evolution of the scattering data for any potential that evolves, according to the KdV equation, from a prescribed initial function. Both discrete and continuous eigenvalues are found to be invariant under such changes, and normalization constants and the reflection and transmission coefficients evolve according to simple exponential laws. The final step of the method deals with the determination of the potential for any time t from the inversion of scattering data. In summary, two distinct steps are involved in the method: (i) the solution of the Schrödinger equation (the Sturm-Liouville problem) for a given initial condition $u(x,0) = u_0(x)$, from which we determine the scattering data $S(t)$, and (ii) the solution of the *Gelfand-Levitan-Marchenko* (GLM) linear integral equation. Even though these two steps involved may not be technically easy to handle, however, in principle, the problem is completely solved. The effectiveness of the method can be best examplified by many simple but nontrivial examples (see Debnath [15, 16]).

The power and success of the inverse scattering method for solving the KdV equation can be attributed to several facts. First, the most remarkable result of the method is the fact that the discrete eigenvalues of the Schrödinger equation for ψ do not change as the potential evolves according to the KdV equation. Second, the method has reduced solving a nonlinear PDE to solving two linear problems: (i) a second-order ordinary differential equation; and (ii) a linear integral equation. Third, the eigenvalues of the ODE are constants and this leads to major simplification in the evolution equation for ψ. Fourth, the time evolution of the scattering data is explicitly determined from the asymptotic form of ψ as $|x| \to \infty$. So, this information allows us to solve the inverse scattering problem and, hence, to obtain the final solution of the KdV equation. The method is presented schematically in Figure 9.7 of Debnath's book [16].

In his seminal paper, Lax [14] developed an elegant formalism for finding isospectral potentials as solutions of a nonlinear evolution equation with all of its integrals. This work deals with some new and fundamental ideas and deeper results and their application to the KdV equation. This study subsequently paved the way to generalizations of the technique as a method for solving other nonlinear partial differential equations. Lax also developed the method of inverse scattering based on abstract formulation of evolution equations and certain properties of operators on a Hilbert space, some of which are familiar in the context of quantum mechanics. His formulation has the special feature of associating certain nonlinear evolution equations with linear equations which are analogs of the Schrödinger equation for the KdV

equation. Subsequently, Zakharov and Shabat [17, 18, 19], Zakharov and Faddeev [20] showed that the KdV equation, which they treated as an infinite-dimensional Hamiltonian system, is completely integrable. All these works can be used as a rather complex mathematical treatment of the inverse scattering method, which allows one to reduce integrable nonlinear problems to linear ones. For the first time, they have extended the Lax formalism for equations with more than one spatial variable. This extension is usually known as *Zakharov and Shabat (ZS) scheme*. Hirota [21, 22, 23] developed an ingenious *direct* method for finding multisoliton solutions of different integrable nonlinear evolution equations. Ablowitz, Kaup, Newell and Segur (AKNS) [24] generalized the ZS scheme so that their method can be applied to solve many other integrable nonlinear evolution equations. In particular, they showed that the sine-Gordon (SG) equation is completely integrable. In 1974-1975, Novikov et al. [25] and his associates formulated a fairly general approach to finding exact periodic solutions of the KdV equation based on deep mathematical results of Riemann, Abel and Jacobi. A further systematic development of these ideas and methods led to deep relations between superstrings and solitons (and with general integrable equations).

In recent years, many important ideas, methods and results in soliton theory are found to be connected with quantum integrability and quantum theory of solitons. Quantum solitons have many applications, especially, in condensed matter physics. It is generally believed that the mathematical theory of solitons seems to be very useful in studies of the fundamental features of nature.

2. FRACTALS, FRACTAL DIMENSION AND FRACTAL GEOMETRY

In the 1970s, Benoit Mandelbrot (1924-) [3, 4, 5] first introduced the concept of fractals, and the fractal dimension based on a definition of Hausdorff (1886-1942) in 1919. He first recognized that many phenomena of nature are so irregular and complex that they cannot be described by the Euclidean geometry. In this context Mandelbrot's [5] quotation seems to be most appropriate to mention: "Why is geometry often described as "cold" and "dry"? One reason lies in its inability to describe the shape of a cloud, a mountain, a coastline, or a tree. Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line." Motivated by the Kolmogorov (1903-1987) theory of turbulence and his 1958 definition of the "capacity" of a geometrical figure, Mandelbrot [26] published a paper entitled, "How long is the Coast of Britain? Statistical self-similarity and fractional dimension," and made

an attempt to explain the energy distribution in intermittent turbulence [27, 28]. His famous book on *The Fractal Geometry of Nature* is the standard reference and contains both the elementary ideas and a wide range of new and advanced topics such as multifractals, dynamical systems and chaotic attractors. He also made a serious attempt to convince the reader that fractal geometry deals with the geometry of a wide variety of irregular phenomena observed in nature. His 1982 book and another new book [29] on fractals and chaos published in 2004 contain a large number of examples of fractals, and many beautiful pictures of fractals which were drawn with the help of computers or graphical analysis.

Mandelbrot [3, 4, 5] introduced the idea of a fractal as a geometrical curve that consists of an identical shape repeating on an ever decreasing scale. He mentioned many common examples of fractals including irregular coastal structures (degree of meandering of coastlines) records of heart beat, Dow Jones industrial averages, variations of traffic flow, electromagnetic fluctuations in galactic radiation noise, textures in images of natural terrain, Weierstrass' and Riemann's everwhere continuous and non-differentiable functions. Some of the curves involved in these examples are highly irregular in shape. Other examples include tree with a trunk that separates into two branches which in turn two smaller side branches, and so on; how to define the speed of the wind during a violent storm, and how to distinguish proper music (good or bad) from noise. With the advent of modern computers and power of simple graphical analysis, fractals and chaos have received widespread attention in recent years. The alluring computer graphics have generated tremendous new interest among mathematicians and scientists in these areas. In recent years, considerable attention has been given to the fields of fractal geometry and chaotic dynamical systems. Current research on fractals and chaos are associated with the names including Cantor, Poincaré, Sierpinski, Julia, Fatou and Mandelbrot.

In mathematics, the topological dimension of a set is traditionally considered as the *natural* dimension. It is defined by a natural number. Thus, point, straight line, plane and volume have topological dimension 0, 1, 2, and 3 respectively. Topological dimension is invariant under homeomorphisms, that is, if the topological dimension of a set S is m, then the topological dimension of $h(S)$ is also m provided h is a homeomorphism. Mandelbrot [5] recognized that the definition of the topological dimension cannot be used to define the dimension of some highly irregular sets (such as natural coastlines). In 1919 Hausdorff pointed out that the topological definition is not suitable for some sets and introduced a new definition of dimension based on the size variations of sets when measured at different scales. Thus, Hausdorff's definitions of Hausdorff measure and dimension in 1919 provided the fundamental

basis of the study of geometric measure theory. The Hausdorff dimension can be defined for any subset (open or closed) of $\mathbb{R}^n$ (see Rogers [30]). Unlike the topological dimension, the Hausdorff dimension is not invariant under homeomorphism. However, the topological dimension S is the infinum of the Hausdorff dimension of its homeomorphic images $h(S)$.

Mandelbrot [5] defined fractals as a set with Hausdorff dimension, D strictly greater than its topological dimension, D_T ($D > D_T$). For example, the set of points on a straight line in ordinary Euclidean space has the topological dimension $D_T = 1$, and the Hausdorff dimension, $D = 1$. Obviously, the line is not a fractal according to Mandelbrot's definition. However, there are many examples where there exist fractal sets for which the Hausdorff dimension is a noninteger. The familiar examples are $D = D_T = 0$ for points, $D = D_T = 1$ for lines, $D = D_T = 2$ for planes and surfaces, and $D = D_T = 3$ for spheres and other volumes.

The *capacity dimension* is a simplification of the Hausdorff dimension that is relatively easy to compute numerically. We consider a bounded set S in $\mathbb{R}^n$ and count the minimum number $N(r)$ of balls of radius r required to cover the set S. Based on the famous experimental diagrams for natural coastlines of West Great Britian and Spanish-Portuguese land frontier of L.F. Richardson (1881-1953), the scaling behaivor is described by the so-called fundamental relation of self-similar fractal as

$$N(r)\, r^D = 1, \tag{18}$$

where $N(r)$ is the number of segments which is plotted against their unitary length r in a bilogarmithic diagram so that it gives straight line whose slope is $-D$. Thus, it follows from (18) that

$$D = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}. \tag{19}$$

The *capacity dimension* D of S generalizes this result and is defined by

$$D = \lim_{r \to 0} \inf \frac{\log N(r)}{\log(1/r)}. \tag{20}$$

The measure of $N(r)$ is then

$$M = \lim_{r \to 0} N(r)\, r^D. \tag{21}$$

It may be finite or infinite. The Hausdorff dimension is a fractal measure that includes all covers of S with balls of radius less than r. It is often equal to the capacity dimension that is called the *fractal dimension*.

It can be shown that the fractal dimension D is in excellent with the Hausdorff dimension for self-similar sets. When the length ℓ of the initiator is not equal to one, the fundamental relation (18) becomes

$$N(r)\, r^D = \ell^D. \tag{22}$$

Thus, the dimension of self-similar fractals is then given by

$$D = \lim_{r\to 0} \frac{\log N(r)}{\log\left(\frac{\ell}{r}\right)}. \tag{23}$$

This result governs the geometrical scaling law at all scales. The nominal length at each iteration can be given by

$$\ell_n = N_n r_n. \tag{24}$$

It then follows from (24) that

$$\ell_n = \left(\frac{1}{r_n}\right)^D \cdot r_n = r_n^{1-D} = \left(\frac{1}{r_n}\right)^{D-1}. \tag{25}$$

For the case of a real fractal set, that is, as $n \to \infty$ (attractor) with $D > 1$, it turns out that

$$\lim_{n\to\infty} \ell_n = \lim_{r_n\to 0} \ell_n = \infty. \tag{26}$$

This clearly implies that fractal sets are *not* measurable by means of integral powers of the length. The Hausdorff dimension provided the clear possibility of taking finite measures of these unusual sets, if the ordinary dimension is replaced by the nonintegral value. It is noted that the above definition defines the *Hausdorff* (or *Hausdorff-Besicovitch*) dimension D as a *local property* in the sense that it measures properties of sets of points in the limit of radius $r \to 0$ of the test function employed to cover the set. The above definition can easily be generalized to higher-dimensional spaces, and includes also the Euclidean sets as special cases.

In applications, the fractal dimension can be interpreted as the *degree of meandering* of a curve. In practice, the length r is used as small step size so that the fraction $\log N(r) / \log\left(\frac{1}{r}\right)$ tends to a fixed value D in the limit. In some cases, this fraction has the same value at each step so that we can write formula (23) more simply as

$$D = \frac{\log N(r)}{\log\left(\frac{1}{r}\right)}. \tag{27}$$

Or, equivalently,

$$N(r) = \left(\frac{1}{r}\right)^{D}. \tag{28}$$

Clearly, the total length $1 = Nr$ can be expressed as

$$L = \left(\frac{1}{r}\right)^{D-1}. \tag{29}$$

This once again clearly shows that the total length measured increases as the measuring unit r decreases.

2.1 The Cantor Middle Third Set

A very simple construction due to Cantor generate fractals sets with a fractal dimension in the range $0 < D < 1$. As shown in Figure 5, we start with a line segment of length $\ell = 1$, called the *initiator*. We then divide the line segment into three equal parts, and delete the open middle part leaving its end points. We then apply similar construction to each of the parts and so on. This procedure leads to extremely small line segments as shown in Figure 5.

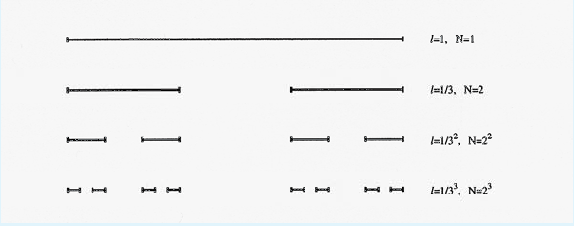


Figure 5: The Cantor Middle-Third Set

In the nth stage, $\ell = \frac{1}{3^n}$, and $N = 2^n$. Clearly, as $\ell \to 0$ as $N \to \infty$. This process generates a Cantor dust set whose topological dimension is $D_T = 0$. We calculate the fractal dimension of the Cantor middle-third set by (23) so that

$$D = \lim_{\ell \to 0} \frac{\log N(r)}{\log\left(\frac{1}{\ell}\right)} = \lim_{n \to 0} -\frac{\log 2^n}{\log 3^{-n}} = \frac{\log 2}{\log 3} = 0.6309.$$

This shows that the fractal dimension of the Cantor set is 0.6309 which is not an positive integer, but a positive fraction less than one and $0 = D_T < D = 0.6309$.

Also, the fractal dimension of this set is less than the topological dimension 1 of its initiator.

2.2 The Koch Triadic Fractal Curve

This fractal curve can be constructed geometrically by successive iteractions. The construction begin with a line segment of length 1 ($L(1) = 1$), called the *initiator*. We divide it into three equal line segments, and replace the middle segment by an equilateral triangle without a base. This completes the first step ($n = 1$) of the construction, giving a curve of four line segments, each of length $\ell = \frac{1}{3}$, and the total length is $L = \frac{4}{3}$. This new shape of the curve is called the generator. The second step ($n = 2$) is obtained by replacing each line segment by a scaled-down version of the generator. Thus, the second-generation curve consists of $N = 4^2$ line segments, each of length $= \ell = \frac{1}{3^2}$, with the total length of the curve $L(\ell) = \left(\frac{4}{3}\right)^2$. Continuing this iteration process successively leads to the famous Koch triadic curve of total length $L(\ell) = \left(\frac{4}{3}\right)^n$ where $\ell = 3^{-n}$ as shown in Figure 6. The name triadic is justified because individual line segments at each step decrease in length by a factor of 3. Obviously, the Koch curve at the end of many iterations ($n \to \infty$) would have a wide range of scales (see Debnath [31]). As the resolution increases microscopically ($n \to \infty$), the length of the Koch curve also increases without limit. This shows a striking contrast to an ordinary curve whose length remains the same for all resolutions. The intrinsic parameter that measures this property is called the *fractal Haussdorff dimension* D which is defined by

$$D = \lim_{\ell \to 0} \frac{\log N(\ell)}{\log\left(\frac{1}{\ell}\right)} = \lim_{\ell \to 0} \frac{\log\left\{\frac{L(\ell)}{\ell}\right\}}{\log\left(\frac{1}{\ell}\right)}, \tag{30}$$

where $L(\ell) = \ell N(\ell) = \ell^{1-D}$ for small number ℓ.

For the triadic Koch curve, $N(\ell) = 4^n$ and $\ell = 3^{-n}$, so that its fractal dimension is given by

$$D = \frac{\log 4}{\log 3} \approx 1.2628 > 1, \tag{31}$$

and is noninteger and greater than one. The reason for this conclusion is due to the convolutedness of the Koch curve, which becomes more and more convoluted as the resolution becomes finer and finer. When the curve is highly convoluted, it effectively covers a two-dimensional area, that is, the one-dimensional curve fills up a space of dimension two. In general, a fractal surface has a dimension greater than two, and its dimension could become as large as three for a very highly convoluted

surface, so that it can essentially cover a three-dimensional volume. This leads to a general result that the fractal Hausdorff dimension of a set is a measure of its space-filling ability.

In terms of the box-counting algorithm in fractal geometry, the minimum $N(\ell) = 4$ boxes of size $\left(\frac{1}{3}\right)$ are needed to cover the line in the Koch curve in Figure 6(b). Similarly, at least $N(\ell) = 4^2$ boxes of size $\ell = \left(\frac{1}{3}\right)^2$ are required to cover the line in Figure 6(c). In general, a minimum of $N(\ell) = 4^n$ boxes of size $\ell = \left(\frac{1}{3}\right)^n$ are needed to cover the Koch curve obtained at the nth step. On the other hand, the total length $L(3^{-n}) = \left(\frac{4}{3}\right)^n$ at the nth iteration is obtained at a finer resolution of 3^{-n}.

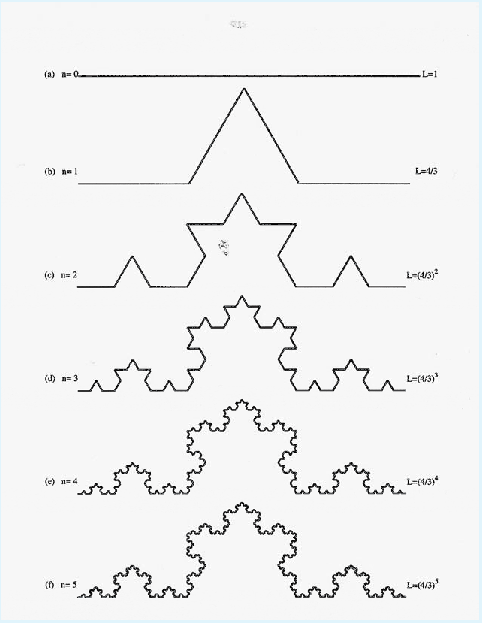


Figure 6: The triadic Koch curve.

2.3 The Mandelbrot and Given Fractal

Mandelbrot and Given [32] recognized that the Koch fractal curve is a beautiful example of fractal structure that is the prototype of a wide variety of fractals. In order to describe percolation processes by fractal structures, they constructed a new fractal curve from a generator which divides a line segment into pieces of length $r = \frac{1}{3}$ and adds a *loop* consisting of three pieces in addition to new branches appended. In each iteration of the process, from one generator of the prefractal to the next,

the generator replaces each line segment in the prefractal by $N = 8$ segments, as shown by the motif in Figure 7, that have been scaled down by the ratio $r = \frac{1}{3}$. This process of iteration leads to a beautiful fractal structure – the so-called the *Mandelbrot and Given fractal* as shown in Figure 2.13 of the book by Feder ([33], page 21). Using the formula (19), the fractal dimension of the Mandelbrot and Given fractal is $D = \log 8/\log 3 = 1.89 > 1$.

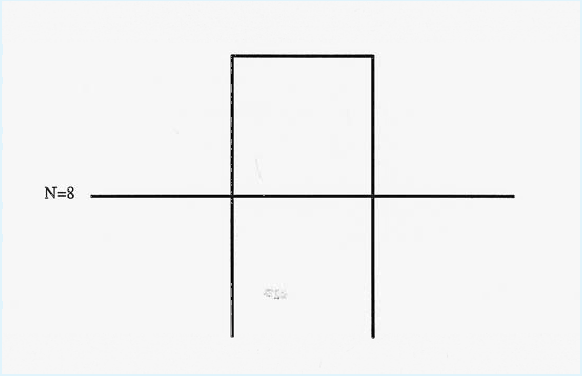


Figure 7: Motif for the Mandelbrot and Given fractal.

In their pioneering work, Sapoval et al. [34] discovered that the diffusion front resulting from the diffusion phenomenon has a fractal structure similar to that of Mandelbrot and Given [32]. This structure is not only closely related to the fractal geometry of percolation, but also creates a new interest in understanding of a wide variety of miraculous geometrical shapes of nature. Indeed, the discovery of Sapoval et al. [34] represents a significant contribution to the fractal nature of a diffusion front and fractal model of percolation clusters.

2.4 The Minkowski Fractal

Based on a similar iteration process of the Koch fractal, this fractal can also be easily be constructed by successive iterations. The construction begins with a line segment of length $L = 1$, called the *initiator* which is divided into 4 equal line segments. We then make two squares on the two middle parts so that this leads to a motif of eight equal line segments $\left(N = 8, r = \frac{1}{4}\right)$ as shown in Figure 8.

Continuing the above iteration process successfully four times leads to a beautiful fractal structure – the so called *Minkowski fractal* as shown in Figure 3.9 on page 38 in the book by Lauwerier [35]. Once again using the formula (19), the fractal

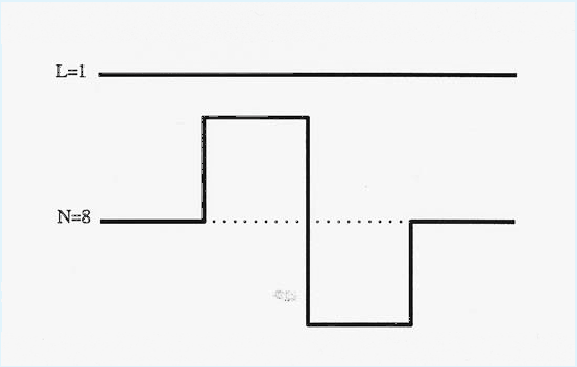


Figure 8: The motif for the Minkowski fractal.

dimension of the Minkowski fractal in $D = 1.5 > 1$.

We close this subsection by adding a comment. There exists a wide variety of beautiful fractal structures in nature including spirals, shells, trees and stars. Some of them represents building blocks of the living world. The cell nucleus consists of a long, spiral structure, the nucleic acid or DNA, carrier of the genetic code, a building scheme for an organism yet to be formed. Spiral like fractal structures can still be found in countless species of shellfish alive now. For more examples and beautiful photographs, the reader is referred to Lauwerier [35] and Peitgen and Richter [36].

2.5 Fractals In Fracture Mechanics

Fractals have recently been used to describe irregular phenomena in many scientific areas. In addition to their use for measuring the length of irregular coastlines, the concept of fractals can be used to study complex shapes of fracture surfaces of materials. In general, fractures, which originate and grow in rocks, metals and concrete, present a ramified and self-similar structure. This means that certain geometrical properties appear at any scale. Mandelbrot et al. [37] introduced fractal character of fracture surfaces of solid materials. In almost all aspects of fracture mechanics of disordered materials, fractal dimensions of fracture surfaces play an important role. So, the fractal geometry is found to be an effective tool in solid mechanics to describe mechanical damage and crack growth phenomena with statistical characteristics. For example, in many disordered materials including concrete, rocks, and ceramics, there are several random parameters, such as the position, size and orientation of the preexisting microcracks. At the beginning of the loading process,

the microcracks can be considered as two-dimensional surfaces and they propagate during the loading process. As the load is increased, these microcracks grow, coalesce and form a fractal set with a dimension between two and three. The fractals dimension measuring the microcrack size distribution increases with propagation of cracks. It turns out that the materials become progressively more disordered with the development crack net, from the initial loading up to the failure state.

In nature there are no ideal geometrical fractals, natural morphologics are generally random multifractals due to the limited size of the heterogenetics. Although ideal fractals have no characterisitics length scale, for *random multifractals*, it is possible to identify small scales at which disorder occurs, and large scales at which order prevails. For fracture surfaces, the microscopic disorder is that of *Brownian surfaces* with fractal dimension equal to 2.5, whereas the macroscopic order is that of Euclidean surfaces with integral dimension equal to 2. It is possible to examine the evolution of the fractal parameter corresponding to different degrees of disorder in sizes of the crack distribution during microcrack propagation.

In order to study the fractal dimension of microcrack net in disordered materials, we assume the probability density function associated with the existence of microcracks of length greater than x is of the form

$$p\,(x) \;=\; C\;N\;x^{-(N+1)}, \tag{32}$$

where C is a constant and N is a measure of the degree of order in crack size distribution. The corresponding cummulative probability distribution is

$$p\,(x) \;=\; \int_{x}^{\infty} p\,(x)\,dx = C\;x^{-N}, \tag{33}$$

so that the crack size distribution can be considered as a kind of fractal with exponent N which can be obtained from (33) as

$$N \;=\; \frac{(\log P\,(x) - \log C)}{\log\left(\frac{1}{x}\right)}. \tag{34}$$

The fractal dimension D of the crack size distribution in disordered materials is defined by Carpinteri and Yang [38] as

$$D \;=\; \left(\frac{N+2}{N+1}\right), \tag{35}$$

where the range of D is in between 1 and 2 provided N assumes positive values based on experimental findings. This parameter N is called the *order parameter* as larger

values of N correspond to more order. In fact, (35) gives $D = 1$ as $N \to \infty$ that corresponds to the perfect order, and $D = 1.5$ as $N \to 1$ leads to the self-similarity due to Carpinteri [39].

Usually, fractal geometry is concerned with highly disordered self-similar morphologies similar to rough interfaces or microcraked structures with statistical characteristics. The common underlying principle involves random self-similarity which implies that statistically similar morphology appears in a wide range of problems in fracture mechanics. From a physical point of view, a scaling behaivor of physical quantities is observed during the experiments on a system. The major assumption of a scaling theory is that these quantities are self-similar functions of the independent variables of physical phenomena. From a mathematical point of view, self-similar scaling implies a *power law* satisfied by independent variables. The generic form of many power laws has the form $y = a\, x^{\alpha}$ which is characterized by two parameters a (the *amplitude*) and α (the *exponent*) where the former depends on the choice for the physical quantities involved, and the latter is characterized by the physical process itself, that is, the self-similar property which governs the scaling. In general, any power-law distribution is mathematically equivalent to a *fractal distribution*, where the exponent involved has *non-integer values*. The upshot of this discussion is how to determine the fractal dimension of microcrack structures and fracture surfaces. In recent years, considerable attention has been given to use fractal geometry in many problems in fracture mechanics and contact mechanics.

Fractal analysis has become a useful tool for investigating scaling behaivor of microscopic inhomogeneous properties in various disordered materials including ceramics, colloids, organic polymer alloys, aerogals, granial and cement metal films, and porous media. Several authors including Muller [40], and Avnir [41] have studied the morphology of disordered materials by multifractal analysis. The fundamental idea behind the multifractal scaling analysis is based on experimental observations that a certain statistical quantity called *generating function scales* as powers of x associated with heterogeneous fractal objects can be expressed as

$$X\,(q, x) \quad \sim \quad x_n^{\tau(q)}, \tag{36}$$

where x_n is the nth generation scale, the exponent τ depends on an arbitrary real parameter q. The main difference between the exponent in (36) and the scaling exponent in the conventional fractal analysis is that, the exponent $\tau\,(q)$ in (36) is a function of q, whereas the Hausdorff dimension is a number.

2.6 Fractals In Turbulence

While Mandelbrot [27, 28] introduced fractals to study of turbulence in the 1970s, he first recognized the fact that fractals are geometrical curves which look the same from nearby and far away. The phrase "look the same from nearby and far away" means self-similar. In other words, fractals are self-similar sets that are built from parts similar to the entire set but on a finer and finer scale. More precisely, a set $S < \mathbb{R}^n$ is called *self-similar* if it is the union of disjoint subsets S_1, S_2, ..., S_n that can be obtained from S with a scaling, translation, and rotation. The self-similarity often implies infinite multiplication of details which generate irregular structures. Mandelbrot [4] constructed such sets by an iterative process using an initial and standard polygon. Subsequently, Hutchinson [42] generalized this process such that fractals can be considered as the fixed points of certain set maps. They are generated by the application of simple transformations such as translations, scalings, rotation and congruences in simple spaces.

Several reliable numerical and experimental results predicted that the Richardson cascade of eddy motions produces a self-similar cascade of wrinkles on the interface of turbulent flows on a wide range of length scales. Since many fractals display some form of self-similarity, Vassillicos [43] introduced fractals to study turbulent flows. It was shown that both locally or globally self-similar interfaces have a power spectrum of the form $\Gamma(k) \sim k^{-p}$ at large wavenumbers k where noninteger p is related to the Kolmogorov capacity D_K of the interface. The value of D_K is in agreement with experimental results. Several fractal models of turbulence have received considerable attention from Vassillicos [43], Sreenivasan and Meneveau [44]. Their analysis revealed some complicated geometric features of turbulent flows. They showed that several features of turbulence can be described approximately by fractals and that fractal dimensions can be calculated. However, these studies can hardly prove that turbulence can be described fully by fractals. Indeed, these models now constitute a problem in themselves. So fractal models of turbulence have not yet been fully successful.

In view of several difficulties with fractal models of turbulence, multifractal approach with a continuous spectrum of fractal dimension D has been developed by several authors including Meneveau and Sreenivasan [45] and Benzi et al. [46]. These models produced scale exponents which are in agreement with experimental results with a single free parameter. However, it is important to point out that both the multifractal models and log-normal models lack true dynamical motivation.

We close this section by adding some comments on the possible development of singularities in turbulence. Mandelbrot [47] has remarked that "the turbulent solutions of the basic equations involve singularities or near-singularities' (approximate singularities valid down to local viscous length scales where the flow is regular) of an entirely new kind." He also stated that "the singularities of the solutions of the Navier-Stokes equations can only be fractals." In his authoritative review, Sreenivasan [48] described the major influence of the fractal and multifractal formalisms in understanding some aspects of turbulence, but he pointed out some inherent problems in these formalisms with the following comment, "However, the outlook for certain other aspects is not so optimistic, unless magical inspiration or breakthrough in analytical tools occur."

Indeed, several theoretical works and experimental observations revealed that turbulence possesses some sort of singularities in the velocity field or vorticity field. Sarkar's [49] analytical treatment confirmed that finite-time cusp singularities always exist for essentially any arbitrary set of initial data, and are shown to be generic. New experimental methods (Hunt and Vassilicos, [50]) also provide evidence of spiraling streamlines and streaklines within eddies, and thin layers of large vorticity grouped together (Schwarz [51]); both of these features are associated with accumulation points in the velocity field. It also follows from solutions of the Navier-Stokes equations (Vincent and Meneguzzi [52] and She et al. [53]), that very large deviations exist in isolated eddies with complicated internal structure. These studies identify regions of intense vorticity so that streamlines form spirals. The Kolmogorov inertial energy spectrum $k^{-5/3}$ also implies that there must be singularities in the derivatives of the velocity field on scales where the rate of energy dissipation is locally very large. It has been suggested by Moffatt [54] that the accumulation points of discontinuities associated with spiral structures could give rise to fractional power laws k^{-2p} with $1 < 2p < 2$. The question also arises whether the self-similarity leading to the Kolmogorov spectrum is local or global. Moffatt's analysis (see Vassilicos [55]) has shown that spiral singularities are responsible for non-integer power of self-similar spectra k^{-2p}. It is also known now that locally self-similar structures have a self-similar high wavenumber spectrum with a non-integer power $2p$. Thus the general conclusion is that functions with the Kolmogorov spectrum have some kind of singularities and accumulation points, unless they are fractal functions with singularites everywhere, since they are everywhere continuous but nowhere differentiable. Thus the upshot of this discussion is that the statistical structure of the small-scale turbulent flows is determined by local regions where the velocity and any other associated scalar functions have very large derivatives or have rapid variations in their magnitude or that of their derivatives. These are regions surrounding points

that are singular. It remains an open question whether the nature of this singularity is due to random fluctuations of the turbulent motions resulting from their chaotic dynamics or to the presence of localized singular structures originating from an internal organization of the turbulent flows. In spite of a lot of progress made in the last two decades, there are still many open questions than answers. Indeed, the problem of turbulence remain unsolved due to many complexities involved in the problem. It would be a challenging problem for the 21st century (see Debnath [56]).

2.7 Fractals, Dynamical Systems And Iterative Mappings

Based on his pioneering work on celestial mechanics, Henri Poincaré (1854-1912) laid the foundation of what is now called *dynamical systems and iterative mappings*. He pointed out that even though Newtonian mechanics is determinitstic, the motion of celestial bodies attracting one another are very complex in the sense that in the long run this behaivor is unpredictable and even chaotic. Both chaos and fractals have received widespread attention in recent years. There are two major features of fractals. First, in many cases, a small number of parameters or invariants can be used to describe a complex fractal structure. Second, many fractals are naturally generated by underlying dynamical system. Such a system can often describe the relationships between different parts or more importantly different scales of fractals. As far as fractals are concerned, iterative mappings in the plane - the so called *Poincaré mappings* - are of special significance. The fact that this is a dynamical system means that these mappings are conservative (or area-conserving) in the sense that an arbitrary circle can be transformed into a closed curve of the same area. More importantly, iterative mappings can, in general, be used to serve as a model for a wide variety of geometrical and physical phenomena.

Two French mathematicians, Gaston Julia (1893-1978) and Pierre Fatou (1878-1929) considered an elementary quadratic transformation $x \to x^2 + c$ by replacing real x by a complex number $z = x + iy$ and real c by a complex $c = a + ib$. Such a replacement seems to be a minor change and gives

$$z \quad \to \quad z^2 + c = (x + iy)^2 + (a + ib)\,, \tag{37}$$

where a and b are arbitrary real numbers. This is an important example of a conformal transformation that leaves angles unchanged. For every value of a and b, fractals are generated by this transformation, and they are called the *Julia fractals* that are visible as beautiful color pictures on the screen of a computer. Fractals, chaos, bifurcations, and Hausdorff dimension have been essential elements in the sturdy of Julia sets as described by many authors including Keen [58] with many

open questions and unsolved problems. Several books by Mandelbrot [5], Lauwerier [35], Peitgen and Richter [36] contain a large number of beautiful color photographs of fractals and many examples of computer graphics.

It is amazing that the computational iteration process of the transformation (37) produces a totally unexpected result. It turns out that computational experiments on this complex quadratic transformation generate a new geometric structure that is very complex and very widely known to be extremely beautiful. These Julia fractals (or Julia sets) bear a remarkable resemblance to a shape called *snowflake curve* that is very similar to the Koch curve discovered by Von Koch in 1904.

It was shown that many more Julia fractals can be obtained easily using complex numbers. The question was raised about the nature and type of Julia fractals $J(a, b)$ of the model (37). It turned out that they can be either totally disconnected or totally connected. Based on a study of Mandelbrot, it turns out that all points for which the Julia fractals $J(a, b)$ is connected constitute the so called *Mandelbrot fractal* whose beautiful photograph is shown in many papers and books including Lauwerier ([35] Figure 7.11, page 150-154), Peitgen and Ritcher [36], and Mandelbrot [57]. The pioneering research work of Douady and Hubbard [59] provided major results and understanding of many aspects of the Mandelbrot set. On the other hand, Branner [60] described many mathematical properties of this set with many beautiful pictures. It turns out that the Mandelbrot set is not only compact in a plane, but also a connected and cellular as it is equal to the intersections of a nested sequence of sets homeomorphic to solid balls. Indeed, the Mandelbrot set was proved to be a self-similar and universal in nature.

On the other hand, in 1969, a French mathematician and astronomer M. Hénon investigated a new iterative mapping defined by

$$x' = x, \qquad y' = -x + 2ay + y^2. \tag{38}$$

This mapping serves a model for a wide range of physical phenomena, from celestial mechanics to particle physics. A wide variety of computer experiments has been performed on this model which reveals different aspects of chaos and self-similarity - indeed fractal-like structures. Poincaré had predicted their existence, and they are now visible with the aid of modern computer technology.

Another Hénon model deals with iterative transformations

$$\left.\begin{aligned} x_{n+1} &= a\,x_n - b\,(y_n - x_n^2) \\ y_{n+1} &= b\,x_n + a\,(y_n - x_n^2) \end{aligned}\right\}, \tag{39}$$

where $a = \cos\theta$ and $b = \sin\theta$. This model was studied for different values of θ which reveals a closed island structure with the origin as a point of stable equilibrium. This study also shows an irregular external orbit - the so called *chaotic orbit*. The structure generated by (39) consists of periodic cycles that are either stable or unstable as stable orbits which fill up a closed curve and unstable chaotic orbits. So this structure is universal and, in general, applies to every area-conserving transformation.

All of the above discussion reveal that the iteration process of different simple mathematical transformations with the aid of modern computer technology has the tremendous ability and power to produce a totally new and unexpected complex fractal structure that is surprisingly very beautiful. In this connection, the reader is referred to an interesting survey article by Blanchard [61] on complex iteration process and complex analytic dynamics on the Riemann sphere.

3. WAVELETS, WAVELET TRANSFORMS AND MULTIRESOLUTION ANALYSIS

The Fourier transform analysis has also been very useful in many areas, including quantum mechanics, wave motion, and turbulence. In these areas, the Fourier transform $\widehat{f}\,(k)$ (see Debnath [31]) of a function $f\,(x)$ is defined by the space and wavenumber domains, where x represents the space variable and k is the wavenumber. One of the important features is that the trigonometric kernel $\exp\,(-ikx)$ in the Fourier transform oscillates indefinitely, and hence, the localized information contained in the signal $f\,(x)$ in the x-space is widely distributed among $\widehat{f}\,(k)$ in the Fourier transform space. Although $\widehat{f}\,(k)$ does not lose any information of the signal $f\,(x)$, it spreads out in the k-space. If there are computational or observational errors involved in the signal $f\,(x)$, it is almost impossible to study its properties from those of $\widehat{f}\,(k)$.

The Fourier transform theory has been very useful for analyzing harmonic signals or signals for which there is no need for local information. In spite of great success, Fourier transform analysis seems to be inadequate for studying the above physical problems for at least two reasons. First, the Fourier transform of a signal does not contain any local information in the sense that it does not reflect the

change of wavenumber with space or of frequency with time. Second, the Fourier transform method enables us to investigate problems either in the time (space) domain or in the frequency (wavenumber) domain, but not simultaneously in both domains. These are probably the major weaknesses of the Fourier transform analysis. It is often necessary to define a single transform of time and frequency (or space and wavenumber) in both time and frequency domains. Such a single transform would give complete time and frequency (or space and wavenumber) information of a signal.

In 1982, Jean Morlet et al. [6, 7] a French geophysical engineer, discovered the idea of the wavelet transform, providing a new mathematical tool for seismic wave analysis. In Morlet's analysis, signals consist of different features in time and frequency, but their high-frequency components would have a shorter time duration than their low-frequency components. In order to achieve good time resolution for the high-frequency transients and good frequency resolution for the low-frequency components, Morlet et al. [6, 7] first introduced the idea of wavelets as a family of functions constructed from translations and dilations of a single function called the "mother wavelet" $\psi(t)$. They are defined by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\,\psi\left(\frac{t-b}{a}\right), \qquad a, b \in \mathbb{R},\ a \neq 0, \tag{40}$$

where a is called a *scaling parameter* which measures the degree of compression or scale, and b a *translation parameter* which determines the time location of the wavelet. If $|a| < 1$, the wavelet (40) is the compressed version (smaller support in time-domain) of the mother wavelet and corresponds mainly to higher frequencies. On the other hand, when $|a| > 1$, $\psi_{a,b}(t)$ has a larger time-width than $\psi(t)$ and corresponds to lower frequencies. Thus, wavelets have time-widths adapted to their frequencies. This is the main reason for the success of the Morlet wavelets in signal processing and time-frequency signal analysis. It may be noted that the resolution of wavelets at different scales varies in the time and frequency domains as governed by the Heisenberg uncertainty principle. At large scale, the solution is coarse in the time domain and fine in the frequency domain. As the scale a decreases, the resolution in the time domain becomes finer while that in the frequency domain becomes coarser.

Morlet first developed a new time-frequency signal analysis using what he called "wavelets of constant shape" in order to contrast them with the analyzing functions in the short-time Fourier transform which do not have a constant shape. It was Alex Grossman, a French theoretical physicist, who quickly recognized the importance of the Morlet wavelet transforms which are somewhat similar to the formalism for

coherent states in quantum mechanics, and developed an exact inversion formula for this wavelet transform. Unlike the Weyl-Heisenberg coherent states, these coherent states arise from translations and dilations of a single function. They are often called *affine coherent states* because they are associated with an affine group (or "$ax+b$" group). From a group-theoretic point of view, the wavelets $\psi_{a,b}(x)$ are in fact the result of the action of the operators $U(a,b)$ on the function ψ so that

$$[U(a,b)\psi](x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right). \tag{41}$$

These operators are all unitary on the Hilbert space $L^2(\mathbb{R})$ and constitute a representation of the "$ax+b$" group:

$$U(a,b)U(c,d) = U(ac, b+ad). \tag{42}$$

This group representation is *irreducible*, that is, for any non-zero $f \in L^2(\mathbb{R})$, there exists no nontrivial g orthogonal to all the $U(a,b)f$. The success of Morlet's numerical algorithms prompted Grossmann to make a more extensive study of the Morlet wavelet transform which led to the recognition that wavelets $\psi_{a,b}(t)$ correspond to a square integrable representation of the affine group. Grossmann was concerned with the *wavelet transform* of $f \in L^2(\mathbb{R})$ defined by (see Debnath [31])

$$\mathcal{W}_\psi[f](a,b) = \langle f, \psi_{a,b}\rangle = \int_{-\infty}^{\infty} f(t)\,\overline{\psi_{a,b}(t)}\,dt, \tag{43}$$

where $\psi_{a,b}(t)$ plays the same role as the kernel $\exp(i\omega t)$ in the Fourier transform and $\langle f, g\rangle$ represents an inner product in the Hilbert space $L^2(\mathbb{R})$. This is called a *continuous wavelet transform* of $f(t)$. Like the Fourier transform, the continuous wavelet transformation $\mathcal{W}_\psi$ is linear. However, unlike the Fourier transform, the continuous wavelet transform is not a single transform, but any transform obtained in this way. The inverse wavelet transform can be defined so that f can be reconstructed by means of the formula

$$f(t) = C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{W}_\psi[f](a,b)\,\psi_{a,b}(t)\left(a^{-2}da\right)db, \tag{44}$$

provided C_ψ satisfies the so called *admissibility condition*

$$C_\psi = 2\pi\int_{-\infty}^{\infty}\frac{\left|\widehat{\psi}(\omega)\right|^2}{|\omega|}d\omega < \infty, \tag{45}$$

where $\widehat{\psi}(\omega)$ is the Fourier transform of the mother wavelet $\psi(t)$.

Grossmann's ingenious work also revealed that certain algorithms that decompose a signal on the whole family of scales, can be utilized as an efficient tool for multiscale analysis. In practical applications involving fast numerical algorithms, the continuous wavelet can be computed at discrete grid points. To do this a general wavelet ψ can be defined by replacing a with a_0^m $(a_0 \neq 0, 1)$, b with $nb_0a_0^m$ $(b_0 \neq 0)$, where m and n are integers, and making

$$\psi_{m,n}(t) = a_0^{-m/2}\,\psi\left(a_0^{-m}t - nb_0\right). \tag{46}$$

The discrete wavelet transform of f is defined as the doubly indexed sequence

$$\tilde{f}(m,n) = \mathcal{W}[f](m,n) = \langle f, \psi_{m,n}\rangle = \int_{-\infty}^{\infty} f(t)\ \overline{\psi}_{m,n}(t)\,dt, \tag{47}$$

where $\psi_{m,n}(t)$ is given by (46). The double series

$$f(t) = \sum_{m,n=-\infty}^{\infty} \tilde{f}(m,n)\,\psi_{m,n}(t) = \sum_{m,n=-\infty}^{\infty} \langle f, \psi_{m,n}\rangle\,\psi_{m,n}(t), \tag{48}$$

is called the *wavelet series* of f, and the functions $\{\psi_{m,n}(t)\}$ are called the *discrete wavelets*, or simply *wavelets*. However, there is no guarantee that the original function f can be reconstructed from its discrete wavelet coefficients in general. The reconstruction of f is still possible if the discrete lattice has a very fine mesh. For very coarse meshes, the coefficients may not contain sufficient information for determination of f from these coefficients. However, for certain values of the lattice parameter (m,n), a numerically stable reconstruction formula can be obtained.

If the set $\{\psi_{m,n}(t)\}$ defined by (46) is complete in $L^2(\mathbb{R})$ for some choice of ψ, a and b, then the set is called an *affine wavelet*. Then any $f(t) \in L^2(\mathbb{R})$ can be completely determined by (48). Such a complete set $\{\psi_{m,n}(t)\}$ in $L^2(\mathbb{R})$ is called a *frame*. A frame does not satisfy the Parseval theorem for the Fourier series, and the expansion in terms of a frame is not unique. In fact, it can be shown that

$$A\|f\|^2 \leq \sum_{m,n=-\infty}^{\infty} |\langle f, \psi_{m,n}\rangle|^2 \leq B\|f\|^2, \tag{49}$$

where A and B are two constants and $\|f\| = \sqrt{\langle f, f\rangle}$ is the norm of f. The set $\{\psi_{m,n}(t)\}$ constitutes a frame if $\psi(t)$ satisfies the admissibility condition and $0 < A < B < \infty$. Considerable attention has been given to find some necessary and sufficient conditions for a system of wavelets to form a frame or orthonormal basis (see Debnath [31]).

For computational efficiency, $a_0 = 2$ and $b_0 = 1$ are commonly used so that results lead to a binary dilation 2^{-m} and a dyadic translation of $n\,2^m$. Therefore, a practical lattice is $a_0 = 2^m$ and $b_0 = n\,2^m$ in (46) so that

$$\psi_{m,n}(t) \;=\; 2^{-\frac{m}{2}}\psi\left(2^{-m}t - n\right). \tag{50}$$

We sketch a typical mother wavelet with a compact support $[-T, T]$ in Figure 9(a). Different values of the parameter b represent the time localization center, and each $\psi_{a,b}(t)$ is localized around the center $t = b$. As scale parameter a varies, wavelet $\psi_{a,b}(t)$ covers different frequency ranges. Small values of $|a|\,(0 < |a| \ll 1)$ result in very narrow windows and correspond to high frequencies or very fine scales $\psi_{a,b}$, as shown in Figure 9(b), whereas very large values of $|a|\,(|a| \gg 1)$ result in very wide windows and correspond to small frequencies or very coarse scales $\psi_{a,b}$ as shown in Figure 9(c). The wavelet transform (43) gives a time-frequency description of a signal f. Different shapes of the wavelets are plotted in Figure 9(b) and 9(c).

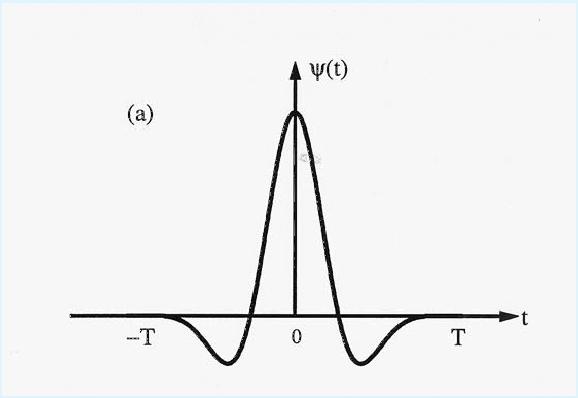


Figure 9(a): Typical mother wavelet.

It follows from the preceding discussion that a typical mother wavelet physically appears as a local oscillation (or wave) in which most of the energy is localized to a narrow region in the physical space. This can be shown that the time resolution σ_t and the frequency resolution σ_ω are proportional to the scale a and a^{-1}, respectively, and $\sigma_t\sigma_\omega \geq 2^{-1}$. When a decreases or increases, the frequency support of the wavelet atom is shifted toward higher or lower frequencies, respectively. Therefore, at higher frequencies, the time resolution becomes finer (better) and the frequency resolution becomes coarser (worse). On the other hand, the time resolution becomes coarser but the frequency resolution becomes finer at lower frequencies. As a function of

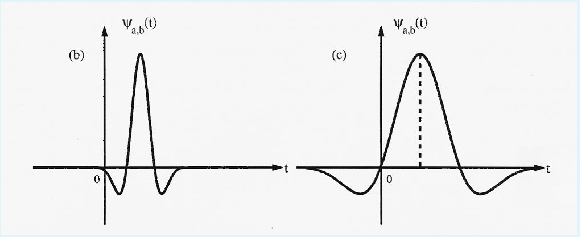


Figure 9(b): Compressed and translated wavelet $\psi_{a,b}(t)$ with $0 < |a| \ll 1$, $b > 0$;

9(c) Magnified and translated wavelet $\psi_{a,b}(t)$ with $|a| \gg 1$, $b > 0$.

b for a fixed scaling parameter a, $\mathcal{W}_\psi[f](a,b)$ represents the detailed information contained in the signal $f(t)$ at the scale a. In fact, this interpretation motivated Morlet et al. [6, 7] to introduce the translated and scaled versions of a single function for the analysis of seismic waves.

We next give a formal definition (see Debnath [31]) of a wavelet.

Definition 1 (Wavelet). A *wavelet* is a function $\psi \in L^2(\mathbb{R})$ which satisfies the condition

$$C_\psi = \int_{-\infty}^{\infty} \frac{\left|\widehat{\psi}(\omega)\right|^2}{|\omega|} d\omega < \infty, \tag{51}$$

where $\widehat{\psi}(\omega)$ is the Fourier transform of $\psi(t)$.

If $\psi \in L^2(\mathbb{R})$, then, a family of wavelets defined by (40), $\psi_{a,b}(t) \in L^2(\mathbb{R})$ for all a, b. For

$$\|\psi_{a,b}(t)\|^2 = |a|^{-1} \int_{-\infty}^{\infty} \left|\psi\left(\frac{t-b}{a}\right)\right|^2 dt = \int_{-\infty}^{\infty} |\psi(x)|^2 dx = \|\psi\|^2. \tag{52}$$

The Fourier transform of $\psi_{a,b}(t)$ is given by

$$\widehat{\psi}_{a,b}(\omega) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-i\omega t} \psi\left(\frac{t-b}{a}\right) dt = |a|^{\frac{1}{2}} e^{-ib\omega} \widehat{\psi}(a\omega), \tag{53}$$

where $\widehat{\psi}(\omega) = \mathcal{F}\{\psi(t)\}$ is the Fourier transform of $\psi(t)$.

Using the Parseval relation of the Fourier transform (see Debnath [31]), it also follows from (43) that

$$\begin{aligned} \mathcal{W}_\psi[f](a,b) &= \langle f, \psi_{a,b}\rangle = \frac{1}{2\pi}\left\langle \widehat{f}, \widehat{\psi}_{a,b}\right\rangle \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\sqrt{|a|}\,\widehat{f}(\omega)\,\overline{\widehat{\psi}}(a\omega)\right\}e^{ib\omega}d\omega, \qquad \text{by (53)}. \end{aligned}$$

This means that

$$\mathcal{F}\{\mathcal{W}_\psi[f](a,b)\} = \int_{-\infty}^{\infty} e^{-ib\omega}\mathcal{W}_\psi[f](a,b)\,db = \sqrt{|a|}\,\widehat{f}(\omega)\,\overline{\widehat{\psi}}(a\omega). \tag{54}$$

Example 1 (The Haar Wavelet). The *Haar wavelet* is one of the classic examples of wavelets. It is defined by

$$\psi(t) = \left\{\begin{array}{ll} 1, & 0 \le t < \frac{1}{2} \\ -1, & \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{array}\right\}. \tag{55}$$

The Haar wavelet has a compact support. It is obvious that

$$\int_{-\infty}^{\infty}\psi(t)\,dt = 0, \qquad \int_{-\infty}^{\infty}|\psi(t)|^2\,dt = 1. \tag{56}$$

This wavelet is very well-localized in the time domain, but it is not continuous. Its Fourier transform $\widehat{\psi}(\omega)$ is given by

$$\widehat{\psi}(\omega) = i\exp\left(-\frac{i\omega}{2}\right)\frac{\sin^2\left(\frac{\omega}{4}\right)}{\left(\frac{\omega}{4}\right)}, \tag{57}$$

and

$$\int_{-\infty}^{\infty}\frac{\left|\widehat{\psi}(\omega)\right|^2}{|\omega|}d\omega = 16\int_{-\infty}^{\infty}|\omega|^{-3}\left|\sin\frac{\omega}{4}\right|^4 d\omega < \infty. \tag{58}$$

Both $\psi(t)$ and $\widehat{\psi}(\omega)$ are plotted in Figure 10. These figures indicate that the Haar wavelet has good time localization but poor frequency localization. The function $\left|\widehat{\psi}(\omega)\right|$ is even, attains its maximum at the frequency $\omega_0 \sim 4.662$, and decays slowly as ω^{-1} as $\omega \to \infty$, which means that it does not have compact support in the

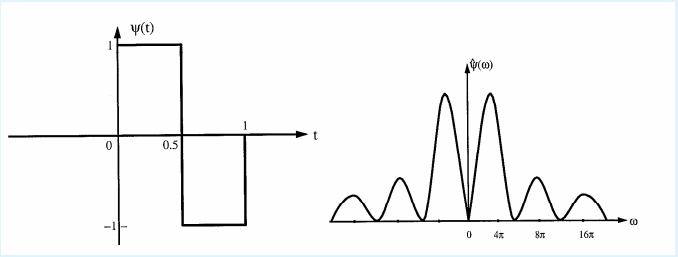


Figure 10: The Haar wavelet and its Fourier transform.

frequency domain. Indeed, the discontinuity of ψ causes a slow decay of $\widehat{\psi}$ as $\omega \to \infty$. Its discontinuous nature is a serious weakness in many applications. However, the Haar wavelet is one of the most fundamental examples that illustrate major features of the general wavelet theory.

Theorem 1. If ψ is a wavelet and ϕ is a bounded integrable function, then the convolution function $\psi * \phi$ is a wavelet.

The proof of this theorem is fairly easy and the reader is referred to Debnath [31].

Example 2. This example illustrate how to generate other wavelets by using Theorem 1. For example, if we take the Haar wavelet and convolute it with the following function

$$\phi(t) = \left\{ \begin{array}{ll} 0, & t < 0 \\ 1, & 0 \leq t \leq 1 \\ 0, & t \geq 1 \end{array} \right\}, \tag{59}$$

we obtain a simple wavelet, as shown in Figure 11.

Example 3. The convolution of the Haar wavelet with $\phi(t) = \exp(-t^2)$ generates a smooth wavelet, as shown in Figure 12.

In order for the wavelets to be useful analyzing functions, the mother wavelet must have certain properties. One such property is defined by the condition (51) which guarantees the existence of the inversion formula for the continuous wavelet

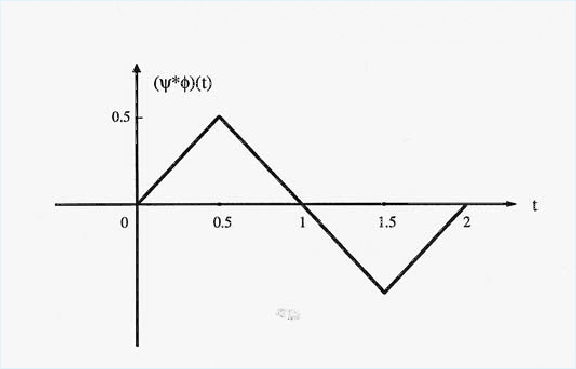


Figure 11: The wavelet $(\psi * \phi)(t)$.

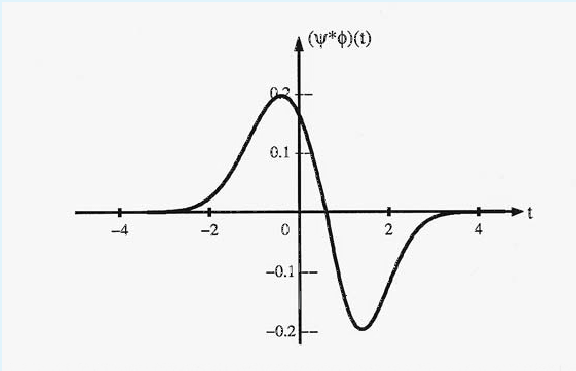


Figure 12: The wavelet $(\psi * \phi)(t)$.

transform. Condition (51) is usually referred to as the *admissibility condition* for the mother wavelet. If $\psi \in L^1(\mathbb{R})$, then its Fourier transform $\widehat{\psi}$ is continuous. Since $\widehat{\psi}$ is continuous, C_ψ can be finite only if $\widehat{\psi}(0) = 0$ or, equivalently, $\int_{-\infty}^{\infty} \psi(t)\, dt = 0$. This means that ψ must be an oscillatory function with zero mean. Condition (51) also imposes a restriction on the rate of decay of $\left|\widehat{\psi}(\omega)\right|^2$ and is required in finding the inverse of the continuous wavelet transform.

In addition to the admissibility condition, there are other properties that may be useful in particular applications. For example, we may want to require that ψ be n times continuously differentiable or infinitely differentiable. If the Haar wavelet is convoluted $(n+1)$ times with the function ϕ given in Example 2, then the resulting

function $\psi * \phi * \ldots * \phi$ is an n times differentiable wavelet. The function in Figure 12 is an infinitely differentiable wavelet. The so-called "Mexican hat wavelet" is another example of an infinitely differentiable (or smooth) wavelet as stated below.

Example 4 (The Mexican Hat Wavelet). The Mexican hat wavelet is defined by the second derivative of a Gaussian function as

$$\psi(t) = \left(1-t^2\right)\exp\left(-\frac{t^2}{2}\right) = -\frac{d^2}{dt^2}\exp\left(-\frac{t^2}{2}\right) = \psi_{1,0}(t), \tag{60}$$

$$\widehat{\psi}(\omega) = \widehat{\psi}_{1,0}(\omega) = \sqrt{2\pi}\,\omega^2\exp\left(-\frac{\omega^2}{2}\right). \tag{61}$$

In contrast to the Haar wavelet, the Mexican hat wavelet is C^∞-function. It has two vanishing moments. The Mexican hat wavelet $\psi_{1,0}(t)$ and its Fourier transform are shown in Figures 13(a) and 13(b). This wavelet has excellent localization in time and frequency domains and clearly satisfies the admissibility condition.

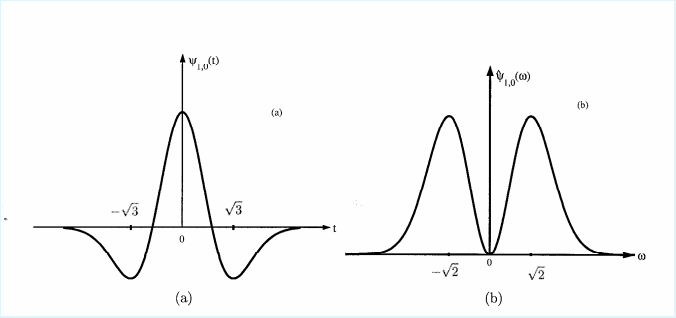


Figure 13: (a) The Mexican hat waveletwavelet $\psi_{1,0}(t)$ and (b) its Fourier transform $\widehat{\psi}_{1,0}(\omega)$.

Two other wavelets, $\psi_{\frac{3}{2},-2}(t)$ and $\psi_{\frac{1}{4},\sqrt{2}}(t)$, from the mother wavelet (60) can be obtained. These three wavelets, $\psi_{1,0}(t)$, $\psi_{\frac{3}{2},-2}(t)$ and $\psi_{\frac{1}{4},\sqrt{2}}(t)$, are shown in Figure 14 (i), (ii), and (iii), respectively.

Example 5 (The Morlet Wavelet). The Morlet wavelet is defined by

$$\psi(t) = \exp\left(i\omega_0 t - \frac{t^2}{2}\right), \tag{62}$$

$$\widehat{\psi}(\omega) = \sqrt{2\pi}\,\exp\left[-\frac{1}{2}(\omega-\omega_0)^2\right]. \tag{63}$$

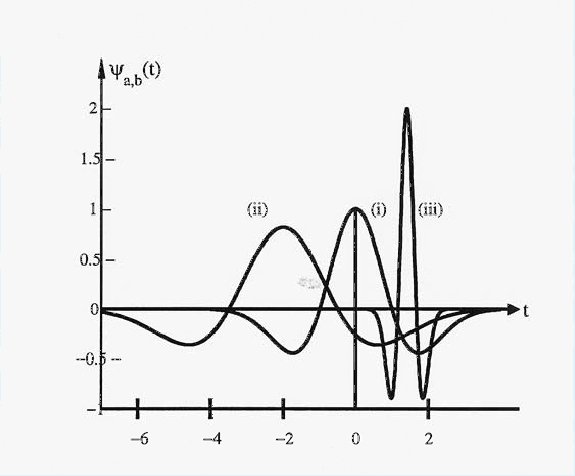


Figure 14: Three wavelets $\psi_{1,0}(t)$, $\psi_{\frac{3}{2},-2}(t)$, and $\psi_{\frac{1}{4},\sqrt{2}}(t)$.

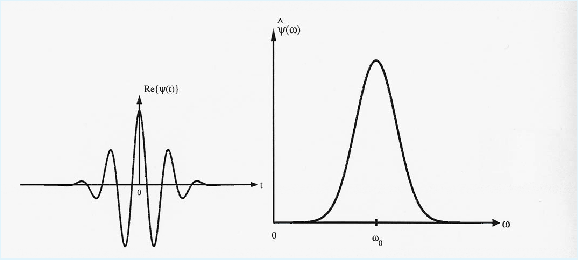


Figure 15: The Morlet wavelet and its Fourier transform

The Morlet wavelet and its Fourier transform are plotted in Figures 15.

3.1 Some Basic Properties of Wavelet Transforms

The following theorem gives several properties of continuous wavelet transforms.

Theorem 2. If ψ and ϕ are wavelets and f, g are functions which belong to $L^2(\mathbb{R})$, then
(i) *(Linearity)*

$$\mathcal{W}_\psi(\alpha f + \beta g)(a,b) = \alpha(\mathcal{W}_\psi f)(a,b) + \beta(\mathcal{W}_\psi g)(a,b), \tag{64}$$

where α and β are any two scalars.
(ii) *(Translation)*

$$\left(\mathcal{W}_\psi\left(T_c f\right)\right)(a,b) \;=\; \left(\mathcal{W}_\psi f\right)(a,b-c)\,, \tag{65}$$

where T_c is the *translation operator* defined by $T_c f\,(t) = f\,(t-c)$.
(iii) *(Dilation)*

$$\mathcal{W}_\psi\left(D_c f\right)(a,b) \;=\; \frac{1}{\sqrt{c}}\left(\mathcal{W}_\psi f\right)\left(\frac{a}{c},\frac{b}{c}\right), \qquad c>0, \tag{66}$$

where D_c is a *dilation operator* defined by and $D_c f\,(t) = \frac{1}{c}\, f\left(\frac{t}{c}\right)$, $c>0$.
(iv) *(Symmetry)*

$$\left(\mathcal{W}_\psi f\right)(a,b) \;=\; \overline{\left(\mathcal{W}_{\widehat{f}}\,\psi\right)\left(\frac{1}{a},-\frac{b}{a}\right)}, \qquad a\neq 0. \tag{67}$$

(v) *(Parity)*

$$\left(\mathcal{W}_{P\psi} P f\right)(a,b) \;=\; \left(\mathcal{W}_\psi f\right)(a,-b)\,, \tag{68}$$

where P is the parity operator defined by $P\,f\,(t) = f\,(-t)$.
(vi) *(Antilinearity)*

$$\left(\mathcal{W}_{\alpha\psi+\beta\phi} f\right)(a,b) \;=\; \overline{\alpha}\left(\mathcal{W}_\psi f\right)(a,b) + \overline{\beta}\left(\mathcal{W}_\phi f\right)(a,b)\,, \tag{69}$$

for any scalars α, β.

$$\text{(vii)} \qquad \left(\mathcal{W}_{T_c\psi} f\right)(a,b) \;=\; \left(\mathcal{W}_\psi f\right)(a,b+ca)\,. \tag{70}$$

$$\text{(viii)} \qquad \left(\mathcal{W}_{D_c\psi} f\right)(a,b) \;=\; \frac{1}{\sqrt{c}}\left(\mathcal{W}_\psi f\right)(ac,b)\,, \qquad c>0. \tag{71}$$

Proofs of the above properties are straightforward and are left as exercises.

Theorem 3 (Parseval's Formula for Wavelet Transforms). If $\psi \in L^2(\mathbb{R})$ and $\left(\mathcal{W}_\psi f\right)(a,b)$ is the wavelet transform of f defined by (43), then, for any functions f, $g \in L^2(\mathbb{R})$, we obtain

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\mathcal{W}_\psi f\right)(a,b)\,\overline{\left(\mathcal{W}_\psi g\right)(a,b)}\,\frac{dbda}{a^2} \;=\; C_\psi\,(f,g)\,, \tag{72}$$

where C_ψ is given by (51).

We refer to Debnath [31] for a detailed proof of this theorem.

Theorem 4 (Inversion Formula). If $f \in L^2(\mathbb{R})$, then f can be reconstructed by the formula

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{W}_\psi f)(a,b)\, \psi_{a,b}(t)\, \frac{dbda}{a^2}, \tag{73}$$

where the equality holds almost everywhere.

Proof. For any $g \in L^2(\mathbb{R})$, we have, from Theorem 3.

$$\begin{aligned} C_\psi \langle f, g \rangle &= \langle \mathcal{W}_\psi f, \mathcal{W}_\psi g \rangle \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{W}_\psi f)(a,b)\, \overline{(\mathcal{W}_\psi g)(a,b)}\, \frac{dbda}{a^2}, \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{W}_\psi f)(a,b)\, \overline{\int_{-\infty}^{\infty} g(t)\, \overline{\psi_{a,b}(t)} dt}\, \frac{dbda}{a^2}, \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{W}_\psi f)(a,b)\, \psi_{a,b}(t)\, \frac{dbda}{a^2}\, \overline{g(t)}\, dt, \\ &= \left\langle \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{W}_\psi f)(a,b)\, \psi_{a,b}(t)\, \frac{dbda}{a^2}, g \right\rangle. \end{aligned} \tag{74}$$

Since g is an arbitrary element of $L^2(\mathbb{R})$, the inversion formula (73) follows.

If $f = g$ in (73), then

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |(\mathcal{W}_\psi f)(a,b)|^2\, \frac{da\, db}{a^2} = C_\psi \|f\|^2 = C_\psi \int_{-\infty}^{\infty} |f(t)|^2\, dt. \tag{75}$$

This shows that, except for the factor C_ψ, the wavelet transform is an isometry from $L^2(\mathbb{R})$ to $L^2(\mathbb{R}^2)$.

3.2 Multiresolution Analysis

The concept of multiresolution (MRA) analysis of a Hilbert space of functions is related to the study of signals or images at different levels of resolution – almost like a pyramid. This is a new and remarkable idea which deals with a general mathematical formalism for construction of an orthogonal basis of wavelets. Indeed, it is central to all constructions of wavelet bases. Mallat's brilliant work [8] has been the major source of many new developments in wavelet analysis and its wide variety of applications. As Ingrid Dubechics said: "The history of the formalism of multiresolution analysis is a beautiful example of applications stimulating theoretical development."

Mathematically, the fundamental idea of multiresolution analysis is to represent a function (or signal) f as a limit of successive approximations, each of which is a finer version of the function f. These successive approximations correspond to different levels of resolutions. Thus, multiresolution analysis is a formal approach to constructing orthogonal wavelet bases using a definite set of rules and procedures. The key feature of this analysis is to describe mathematically the process of studying signals or images at different scales. The basic principle of the MRA deals with the decomposition of the whole function space into individual subspaces $V_n \subset V_{n+1}$ so that the space V_{n+1} consists of all rescaled functions in V_n. This essentially means a decomposition of each function (or signal) into components of different scales (or frequencies) so that an individual component of the original function f occurs in each subspace. These components can describe finer and finer versions of the original function f. For example, a function is resolved at scales $\triangle t = 2^0, 2^{-1}, \ldots, 2^{-n}$. In audio signals, these scales are basically *octaves* which represent higher and higher frequency components. For images and, indeed, for all signals, the simultaneous existence of a multiscale may also be referred to as *multiresolution*. From the point of view of practical applications, MRA is really an effective mathematical framework for hierarchical decomposition of an image (or signal) into components of different scales (or frequencies).

In general, frames have many of the properties of bases, but they lack a very important property of orthogonality. If the condition of orthogonality

$$\langle \phi_{k.\ell}, \phi_{m,n} \rangle \;=\; 0 \;\text{ for all }\; (k,\ell) \neq (m,n)\,, \tag{76}$$

is satisfied, the reconstruction of the function f from $\langle f, \phi_{m,n} \rangle$ is much simpler and, for any $f \in L^2(\mathbb{R})$, we have the following representation

$$f \;=\; \sum_{m,n=-\infty}^{\infty} \langle f, \phi_{m,n} \rangle \, \phi_{m,n}, \tag{77}$$

where

$$\phi_{m,n}(x) \;=\; 2^{-m/2} \phi\left(2^{-m}x - n\right), \tag{78}$$

is an orthonormal basis of V_m.

Definition 3.2.1 (Multiresolution Analysis). A multiresolution analysis (MRA) consists of a sequence $\{V_m : m \in \mathbb{Z}\}$ of embedded closed subspaces of $L^2(\mathbb{R})$ that satisfy the following conditions:
(i) $\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots V_m \subset V_{m+1} \ldots$,

(ii) $\bigcup_{m=-\infty}^{\infty} V_m$ is dense in $L^2(\mathbb{R})$, that is, $\overline{\bigcup_{m=-\infty}^{\infty} V_m} = L^2(\mathbb{R})$,

(iii) $\bigcap_{m=-\infty}^{\infty} V_m = \{0\}$,

(iv) $f(x) \in V_m$ if and only if $f(2x) \in V_{m+1}$ for all $m \in \mathbb{Z}$,

(v) there exists a function $\phi \in V_0$ such that $\{\phi_{0,n} = \phi(x-n),\ n \in \mathbb{Z}\}$ is an orthonormal basis for V_0, that is,

$$\|f\|^2 = \int_{-\infty}^{\infty} |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |\langle f, \phi_{0,n}\rangle|^2 \qquad \text{for all } f \in V_0.$$

The function ϕ is called the *scaling function* or *father wavelet*. If $\{V_m\}$ is an multiresolution of $L^2(\mathbb{R})$ and if V_0 is the closed subspace generated by the integer translates of a single function ϕ, then we say that ϕ generates the multiresolution analysis.

Sometimes, condition (v) is relaxed by assuming that $\{\phi(x-n), \quad n \in \mathbb{Z}\}$ is a Riesz basis for V_0, that is, for every $f \in V_0$, there exists a unique sequence $\{c_n\}_{n=-\infty}^{\infty} \in \ell^2(\mathbb{Z})$ such that

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \phi(x-n),$$

with convergence in $L^2(\mathbb{R})$ and there exist two positive constants A and B independent of $f \in V_0$ such that

$$A \sum_{n=-\infty}^{\infty} |c_n|^2 \leq \|f\|^2 \leq B \sum_{n=-\infty}^{\infty} |c_n|^2,$$

where $0 < A < B < \infty$. In this case, we have a multiresolution analysis with a Riesz basis.

Note that condition (v) implies that $\{\phi(x-n), \quad n \in \mathbb{Z}\}$ is a Riesz basis for V_0 with $A = B = 1$.

Since $\phi_{0,n}(x) \in V_0$ for all $n \in \mathbb{Z}$, further, if $n \in \mathbb{Z}$, it follows from (iv) that

$$\phi_{m,n}(x) = 2^{m/2}\phi(2^m x - n), \quad m \in \mathbb{Z}, \tag{79}$$

is an orthonormal basis for V_m.

Consequences of Definition 3.2.1.

1. A repeated application of condition (iv) implies that $f \in V_m$ if and only if $f\left(2^k x\right) \in V_{m+k}$ for all $m,\ k \in \mathbb{Z}$. In other words, $f \in V_m$ if and only if $f\left(2^{-m}x\right) \in V_0$ for all $m \in \mathbb{Z}$.
   This shows that functions in V_m are obtained from those in V_0 through a scaling 2^{-m}. If the scale $m = 0$ is associated with V_0, then the scale 2^{-m} is associated with V_m. Thus, subspaces V_m are just scaled versions of the central space V_0, which is invariant under translation by integers, that is, $T_n V_0 = V_0$ for all $n \in \mathbb{Z}$.

2. It follows from Definition 3.2.1 that a multiresolution analysis is completely determined by the scaling function ϕ but not conversely. For a given $\phi \in V_0$, we first define

$$V_0 \;=\; \left\{ f(x) = \sum_{n=-\infty}^{\infty} c_n\, \phi_{0,n} = \sum_{n=-\infty}^{\infty} c_n\, \phi(x-n) : \{c_n\} \in \ell^2(\mathbb{Z}) \right\}.$$

   Condition (iv) implies that V_0 has an orthonormal basis $\{\phi_{0,n}\} = \{\phi(x-n)\}$. Then, V_0 consists of all functions $f(x) = \sum_{n=-\infty}^{\infty} c_n\, \phi(x-n)$ with finite energy $\|f\|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2 < \infty$. Similarly, the space V_m has the orthonormal basis $\phi_{m,n}$ given by (79) so that $f_m(x)$ is given by

$$f_m(x) \;=\; \sum_{n=-\infty}^{\infty} c_{mn}\, \phi_{m,n}(x), \tag{80}$$

   with the finite energy

$$\|f_m\|^2 \;=\; \sum_{n=-\infty}^{\infty} |c_{mn}|^2 < \infty.$$

   Thus, f_m represents a typical function in the space V_m. It builds in self-invariance and scale invariance through the basis $\{\phi_{m,n}\}$.

3. Conditions (ii) and (iii) can be expressed in terms of the orthogonal projections P_m onto V_m, that is, for all $f \in L^2(\mathbb{R})$,

$$\lim_{m\to-\infty} P_m f \;=\; 0 \text{ and } \lim_{m\to+\infty} P_m f = f. \tag{81ab}$$

The projection $P_m f$ can be considered as an approximation of f at the scale 2^{-m}. Therefore, the successive approximations of a given function f are defined as the orthogonal projections P_m onto the space V_m:

$$P_m f = \sum_{n=-\infty}^{\infty} \langle f, \phi_{m,n} \rangle \, \phi_{m,n}, \tag{82}$$

where $\phi_{m,n}(x)$ given by (79) is an orthonormal basis for V_m.

4. Since $V_0 \subset V_1$, the scaling function ϕ that leads to a basis for V_0 is also V_1. Since $\phi \in V_1$ and $\phi_{1,n}(x) = \sqrt{2}\, \phi(2x - n)$ is an orthonormal basis for V_1, ϕ can be expressed in the form

$$\phi(x) = \sum_{n=-\infty}^{\infty} c_n \, \phi_{1,n}(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} c_n \, \phi(2x - n), \tag{83}$$

where

$$c_n = \langle \phi, \phi_{1,n} \rangle \text{ and } \sum_{n=-\infty}^{\infty} |c_n|^2 = 1.$$

Equation (83) is called the *dilation equation*. It involves both x and $2x$ and is often referred to as the *two-scale equation* or *refinement equation* because it displays $\phi(x)$ in the refined space V_1. The space V_1 has the finer scale 2^{-1} and it contains $\phi(x)$ which has scale 1.

All of the preceding facts reveal that multiresolution analysis can be described at least three ways so that we can specify
(a) the subspaces V_m,
(b) the scaling function ϕ,
(c) the coefficients c_n in the dilation equation (83).

The real importance of a multiresolution analysis lies in the simple fact that it enables us to construct an orthonormal basis for $L^2(\mathbb{R})$. In order to prove this statement, we first assume that $\{V_m\}$ is a multiresolution analysis. Since $V_m \subset V_{m+1}$, we define W_m as the orthogonal complement of V_m in V_{m+1} for every $m \in \mathbb{Z}$ so that

we have

$$\begin{aligned} V_{m+1} &= V_m \bigoplus W_m \\ &= \left(V_{m-1} \bigoplus W_{m-1}\right) \bigoplus W_m \\ &= \dots \\ &= V_0 \bigoplus W_0 \bigoplus W_1 \bigoplus \dots \bigoplus W_m \\ &= V_0 \bigoplus \left(\bigoplus_{m=0}^{m} W_m \right) \end{aligned} \tag{84}$$

and $V_n \perp W_m$ for $n \neq m$.

Since $\bigcup_{m=-\infty}^{\infty} V_m$ is dense in $L^2(\mathbb{R})$, we may take the limit as $m \to \infty$ to obtain

$$V_0 \bigoplus \left(\bigoplus_{m=0}^{\infty} W_m \right) = L^2(\mathbb{R}).$$

Similarly, we may go in the other direction to write

$$\begin{aligned} V_0 &= V_{-1} \bigoplus W_{-1} \\ &= \left(V_{-2} \bigoplus W_{-2}\right) \bigoplus W_{-1} \\ &= \dots \\ &= V_{-m} \bigoplus W_{-m} \bigoplus \dots \bigoplus W_{-1}. \end{aligned}$$

We may again take the limit as $m \to \infty$. Since $\bigcap_{m \in Z} V_m = \{0\}$, it follows that $V_{-m} = \{0\}$. Consequently, it turns out that

$$\bigoplus_{m=-\infty}^{\infty} W_m = L^2(\mathbb{R}). \tag{85}$$

Example 3.2.1 (Characteristic Function). If $\phi = \chi_{[0,1]}$ is the characteristic function of the interval $[0,1]$, the spaces V_m defined by

$$V_m = \left\{ \sum_{k=-\infty}^{\infty} c_k \, \phi_{m,k} : \{c_k\} \in \ell^2(\mathbb{Z}) \right\}, \tag{86}$$

where

$$\phi_{m,n}(x) = 2^{-m/2} \phi\left(2^{-m} x - n\right), \tag{87}$$

satisfy all conditions of Definition 3.2.1. So, $\{V_m\}$ is a multiresolution analysis.

Example 3.2.2 (Piecewise Constant Function). Consider the space V_m of all functions $L^2(\mathbb{R})$ which are constant on intervals $[2^{-m}n, 2^{-m}(n+1)]$, where $n \in \mathbb{Z}$. This space V_m constitutes a multiresolution analysis with the scaling function $\phi = \chi_{[0,1]}$. Moreover, ϕ satisfies the dilation equation

$$\phi(x) = \phi(2x) + \phi(2x-1). \tag{88}$$

This means that $\phi(x)$ is a linear combination of the even and odd translates of ϕ as shown in Figure 7.4 on page 432 of Debnath [31].

It can be shown that the Haar mother wavelet (55) can be obtained as a simple *two-scale relation*

$$\psi(x) = \phi(2x) - \phi(2x-1) = \chi_{[0,.5]}(x) - \chi_{[.5,1]}(x). \tag{89}$$

Example 3.2.3 (Cardinal B-Splines and Spline Wavelets). The cardinal B-splines (basic splines) consists of functions in $C^{n-1}(\mathbb{R})$ with equally spaced integer knots that coincide with polynomials of degree n on the interval $[2^{-m}k, 2^{-m}(k+1)]$. These B-splines of order n with compact support generate a linear space V_0 in $L^2(\mathbb{R})$. This leads to a multiresolution analysis $\{V_m, m \in \mathbb{Z}\}$ by $f(x) \in V_m$ if and only if $f(2x) \in V_{m+1}$.

The cardinal B-splines $B_n(x)$ of order n are defined by the following convolution product

$$B_1(x) = \chi_{[0,1]}(x), \tag{90}$$
$$B_n(x) = B_1(x) * B_1(x) * \ldots * B_1(x) = B_1(x) * B_{n-1}(x), \quad (n \geq 2), \tag{91}$$

where n factors are involved in the convolution product.

We state a fundamental result in the following:

Theorem 5. If $\{V_n\}$, $n \in \mathbb{Z}$ is a multiresolution analysis with the scaling funtion ϕ, then there is a mother wavelet ψ given by

$$\psi(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} (-1)^{n-1} \overline{c}_{-n-1} \phi(2x-n), \tag{92}$$

where the coefficients c_n are given by

$$c_n = \langle \phi, \phi_{1,n} \rangle = \sqrt{2} \int_{-\infty}^{\infty} \phi(x) \overline{\phi(2x-n)}\, dx. \tag{93}$$

That is, the system $\{\psi_{m,n}(x) : m,\ n \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

The reader is referred to Debnath [31] for the detailed proof.

Example 3.2.4 (Daubechies' Orthonormal Wavelet). This example illustrates one of the compactly supported orthonormal wavelets first discovered by Daubechies [62]. Making reference to Daubechies [63] or Debnath [31], we outline the construction of this wavelet using the dilation equation for the scaling function ϕ as

$$\phi(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} c_n \phi(2x - n), \tag{94}$$

where $c_n = \langle \phi, \phi_{1,n} \rangle$ and $\sum_{n=-\infty}^{\infty} |c_n|^2 \leq \infty$.

If the scaling function ϕ has compact support, then only a finite number of c_n have nonzero values. The associated generating function $\widehat{m}(\omega)$ is

$$\widehat{m}(\omega) = \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} c_n \exp(-i\,\omega\, n), \tag{95}$$

is a trigonometric polynomial which satisfies the orthogonality condition

$$|\widehat{m}(\omega)|^2 + |\widehat{m}(\omega + \pi)|^2 = 1 \qquad \text{a.e.} \tag{96}$$

with special values $\widehat{m}(0) = 1$ and $\widehat{m}(\pi) = 0$. If coefficients c_n are real, then the corresponding scaling function as well as the mother wavelet ψ will also be real-valued. The Fourier transform $\widehat{\psi}(\omega)$ of $\psi(x)$ corresponding to ϕ is given by the formula

$$\widehat{\psi}(\omega) = \exp\left(\frac{i\omega}{2}\right) \overline{\widehat{m}}\left(\frac{\omega}{2} + \pi\right) \widehat{\phi}\left(\frac{\omega}{2}\right), \tag{97}$$

with $\left|\widehat{\phi}(0)\right| = 1$.

The Fourier transform $\widehat{\psi}(\omega)$ of order N is N-times continuously differentiable, and it satisfies the moment condition

$$\left[\psi^{(k)}(\omega)\right]_{\omega=0} = 0 \qquad \text{for } k = 1, 2, \ldots, m. \tag{98}$$

It follows that $\psi \in C^m$ implies that $\widehat{m}_0(\omega)$ has a zero at $\omega = \pi$ of order $(m+1)$. In other words,

$$\widehat{m}_0(\omega) = \left(\frac{1 + e^{-i\omega}}{2}\right)^{m+1} \widehat{L}(\omega), \tag{99}$$

where $\widehat{L}(\omega)$ is a trigonometric polynomial.

In addition to the orthogonality condition (96), we assume

$$\widehat{m}_0(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^N \widehat{L}(\omega), \tag{100}$$

where $\widehat{L}(\omega)$ is 2π-periodic and $\widehat{L} \in C^{N-1}$. It turns out that

$$|\widehat{m}_0(\omega)|^2 = \left(\cos^2\frac{\omega}{2}\right)^N \left|\widehat{L}_0(\omega)\right|^2, \tag{101}$$

where $\left|\widehat{L}_0(\omega)\right|^2 = Q(\cos\omega)$ is a polynomial in $\cos\omega$ so that $Q(\cos\omega) = Q(1-2x)$ with $x = \sin^2\frac{\omega}{2}$. Consequently, (101) becomes

$$|\widehat{m}_0(\omega)|^2 = (1-x)^N P(x), \tag{102}$$

where $P(x)$ is a polynomial in x.

Using the orthogonality condition (96) and argument of Daubechies [63], it turns out that there exists a unique polynomial $P_N(x)$ of degree $\leq N-1$

$$P_N(x) = \sum_{k=0}^{N-1} \left(\begin{array}{c} N+k-1 \\ k \end{array} \right) x^k, \tag{103}$$

which is positive in $0 < x < 1$ so that $P_N(x)$ is at least a possible candidate for $\left|\widehat{L}(\omega)\right|^2$.

Finally, it turns out for $N = 2$ that

$$\widehat{m}_0(\omega) = \frac{1}{8}\left[\left(1+\sqrt{3}\right) + \left(3+\sqrt{3}\right)e^{-i\omega} + \left(3-\sqrt{3}\right)e^{-2i\omega} + \left(1-\sqrt{3}\right)e^{-3i\omega}\right], \tag{104}$$

where $\widehat{m}_0(0) = 1$.

Comparing the coefficients of (104)with (94) gives c_n as

$$c_0 = \frac{1}{4\sqrt{2}}\left(1+\sqrt{3}\right), \quad c_1 = \frac{1}{4\sqrt{2}}\left(3+\sqrt{3}\right), \quad c_2 = \frac{1}{4\sqrt{2}}\left(3-\sqrt{3}\right)$$

and $$c_3 = \frac{1}{4\sqrt{2}}\left(1-\sqrt{3}\right).$$

Consequently, the Daubechies scaling function ${}_2\phi(x)$ takes the form, dropping the subscript and deleting the factor $\frac{1}{\sqrt{2}}$,

$$\phi(x) = c_0\,\phi(2x) + c_1\,\phi(2x-1) + c_2\,\phi(2x-2) + c_3\,\phi(2x-3). \tag{105}$$

In view of the Theorem 5 with the factor $\frac{1}{\sqrt{2}}$ deleted, it turns out that the corresponding mother wavelet is given by

$$\psi(\omega) = [-c_3\,\phi(2x) + c_2\,\phi(2x-1) - c_1\,\phi(2x-2) + c_0\,\phi(2x-3)], \quad (106)$$

where the coefficients in (106) are the same as for the scaling function $\phi(x)$ and with alternate terms having their signs changed from plus to minus.

It is often referred to as the Daubechies D4 wavelet as it is generated by four coefficients. The reader is referred to Daubechies [62] or Debnath [31] for the Figures of both Daubechies' wavelet $\psi(x)$ and Daubechies' scaling function $\phi(x)$.

4. COMPACTONS AND INTRINSIC LOCALIZED MODES

Rosenau and Hyman [9] first discovered a new class of solitary waves with compact support which are called *compactons*. This new class of solutions is governed by a two-parameter family of strongly dispersive nonlinear equations, denoted by $K(m,n)$,

$$u_t \pm a\,(u^m)_x + b\,(u^n)_{xxx} = 0, \qquad m > 0, \quad 1 < n \leq 3, \qquad (107ab)$$

for certain values of m and n, where a and b are positive real constants. Thus, compactons are defined as solitons with a compact support. In other words, they are solitons with finite wavelength or solitons that are free from exponential trails or wings. Unlike the standard KdV soliton which narrows as the amplitude (speed) increases, the width of a compacton is independent of the amplitude, but its speed depends on its height. Since dispersion increases with amplitude, at high amplitudes, dispersion is more dominant than that of the KdV equation, and hence, it can more effectively counterbalance the effect of nonlinear steepening. Numerous numerical experiments of Rosenau and Hyman [9] confirmed that, when two or more compactons collide, they undergo a nonlinear elastic interaction according to (107ab) and emerge from the interaction with the original form unchanged.

Equation (107a) with $(+a)$ is called the *focusing branch* and admits traveling solitary wave solutions. On the other hand, equation (107b) with $(-a)$ is referred to as the *defocusing branch* and admits solitary wave solutions with cusps or infinite slopes. Thus, equations (107ab) represent two nonlinear models with entirely different physical structures.

We follow Rosenau and Hyman [9] to find the solution of $K(2,2)$ with $a = b = 1$, that is, the equation

$$u_t + \left(u^2\right)_x + \left(u^2\right)_{xxx} = 0. \qquad (108)$$

We seek a traveling wave solution $u = u(\xi)$, $\xi = x - ct$ of (108) and integrate the resulting equation twice to obtain the following nonlinear ordinary differential equation

$$\left(\frac{\partial u}{\partial \xi}\right)^2 + \frac{1}{4}u^2 - \frac{1}{3}c\,u + \frac{c_1}{u^2} \;=\; c_2, \tag{109}$$

where c_1 and c_2 are integrating constants. Putting $c_1 = c_2 = 0$ leads to the solution

$$u(x,t) \;=\; \left\{ \begin{array}{cc} \left(\frac{4c}{3}\right)\cos^2\left[\frac{1}{4}(x-ct)\right], & |x-ct| \le 2\pi \\ \\ 0, & \text{otherwise} \end{array} \right\}. \tag{110}$$

This solution is referred to as *compacton* and is shown in Figure 16.

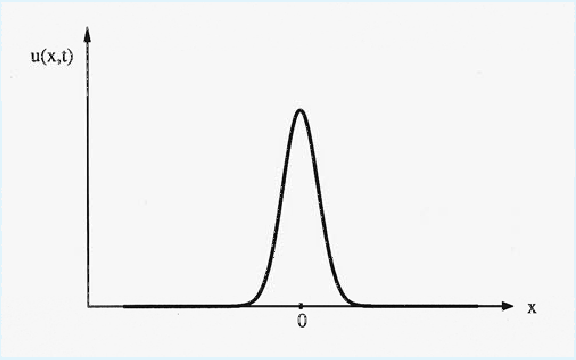


Figure 16: A Compacton.

Although the second derivative of the compacton solution is discontinuous at its edges, it represents a solitary wave with compact support because the third derivative acts on u^2, which has smooth derivatives everywhere including the edges. It has already been indicated that dispersion increases with amplitude, and is more dominant at higher amplitude than the KdV soliton, and hence, it can more effectively counterbalance the steepening effects of nonlinearity so the result is a solitary wave with compact support or compacton.

In general, there are three distinct traveling wave solutions of (109). When $c_1 \neq 0$, the solutions represent waves that can be described by elliptic functions. When $c_1 = 0$, there exists a singular trajectory that describes trigonometric wave solution with period 4π and its amplitude depends on the constant c_2. For $c_2 = 0$, the solution $u(x,t)$ is non-negative and represents a series of compactons. In view of

degeneracy of $K(2,2)$ at $u = 0$, these compactons do not interact with each other, and therefore, can be separated.

It was also shown by Rosenau and Hyman [9] that, for a class of general $K(m,n)$ equations, the compacton solution exists only for $1 < n \leq 3$, and the singular dispersion at $u = 0$ plays a major role on the compactification. The upper limit ($n \leq 3$) is necessary for the existence of compacton solutions in the classical sense.

Based on hundreds of numerical experiments, Rosenau and Hyman [9] have confirmed that, like solitons, two or more compactons physically interact with each other, and they always remain unchanged after collision except for a slight phase shift. Figure 17 exhibits the interaction of three compactons with speeds $c = 2$, 1.5, and 1 and their identities before and after collision.

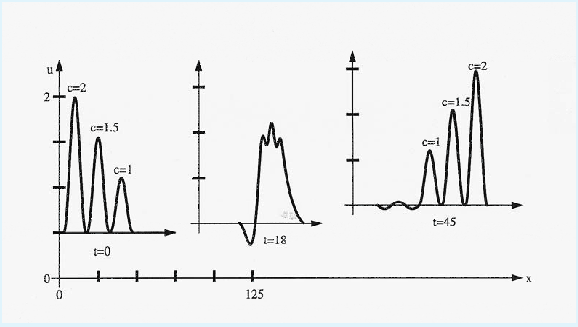


Figure 17: The interaction of three $K(2,2)$ compactons with speeds $c = 2$, 1.5, and 1 starting with centers at $x = 10$, 15, and 40. (Rosenau and Hyman [9]).

It was shown by Oron and Rosenau [64] that $K(m,n)$ type equations arise in the study of nonlinear dispersion in the formation of localized patterns in liquid drops. In their study of a nonlinear model describing new modes of motion of the free surface of a liquid, Ludu and Draayer [65] demonstrated the existence of localized multiple patterns and nonlinear oscillations which include compactons, solitons and cnoidal waves as traveling non-axially symmetric shapes. Subsequently, Ludu et al. [66] proposed a generalized similarity analysis of nonlinear dispersive equations to find a qualitative description of localized solutions. Their study reveals that compactons fulfill both characteristics of solitons and wavelets with possible new applications to the physics of droplets, bubbles, traveling patterns, fragmentation, fission and inertial fusions. Dusuel et al. [67] made an interesting analytical, numerical and experimental study of physical systems modeled by a nonlinear Klein-Gordon equation

with anharmonic coupling, and showed the existence of compactons. In a real physical system, they have also investigated the existence and stability of compactons and kinks consisting of a chain of identical pendulums that are nonlinearly coupled and experience a double-well on site potential.

In general, compacton solutions of $K(m,n)$ equations for any $m \neq n$ are not yet known. We closely follow the method of solution due to Rosenau and Hyman [9] and assume the general solution of $K(n,n)$ equation given by (107a) in the form

$$u(x,t) = A\left[\sin\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}} \tag{111}$$

or, of the form

$$u(x,t) = A\left[\cos\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}}, \tag{112}$$

where A and k are constants to be determined.

Substituting these solutions into (107a) and solving the resulting equations for A and k yields

$$A = \begin{cases} \left(\frac{2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & n \text{ is even} \\ \\ \pm\left(\frac{2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & n \text{ is odd} \end{cases} \tag{113ab}$$

and

$$k = \pm\frac{(n-1)}{2n}\sqrt{\frac{a}{b}}. \tag{114}$$

Consequently, the general compacton solutions are:
(i) For even n

$$u(x,t) = \left[\sqrt{A}\sin\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}} H\left(k|x-ct|-2n\pi\right), \tag{115}$$

and

$$u(x,t) = \left[\sqrt{A}\cos\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}} H\left(k|x-ct|-n\pi\right), \tag{116}$$

where $H(|x|-a)=1$, for $|x|\leq a$, and zero, for $|x|>a$.
(ii) For odd n

$$u(x,t) = \pm\left[\sqrt{A}\sin\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}} H\left(k|x-ct|-2n\pi\right), \tag{117}$$

and

$$u(x,t) = \pm\left[\sqrt{A}\cos\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}} H\left(k|x-ct|-n\pi\right), \tag{118}$$

Similarly, we seek a solution of the one-dimensional *defocusing branch* of $K(n,n)$ equation (107b) in the form

$$u(x,t) = A\left[\sinh\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}}, \tag{119}$$

or

$$u(x,t) = A\left[\cosh\left\{k(x-ct)\right\}\right]^{\frac{2}{n-1}}, \tag{120}$$

where A and k constants to be determined.

Substituting these solutions in (107b) and solving the resulting equations for A and k gives the solutions for the sinh-profile where A and k are given by

$$A = \begin{cases} \left(\frac{2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & \text{where } n \text{ is even,} \\ \\ \pm\left(\frac{2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & \text{where } n \text{ is odd,} \end{cases} \tag{121ab}$$

and

$$k = \pm\frac{(n-1)}{2n}\sqrt{\frac{a}{b}}. \tag{122}$$

For the cosh-profile, we obtain

$$A = \begin{cases} -\left(\frac{2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & n \text{ is even,} \\ \\ \pm\left(\frac{-2nc}{a(n+1)}\right)^{\frac{1}{n-1}}, & n \text{ is odd.} \end{cases} \tag{123ab}$$

Consequently, the general solutions are given as follows:
(i) For even n

$$u(x,t) = \left[\sqrt{|A|}\sinh\left\{|k|(x-ct)\right\}\right]^{\frac{2}{n-1}}, \tag{124}$$

and

$$u(x,t) = -\left[\sqrt{|A|}\,\cos h\left\{|k|(x-ct)\right\}\right]^{\frac{2}{n-1}}. \tag{125}$$

(ii) For odd n

$$u(x,t) = \pm\left[\sqrt{|A|}\sinh\left\{|k|(x-ct)\right\}\right]^{\frac{2}{n-1}}, \qquad c>0, \tag{126}$$

and

$$u(x,t) = \pm\left[\sqrt{-|A|}\cosh\left\{|k|(x-ct)\right\}\right]^{\frac{2}{n-1}}, \quad c<0. \tag{127}$$

With regard to the compactons, it has been shown by Rosenau and Hyman [9] and Rosenau [68] that equations $K(m,n)$ for $m,n=2,3$ admit a finite number of local conservation laws. Extensive numerical experiments for $m=n=2,3$ reveal that many of these compactons have a remarkable particle-like robustness that goes far beyond that could be expected from four local conservation laws. Probably, there exists nonlocal conservation laws which play an important role in compacton dynamics.

From a physical point of view, it is evident that nonlinearity produces the steepening effects which are counterbalanced by the smoothing effects of dispersion. These effects play a major role in wave peaking and breaking, and other physical features of wave phenomena including a variety of weakly singular patterns. In order to understand the major role of these effects, several strongly nonlinear and dispersive models have been developed without a full resolution of the problems, in spite over 150 years of progress.

As an example of application of compactons, we consider a vibration of an anharmonic mass-spring system consisting of N initially equally spaced ($h \ll 1$) mass points m. The potential part of the associated Hamiltonian is

$$H = \sum_{n=1}^{N} \frac{1}{h}\left(y_{n+1}-y_n\right) P_n(y), \tag{128}$$

where $P_n(y) = \frac{1}{N}\alpha_N y^N$, α_N is an anharmonic parameter. For a mixed potential $P(y) = \frac{1}{2}\alpha_2 y^2 + \frac{1}{3}\alpha_3 y^3$, where α_2 and α_3 are anharmonic parameters with small α_3. For the purely quartic potential, Rosenau [69] obtained the nonlinear Boussinesq equation of motion in the continuum limit with $y_x = u$, $\varepsilon = \frac{1}{12}h^2$,

$$u_{tt} = \left(\alpha_3 u + \alpha_3 u^2\right)_{xx} + \varepsilon\,\alpha_2 u_{xxxx} + 2\,\varepsilon\,\alpha_3\left[q\left(\frac{1}{2}\right)\right]_{xx}, \tag{129}$$

where

$$q(\omega) = u^{1-\omega}\left(u^{\omega}u_x\right)_x. \tag{130}$$

Rosenau [64] showed that equation (129) admits both compacton and usual soliton solutions. For the purely quartic potential in normalized units, the equation of motion becomes

$$u_{tt} \;=\; \left(u^3\right)_{xx} + \left[u\left(u^2\right)_{xx}\right]_{xx}. \tag{131}$$

This is clearly a purely cubic nonlinear dispersion equation and fundamentally different from the weakly nonlinear models, in that it is nonlinear in the highest order derivatives, $2\,u^2\,u_{xxxx}$. Among other features, this equation also admits compacton solutions in the form $\sqrt{2}\,c\cos\left(x-ct\right)$. In addition, it also supports *compact breathers* of the form $u = Q\left(t\right)Z\left(x\right)$, where $Q\left(t\right)$ satisfies the nonlinear ordinary differential equation in the form

$$Q''\left(t\right) + \kappa^2\,Q^3\left(t\right) \;=\; 0, \tag{132}$$

where κ is a separation constant. This equation gives the periodic Jacobi elliptic function solution

$$Q\left(t\right) \;=\; cn\left(\kappa t,\; \frac{1}{\sqrt{2}}\right). \tag{133}$$

The function $Z\left(x\right)$ satisfies the equation

$$\left[Z\left(Z^2\right)_{xx}\right]_{xx} + \left(Z^3\right)_{xx} + \kappa^2 Z \;=\; 0, \tag{134}$$

which admits the following compacton solution

$$Z\left(x\right) \;=\; \left\{\begin{array}{cc} \sqrt{8}\,\kappa\cos\left(\frac{1}{2}x\right), & |x| \le \pi \\ \\ 0, & \text{otherwise} \end{array}\right\}. \tag{135}$$

While similar to this particular solution is not known, extensive numerical studies indicate that compacton's smoothness at the edge is not informative of their stability. These numerical experiments also show that the low order dispersion is unable to stabilize the compacton which decomposes immediately into a series of waves.

The nonlinear model equation

$$u_t + \left[\delta\,u + \frac{3}{2}\,\gamma\,u^2 + q\left(\omega\right)\right]_x + \nu\,u_{txx} \;=\; 0, \tag{136}$$

where δ, γ, ω and ν are constants, admits compacton solutions, and, for $2\omega = \nu\gamma = 1$, it has a bi-Hamiltonian structure. Rosenau [64] also proved that the infinite sequence of commuting flows generates an integrable, compacton's supporting variant

of the Harry Dym equation. In summary, the equation governing the motion of a mass-spring system is a prototype of compacton generating equations. With appropriate scalings, the resulting nonlinear model can be applied to study the motion of ion-acoustic waves, and a flow of a two-layer liquid. This model also admits compacton solutions.

We next discuss physical solid models that are inherently discrete where the lattice spacing represents a fundamental physical parameter. Such descrete models admit compacton solutions, that is, soliton solutions with finite wavelength. Soliton-type equations can be derived from such discrete models in which expansions of the wave amplitude and the inverse pulse width that normally require a scaling procedure. In other words, the continuum limit approach produces the condition of the slowly varying wave envelope which is consistent with the effect of weak dispersion balanced by a weak nonlinearly. As soon as we deal with compactons instead of typical solitons, the continuum limit approximation can hardly be justified because higher-order derivative terms are numerically small.

4.1 Intrinsic Localized Modes in Anharmonic Crystals

We closely follow Kivshar [70] to consider a one-dimensional lattice model in which each atom interacts with the nearest neighbors by purely *anharmonic forces*. If $u_n(t)$ is the nondimensional displacement function of the nth atom from its equilibrium position, and the atoms interact through quartic anharmonic potentials, the equation of motion for the nth atom is given by

$$\frac{d^2u_n}{dt^2} = \left[(u_{n+1}-u_n)^3+(u_{n-1}-u_n)^3\right], \tag{137}$$

where nondimensional units are employed.

In the continuum limit, the particle number is treated as a continuous variable, the long wavelength excitation of the nonlinear model equation (137) can be written as

$$v_{tt} = \left(v^3\right)_{xx}+\ldots, \tag{138}$$

where $x=av$, $a\,(=1)$ is the space of the lattice, and $v=(u_{n+1}-u_n)$ is assumed to be a slowly varying function. For short wavelength excitations, the continuum limit approximation can be used to the wave envelope $\phi_n(x,t)$ defined by the relation $u_n=(-1)^n\phi_n(x,t)$ so that equation (137) takes the form

$$\phi_{tt}+16\phi^3+6\phi\left(\phi^2\right)_{xx}+\ldots = 0. \tag{139}$$

Using the method of solution due to Rosenau and Hyman [9], equations (138) and (139) can be solved to describe compacton solution properties. However, these nonlinear evolution equation have higher order dispersive terms that can be neglected because these terms are numerically small for constant-width solutions. Thus, this nonlinear discrete models seem to be natural models for description of compacton solutions. We assume that ϕ_n is independent of time t and then seek for standing oscillatory solutions of (137) in the form

$$u_n(t) = (-1)^n \phi_n F(t). \tag{140}$$

Substituting (140) into (137) gives two separable nonlinear equations in the form

$$\frac{d^2F}{dt^2} + a\, F^3 = 0, \tag{141}$$

$$(\phi_{n+1} + \phi_n)^3 + (\phi_{n-1} + \phi_n)^3 = a\, \phi_n, \tag{142}$$

where a is a separation constant. Clearly, equation (141) admits the Jacobi elliptic function solution in the form

$$F(t) = A\, cn(\omega t, k), \tag{143}$$

where $\omega = \sqrt{a}\, A$, A is the amplitude, and $k = \frac{1}{\sqrt{2}}$.

Assuming a quasilinear solution with finite wavelength, the method of Rosenau and Hyman [9] can be used to seek a solution of (142) in the form

$$\phi_n = \begin{cases} \cos\{\theta(n - n_0)\}, & |(n - n_0)\,\theta| < \frac{\pi}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{144}$$

Substituting (144) into (143) gives two relations

$$\tan^2\left(\frac{\theta}{2}\right) = \frac{1}{3}, \quad \text{that is, } \theta = \frac{\pi}{3}, \quad \text{and } a = \frac{27}{4}. \tag{145}$$

Consequently, the general compacton solution of the lattice equation (137) is given by

$$u_n(t) = \begin{cases} (-1)^n A \cos\{\theta(n - n_0)\}\, cn\left(\omega t, \frac{1}{\sqrt{2}}\right), & |n - n_0| < \frac{3}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{146}$$

If the amplitude of the compacton is taken as an independent parameter, the frequency ω of the compacton can be defined in terms of amplitude A by

$$\omega^2 \;=\; a\,A^2. \tag{147}$$

This is identified as the *nonlinear dispersion relation.*

It is evident that the arbitrary parameter n_0 represents the center of the compacton (146) so the $n_0 = 0$ corresponds to the compacton center at the particle site $(n = 0)$ and the corresponding compacton pattern is shown in Fig. 1(b) given by Kivshar [70]. With only three lattice spacings, the compacton mode involves only three neighboring particles oscillating with the opposite phases. At $n_0 = 0$, the solution (146) can be rewritten as

$$u_n\,(t) \;=\; A\left(\ldots,0,-\frac{1}{2},1,-\frac{1}{2},0,\ldots\right) cn\left(\omega t,\frac{1}{\sqrt{2}}\right). \tag{148}$$

This describes the mode pattern through the amplitude of the oscillating particles. On the other hand, when the compacton is centered just between the neighboring particle sites, that is, at $n_0 = \frac{1}{2}$, only two neighboring particles oscillate and the other remains at rest as shown in Figure 1(b) by Kivshar [70]. The mode pattern solution is obtained in the form

$$u_n\,(t) \;=\; \frac{\sqrt{3}}{2}A\,(\ldots,0,-1,1,0,\ldots)\,cn\left(\omega t,\frac{1}{\sqrt{2}}\right), \tag{149}$$

where $\frac{\sqrt{3}}{2}A$ is used as a renormalized amplitude of this mode in order to conserve the total energy. Indeed, solution (146) describes an infinite family of different *localized modes* that are characterized by a particular value of $n_0 \in \left(0,\frac{1}{2}\right)$. Such a compacton solution has been discovered for a chain of particles with quartic interatomic potentials, and can naturally be used to explain the existence of new intrinsic localized modes in anharmonic crystals. Indeed, in their pioneering work, Sievers and Takeno [10] and Page [11] discovered these new modes based on the rotating-wave approximation (RWA) in which only a single frequency component was included in the time dependence. More precisely, the model is described by the equation

$$m\,\ddot{w}_n \;=\; k_2\,(w_{n+1}+w_{n-1}-2w_n)+k_4\left[(w_{n+1}-w_n)^3+(w_{n-1}-w_n)^3\right], \tag{150}$$

where k_2 and k_4 are the nearest neighbor harmonic and anharmonic force constants. Using the RWA approximation with only the first harmonic contribution, Sievers and Takeno [10] obtained the so-called odd-parity s mode with the displacement function

$$w_n\,(t) \;=\; A\left(\ldots,0,-\frac{1}{2},1,-\frac{1}{2},0,\ldots\right)\cos\Omega t, \tag{151}$$

where A is the amplitude and Ω is the frequency of the mode above the cutoff frequency $\Omega_m^2 = (4k_2/m)$ of the linear spectrum band. The solution (151) is, indeed, the approximate solution of (150) in the limit as $(k_4A^2/k_2) \to \infty$. Subsequently, Page [11] discovered another type intrinsic localized mode, the so-called even-parity *p mode* with the displacement function

$$w_n(t) \;=\; A(\ldots, 0, -1, 1, 0, \ldots)\cos\Omega t. \tag{152}$$

In above limiting case $k_4A^2 \gg k_2$, the contribution of the nonlinear interaction between particles in the model (150) is much more significant than that of a linear coupling term so that this model can be treated as model (137) for the displacement function $u_n = w_n\sqrt{k_4/m}$ which is perturbed by small linear coupling term. That is why the approximate solutions (151) and (152) are very close to the exact solutions (148) and (149) respectively. It is pertinent to point out another striking feature of the localized modes in the model (150) compared to the compacton solution (146) for purely anharmonic lattice model described by (137). Based on a perturbation theory, Sandusky et al. [71] have demonstrated that the odd-parity s-mode is unstable against certain velocity and displacement perturbations, whereas even-parity p mode is absolutely stable against similar perturbations. For positive anharmonicity, both of these modes have amplitude - dependent frequencies above the maximum phonon frequency. Furthermore, the unstable odd-parity mode is observed to evolve into several different kinds of moving localized modes. For certain perturbations the odd-parity mode evolves into a mode which smoothly travels from site to site with a constant speed. These traveling modes exist over a wide range of anharmonicity and can become trapped as the anharmonicity increases. As they travel, these modes have a nonconstant phase difference between adjacent relative displacements. Based on the phenomenon of the so-called *Peierls-Nabarro potential* to the localized mode, Chaude et al. [72] explained this instability of the s-mode. On the other hand, the existence of the exact compacton solution (146) with arbitrary n_0 suggests that the Peierls-Nabarro potential is absent for the compactons and they, therefore, move freely in the lattice provided the interatomic coupling is purely anharmonic in nature.

As has been demonstrated by Sievers and Takeno [10], for sufficiently strong anharmonicity, stable *odd-parity* localized excitations are possible at any lattice site with a frequency given by

$$\omega^2 \;\approx\; \frac{3}{m}\left(k_2 + \frac{27}{16}\,k_4A^2\right), \tag{153}$$

where m is the mass of the each atom and A is the amplitude of oscillations of the central atom in the mode pattern (151). The above analysis also reveals that

anharmonicity is fully responsible for the existence of the new intrinsic localized modes in anharmonic quantum crystals at finite temperature. Furthermore, the general compacton solution can describe well two new intrinsic localized modes obtained in the framework of the RWA approximation. Indeed, the compacton (146) gives the s-mode pattern when it is centered at the particle site, and it reproduces the p-mode pattern when the compacton (146) is centered in between the nearest particle sites.

We close this section by stating higher dimensional focusing branches $(+a)$ and defocusing branches $(-a)$ of $K(n,n)$ equations:

$$u_t + a\left(u^n\right)_x + b\left(u^n\right)_{xxx} + c\left(u^n\right)_{yyy} = 0, \quad a > 0, \tag{154}$$

$$u_t + a\left(u^n\right)_x + b\left(u^n\right)_{xxx} + c\left(u^n\right)_{yyy} + d\left(u^n\right)_{zzz} = 0, \quad a > 0, \tag{155}$$

$$u_t - a\left(u^n\right)_x + b\left(u^n\right)_{xxx} + c\left(u^n\right)_{yyy} = 0, \quad a > 0, \tag{156}$$

$$u_t - a\left(u^n\right)_x + b\left(u^n\right)_{xxx} + c\left(u^n\right)_{yyy} + d\left(u^n\right)_{zzz} = 0, \quad a > 0, \tag{157}$$

where $n > 1$.

Some of these equations have not yet completely solved and their solutions may provide new information about the properties of higher dimensional compactons.

ACKNOWLEDGEMENT.

This paper is an expanded version of an invited lecture delivered at the International Conference on Mathematics and its Applications organized by Kuwait University in April 2004. The author expresses his grateful thanks to the Department of Mathematics and Computer Science at Kuwait University for their financial support and cordial hospitality. The author is also grateful to Dr. Andras Balogh for drawing the figures of this paper.

REFERENCES

[1] Zabusky, N.J. and Kruskal, M.D., Interaction of solitons in a collissionless plasma and the recurrence of initial states, *Phys. Rev. Lett.* 15 (1965), 240-243.

[2] Gardner, C.S., Greene, J.M., Kruskal, M.D., and Muira, R.M., Method for solving the Korteweg-de Vries equation , *Phys. Rev. Lett.* 19 (1967), 1095-1097.

[3] Mandelbrot, B.B. *Les Objets Fractals: Forme, Hasard et Dimension*, Flammarion, Paris, 1975.

[4] Mandelbrot, B.B. *Fractals: Form, Chance, and Dimension*, W.H. Freeman, New York, 1977.

[5] Mandelbrot, B.B. *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.

[6] Morlet, J., Arens, G., Fourgeau, E., and Giard, D., Wave Propagation and Sampling Theory, Part I: Complex signal land scattering in multilayer media, *J. Geophys.* 47 (1982), 203-221.

[7] Morlet, J., Arens, G., Fourgeau, E., and Giard, D., Wave Propagation and Sampling Theory, Part II: Sampling Theory and Complex Waves, *J. Geophys.* 47 (1982), 222-236.

[8] Mallat, S. Multiresolution approximations and wavelet orthonormal basis of $L^2(\mathbb{R})$, *Trans. Amer. Math Soc.* 315 (1988), 69-88.

[9] Rosenau, P. and Hyman, J.M., Compactons: Solitons with finite wavelength, *Phys. Rev. Lett.* 70 (1993), 564-567.

[10] Sievers, A.J. and Takeno, S., Intrinsic localized modes in anharmonic crystals, *Phys. Rev. Lett.* 61 (1988), 970-973.

[11] Page, B. Asymptotic solutions for localized vibrational modes in strongly anharmonic periodic systems, *Phys. Rev.* B41 (1990), 7835-7837.

[12] Korteweg, D.J. and de Vries G., On the change of form of long waves advancing in a rectagular canal, and on a new type of long stationary waves, *Phil. Mag.* (5), 39 (1895), 422-443.

[13] Zabusky, N.J. A synergetic approach to problems of nonlinear dispersive wave propagation and interaction, *Proc. Symp. on Nonlinear Partial Differential Equations* (ed. W.F. Ames), Academic Press, Boston, (1967), 223-258.

[14] Lax, P.D. Integrals of nonlinear equations of evolution and solitary waves, *Comm. Pure Appl. Math.,* 21 (1968), 467-490.

[15] Debnath, L., *Nonlinear Water Waves,* Academic Press, Boston 1994.

[16] Debnath, L., *Nonlinear Partial Differential Equations for Scientists and Engineers*, (Second Edition) *Birkhauser Verlag, Boston,* 2004.

[17] Zakharov, V.E. and Shabat, A.B., Exact theory of two-dimensional self focusing and one-dimensional self modulation of waves in nonlinear media, *Soviet Phys. J.E.T.P.* 34 (1972), 62-69.

[18] Zakharov, V.E. and Shabat, A.B., Interaction between solitons in a stable medium, *Sov. Phys. JETP,* 37 (1973), 823-828.

[19] Zakharov, V.E. and Shabat, A.B., A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem, *Funct. Anal. Appl.* 8 (1974), 226-235.

[20] Zakharov, V.E. and Faddeev, L.D., Korteweg-de Vries equation, a completely integrable Hamiltonian system, *Funct. Anal. Appl.* 5 (1971), 280-287.

[21] Hirota, R. Exact solution of the Korteweg-de Vries equation for multiple collisions of solitons, *Phys. Rev. Lett.* 27 (1971), 1192-1194.

[22] Hirota, R. Exact envelope-soliton solutions of a nonlinear wave equation, *J. Math. Phys.* 14 (1973), 805-809.

[23] Hirota, R. Exact N-Solutions of the wave equation of long waves in shallow water and in nonlinear lattices, *J. Math. Phys.* 14 (1973), 810-814.

[24] Ablowitz, M.J., Kaup., D.J., Newell, A.C. and Segur, H., The inverse scattering transform - Fourier analysis for nonlinear problems, *Stud. Appl. Math.* 53 (1974), 249-315.

[25] Novikov, S.P., Manakov, S.V., Pitaevskii, L.P., and Zakharov, V.E., *Theory of solitons; The Inverse Scattering Method*, Plenum, New York, 1984.

[26] Mandelbrot, B.B. How long is the coast of Britain? Statistical self-similarity and fractional dimension, *Science,* 155 (1967), 636-638.

[27] Mandelbrot, B.B., Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier, *J. Fluid Mech.* 62 (1974), 331-358.

[28] Mandelbrot, B.B., On the geometry of homogeneous turbulence with stress on the fractal dimension of the iso-surfaces of scalars, *J. Fluid Mech.* 72 (1975), 401-416.

[29] Mandelbrot, B.B. *Fractals and Chaos, The Mandelbrot Set and Beyond*, Springer Verlag, New York, 2004.

[30] Rogers, C.A. *Hausdorff measure*, Cambridge University Press, Cambridge, 1970.

[31] Debnath, L. *Wavelet Transforms and Their Applications*, Birkhauser Verlag, Boston 2002.

[32] Mandelbrot, B.B., and Given, J.A., Physical properties of a new fractal model of percolation clusters., *Phys. Rev. Lett.* 52 (1984) 1853-1856.

[33] Feder, J. *Fractals*, Plenum Press, New York, 1988.

[34] Sapoval, B., Rosso, M., and Gouyet, J.F., The fractal nature of a diffusing front and the relation to percolation, *J. Phys. Lett.* 46 (1985) L325-L328.

[35] Lauwerier, H. *Fractals*, *Endlessly Repeated Geometrical Figures*, Princeton University Press, Princeton, 1991.

[36] Peitgen, H.O., and Richter, P.H., *The Beauty of Fractals*, Springer Verlag, New York, 1986.

[37] Mandelbrot, B.B., Passoja, D.E., and Paullay, A.J., Fractal character of fracture surfaces of metals, *Nature* 308 (1984), 721-722.

[38] Carpinteri, A. and Yang, G.P., Fractal dimension evolution of microcrack net in disordered materials, *Theoretical and Applied Fracture Mechanics*, 25 (1996), 73-81.

[39] Carpinteri, A. *Mechanical Damage and Crack Growth in Concrete: Plastic Collapse to Brittle Fracture*, Martinus Nijhoff Publishers, Dordrecht, 1986.

[40] Muller, J. Morphology of Disordered Materials studied by multifractals, In *Wavelets, Fractals, and Fourier Transforms*, (Ed. M. Farge, J.C.R. Hunt, and J.C. Vassilicos), Oxford University Press, Oxford (1993), 397-403.

[41] Avnir, D. *The Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers*. John Wiley, New York, 1989.

[42] Hutchinson, J.E., Fractals and Self-Similarity, *Indiana University Math. J.*, 30 (1981), 713-747.

[43] Vasillicos, J.C. Fractals in Turbulence, In *Wavelets, Fractals and Fourier Transforms*, (Ed. Farge, M., Hunt, J.C.R., and Vasillicos, J.C.), Oxford University Press, Oxford, (1993) 325-340.

[44] Sreenivasan, K.R. and Meneveau, C., The fractal facets of turbulence, *J. Fluid Mech.* 173 (1986), 357-386.

[45] Meneveau, C. and Sreenivasan, K.R., Simple multifractal cascade model for fully developed turbulence, *Phys. Rev. Lett.* 59 (1987), 1424-1427.

[46] Benzi, R., Paladin, G., Parisi, G., and Vulpiani, A., On the multifractal nature of fully developed turbulence and chaotic systems, *J. Phys.* A17 (1984), 3521-3531.

[47] Mandelbrot, B.B. Géométrie fractle de la turbulence. Dimension de Hausdorff, dispersion et nature des singularités du mouvement des fluides, *Comptes. Rendus,* Paris, 282A (1976), 119-120.

[48] Sreenivasan, K.R. Fractals and multifractals in fluid turbulence, *Ann. Rev. Fluid Mech.* 23 (1991), 539-600.

[49] Sarkar, S.K. Generalization of singularities in nonlocal dynamics, *Phy. Rev.* A31 (1985), 3468-3472.

[50] Hunt, J.C.R. and Vassilicos, J.C., Kolmogorov's contributions to the physical and geometrical understanding of small-scale turbulence and recent developments, *Proc. Roy. Soc. London* A434 (1991), 183-210.

[51] Schwarz, K.W. Evidence for organized small-scale structure in fully developed turbulence, *Phys. Rev. Lett.* 64 (1990), 415-418.

[52] Vincent, A. and Meneguzzi, M., The spatial structure and statistical properties of homogeneous turbulence, *J. Fluid Mech.* 225 (1991), 1-20.

[53] She, Z.S., Jackson, E. and Orszag, S.A., Structure and dynamics of homogeneous turbulence, models and simulations, *Proc. Roy. Soc. London* A434 (1991), 101-124.

[54] Moffatt, H.K. Some topological aspects of turbulent vorticity dynamics, in *Turbulence and Chaotic Phenomena in Fluids* (Ed. T. Tatsumi), Elsevier, New York (1984), 223-230.

[55] Vassilicos, J.C. The multi-spiral model of turbulence and intermittency, in *Topological Aspects of the Dynamics of Fluids and Plasmas*, Klewer Amsterdam, (1992), 427-442.

[56] Debnath, L. Wavelet transforms, fractals, and turbulence, *In Nonlinear Instability, Chaos and Turbulence* (Ed. L. Debnath and D.N. Riahi) volume 1, (1998), 129-195, WIT Press, Southampton.

[57] Mandelbrot, B.B. Fractals and the rebirth of iteration theory in *The Beauty of Fractals* by Peitgen, H.O. and Ritcher, P.H., Springer Verlag, New York, (1986), 151-160.

[58] Keen, L. Julia sets In *Chaos and Fractals* (Ed. Devaney, R.L. and Keen, L.) *Proc. Sympos. Appl. Math.* 39 (1989), 57-74.

[59] Douady, A., and Hubbard, J.H., Itération des polynômes quadratiques complexes, *C.R. Adad. Sci.* 294 (1982), 123-126.

[60] Branner, B. The Mandelbrot set, In *Chaos and Fractals* (Ed. Devaney, R.L., and Keen, L.), *Proc. Sympos. Appl. Math.* 39 (1989), 75-105.

[61] Blanchard, P. Complex analytic dynamics on the Riemann sphere, *Bull. Amer. Math. Soc. (N.S.)* 11 (1984), 85-141.

[62] Daubechies, I. Orthogonal bases of Compactly supported wavelets, *Commun. Pure and Appl. Math.*, 41, 1988.

[63] Daubechies, I. *Ten Lectures on Wavelets*, SIAM Publications, Philadelphia, 1992.

[64] Oron, A. and Rosenau, P., Evolution of the coupled Bernard-Marangoni Convection, *Phys. Rev.* A39, (1989), 2063-2069.

[65] Ludu, A. and Draayer, J.P. Patterns on liquid surfaces: Cnoidal waves, compactons and scaling, *Physica*, D123, (1998), 82-91.

[66] Ludu, A., Stoitcheva, G., and Draayer, J.P., Similarity analysis of nonlinear equations and bases of finite wavelength solitons, *Internat. J. Mod. Phys. E* 9, (2000), 263-278.

[67] Dusuel, S., Michaux, P., and Remoissnet, M., From Kinks to Comfactionlike Kinks, *Phys. Rev. E* 57, (1998), 2320-2326.

[68] Rosenau, P. On nonanalytical solitary waves formed by a nonlinear dispersion, *Phys. Lett.* A230, (1997), 305-318.

[69] Rosenau, P. Nonlinear Dispersion and Compact Structures, *Phys. Rev. Lett.* 73, (1994), 1737-1741.

[70] Kivshar, Y. Intrinsic localized modes as solitons with a compact support, *Phys. Rev.* B48, (1993), R43-R45.

[71] Sandusky, K.W., Page, J.B., Schmidt, K.E., Stability and motion of intrinsic localized modes in nonlinear periodic lattices, *Phys. Rev.* B46, (1992), 6161-6170.

[72] Claude, Ch., Kivshar, Yu. S., Kluth, O., and Spataschek, K.H., Liapunov Stability of Generalized Langmuir Solitons, *Phys. Rev.* 47, (1993), 228-234.

THE GENERALISED COUPLED ALTERNATING GROUP EXPLICIT (CAGE) METHOD

D. J. Evans
Parallel Algorithms Research Centre
University of Technology, Loughborough, Leics., U.K.

Abstract

In this paper generalizations of the Coupled Alternating Group Explicit (CAGE) method (1990) are presented and compared with the AGE methods for solving tridiagonal linear systems.

KEYWORDS: ADI and AGE methods, CAGE method, SMAGE method.

1. THE AGE METHOD

Here we summarize briefly the AGE method (Evans, 1985), in Peaceman-Rachford form (1956) and introduce the forms introduced by Douglas-Rachford (1956), Douglas (1962) and Guittet (1967).

Consider the linear system

$$Au = b, \tag{1}$$

where u and b are N-dimensional vectors and $\mathbf{A}$ is given as

$$\mathbf{A} = \begin{bmatrix} 2g_1 & c_1 & & \\ a_2 & 2g_2 & c_2 & \mathbf{O} \\ \ddots & \ddots & \ddots & \\ & a_{N-1} & 2g_{N-1} & c_{N-1} \\ \mathbf{O} & & a_N & 2g_N \end{bmatrix} \tag{2}$$

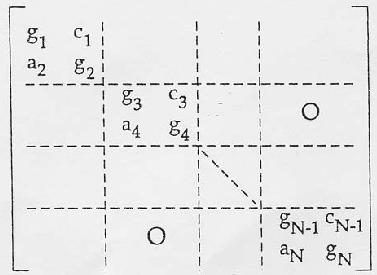
The basic principle of the AGE iterative method consists of splitting the matrix A into the form,

$$A = G_1 + G_2, \tag{3}$$

where,

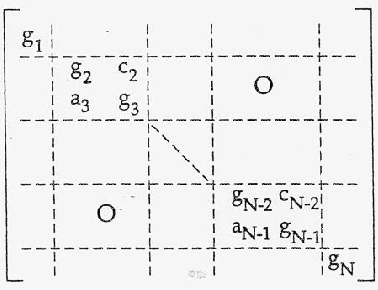
for N is even. Let us assume that all the eigenvalues of G_1 and G_2 are real and positive, i.e., $g_1 > \frac{1}{2}(a_1 + c_1), i = 1, 2, ..., N$.

$G_1 =$



(4)

$G_2 =$



(5)

By applying the Peaceman Rachford form to (3), then for any iteration parameter $r > 0$, the AGE-PR(1) iterative method can be written as

$$(rI + G_1)u^{(k+1/2)} = b + (rI - G_2)u^{(k)}, \tag{6}$$

$$(rI + G_2)u^{(k+1)} = b + (rI - G_1)u^{(k+1/2)}, \tag{7}$$

or explicitly as

$$u^{(k+1/2)} = (rI + G_1)^{-1}[b + (rI - G_2)u^{(k)}], \tag{8}$$

$$u^{k+1} = (rI + G_2)^{-1}[b + (rI - G_1)u^{(k+1/2)}]. \tag{9}$$

It has been shown earlier that the AGE-PR(1) is convergent (1985). Now let us consider the modification of equation (7) of the AGE-PR(1) method. Then, for any $r > 0$, the AGE-PR(2) scheme can be written as

$$(rI + G_1)u^{(k+1/2)} = b + (rI - G_2)u^{(k)}, \tag{10}$$
$$(rI + G_2)u^{(k+1)} = 2ru^{(k+1/2)} - (rI - G_2)u^{(k)}, \tag{11}$$

after using equation (6) to express $G_1u^{(k+1/2)}$ in terms of $G_2u^{(k)}$, thereby saving on the evaluation of the right hand side sectors. The matrices G_1 and G_2 are as given in (4) - (5). In explicit form the AGE-PR(2) scheme can be written as

$$u^{(k+1/2)} = (rI + G_1)^{-1}[b + (rI - G_2)u^{(k)}], \tag{12}$$
$$u^{(k+1)} = (rI + G_2)^{-1}[2ru^{(k+1/2)} - (rI - G_2)u^{(k)}], \tag{13}$$

where the iteration matrix T_r is given by

$$T_r = (rI + G_2)^{-1}[2r(rI + G_1)^{-1} - I](rI - G_2). \tag{14}$$

It is obvious that, after dropping all the superscripts, the AGE-PR(2) scheme is consistent.

Now let us introduce a parameter ω into equation (10). Then, the new set of equations become

$$(rI + G_1)u^{(k+1/2)} = b + (rI - G_2)u^{(k)}, \tag{15}$$
$$(rI + G_2)u^{(k+1)} = (2 - \omega)ru^{(k+1/2)} - [r(1 - \omega)I - G_2]u^{(k)}, \tag{16}$$

or in explicit form,

$$u^{(k+1/2)} = (rI + G_1)^{-1}[b + (rI - G_2)u^{(k)}], \tag{17}$$
$$u^{(k+1)} = (rI + G_2)^{-1}[(2 - \omega)ru^{(k+1/2)} - [r(1 - \omega)I - G_2]u^{(k)}], \tag{18}$$

resulting in the generalized AGE scheme (GAGE).

Putting $\omega = 0$, we then have the AGE scheme (10) - (11). For $\omega = 1$, the scheme is analogous to the one given by Douglas and Rachford (1956). We call this scheme as the AGE method in Douglas-Rachford form (AGE-Dr(1)). This AGE-DR(1) can be written as

$$u^{(k+1/2)} = (rI + G_1)^{-1}[b + (rI - G_2)u^{(k)}], \tag{19}$$
$$u^{(k+1)} = (rI + G_2)^{-1}[ru^{(k+1/2)} + G_2u^{(k)}]. \tag{20}$$

The important feature for the new generalized scheme, $\omega \neq 1$, is that it can be applied to solve the boundary value problems with two or more variables, i.e., the two and three dimensional problems.

Another modification of equation (19) of the AGE-DR(1) scheme is as follows. Then, for any $r > 0$, the new scheme can be written in explicit form as

$$\begin{aligned} u^{(k+1/2)} &= (rI + G_1)^{-1}[b - Au^{(k)} + (rI + G_1)u^{(k)}], \quad (21) \\ u^{(k+1)} &= (rI + G_2)^{-1}[ru^{(k+1/2)} + G_2u^{(k)}], \quad (22) \end{aligned}$$

or

$$\begin{aligned} u^{(k+1/2)} &= (rI + G_1)^{-1}[b + \{(rI + G_1) - A\}u^{(k)}], \quad (23) \\ u^{(k+1)} &= (rI + G_2)^{-1}[ru^{(k+1/2)} + G_2u^{(k)}], \quad (24) \end{aligned}$$

with the iteration matrix given by,

$$\begin{aligned} T_r &= (rI + G_2)^{-1}[r\{I - (rI + G_1)^{-1}A\} + G_2], \\ &= I - r(rI + G_2)^{-1}(rI + G_1)^{-1}A, \end{aligned}$$

which simplifies to,

$$T_r = I - r\prod_{i=2}^{1}(rI + G_i)^{-1}A, \quad (25)$$

and it is obvious that the scheme (23)-(24) is consistent.

We now introduce a parameter ω in equation (23). The new set of equations then becomes

$$u^{(k+1/2)} = (rI + G_1)^{-1}[\omega b + \{(rI + G_1) - \omega A\}u^{(k)}], \quad (26)$$

$$u^{(k+1)} = (rI + G_2)^{-1}[ru^{(k+1/2)} + G_2u^{(k)}]. \quad (27)$$

This is another important feature, since the generalized AGE method (26)-(27) is applicable to solve problems with higher dimensions. Putting $\omega = 1$, we have the scheme which is similar to AGE-DR(1) and denote this scheme as AGE-DR(2). For $\omega = 2$, the scheme is analogous to one given by Douglas (1956). We denote this scheme as the AGE method in Douglas for (AGE-DG). Hence, the AGE-DG scheme is given by

$$\begin{aligned} u^{(k+1/2)} &= (rI + G_1)^{-1}[2b + \{(rI + G_1) - 2A\}u^{(k)}], \quad (28) \\ u^{(k+1/2)} &= (rI + G_1)^{-1}[ru^{(k+1/2)} + G_2u^{(k)}], \quad (29) \end{aligned}$$

with the iteration matrix given by,

$$T_r = I - 2r\prod_{i=2}^{1}(rI + G_i)^{-1}A, \quad (30)$$

and obviously, the AGE-DG scheme is consistent

In general, the generalized AGE scheme (26)-(27) with the values if ω in [1,2] will have the iteration matrix as

$$T_r = I - \omega r \prod_{i=2}^{1} (rI + G_1)^{-1} A, \tag{31}$$

and the scheme is convergent.

It is clear that the generalized AGE scheme (26)-(27) will give the AGE-DG scheme when $\omega = 2$ and the AGE-DR(2) when $\omega = 1$. The AGE-DG scheme has also been shown to achieve a similar rate of convergence as the AGE-PR(2) scheme.

Now we present the algorithm for the AGE-DG scheme in computational form by using equations (28)-(29).

Algorithm 1.1 The computational form of the AGE-DG scheme.
Set: $u_1^{(k)} = 0, i = 0, ..., N+1, a_1 = 0, c_N = 0.$

Step 1: To compute $u^{(k+1/2)}$. Set $i = 1$. while $i \leq N-1$, compute

$$\begin{aligned}
d &= 1/(\alpha_1\alpha_{i+1} - a_{i+1}c_i), \\
A &= -2da_1\alpha_{i+1}, D = 2dc_ic_{i+1} \\
B &= 1 + 2d(a_{i+1}c_i - 2g_i\alpha_{i+1}), P = 2da_ia_{i+1}, \\
C &= 2dc_i(2g_{i+1} - \alpha_{i+1}), \\
E &= 2d(\alpha_{i+1}b_i - c_ib_{i+1}), \\
Q &= 2da_{i+1}(2g_i - \alpha_i), \\
T &= 2d(\alpha_ib_{i+1} - a_{i+1}b_i), \\
R &= 1 + 2d(a_{i+1}c_i - 2g_{i+1}\alpha_i), S = -2dc_{i+1}\alpha_i, \\
u_i^{(k+1/2)} &= Au_{i-1}^{(k)} + Bu_i^{(k)} + Cu_{i+1}^{(k)} + Du_{i+1}^{(k)} + E, \\
u_{i+1}^{(k+1/2)} &= Pu_{i-1}^{(k)} + Qu_i^{(k)} + Ru_{i+1}^{(k)} + Su_{i+1}^{(k)} + T, \\
i &= i + 2.
\end{aligned}$$

Step 2: To compute u^{k+1}. Set $i = 2$.
$u_1^{(k+1)} = (ru_1^{(k+1/2)} + g_1u_1^{(k)})/\alpha_1$

while $i \leq N-2$, compute

$$\begin{aligned}
d &= 1/(\alpha_1\alpha_{i+1} - a_{i+1}c_i), A = dr\alpha_{i+1}, B = -drc_i, \\
C &= d(\alpha_{i+1}g_i - a_{i+1}c_i), D = dc_i(\alpha_{i+1} - g_{i+1}), \\
P &= -dra_{i+1}, Q = dr\alpha_i, R = da_{i+1}(\alpha_i - g_i), \\
S &= d(\alpha_i g_{i+1}, Q = dr\alpha_i, R = da_{i+1}(\alpha_i - g_i), \\
S &= d(\alpha_i g_{i+1} - a_{i+1}c_i), \\
u_i^{(k+1)} &= Au_i^{(k+1/2)} + Bu_{i+1}^{(k+1/2)} + Cu_i^{(k)} + Du_{i+1}^{(k)}, \\
u_{i+1}^{(k+1)} &= Pu_i^{(k+1/2)} + Qu_{i+1}^{(k+1/2)} + Ru_i^{(k)} + Su_{i+1}^{(k)}, \\
i &= i+2 \\
u_N^{(k+1)} &= (ru_N^{(k+1/2)} + g_N u_N^{(k)})\alpha_N.
\end{aligned}$$

Step 3: Repeat Step 1 and Step 2 until convergence is achieved.

The computational molecules for the evaluation of $u_i^{(k+1/2)}, i = 1, 3, .., N-1$, are given by Figure 1.

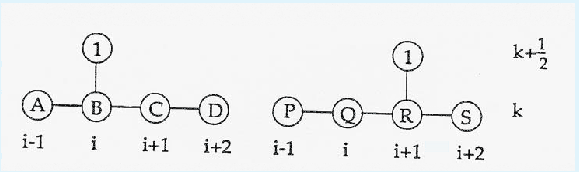


Figure 1: The computational molecule for $u^{(k+1/2)}$, AGE-DG

For the evaluation of $u_i^{(k+1)}, i = 2, , ..., N$, the computational molecules can be presented in Figure 2.

Guittet (1967) has considered another generalized form to solve the PDE problems with higher dimensions. Analogous to the one dimensional problem, the AGE method in Guittet's form (AGE-GT) may be written as

$$(rI + G_1)u^{(k+1/2)} = \omega r[b - Au^{(k)}] + \prod_{2}^{i=1}(rI + G_i)u^{(k)}, \tag{32}$$

$$(rI + G_2)u^{(k+1)} = u^{k+1/2)}, \tag{33}$$

where the iteration matrix can be shown to have the form

$$T_r = I - \omega r \prod_{1}^{i=2}(rI + G_i)^{-1}A. \tag{34}$$

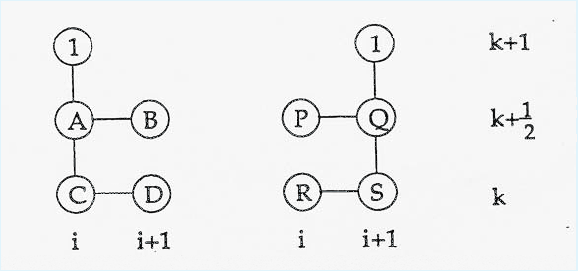


Figure 2: The computational molecules for $u^{(k+1)}$, AGE-DG

Thus the scheme is also convergent and can be shown to be consistent.

2. COUPLED AGE (CAGE) FORM OF PR, DOUGLAS AND GUITTET (ADI) METHODS

The Alternating Group Explicit (AGE) schemes have previously used two stages, i.e., the first to solve for $u^{(k+1/2)}$ followed by the solution for $u^{(k+1)}$. In this section, we will show that the two stage schemes can be combined into a single layer coupled AGE(CAGE) method (1990).

Let us recall the four schemes that have been discussed in Section 1. The respective CAGE formulation for these schemes can be written as follows. Let μ and ν be the respective eigenvalues of G_1 and G_2.

The CAGE-PR(2) Scheme

$$\begin{aligned} u^{(k+1)} &= (rI+G_2)^{-}[2r(rI+G_1)^{-1}-I](rI-G_2)u^{(k)} \\ &+ 2r(rI+G_2)^{-1}(rI+G_1)^{-1}b. \end{aligned} \tag{35}$$

The iteration matrix, T_r is given by,

$$T_r = (rI+G_2)^{-1}[2r(rI+G_1)^{-1}-I](rI-G_2).$$

We now show that the scheme is convergent. Since,

$$\begin{aligned} \| T_r \|_2 &= \| (rI+G_2)^{-1}[2r(rI+G_1)^{-1}-I](rI-G_2) \|_2 \\ &= | \frac{1}{r+\nu}\left[\frac{2}{r+\mu}-1\right](r-\nu) | \\ &= | \frac{r-\nu}{r+\nu} | \quad | \frac{r-\mu}{r+\mu} | < 1. \end{aligned}$$

Hence the scheme is convergent

The CAGE-DG Scheme

$$\begin{aligned} u^{k+1} &= (rI - G_2)^{-1}[r\{I - 2(rI + G_1)^{-1}A\} + G_2]u^{(k)} \\ &+ 2r(rI + G_2)^{-1}(rI + G_1)^{-1}b \\ &= (rI + G_2)^{-1}[(rI + G_2) - 2r(rI + G_1)^{-1}A]u^{(k)} \\ &+ 2r(rI + G_2)^{-1}(rI + G_1)^{-1}b \\ &= [I - 2r(rI + G_2)^{-1}(rI + G_1)^{-1}A]u^{(k)} \\ &+ 2r(rI + G_2)^{-1}(rI + G_1)^{-1}b. \end{aligned} \tag{36}$$

The iteration matrix, T_r is given by,

$$T_r = I - 2r(rI + G_2)^{-1}(rI + G_1)^{-1}A.$$

For convergence, we need $\| T_r \|_2 < 1$. Since,

$$\begin{aligned} \| T_r \|_2 &= \| I - 2r(rI + G_2)^{-1}(rI + G_1)^{-1}A \|_2 \\ &= | 1 - \frac{2r}{(r+\mu)(r+\nu)}(\mu + \nu) | \\ &= | \frac{r^2 - r(\mu+\nu) + \mu\nu}{(r+\mu)(r+\nu)} | \\ &= | \frac{r-\nu}{r+\nu} || \frac{r-\mu}{r+\mu} | < 1. \end{aligned}$$

Hence the scheme is convergent.

The CAGE-GT Scheme

$$\begin{aligned} u^{k+1} &= [I - 2r(rI + G_2)^{-1}(rI + G_1)^{-1}A]u^{(k)} \\ &+ 2r(rI + G_2)^{-1}(rI + G_1)^{-1}b. \end{aligned} \tag{37}$$

The iteration matrix, T_r is given by,

$$T_r = I - 2r(rI + G_2)^{-1}(rI + G_1)^{-1}A.$$

Since the iteration matrix of the CAGE-GT scheme is similar to the iteration matrix of the CAGE-DG scheme, thus the scheme converges.

THE CAGE-PR(1) Scheme

$$\begin{aligned} u^{k+1} &= (rI + G_2)^{-1}(rI - G_1)(rI + G_1)^{-1}(rI - G_2)u^{(k)} \\ &+ (rI + G_2)^{-1}[I + (rI - G_1)(rI + G_1)^{-1}]b. \end{aligned} \tag{38}$$

The iteration matrix, T_r is given by,

$$T_r = (rI + G_2)^{-1}(rI - G_1)(rI + G_1)^{-1}(rI - G_2).$$

We now show that the scheme converges. Since

$$\begin{aligned} \| T_r \|_2 &= \| (rI + G_2)^{-1}(rI - G_1)(rI + G_1)^{-1}(rI - G_2) \|_2 \\ &= | \frac{1}{r+\nu}(r-\mu)\frac{1}{r+\mu}(r-v) | \\ &= | \frac{r-\nu}{r+\nu} || \frac{r-\mu}{r+\mu} | < 1. \end{aligned}$$

hence the scheme is convergent.

In general, the CAGE method is of the form,

$$u^{(k+1)} = T_r u^{(k)} + Cb, \tag{39}$$

with T_r, a matrix of the form given by,

$$T_r = \begin{bmatrix} X & X & X & & & & & & \\ X & X & X & X & X & & & & \\ X & X & X & X & X & & & \mathbf{O} & \\ & X & X & X & X & X & X & & \\ & X & X & X & X & X & X & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & X & X & X & X & X & X & \\ & & X & X & X & X & X & X & \\ & & & & X & X & X & X & X \\ & \mathbf{O} & & & X & X & X & X & X \\ & & & & & & X & X & X \end{bmatrix} \tag{40}$$

and $\mathbf{C}$ is given as

$$\mathbf{C} = \begin{bmatrix} X & X & & & & & & \\ X & X & X & X & & & & \\ X & X & X & X & & & \mathbf{O} & \\ & X & X & X & X & & & \\ & X & X & X & X & & & \\ & \ddots & \ddots & \ddots & \ddots & & & \\ & & X & X & X & X & & \\ & & X & X & X & X & & \\ & & & & X & X & X & X \\ & \mathbf{O} & & & X & X & X & X \\ & & & & & & X & X \end{bmatrix}, \tag{41}$$

where X represents a non-zero element.

3. THE CAGE ALGORITHM

The coefficient for each non-zero element X can be determined by solving the iteration matrix, T_r, and matrix $\mathbf{C}$ from each scheme of the CAGE method.

From (26) and (27) we now present the algorithm for the CAGE method to compute $u_i^{(k+1)}$.

The following algorithm is for the CAGE-DG and CAGE-GT schemes, since both schemes have similar iteration matrix T_r and matrix $\mathbf{C}$.

Algorithm 3.1: The CAGE-DG scheme, equation (22) and the CAGE-GT scheme, equation (23)

Set $u_i^k = 0, i = 0, ..., N+1.$
Step 1: To compute u^{k+1}. Set $i = 2$

$$\begin{aligned} d &= 1/(\alpha_1\alpha)_2 - a_2c_1), w = 2rd/\alpha_1, E = -wc_1, C = -Ec_2, \\ A &= 1 + w(a_2c_1 - 2g_1\alpha_2), B = -E(2g_2 - \alpha_2), D = w\alpha_2, \\ u_1^{(k+1)} &= Au_1^{(k)} + Bu_2^{(k)} + Cu_3^{(k)} + Db_1 + Eb_2 \end{aligned}$$

while $i \leq N-2$, compute

$$
\begin{aligned}
d_1 &= 1/(\alpha_{i-1}\alpha_i - a_i c_{i-1}), d_2 = 1/(\alpha_i\alpha_{i+1} - a_{i+1}c_i),\\
d_3 &= 1/(\alpha_{i+1}\alpha_{i+2} - a_{i+2}c_{i+1}), w_1 = 2rd_1d_2,\\
w_2 &= 2rd_2d_3, A_1 = w_1 a_i \alpha_{i+1}, K = w_2 c_i c_{i+1},\\
A &= A_i a_{i-1}, B = A_1(2g_{i-1} - \alpha_{i-1}),\\
C &= 1 + w_2 a_{i+1} c_i \alpha_{i+2} + w_1\alpha_{i+1}(a_i c_{i-1} - 2\alpha_{i-1}g_i),\\
D &= c_i[w_2(2\alpha_{i+2}g_{i+1} = a_{i+2}c_{i+1}) - w_1\alpha_{i-1}\alpha_{i+1}],\\
E &= K(\alpha_{i+2} - 2g_{i+2}), F = -Kc_{i+2}, G = -w_1 a_i \alpha_{i+1},\\
H &= w_1\alpha_{i-1}\alpha_{i+1}, J = -w_2 c_i \alpha_{i+2}, P_1 = w_1 a_i \alpha_{i+1},\\
P &= -P_i a_{i-1}, Q = P_1(\alpha_{i-1} - 2g_{i-1}),\\
R &= a_{i+1}[w_1(2\alpha_{i-1}g_i - a_i c_{i-1}) - w_2\alpha_i a_{i+2}],\\
S &= 1 + w_1 c_i a_{i+1}\alpha_{i-1} + w_2\alpha_1(a_{i+2}c_{i+1} - 2g_{i+1}\alpha_{i+2}),\\
Y &= -w_2 c_{i+1}\alpha_i, T = -Y(2g_{i+2} - \alpha_{i+2}), U = -Yc_{i+2},\\
V &= w_1 a_i a_{i+1}, W = -w_1 a_{i+1}\alpha_{i-1}, X = w_2\alpha_i\alpha_{i+2},\\
u_i^{(k+1)} &= Au_{i-2}^{(k)} + Bu_{i-1}^{(k)} + Cu_i^{(k)} + Du_{i+1}^{(k)} + Eu_{i+2}^{(k)} + Fu_{i+3}^{(k)}\\
&+ Gb_{i-1} + Hb_i + Jb_{i+1}Kb_{i+2},\\
u_{i+1}^{(k+1)} &= Pu_{i-2}^{(k)} + Qu_{i-1}^{(k)} + Ru_i^{(k)} + Su_{i+1}^{(k)} + Tu_{i+2}^{(k)} + Uu_{i+3}^{(k)}\\
&+ Vb_{i-1} + Wb_i + Xb_{i+1}Yb_{i+2},\\
i &= i + 2.\\
d_1 &= 1/(\alpha_{N-1}\alpha_N - a_N c_{N-1}), w = 2rd_1/\alpha_N, D = -wa_N,\\
A &= -Da_{N-1}, B = -D(2g_{N-1} - \alpha_{N-1}),\\
C &= 1 + w(a_N c_{N-1} - 2g_N\alpha_{N-1}), E = w\alpha_{N-1},\\
u_N^{(k+1)} &= Au_{N-2}^{(k)} + Bu_{N-1}^{(k)} + Cu_N^{(k)} + Db_{N-1} + Eb_N.
\end{aligned}
$$

Step 2: Repeat Step 1 until convergence is achieved.

The presented algorithms for the CAGE-DG and CAGE-GT schemes show that in order to determine the coefficient for each node and element b_i, we need to compute many intermediate values. Also extra work is needed in each iteration if g_i (for a given problem) depends on the solution vector u_i. However, in the case where g_i is independent of u_i, all the intermediate values can be computed outside the loop and thereby save time in each iteration. Moreover, from these algorithms, it can be deduced that the computational molecule for the CAGE method for large N, is given by the 6 nodal formulae, i.e.,

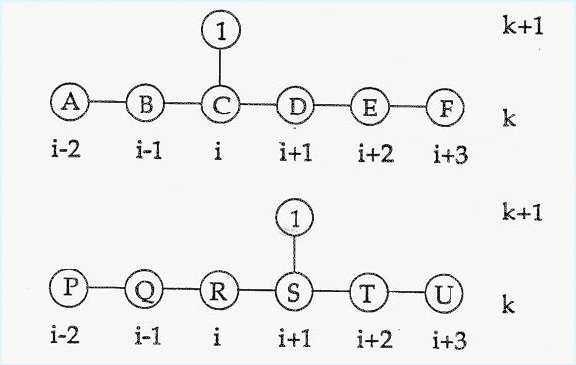


Figure 3: Computational molecule for the one step CAGE method.

Table 1 below shows the number of nodes for each scheme presented in Section 2 compared to the scheme derived in the form of the CAGE method above

Comp.Molec. Scheme	**No. of Nodes**	**The CAGE method**	**No. of Nodes**
AGE-PR(2)	8	CAGE-PR(2)	6
AGE-DG	8	CAGE-DG	6
AGE-GT	6	CAGE-GT	6
AGE-PR(1)	8	CAGE-PR(1)	6

TABLE 1: The number of nodes in the computational molecules.

Table 1 indicates that there is a 25% saving in computation work in the CAGE method over the AGE schemes, except the case of the CAGE-GT scheme. Although, many intermediate values are needed prior to calculating the coefficients for the solution vector u, these gains make the CAGE method better than the two step AGE schemes.

Finally, it can be seen that in the CAGE-PR(2) scheme, the evaluation of the coefficients A, B, C,... etc. of the matrix T_r is more difficult than in the CAGE-DG scheme. Thus, at this stage, we may consider the CAGE-DG or CAGE-GT schemes to be the best choice.

4. THE SMART AGE (SMAGE) METHOD

Finally an alternative approach of evaluating the AGE method based on the AGE-PR(2) scheme is considered. This new scheme is called Smart AGE (SMAGE), and is predicted to save time as the idea involved is to eliminate evaluating two similar terms on the right hand sides of the AGE-PR(2) scheme.

If we recall the AGE-PR(2) scheme in explicit form, we have

$$u^{(k+1/2} = (rI + G_1)^{-1}[b + (rI - G_2)u^{(k)}], \tag{42}$$

$$u^{(k+1)} = (rI + G_2)^{-1}[2ru^{(k+1/2)} - (rI - G_2)u^{(k)}]. \tag{43}$$

The two similar terms in these equations is $(rI - G_2)u^{(k)}$ and we let this term be ϕ. The evaluation and saving of ϕ depends upon the problems, i.e., either linear or nonlinear. The matrix A derived from the linear problems give either a constant or variable diagonal element, where is for nonlinear problems this element is variable.

The SMAGE scheme for linear problems is envisaged to save 2 multiplications and 1 addition for every iteration whilst, for nonlinear problems, it is expected to save 1 multiplication and 1 addition for every iteration.

The algorithm for the linear schemes is presented as follows.

Algorithm 4.1: The SMAGE scheme

Set: $u_i^{(k)} = 0, i = 0, ...N + 1, a_1 = 0, c_N = 0,$
$\alpha_i = r + g_i, \beta_i - r - g_i, i = 1, 2, ..., N.$

Step 1: To compute $\phi = (rI - G_2)u^{(0)}$. Set $i = 1$
while $i \leq N - 1$, compute

$$\begin{aligned} \phi_i &= -c_i u_{i-1}^{(0)} + \beta_i u_i^{(0)} \\ \phi_i &= \beta_{i+1} u_{i+1}^{(0)} - a_{i+1} u_{i+2}^{(0)} \end{aligned}$$

Step 2: To compute $u^{(k+1/2)}$. Set $i = 1$
while $i \leq N - 1$, compute

$$\begin{aligned} r_1 &= b_1 + \phi_i, r_2 = b_{i+1} + \phi_{i+1} \\ d &= /(\alpha_i \alpha_{i+1} - a_{i+1} c_i) \\ u_i^{(k+1/2)} &= (\alpha_{i+1} r_1 - c_i r_2) d \\ u_{i+1}^{(k+1/2)} &= (-a_{i+1} r_1 + \alpha_i r_2) d \\ i &= i + 2. \end{aligned}$$

Step 3: For i=1,2,...,N, compute $\phi_i = -\phi_i + 2ru_i^{(k+1/2)}$.

Step 4: To compute $u^{(k+1)}$

$u_1^{(k+1)} = \phi_1/\alpha_1$

$i \leq N-2$, compute

$$\begin{aligned} d &= 1/(\alpha_i\alpha_{i+1} - a_{i+1}c_i \\ u_i^{(k+1)} &= (\alpha_{i+1}\phi_i - c_i\phi_{i+1})d \\ u_{i+1}^{(k+1)} &= (-a_{i+1}\phi_i + \alpha_i\phi_{i+1})d \\ i &= i+2. \\ u_N^{(k+1)} &= \phi_N/\alpha_N. \end{aligned}$$

Step 5: For $i = 1, 2, ..., N$, compute $\phi_i = -\phi_i + 2ru_i^{(k+1)}$.

Step 6: Repeat Step 2 to Step 5 until convergence is achieved.

5. EXPERIMENTAL RESULTS

The AGE, CAGE and SMAGE schemes of sections 1, 2 and 3 were investigated experimentally on the following linear problem and the computational complexity and the speed (CPU time) for each of the schemes presented in Tables 2 and 3. The time is measured initially from the initialization of $u^{(0)}$ until the solution converges to $u^{(k)}$, where k is the number of iterations.

Problem 1- A linear problem

$$\begin{aligned} -U'' + \rho U &= (\rho+1)(\sin x + \cos x), \quad 0 \leq x \leq \frac{\pi}{2}, \\ U(0) &= 1, \quad U(\frac{\pi}{2}) = 1, \quad h = \pi/2(N+1). \end{aligned}$$

The exact solution is

$$U(x) = \sin x + \cos x.$$

The matrix A is given by

$$\mathbf{A} = \begin{bmatrix} 2g & -1 & & \\ -1 & 2g & -1 & \mathbf{O} \\ \ddots & \ddots & \ddots & \\ \mathbf{O} & -1 & 2g & -1 \\ & & -1 & 2g \end{bmatrix},$$

where $g = 1 + 0.5\rho h^2$.

1) The computational complexity

Since g is independent of the solution vector u and is a constant, all intermediate calculations may be computed outside the iteration loop. The evaluation of vector b, at each point, can be assigned as an array so that it will save 1 addition and 2 multiplication per iteration, Thus, the work involves only one addition and multiplication for each node and the addition of an element of an array.

It should be noticed that for the three CAGE schemes, the amount of computational work is the same, i.e., 6 multiplications and 6 additions. Thus, it is sufficient to tabulate the amount of work for the CAGE-PR(2) scheme only in Table 2. The total operations for the other schemes may be derived in the same way.

The results for large N are tabulated in Table 2.

The Scheme	**Multiplication**	**Addition**	**Total**
AGE-PR(2)	8N	7N	15N
AGE-DG	8N	7N	15N
AGE-GT	6N	5N	11N
AGE-PR(1)	8N	8N	16N
CAGE-PR(2)	6N	6N	12N
SMAGE	7N	6N	13N

TABLE 2: Problem 1: the amount of work per iteration.

2) The CPU time

Table 3 shows the times taken for each prescribed scheme for solving Problem 1.

		$\rho = 70$	The Schemes with times taken in sec.					
N	r	iter.	[1]	[2]	[3]	[4]	[5]	[6]
20	0.60	10	0.04	0.04	0.03	0.04	0.03	0.04
40	0.30	19	0.15	0.15	0.11	0.16	0.12	0.14
80	0.11	37	0.59	0.57	0.44	0.61	0.45	0.56
160	0.05	73	2.30	2.24	1.75	2.43	1.78	2.15
320	0.02	140	8.74	8.63	6.69	9.21	6.80	8.30

TABLE 3: Problem 1: the CPU time taken for each scheme.

Notation: *[1]: AGE-PR(2),* *[2]:AGE-DG,* *[3]: AGE-GT,*
[4]: AGE-PR(1), *[5]: CAGE-PR(2),* *[6]: SMAGE.*

5. CONCLUSIONS

The standard form of PR, i.e., the AGE-PR(1) scheme is known to be limited as it will only serve to solve the one dimensional problem. This equation, when rewritten in generalised form results in the AGE-PR(2) scheme.

This approach, together with the strategy suggested by Douglas and Guittet, present new schemes which were analyzed theoretically and tested for convergence. The results show that all the schemes give a similar rate of convergence when $\omega = 2$, i.e., all the schemes are identical.

By comparing the computational complexity for each scheme, the AGE-PR(1) scheme has less computations but is limited in application. The AGE-DG scheme however can be extended to solve the two and three dimensional problems. The scheme has been shown to have slightly less computational work over the AGE-GT scheme and is easier to implement. Thus, we might consider the AGE-DG scheme is the best choice.

The formulae above, when rewritten in a single stage or coupled form, produce the CAGE formula. Based on the fact that the CAGE method for the schemes give a similar matrix, one would expect that the CPU times would be the same. Experimentally, this is not true and the CAGE-PR(2) scheme is shown to give a better time because it uses less intermediate variables.

The results conclude the CAGE method is not competitive to solve the one dimensional problem, by a single parameter. However, the CAGE-DG and CAGE-GT schemes can easily be combined with other methods such as the Richardson method to form a good second order method.

Although the SMAGE scheme shows a significant improved CPU time, it can only be considered for solving the one dimensional problem. The scheme which is based on the AGE-PR(2) scheme cannot be extended to solve problems with higher dimensions and is not suitable to be used with a second order method.

Hence, in conclusion, we may consider the AGE-DG or SMAGE scheme for solving the one dimensional problem with a single parameter, while for the solution

of higher dimensional problems using multiparameters the CAGE-DG or CAGE-GT schemes are recommended to be used with the second order methods.

REFERENCES

[1] Douglas J. Alternating Direction Methods for Three Space Variables *Numer. Math.,* Vol. 4, (1962) pp. 41-63.

[2] Douglas J. and Rachford, H. H., On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables, *Trans. Amer. Maths. Soc.,* Vol. 82, (1956) pp. 421-439.

[3] Evans D. J. Group Explicit Iterative Methods for Solving Large Linear Systems, *Int. J. Comp. Maths.,* Vol. 17, (1985) pp. 81-108.

[4] Evans D. J. The Solution of Periodic Parabolic Equations by the Coupled Alternating Group Explicit (CAGE) Iterative Method, *Int. J. Comp. Maths.,* Vol. 34, (1990) pp. 227-235.

[5] Guittet J. Une Nouvelle Methode De Directions Alternees a q Variables, *Journ. Math. Anal. and App.,* Vol. 17, (1967) pp. 199-213.

[6] Peaceman D. W. and Rachford, H. H., The Numerical Solution of Parabolic and Elliptic Differential Equations, J. *Soc. Indus. Appl. Math.,* Vol. 3, (1955) pp. 28-41.

DYADIC FRACTIONAL DIFFERENTIATION AND INTEGRATION OF WALSH TRANSFORM

B. I. Golubov
Department of Higher Mathematics,
Moscow Engineering Physics Institute (State University)
115409, Moscow, Kashirskoe shosse, 31
e-mail: golubov@mail.mipt.ru

INTRODUCTION

Following the concept of J. E. Gibbs [1] P.L. Butzer and H.J. Wagner [2] defined the notion of a dyadic strong derivative D. After that they introduced the dyadic strong integral I and dyadic pointwise derivative d (see [3]–[5]). Their definitions concerns functions defined on dyadic group G or dyadic field K. The dyadic group G and dyadic field K are isomorphic to modified segment $[0,1]^*$ and modified positive half-line $R_+^* = [0,+\infty)^*$ respectively. The characters of dyadic group G and dyadic field K are Walsh-Paley functions $w_n(o)$, $n \in Z_+ = \{0,\ 1,\ 2, \dots\}$ and generalized Walsh functions $\psi_y(o)$, $y \in R_+$ respectively. P.L. Butzer and H.J. Wagner proved the equalities $D\, w_n = n\, w_n$ and $d\, w_n(x) = n\, w_n(x)$ for $n \in Z_+$, $x \in G$ and $d\, \psi_y(x) = |y|\psi_y(x)$ for $x, y \in K$. In [3] for the functions $f \in L(R+)$ the equality $(D(f))\tilde{}\,(x) = x\tilde{f}(x)$ is proved, where $\tilde{f}$ is the Walsh transform of the function f.

C.W. Onneweer [6] introduced modified pointwise and strong dyadic derivatives for functions defined on dyadic group G or dyadic field K. He proved that the characters of dyadic group G or dyadic field K are differentiable in his sense and they are eigenfunctions of modified differential operator δ. He proved the equalities $\delta(w_0)(x) \equiv 0$, $\delta(w_n)(x) = 2^k w_n(x)$, $2^k \le n < 2^{k+1}$, $k \in Z_+$, $x \in D$. In another article [7] C.W. Onneweer introduced modified fractional differentiation and integration on compact Vilenkin groups G_p of order $p \ge 2$.

Some results on modified dyadic derivatives and integrals were proved in our papers [8] - [13]. J. Pal [14] proved that if $f \in L(R_+)$ and $xf(x) \in L(R_+)$, then Walsh transform $\tilde{f}$ has dyadic pointwise derivative in the sense of Butzer-Wagner and $d(\tilde{f})(x) = (tf(t))\tilde{}\,(x)$ at each point $x \in R_+$.

In this paper we define modified dyadic strong and pointwise integral and derivative of fractional order on R_+ and prove dyadic analogues of the following classical

formulas

$$\frac{d}{dx}\tilde{f}(x) = (-itf(t))\hat{}(x), \quad \int_0^x \hat{f}(t)dt = \int_R f(t)\frac{\exp(-itx)-1}{-it}dt \ \text{ for } \ x \in R,$$

where

$$\hat{f}(x) = \int_R f(t)\exp(-itx)dt$$

is the Fourier transform of the function $f \in L(R)$.

1. NOTATIONS AND DEFINITIONS

For a number $x \in R_+ \equiv [0, +\infty)$ we consider dyadic expansion $x = \sum\limits_{n=-\infty}^{+\infty} 2^{-n-1}x_n$, where x_n equals to 0 or 1. Note that $x_n = 0$ for $n \le n(x)$, where $n(x) \in Z = \{0, \ \pm 1, \ \pm 2, \dots\}$. If x is dyadic rational, then we take its finite expansion, i.e. $x_n = 0$ for $n \ge n_0(x) > -\infty$. We define dyadic sum of two numbers $x, y \in R_+$ by the operation $\oplus$ as follows: $x \oplus y = z$, where $z_n = x_n + y_n \pmod 2$ for all $n \in Z$. Let us set $t(x,y) = \sum\limits_{n=-\infty}^{+\infty} x_n y_{-n-1}$ and define the generalized Walsh functions $\psi(x,y) \equiv \psi_y(x) = (-1)^{t(x,y)}$ for $(x,y) \in R_+ \times R_+$. These were introduced by N. J. Fine [15]. It is evident that $\psi(x,y) = \psi(y,x), \psi(x,y) = \pm 1$ for $x, y \in R_+$. Let us note that the equality

$$\psi(x \oplus y, t) = \psi(x,t)\psi(y,t), \tag{1}$$

holds, if $t, x, y \in R_+$ and $x \oplus y$ is not dyadic rational. Hence for fixed t and x the equality (1.1) for all $y \in R_+$ excepting the countable set is valid.

The function $w_n(x) \equiv \psi(x,n), \ \ n \in Z_+$, are called the Walsh-Paley functions. They are 1-periodic on R_+.

For the function $f \in L(R_+)$ N.J. Fine [15] (see also [16], chapter 1 or [17], chapter 9) introduced its Walsh transform by the equality

$$\tilde{f}(x) = \int\limits_{R_+} \psi(x,y)f(y)\,dy. \tag{2}$$

For a function $f \in L^p(R_+)$, $1 < p \le 2$, then its Walsh transform is defined as the limit as $n \to +\infty$ of the sequence $\int_0^{2^n} f(y)\psi(x,y)\,dy$ in the norm of the space $L^q(R_+)$, where $\frac{1}{p} + \frac{1}{q} = 1$.

For $f \in L(R_+)$, $g \in L^p(R_+)$, $1 \le p \le +\infty$, we set $(f * g)(x) = \int_{R_+} f(x \oplus y)g(y)\,dy$, $x \in R_+$,

i.e. $f * g$ is dyadic convolution of f and g. Let us note that $f * g \in L^p(R_+)$, $(f * g)\ \widetilde{(f * g)} = \tilde{f}\tilde{g}$, if $1 \le p \le 2$.

2. LEMMAS

For $x > 0$ we set

$$h(x) = 2^{-n}, 2^n \le x < 2^{n+1}, n \in Z. \tag{3}$$

It is evident that $x^{-1} \le h(x) < 2x^{-1}$.

Lemma 1. *If $\alpha > 0$ and $n \in Z$, then for each $x > 0$ the following limit*

$$W_n^\alpha(x) = \lim_{m \to +\infty} \int_{2^{-n}}^{2^m} (h(t))^\alpha \psi(x,t)\,dt \tag{4}$$

exists and is finite. More precisely, $W_n^\alpha(x) = -2^{(\alpha-1)n}$ for $2^{n-1} \le x < 2^n$, $W_n^\alpha(x) = -2^{(\alpha-1)^n} + 2(1-2^{-\alpha}) \sum_{i=0}^{k} 2^{(n-i)(\alpha-1)}$ for $2^{n-k-2} \le x < 2^{n-k-1}, k = 0, 1, ...$ and $W_n^\alpha(x) = 0$ for $x \ge 2^n$.

Proof. From (3) and (4) we have

$$\begin{aligned} W_n^\alpha(x) &= \sum_{k=-n}^{+\infty} 2^{-k\alpha} \left(\int_0^{2^{k+1}} \psi(x,y)dy - \int_0^{2^k} \psi(x,y)dy \right) \\ &= \sum_{k=-n}^{+\infty} 2^{-k\alpha} (D_{2^{k+1}}(x) - D_{2^k}(x)), \end{aligned} \tag{5}$$

where

$$D_y(x) = \int_0^y \psi(x,t)dt, \quad x, y \in R_+$$

It is known that the equality

$$D_{2_k}(x) = 2^k X_{[0,2^{-k})}(x) \tag{6}$$

holds (see [17], p. 428). From (5) and (6) it follows that $W_n^\alpha(x) = 0$ for $x \ge 2^n$. for each $x > 0$ the series in the right-hand side of the equality (5) is actually a finite sum, because only a finite number of its members are not equal to zero. Using Abel

transform we have from (5)

$$W_n^{\alpha}(x) = -2^{n\alpha}D_{2^{-n}}(x) + \sum_{i=0}^{+\infty}\left(2^{(n-i)\alpha} - 2^{(n-i-1)\alpha}\right)D_{2^{-n+i+1}}(x).$$

$$= -2^{n(\alpha-1)}X_{[0,2^n)}(x) + 2(1-2^{-\alpha})\sum_{i=0}^{+\infty}2^{(n-i)(\alpha-1)}X_{[0,2^{n-i-1})}(x). \quad (7)$$

From (7) for $2^{n-1} \le x < 2^n$ we have $W_n^{\alpha}(x) = -2^{(\alpha-1)n}$. If $2^{n-k-2} \le x < 2^{n-k-1}, k = 0, 1, ...,$ then

$$W_n^{\alpha}(x) = -2^{(\alpha-1)n} + 2(1-2^{-\alpha})\sum_{i=0}^{k}2^{(n-i)(\alpha-1)}.\triangleright$$

We shall write below $f(x) \approx g(x), x \to a$, if $f(x) = O(g(x)), x \to a$ and simultaneously $g(x) = O(f(x)), x \to a$. Then from the Lemma 1 we have the following corollary.

Corollary 1.

1) If $0 < \alpha < 1, n \in Z$, then $W_n^{\alpha}(x) \approx x^{\alpha-1}, \quad x \to +0$;

2) $W_n^1(x) \approx \log_2(x^{-1}), x \to +0$;

3) if $\alpha > 1$, then $W_n^{\alpha}(x)$ is bounded on R_+

From the Corollary 1 and the equality $W_n^{\alpha}(x) = 0, x \ge 2^n$ we obtain

Corollary 2. *For $\alpha > 0$ and $n \in Z$ the inclusion $W_n^{\alpha} \in L(R_+)$ is valid.*

Remark 1. Let us note that the Corollary 2 can be also obtained from the equality (5). Indeed by using (5) and (6), we obtain easily the inequality $\| W_n^{\alpha} \|_{L(R_+)} \le 2^{n\alpha+1}$.

Lemma 2. *If $\alpha > 0$ and $n \in Z$ then $\tilde{W}_n^{\alpha}(x) = \varphi_n^{\alpha}(x)$ for all $x \in R_+$,* where

$$\varphi_n^{\alpha}(x) = (h(x))^{\alpha}X_{[2^{-n}+\infty}(x). \quad (8)$$

Proof. According to (5) we have

$$\tilde{W}_n^{\alpha}(x) = \sum_{k=-n}^{+\infty}2^{-ka}\left[\tilde{D}_{2^{k+1}}(x) - \tilde{D}_{2^k}(x)\right]. \quad (9)$$

Using the equality $\tilde{D}_{2^k} = X_{[0,2^k)}, k \in Z$ (see [17], p. 435), (9) and (8) we obtain

$$\tilde{W}_n^{\alpha}(x) = \sum_{k=-n}^{+\infty} 2^{-ka}\left[X_{[0,2^{k+1})}(x) - X_{[0,2^k)}(x)\right] = \sum_{k=-n}^{+\infty} 2^{-k\alpha} X_{[2^k,2^{k+1})}(x) = \varphi_n^{\alpha}(x), \;\; x \in R_+. \triangleright$$

Corollary 3. *For $\alpha > 0$ and $n \in Z$ the equality $\tilde{W}_n^{\alpha}(0) = 0$ is valid.*

For $\alpha > 0$ we set

$$\Lambda_n^{\alpha}(x) = \int_0^{2n} (h(t))^{-\alpha}\psi(x,t)dt, x \in R_+. \tag{10}$$

Lemma 3. *If $\alpha > 0$ and $n \in Z$, then we have*

a) $\Lambda_n^{\alpha}(x) = \frac{2^{n(\alpha+1)}}{2^{\alpha+1}-1}$ for $0 \le x < 2^{-n}$;

b) $\Lambda_n^{\alpha}(x) = C_n^{\alpha} 2^{-i(\alpha+1)}$ for $2^{-n+1} \le x, 2^{-n+i+1}, i \in Z_+$, where

$$C_n^{\alpha} = -\frac{1-2^{-\alpha}}{1-2^{-\alpha-1}} 2^{(n-1)(\alpha+1)}. \tag{11}$$

Proof. Taking into account the definition (3) of the function $h(x)$, similarly to (5) we obtain from (10)

$$\Lambda_n^{\alpha}(x) = \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha}\left[D_{2^{-k}}(x) - D_{2^{-k-1}}(x)\right], \tag{12}$$

where $D_{2^k}(x)$ is defined by the equality (6). Therefore

$$\Lambda_n^{\alpha}(x) = \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha}\left[2^{-k}X_{[0,2^k)}(x) - 2^{-k-1}X_{[0,2^{k+1})}(x)\right]. \tag{13}$$

Using Abel transform, from (13) we obtain

$$\Lambda_n^{\alpha}(x) = 2^{(n-1)\alpha}2^n X_{[0,2^{-n})}(x) + (2^{-\alpha}-1)\sum_{k=-n+1}^{+\infty} 2^{-k(\alpha+1)} X_{[0,2^k)}(x). \tag{14}$$

Hence

$$\Lambda_n^{\alpha}(x) = 2^{(n-1)\alpha}2^n + (2^{-\alpha}-1)\sum_{k=-n+1}^{+\infty} 2^{-k(\alpha+1)} = \frac{2^{n(\alpha+1)}}{2^{\alpha+1}-1} \;\text{ for }\; x \in [0, 2^{-n})$$

and the assertion a) of the lemma is proved

If $2^{-n+1} \le x < 2^{-n+i+1}, i \in Z_+$, then from (14) we obtain

$$\Lambda_n^\alpha(x) = (2^{-\alpha} - 1) \sum_{k=-n+1+i}^{+\infty} 2^{-k(\alpha+1)} = C_n^\alpha \, 2^{-i(\alpha+1)},$$

where C_n^α is defined by the equality (11). $\square$

Corollary 4. *For $\alpha > 0, n \in Z$ the function Λ_n^α is bounded on R_+ and $\Lambda_n^\alpha(x) \approx x^{-\alpha-1}, x \to +\infty$.*

Corollary 5. *For $\alpha > 0, n \in Z$ the inclusion $\Lambda_n^\alpha \in L(R_+)$ is valid.*

Remark 2. Let us note that the Corollary 5 can be obtained also from the equality (12). Indeed,

$$|\Lambda_n^\alpha(x)| \le \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha}[D_{2^{-k}}(x) + D_{2^{-k-1}}(x)].$$

From this inequality we easily deduce that $\| \Lambda_n^\alpha \|_{L(R_+)} \le 2.2^{(n-1)\alpha}/(1 - 2^{-\alpha})$.

Lemma 4. *If $\alpha > 0$ and $n \in Z$, then $\tilde{\Lambda}_n^\alpha(x) = \psi_n^\alpha(x)$ for all $x \in R_+$, where $\psi_n^\alpha(x) = (h(x))^{-\alpha} X_{[0,2^n)}(x)$ for $x > 0$ and $\psi_n^\alpha(0) = 0$.*

Proof. From (12) using Corollary 5 we have

$$\tilde{\Lambda}_n^\alpha(x) = \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha}[\tilde{D}_{2^{-k}}(x) - \tilde{D}_{2^{-k-1}}(x)]. \tag{15}$$

As we mentioned above $\tilde{D}_{2^k} = X_{[0,2^k)}, k \in Z$. Hence from (15) it follows

$$\begin{aligned} \tilde{\Lambda}_n^\alpha(x) &= \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha} \left[X_{[0,2^{-k})}(x) - X_{[0,2^{-k-1})}(x)\right] \\ &= \sum_{k=-n}^{+\infty} 2^{-(k+1)\alpha} X_{[2^{-k-1},2^{-k})}(x) = \psi_n^\alpha(x), x \in R_+. \triangleright \end{aligned}$$

3. MAIN RESULTS

Definition 1. *If $\alpha > 0$ and for the function $f \in L^\infty(R_+) \cup L(R_+)$ the following limit $d^{(\alpha)}(f)(x) = \lim\limits_{n\to\infty} (f * \Lambda_n^\alpha)(x)$ exists and is finite at the point $x \in R_+$, then the*

number $d^{(\alpha)}(f)(x)$ is called the dyadic derivative of order α of the function f at the point x.

Theorem 1. *Let us assume that $\alpha > 0$ and $f, h^{\alpha} f \in L(R_+)$. Then the Walsh transform $\tilde{f}$ of the function f has a dyadic derivation of order α at each point $x \in R_+$ and the equality $d^{(\alpha)}(\tilde{f})(x) = (h^{\alpha}\tilde{f})(x)$ holds.*

Proof. For each $x \in R_+$ and $n \in Z_+$, by definition of the Walsh transform we have

$$\begin{aligned}(\tilde{f} * \Lambda_n^{\alpha}(x) &= \int_{R_+} \tilde{f}(u)\Lambda_n^{\alpha}(x \oplus u)du \\ &= \int_{R_+} \left\{ \int_{R_+} f(t)\psi(u,t)dt \right\} \Lambda_n^{\alpha}(x \oplus u)du. \end{aligned} \tag{16}$$

Taking into account that $|\psi(u,t)| \equiv 1, f \in L(R_+), \Lambda_n^{\alpha} \in L(R_+)$ (see the Corollary 5) and applying Fubini's theorem, we can interchange the order of integration in the right-hand side of (16). After that we have

$$\begin{aligned}(\tilde{f} * \Lambda_n^{\alpha}(x) &= \int_{R_+} \left\{ \int_{R_+} \Lambda_n^{\alpha}(x \oplus u)\psi(u,t)du \right\} f(t)dt \\ &= \int_{R_+} \left\{ \int_{R_+} \Lambda_n^{\alpha}(u)\psi(x \oplus u,t)du \right\} f(t)dt. \end{aligned}$$

Hence using the equality (1) we obtain

$$\begin{aligned}(\tilde{f} * \Lambda_n^{\alpha})(x) &= \left\{ \int_{R_+} \Lambda_n^{\alpha}(u)\psi(u,t)du \right\} f(t)\psi(x,t)dt \\ &= \int_{R_+} \tilde{\Lambda}_n^{\alpha}(t)f(t)\psi(x,t)dt. \end{aligned} \tag{17}$$

By the Lemma 4, the equality $\tilde{\Lambda}_n^{\alpha}(x) = \psi_n^{\alpha}(x)$. holds. Hence we can write the equality (17)in the form

$$(\tilde{f} * \Lambda_n^{\alpha})(x) = \int_0^{2^n} (h(t))^{-\alpha} f(t)\psi(x,t)dt. \tag{18}$$

Since $h^{-\alpha} f \in L(R_+)$, then the right-hand side of the equality (18) has a finite limit $(h^{-\alpha}\tilde{f})(x)$ as $n \to +\infty$. But the limit of the left-hand side of (18) by definition is equal to $d^{\alpha}(\tilde{f})(x)$.$\triangleright$

Lemma 5. *The generalized Walsh function $\psi(o,y) \equiv \psi_y(o)$ has modified dyadic derivative of any order $\alpha > 0$ at each point $x \in R_+$. More precisely $d^{(\alpha)}(\psi_0)(x) \equiv 0$ on R_+ and $d^{(\alpha)}(\psi_y)(x) = (h(y))^{-\alpha}\psi_y(x)$ for $y > 0, x \in R_+$.*

Proof. Since $\psi_0(x) \equiv 1$ on R_+ then by the Lemma 4 we have

$$\left(\Lambda_n^\alpha * \psi_0\right)(x) = \int_{R_+} \Lambda_n^\alpha(t)dt = \tilde{\Lambda}_n^\alpha(0) = 0, x \in R_+.$$

From this equality as $n \to +\infty$ it follows $d^{(\alpha)}(\psi_0)(x) \equiv 0$ on R_+. For $y > 0$ using equality (1) and Lemma 4 we obtain

$$\begin{aligned}(\Lambda_n^\alpha * \psi_y)(x) &= \\ &\int_{R_+} \Lambda_n^\alpha(t)\psi_y(x \oplus t)dt \\ &= \psi_y(x)\int_{R_+} \Lambda_n^\alpha(t)\psi_y(t)dt = \psi_n^\alpha(y)\psi_y(x).\end{aligned}$$

Taking the limit as $n \to +\infty$ we have $d^{(\alpha)}(\psi_y)(x) = (h(y))^{-\alpha}\psi_y(x)$ for $x \in R_+$.▷

Remark 3. Let us note that the relation $d^{(\alpha)}(\tilde{f})(x) = (h^{-\alpha}f\tilde{)}(x)$ can be formally obtained from the equation (2) by dyadic differentiation of order $\alpha > 0$ if we take into account the Lemma 5.

Definition 2. *If $\alpha > 0$ and for a function $f \in L^\infty(R_+)$ the limit $j_\alpha(f)(x) = \lim\limits_{n\to+\infty}(f * W_n^\alpha)(x)$ exists and is finite at the point $x \in R_+$, then the number $j_\alpha(f)(x)$ is called the modified dyadic integral of order α of the function f at the point x.*

Theorem 2. *If $\alpha > 0, f \in R(R_+)$ and $h^\alpha f \in L(R_+)$, then the Walsh transform of the function f has a modified dyadic integral of order α at each point $x \in R_+$ and the equality $j_\alpha(\tilde{f})(x) = (h^\alpha f\tilde{)}(x)$ holds.*

Proof. For each $x \in R_+$ and $n \in Z_+$ we have

$$\begin{aligned}(\tilde{f} * W_n^\alpha)(x) &= \int_{R_+} \tilde{f}(u)W_n^\alpha(x \oplus u)du \\ &= \int_{R_+}\left\{\int_{R_+} f(t)\psi(u,t)dt\right\}W_n^\alpha(x \oplus u)du. \qquad (19)\end{aligned}$$

Taking into account that $|\psi(u,t)| \equiv 1, f \in L(R_+), W_n^\alpha \in L(R_+)$ (see the Corollary 2) and applying Fubini's theorem, we can interchange the order of integration in the right-hand side of (19). After that we have

$$\begin{aligned}(\tilde{f} * W_n^\alpha)(x) &= \int_{R_+}\left\{\int_{R_+} W_n^\alpha(x \oplus u)\psi(u,t)du\right\}f(t)dt \\ &= \int_{R_+}\left\{\int_{R_+} W_n^\alpha(u)\psi(x \oplus u,t)du\right\}f(t)dt.\end{aligned}$$

Hence using the equality (1) we obtain

$$(\tilde{f} * W_n^\alpha)(x) = \int_{R_+} \left\{ \int_{R_+} W_n^\alpha(u)\psi(u,t)du \right\} f(t)\psi(x,t)dt$$
$$\int_{R_+} \tilde{W}_n^\alpha(t) f(t)\psi(x,t)dt. \tag{20}$$

By the Lemma 2 the equality $\tilde{W}_n^\alpha = \varphi_n^\alpha$ holds. Hence we can write the equality (20) in the form

$$(\tilde{f} * W_n^\alpha)(x) = \int_{2^{-n}}^{+\infty} (h(t))^\alpha f(t)\psi(x,t)dt. \tag{21}$$

Since $h^\alpha f \in L(R_+)$, then the right-hand side of the equality (21) has finite limit $(h^\alpha f\tilde{)}(x)$ as $n \to +\infty$. But the limit of the left-hand side of (21) by definition is equal to $j_\alpha(\tilde{f})(x)$.▷

Lemma 6. *The generalized Walsh function $\psi(o,y) \equiv \psi_y(o)$ has a modified dyadic integral of any order $\alpha > 0$ at each point $x \in R_+$. More precisely $j_\alpha(\psi_0(x) = 0$ on R_+ and $j_\alpha(\psi_y)(x) = (h(y))^\alpha \psi_y(x)$ for $y > 0$, $x \in R_+$.*

Proof. Since $\psi_0(x) \equiv 1$ on R_+ then the Lemma 2 and Corollary 3 we have $(W_n^\alpha * \psi_0)(x) = \tilde{W}_n^\alpha(0) = 0$. From this equality as $n \to +\infty$ it follows $j_\alpha(\psi_0(x) = 0$ on R_+. For $y > 0$ using the equality (1) and Lemma 2 we obtain

$$(W_n^\alpha * \psi_y)(x) = \int_{R_+} W_n^\alpha(t)\psi_y(x \oplus T)dt$$
$$= \psi_y(x) \int_{R_+} W_n^\alpha(t)\psi_y(t)dt = \varphi_n^\alpha(y)\psi_y(x).$$

Taking the limit as $n \to +\infty$, we deduce that $j_\alpha(\psi_y)(x) = (h(y))^\alpha \psi_y(x)$ for $x \in R_+$.▷

Remark 4. Let us note that the equality $j_\alpha(\tilde{f})(x) = (h^\alpha f\tilde{)}(x)$ can be formally obtained from the relation (2) by dyadic integration of order $\alpha > 0$ if we take into account the Lemma 6.

Example 1. For the function $\varphi = X_{[0,1)}$ and $\alpha > 0$ we have

a) If $x \in [0,1)$, then $d^{(\alpha)}(\varphi)(x) = (2^{\alpha+1} - 1)^{-1}$

b) if $x \geq 1$, then $d^{(\alpha)}(\varphi)(x) = -(1 - 2^{-\alpha}) \sum\limits_{k=1}^{+\infty} 2^{-k(\alpha+1)} X_{[0,2^k)}(x)$.

Example 2. For the function $\varphi = X_{[0,1)}$ we have:

a) if $\alpha \in (0,1)$ then the function φ has a modified dyadic integral of order α at each point $x \in R_+$;

b) for $\alpha \geq 1$ the function φ does not have a modified dyadic integral of order α at any given point $x \in R_+$.

ACKNOWLEDGEMENT

I would like to express my gratitude to the Organizing Committee of the International Conference "Mathematics and its applications, Kuwait University, 2004" for their invitation to participate in the conference.

This work was supported by the Russian Foundation for Basic Research under Grant 02-01-00428.

I would like to express my gratitude to the referee for many useful remarks.

REFERENCES

[1] Gibbs, G. E. Walsh spectrometry, a form of spectral analysis well suited to binary digital computation Nat. Phys. Lab., Teddington, UK, 24 p., 1967.

[2] Butzer, P. L. and Wagner, H. J., Walsh series and the concept of a derivative Applicable Analysis 3, No. 1 (1973), 29-46.

[3] Butzer, P. L. and Wagner, H. J., A calculus for Walsh functions defined on R_+ Proc. Symp. Naval Res. Laboratory, Washington, D. C., April 18–20 (1973), p. 75-81.

[4] Butzer, P. L. and Wagner, H. J., On dyadic analysis based on pointwise dyadic derivative Anal. Math. 1. (1975), 171-196.

[5] Wagner, H. J. On dyadic calculus for functions defined on R_+ "Theory and applications of Walsh functions", Proc. Symp., Hatfield Polytechnic (1975), p.101-129.

[6] Onneweer, C.W. On the definition of dyadic differentiation Applicable Anal., 9 (1979), 267–278.

[7] Onneweer, C.W. Fractional differentiation on the group of integers of a p-adic or p-series field Anal. Math., 3 (1977), 119–130.

[8] Golubov, B. I. An analogue of a theorem of Titchmarsh for Walsh-Fourier transform (in Russian), Mat. Sbornik, 189, No 5 (1998), 69-86.

[9] Golubov, B. I., On dyadic analogues of the operators of Hardy and Hardy Littlewood (in Russian), Siberian Math. J., 40, No 6 (1999), 1244-1252.

[10] Golubov, B. I. On boundedness of dyadic Hardy and Hardy-Littlewood operators in dyadic space H BMO (in Russian), Anal. Math., 26 (2000), 287-298.

[11] Golubov, B.I. On an analogue of Hardy's inequality for the Walsh-Fourier transform (in Russian) Izvestiya RAN (Proc. Rus. Acad. Sci.), Ser. Math., 65, No. 3 (2001), 3-14.

[12] Golubov, B.I. On Modified strong dyadic integral and derivative (in Russian) Mat. Sbornik, 193, No. 4 (2002), 37–60.

[13] Golubov, B. I. Dyadic analogue of the Tauberian theorem of Wiener (in Russian) Izvestiya RAN (Proc. Rus. Acad. Sci.), Ser, Math., 67, No 1(2003), 33-58.

[14] Pal J. On the connection between the concept of a derivative defined on the dyadic field and the Walsh-Fourier transform, Annales Sci. Univ. Budapest. Sect. Math., 18 (1975), 49-54.

[15] Fine, N. J. The generalized Walsh functions, Trans. Amer. Math. Soc., 69 (1950), 66-67.

[16] Golubov B., Efimov A. and Skvortsov V., Walsh series and transforms. Theory and applications, Kluwer Academic Publishers, Dordrecht-Boston-London, 1991.

[17] Schipp, F., Wade W. R., Simon P. and Pal, J., Walsh series. An introduction to dyadic harmonic analysis, Akademiai Kiado, Budapest, 1990.

FAST "SPLIT" ALGORITHMS FOR TOEPLITZ AND TOEPLITZ-PLUS-HANKEL MATRICES WITH ARBITRARY RANK PROFILE

G. Heinig[1] and K. Rost[2]
[1]Dept. of Mathematics & Computer Science
Kuwait University, P. O. Box 5969, Safat 13060, Kuwait
e-mail: georg@sci.kuniv.edu.kw

[2] Dept. of Mathematics,
University of Chemnitz, D-09107 Chemnitz, Germany
e-mail: karla.rost@mathematik.tu-chemnitz.de

1. INTRODUCTION

This paper is a comprehensive account of the authors' work on so called split algorithms for solving Toeplitz and Toeplitz-plus-Hankel systems of equations during the last two years. The main aim of this work was to remove additional constraints which are contained in the classical versions of the algorithms, so that the algorithms are applicable to any nonsingular matrix in the given class.

To begin with let us give a brief introduction into the subject. The solution of a linear system $A\mathbf{x} = \mathbf{b}$ with a nonsingular $n \times n$ coefficient matrix A using a standard direct algorithm requires $O(n^3)$ operations. In case that A has a certain structure this amount can be reduced to $O(n^2)$ or less for some important classes (see [17], [32], [14] and references therein). For example, if $A = T_n$ is a Toeplitz matrix $T_n = [\, a_{i-j} \,]_{i,j=1}^n$, then two well known algorithms do this job: the Levinson (also called Levinson-Durbin) algorithm and the Schur (also called Schur-Bareiss) algorithm. Fast algorithms of these two types also exist for Toeplitz-plus-Hankel (briefly T+H) matrices $C = [\, a_{i-j} + s_{i+j-1} \,]$ (see [37], [36], [16] and [26] and references in [26]). The problem with all these algorithms is that they are applicable in their original form only under some conditions. For example, both the classical Levinson and the Schur algorithms require that the leading principal submatrices are nonsingular. A matrix with this property is called *strongly nonsingular*. In order to avoid breakdowns in computer programmes it is desirable to have modifications of the algorithms that are applicable without any restriction. A modification of the Levinson algorithm working for Toeplitz matrices with any rank profile was first described in [10]. Schur-type algorithms and algorithms for block Toeplitz matrices were presented in [17], [6], [38], [40], [11], [15], and other papers. Fast algorithms that are applicable to any T+H matrix were presented in the recent paper [14].

It was observed in [4] and [5] that in the case of a symmetric Toeplitz matrix the number of multiplications in the Levinson and Schur algorithms can be reduced by 50% while keeping the number of additions if symmetry properties are exploited. The decisive feature that allows the computational reduction for a symmetric Toeplitz matrix is not its symmetry but rather its centrosymmetry. A matrix $[\, a_{ij} \,]_{i,j=1}^{n}$ is called *centrosymmetric* if $a_{n+1-i,n+1-j} = a_{ij}$. The subspaces of symmetric and skewsymmetric vectors are invariant subspaces of any centrosymmetric matrix. Thus linear systems can be "splitted" – therefore the name "split" algorithm – into a symmetric and a skewsymmetric part, so that most of the calculations in the algorithm include operations on symmetric and skewsymmetric vectors. This leads to the complexity reduction. The split Levinson algorithm for symmetric Toeplitz matrices was slightly improved in [35] (see also references therein) and [13]. The latter paper also contains the corresponding Schur counterpart and generalizations to centrosymmetric Toeplitz-plus-Hankel matrices. Split algorithm for skewsymmetric Toeplitz matrices were designed in [23]. Note that the classical Levinson and Schur algorithms do not work for skewsymmetric Toeplitz matrices. Analogous algorithms for hermitian Toeplitz matrices were presented in [33], [34], [3]. We call them also "split algorithms" despite they are not "split" in the original sense. Split algorithms for centrosymmetric T+H matrices can be found in [13] and [25]. In the latter paper also centro-skewsymmetric T+H matrices were considered. In our paper [26] it was shown that the approach for the construction of split algorithms can be generalized to general T+H matrices. That leads to methods which are more efficient than previous ones, but the gain is less than in the pure Toeplitz or centrosymmetric T+H case.

All split algorithms designed in the papers mentioned above work only under some conditions. In most cases it is required that the central submatrices are nonsingular. A matrix with this property is called *centrononsingular*. The authors took the challenge to design split algorithms without extra conditions. In this paper we present such algorithms for symmetric, skewsymmetric and hermitian Toeplitz matrices, and for centrosymmetric T+H matrices. Note that it is an open question how to overcome the restriction of centrononsingularity in the split algorithms for general T+H matrices that are presented in [26].

The classical Levinson-type and Schur-type algorithms are related to triangular factorization. The Schur algorithm produces an LU-factorization of the matrix whereas the Levinson-type algorithms produce a UL-factorization of its inverse. It was observed in [2] that the split Levinson algorithm for symmetric Toeplitz matrices produces a WZ-factorization of the inverse matrix (for the definition of this concept

see Section 5). Likewise, the split Schur algorithm provides a ZW-factorization of the matrix itself. Originally the concept of WZ-factorization was introduced and studied by D.J. Evans and his coworkers [8] in connection with parallel solution of tridiagonal systems. The ZW- and WZ-factorizations for skewsymmetric Toeplitz, hermitian Toeplitz, centrosymmetric T+H, and general T+H matrices and their inverses were investigated in [23], [29], [25], and [26], respectively.

Let us explain the methodology of our approach to design algorithms for matrices with any rank profile. There are, in principle, two possibilities to deal with singular principal submatrices. The first one is based on a look-ahead strategy, which results in jumping from one nonsingular principal submatrix to the next one. This approach, however, is not applicable to block Toeplitz matrices and general T+H matrices. The second approach is based on the concept of a fundamental system of the matrix, which is a system of a few vectors (in the Toeplitz case 2, in the T+H case 4) containing all information about the matrix, no matter whether the matrix is nonsingular or singular. In the present paper we focus our attention to the look-ahead approach. Note that the fundamental system approach for skewsymmetric Toeplitz matrices is discussed in the Thesis [2].

Throughout the paper we consider matrices with entries from a field $\mathbb{F}$ with a characteristic different from 2. Only in the sections concerning hermitian matrices the underlying field will be the field of complex numbers $\mathbb{C}$. By $\mathbf{e}_k$ we denote the kth vector in the standard basis of $\mathbb{F}^n$.

2. INVERSION FORMULAS

A common tool for solving a structured system of equations is to use special matrix representations for the inverse matrix. The system is solved then by fast matrix-vector multiplication. In this section we discuss inversion formulas for special classes that are adapted to the algorithms which will be presented in the forthcoming sections.

2.1. General Toeplitz Matrices

To begin with we recall some facts concerning inverses of Toeplitz matrices. It is well known that inverses of Toeplitz matrices are, in general, not Toeplitz matrices again but so-called Toeplitz Bezoutians. We give the definition in terms of the generating function of a matrix.

If $A = [\, a_{ij} \,]_{i,j=1}^{n}$ is a matrix, then the generating function is, by definition, the

bivariate polynomial $A(t,s) = \sum_{i,j=1}^{n} a_{ij} t^{i-1} s^{j-1}$. In the same spirit the polynomial $\mathbf{x}(t)$ is defined for a vector $\mathbf{x}$. Let $\mathbf{p}, \mathbf{q} \in \mathbb{F}^{n+1}$ and let

$$J_n = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}$$

be the $n \times n$ matrix of the counteridentity. Then the (Toeplitz) Bezoutian of $\mathbf{p}$ and $\mathbf{q}$ is defined as the $n \times n$ matrix $B = \mathrm{Bez}\,(\mathbf{p}, \mathbf{q})$ with the generating function

$$B(t,s) = \frac{\mathbf{p}(t)\widehat{\mathbf{q}}(s) - \mathbf{q}(t)\widehat{\mathbf{p}}(s)}{1 - ts}\,,$$

where $\widehat{\mathbf{p}} = J_{n+1}\mathbf{p}$. Originally, Bezoutians were introduced in connection with root separation problems (see [17] and references therein). The entries of the matrix B can be constructed recursively from $\mathbf{p}$ and $\mathbf{q}$ in $O(n^2)$ operations (see [17]). More important are "global" matrix representations of Bezoutians like the Gohberg-Semencul formula

$$\mathrm{Bez}\,(\mathbf{p}, \mathbf{q}) = \begin{bmatrix} p_0 & & \\ \vdots & \ddots & \\ p_{n-1} & \cdots & p_0 \end{bmatrix} \begin{bmatrix} q_n & \cdots & q_1 \\ & \ddots & \vdots \\ & & q_n \end{bmatrix} - \begin{bmatrix} q_0 & & \\ \vdots & \ddots & \\ q_{n-1} & \cdots & q_0 \end{bmatrix} \begin{bmatrix} p_n & \cdots & p_1 \\ & \ddots & \vdots \\ & & p_n \end{bmatrix},$$

where $\mathbf{p} = (p_i)_{i=0}^{n}$, $\mathbf{q} = (q_i)_{i=0}^{n}$.

In the case where $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$ this and other representations allow matrix-vector multiplication by Bezoutians with a computational complexity of $O(n \log n)$ if FFT or fast real trigonometric transformations are used. Note that more efficient formulas than the Gohberg-Semencul formula involving circulant and skewcirculant matrices were presented in [8], [9], [12], and other papers. Representations that involve only diagonal matrices and discrete Fourier or real trigonometric transformations can be found in [18], [20], [21], and [41]. The following fact is well known (see [17]).

Proposition 2.1 The inverse of a nonsingular $n \times n$ Toeplitz matrix T_n admits the representation

$$T_n^{-1} = \mathrm{Bez}\,(\mathbf{p}, \mathbf{q})\,,$$

where

$$\mathbf{p} = \begin{bmatrix} \mathbf{p}' \\ 0 \end{bmatrix},\ \mathbf{p}' = T_n^{-1}\mathbf{e}_1, \quad \mathbf{q} = \begin{bmatrix} \mathbf{q}' \\ 1 \end{bmatrix},\ \mathbf{q}' = T_n^{-1}\mathbf{g},\ \ \mathbf{g} = (-a_{i-n})_{i=0}^{n-1},$$

and $a_{-n} \in \mathbb{F}$ is arbitrary.

A disadvantage of this formula is that if T_n has a symmetry property, then this is not reflected in the formula. But this is desirable in order to design efficient algorithms exploiting the symmetry. Next we discuss formulas that reflect symmetry properties of the Toeplitz matrix.

2.2. Symmetric Toeplitz Matrices

In the case of a symmetric Toeplitz matrix $T_n = [\, a_{|i-j|} \,]_{i,j=1}^n$ the classical Gohberg-Semencul formula reflects the symmetry of the matrix (see [17]). The parameters in the formula are the components of the first column of the inverse of an $(n+1) \times (n+1)$ Toeplitz extension T_{n+1} of T_n. It can be shown that almost all extensions T_{n+1} are nonsingular. In connection with split algorithms it is convenient to consider a formula that involves the symmetric and skewsymmetric part of this vector. Let $\mathbf{x}_{n+1}^{\pm}$ be the solution of

$$T_{n+1}\mathbf{x}_{n+1}^{\pm} = \mathbf{e}_1 \pm \mathbf{e}_{n+1}.$$

Note that $\mathbf{x}_{n+1}^{+}$ is symmetric and $\mathbf{x}_{n+1}^{-}$ is skewsymmetric. Then Proposition 2.1 leads to the following (see [19]).

Theorem 2.2 The inverse of a nonsingular symmetric $n \times n$ Toeplitz matrix $T_n = [\, a_{|i-j|} \,]_{i,j=1}^n$ is given by

$$T_n^{-1}(t,s) = \frac{1}{\gamma} \frac{\mathbf{x}_{n+1}^{+}(t)\mathbf{x}_{n+1}^{-}(s) + \mathbf{x}_{n+1}^{-}(t)\mathbf{x}_{n+1}^{+}(s)}{1 - ts},$$

where $\gamma = \mathbf{x}_{n+1}^{+}(0) + \mathbf{x}_{n+1}^{-}(0)$.

The algorithms described below compute the (symmetric) solutions of the equations $T_k\mathbf{x}_k = \mathbf{e}_1 \pm \mathbf{e}_k$ for $k = n$ and $k = n+2$, where T_{n+2} is an $(n+2) \times (n+2)$ nonsingular symmetric Toeplitz extension of T_{n+1}. We show now how $\mathbf{x}_{n+1}^{\pm}$ can be computed from $\mathbf{x}_n$ and $\mathbf{x}_{n+2}$ (see [28]).

Proposition 2.3 The polynomials $\mathbf{x}_{n+1}^{\pm}(t)$ are given by

$$\mathbf{x}_{n+1}^{\pm}(t) = \frac{t\mathbf{x}_n(t) - c_{\pm}\mathbf{x}_{n+2}(t)}{t \pm 1},$$

where $\mathbf{x}_{n+2}(1) \neq 0$ and $c_- = \frac{\mathbf{x}_n(1)}{\mathbf{x}_{n+2}(1)}$. If n is odd, then $\mathbf{x}_{n+2}(-1) \neq 0$ and $c_+ = -\frac{\mathbf{x}_n(-1)}{\mathbf{x}_{n+2}(-1)}$. If n is even, then $\mathbf{x}_{n+2}(-1) = 0$ and c_+ is not determined by $\mathbf{x}_n$ and $\mathbf{x}_{n+2}$ alone.

If n is even, then the constant c_+ can be obtained by applying a test functional, which could be the multiplication by any row of T_{n+1}.

2.3. Skewsymmetric Toeplitz Matrices

Let $T_n = [\, a_{i-j} \,]_{i,j=1}^n$ be a nonsingular skewsymmetric Toeplitz matrix, i.e. $a_{-i} = -a_i$. Since any skewsymmetric matrix of odd order is singular, n must be even. We extend T_n to a skewsymmetric Toeplitz matrix T_{n+1}. Clearly, the kernel of this matrix as well as the kernel of T_{n-1} have dimension one. Moreover, it can be shown (see [23]) that these kernels consist only of symmetric vectors. Let $\mathbf{u}, \mathbf{u}'$ be vectors spanning $\ker T_{n+1}$ and $\ker T_{n-1}$, respectively, that are normalized according to $\mathbf{e}_1^T \mathbf{u} = 1$ and $[\, a_{n-1} \;\; \ldots \;\; a_1 \,]\, \mathbf{u}' = 1$. The following consequence of Proposition 2.1 was proved in [23].

Theorem 2.4 The inverse of a nonsingular skewsymmetric Toeplitz matrix T_n is given by

$$T_n^{-1}(t,s) = \frac{\mathbf{u}(t)\mathbf{x}(s) - \mathbf{x}(t)\mathbf{u}(s)}{1 - ts} \,.$$

where $\mathbf{x}(t) = t\mathbf{u}'(t)$.

Note that in the formula for the inverse of a skewsymmetric Toeplitz matrix only symmetric vectors are involved, whereas in the corresponding formula for the symmetric case we have one symmetric and one skewsymmetric vector. Some explanation of this surprising fact is given in [22]. As a consequence of this, the computation for skewsymmetric Toeplitz matrices is somehow simpler than for symmetric matrices. There is no need for after processing calculations like in Proposition 2.3.

2.4. Hermitian Toeplitz Matrices

Let $\mathbb{F} = \mathbb{C}$ and $T_n = [\, a_{i-j} \,]_{i,j=1}^n$ be an hermitian Toeplitz matrix, i.e. $a_{-i} = \overline{a}_i$. Since the splitting of hermitian Toeplitz matrices is different to that of symmetric Toeplitz matrices, we have a different kind of algorithms. As a consequence of this we need an inversion formula that does not involve the vectors $\mathbf{x}_{n+1}^{\pm}$ but the solutions of equations

$$T_k \mathbf{q}_k = \mathbf{1} \;\; (k = n, n+1) \,.$$

Here T_{n+1} is a nonsingular $(n+1) \times (n+1)$ hermitian Toeplitz extension of T_n and $\mathbf{1}$ denotes the vector all components of which are equal to 1. The following formula can be found in [30].

Theorem 2.5 The inverse T_n^{-1} is given by

$$T_n^{-1}(t,s) = \frac{\mathrm{i}}{c}\,\frac{\mathbf{w}(t)\overline{\mathbf{q}}_{n+1}(s) + \mathbf{q}_{n+1}(t)\overline{\mathbf{w}}(s)}{1-ts} - \frac{1}{c}\,\mathbf{q}_n(t)\overline{\mathbf{q}}_n(s)\ , \tag{1}$$

where $\mathbf{w}(t) = \mathrm{i}(t-1)\mathbf{q}_n(t)$, $c = \mathbf{q}_{n+1}(1) - \mathbf{q}_n(1)$ and $\mathrm{i} = \sqrt{-1}$.

Note that the vectors $\mathbf{q}_k$ and $\mathbf{w}$ are conjugate-symmetric, which means that $J_k\mathbf{q}_k = \overline{\mathbf{q}}_k$ and $J_{n+1}\mathbf{w} = \overline{\mathbf{w}}$. It is important to mention that $T_{n+1}\mathbf{w} = \rho\mathbf{e}_{n+1} + \overline{\rho}\mathbf{e}_1$ for some $\rho \in \mathbb{C}$ and $\mathbf{w}(1) = 0$. The vector $\mathbf{w}$ is characterized by these two properties up to a real, nonzero factor.

We give another interpretation of this fact which is convenient for our purposes. For this we introduce the matrices

$$\partial T_k = [\, a_{i-j} \,]_{i=1,\,j=0}^{k-1,\;k} \quad (k = n, n+1)$$

obtained from T_k by deleting its first row and adding a column to the right compatible with its Toeplitz structure. Let $\mathcal{C}_k$ denote the subspace over the reals of all conjugate-symmetric vectors $\mathbf{u}$ in the kernel of ∂T_k satisfying $\mathbf{u}(1) = 0$. It can be shown that $\mathcal{C}_k$ is one-dimensional (over the reals) if T_k is nonsingular. In particular, the vector $\mathbf{w}$ spans $\mathcal{C}_n$. Furthermore, the polynomial $\mathbf{q}_{n+1}(t)$ can be obtained, up to a real factor, from a vector $\mathbf{w}_{n+1}$ spanning $\mathcal{C}_{n+1}$ by dividing $\mathbf{w}_{n+1}(t)$ by $\mathrm{i}(t-1)$.

2.5. Centrosymmetric Toeplitz-plus-Hankel Matrices

We consider now T+H matrices $C_n = [\, a_{i-j} + s_{i+j-1} \,]_{i,j=1}^n$ that are centrosymmetric, which means that $J_nC_nJ_n = C_n$. As it was shown in [24], these matrices have some remarkable representation. To present it we introduce some notation. Let $\mathbb{F}^n_+$, $\mathbb{F}^n_-$ be the subspaces of $\mathbb{F}^n$ consisting of all symmetric or skewsymmetric vectors, respectively. Then $P_n^{\pm} = \frac{1}{2}(I_n \pm J_n)$ are the projections onto $\mathbb{F}^n_{\pm}$, respectively. Since C_n is assumed to be centrosymmetric, the subspaces $\mathbb{F}^n_{\pm}$ are invariant under C_n.

A centrosymmetric T+H matrix C_n can be represented in the form

$$C_n = T_n^+P_n^+ + T_n^-P_n^-,$$

where $T_n^{\pm}$ are the symmetric Toeplitz matrices $T_n^{\pm} = [\, c^{\pm}_{|i-j|} \,]_{i,j=1}^n$, $c_i^{\pm} = a_i \pm s_{i+n}$. Conversely, each matrix of this form is a centrosymmetric T+H matrix. For details we refer to [24]. Thus a linear system $C_n\mathbf{f} = \mathbf{b}$ is equivalent to the two symmetric Toeplitz systems

$$T_n^{\pm}\mathbf{f}_{\pm} = P_{\pm}\mathbf{b},$$

where $\mathbf{f} = \mathbf{f}_+ + \mathbf{f}_-$. Furthermore, in case that the matrices $T_n^{\pm}$ are nonsingular, the Toeplitz inversion formula provides an inversion formula for C_n. However, the matrices $T_n^{\pm}$ in the representation of C_n might be singular (see [24]), so that some specific considerations are necessary.

It turned out that the inverse of C_n can be represented in terms of another kind of Bezoutians, which will be called *T+H Bezoutians* and defined next. Let $\mathbf{u}, \mathbf{v} \in \mathbb{F}^{n+2}$ be both either symmetric or skewsymmetric vectors. The T+H Bezoutian of $\mathbf{u}$ and $\mathbf{v}$ is, by definition, the $n \times n$ matrix $\widetilde{B} = \widetilde{B}(\mathbf{u}, \mathbf{v})$ with the generating function

$$\widetilde{B}(t,s) = \frac{\mathbf{u}(t)\mathbf{v}(s) - \mathbf{v}(t)\mathbf{u}(s)}{(t-s)(1-ts)}\,.$$

Besides C_n we consider a nonsingular $(n+2)\times(n+2)$ extension C_{n+2} obtained from C_n by extending the Toeplitz matrices $T_n^{\pm}$ to symmetric $(n+2)\times(n+2)$ Toeplitz matrices. The following theorem is presented in [25] as a modification of a result in [24]. The superscript + at a vector indicates that the vector is symmetric and − that the vector is skewsymmetric.

Theorem 2.6 The equations

$$\begin{aligned} T_n^+\mathbf{x}_n^+ &= P_n^+\mathbf{e}_n\,, \quad T_{n+2}^+\mathbf{x}_{n+2}^+ = P_{n+2}^+\mathbf{e}_{n+2}\,, \\ T_n^-\mathbf{x}_n^- &= P_n^-\mathbf{e}_n\,, \quad T_{n+2}^-\mathbf{x}_{n+2}^- = P_{n+2}^-\mathbf{e}_{n+2} \end{aligned} \tag{2}$$

have unique symmetric or skewsymmetric solutions $\mathbf{x}_n^{\pm}$ and $\mathbf{x}_{n+2}^{\pm}$, respectively, and

$$C_n^{-1} = \frac{1}{r_+}\,\widetilde{B}(\mathbf{x}_{n+2}^+, \widetilde{\mathbf{x}}_n^+) + \frac{1}{r_-}\,\widetilde{B}(\mathbf{x}_{n+2}^-, \widetilde{\mathbf{x}}_n^-),$$

where $r_\pm$ is the last component of $\mathbf{x}_{n+2}^{\pm}$, and $\widetilde{\mathbf{x}}_n^{\pm} \in \mathbb{F}^{n+2}$ is the vector obtained from $\mathbf{x}_n^{\pm} \in \mathbb{F}_{\pm}^{n}$ by adding a zero at the top and the bottom.

3. SPLIT LEVINSON ALGORITHMS

We show in this section how the data in the inversion formulas can be computed in an efficient way.

3.1. Symmetric Toeplitz and Centrosymmetric T+H Matrices

To begin with we introduce some notations. For a vector $\mathbf{u} = (u_i)_{i=1}^l$, let $M_k(\mathbf{u})$

denote the $(k+l-1)\times k$ matrix

$$M_k(\mathbf{u}) = \left.\begin{bmatrix} u_1 & & 0 \\ \vdots & \ddots & \\ u_l & & u_1 \\ & \ddots & \vdots \\ 0 & & u_l \end{bmatrix}\right\} k+l-1\,.$$

It is easily checked that, for $\mathbf{x}\in\mathbb{C}^k$, $(M_k(\mathbf{u})\mathbf{x})(t) = \mathbf{u}(t)\mathbf{x}(t)$, i.e. $M_k(\mathbf{u})$ is the matrix of the operator of multiplication by $\mathbf{u}(t)$. For a given symmetric Toeplitz matrix $T_n = [\,a_{i-j}\,]_{i,j=1}^n$, we denote by T_k the kth leading principal submatrix of T_n. Note that T_k is also a central submatrix of T_n if $n-k$ is even. We denote by T_k^+ the restriction of T_k (as linear operator) to the subspace of symmetric vectors. Let $n_1 < \ldots < n_\ell = n$ be the integers $j\in\{\,1,2,\ldots,n\,\}$ for which T_j^+ is invertible and $n-j$ is even, $n_{\ell+1} = n+2$. We set $d_k = \frac{1}{2}(n_k - n_{k-1})$. Let $\mathbf{x}^{(k)}\in\mathbb{F}_+^{\,n_k}$ be the solution of $T_{n_k}^+\mathbf{x}^{(k)} = \mathbf{e}_1^+$. Then it can be shown (see [28]) that $\mathbf{x}^{(k)}$ has the form

$$\mathbf{x}^{(k)} = \frac{1}{\rho_k}\begin{bmatrix} \mathbf{0}_{d_k-1} \\ \mathbf{u}^{(k)} \\ \mathbf{0}_{d_k-1} \end{bmatrix} \tag{3}$$

for some $\rho_k\in\mathbb{F}$ and monic $\mathbf{u}^{(k)}\in\mathbb{F}_+^{\,n_{k-1}+2}$. We define the *residuals* $r_i^{(k)}$ by

$$r_i^{(k)} = [\,a_{i+n_{k-1}}\ \ldots\ a_{i-1}\,]\,\mathbf{u}^{(k)} \tag{4}$$

for $i = 1,\ldots,n-n_{k-1}+1$ and consider the $(d_k+1)\times(d_k+1)$ triangular Toeplitz matrices

$$R^{(k)} = \begin{bmatrix} r_{d_k}^{(k)} & & 0 \\ \vdots & \ddots & \\ r_{2d_k}^{(k)} & \cdots & r_{d_k}^{(k)} \end{bmatrix}\,. \tag{5}$$

The split Levinson algorithm computes the monic vectors $\mathbf{u}^{(k+1)}$ and the integers d_k from $\mathbf{u}^{(k)}$ and $\mathbf{u}^{(k-1)}$. This will give us also $\mathbf{x}^{(k)}$ via (3), where $\rho_k = r_{d_k}^{(k)}$. We start with

$$\mathbf{u}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ n_0 = 0 \quad\text{or}\quad \mathbf{u}^{(1)} = [\,1\,],\ n_0 = -1\,,$$

depending on whether n is even or odd. We trace the residuals $r_i^{(1)}$ for $i = 1,2,\ldots$ until we get a nonzero number. Let $r_d^{(1)}\neq 0$ and $r_i^{(1)} = 0$ for $1\le i < d$. Then $d_1 = d$ and $\mathbf{x}^{(1)}$ is given by (3) for $k=1$, where $\rho_1 = r_d^{(1)}$. Furthermore, $n_1 = 2d$ if N is even and $n_1 = 2d-1$ if N is odd. We show how to find $\mathbf{u}^{(2)}$. For this we

form $R^{(1)}$ and solve the system $R^{(1)}\mathbf{c} = \mathbf{e}_1 - \mathbf{e}_2$ if N is even or $R^{(1)}\mathbf{c} = \mathbf{e}_1 - \mathbf{e}_3$ if N is odd and $d > 1$. If N is odd and $d = 1$, then $\mathbf{c}$ is a solution of $[\,2a_1\ a_0\,]\,\mathbf{c} = 0$. Now we form the symmetric vector $\mathbf{p}^{(1)} = \frac{1}{\gamma}\begin{bmatrix}\mathbf{c}\\ *\end{bmatrix} \in \mathbb{F}^{\,2d+1}$, where γ is the first component of $\mathbf{c}$ and the asterisk indicates the symmetric extension. Now $\mathbf{u}^{(2)}$ is given by $\mathbf{u}^{(2)} = M_{2d+1}(\mathbf{u}^{(1)})\mathbf{p}^{(1)}$.

Now we assume that $\mathbf{u}^{(k)}$ and $\mathbf{u}^{(k-1)}$ are given. We compute $r_i^{(k)}$ for $i = 1, 2, \ldots$ by (4) until we get a nonzero number. Suppose that $r_d^{(k)} \neq 0$ and $r_i^{(k)} = 0$ for $1 \leq i < d$. Then $d_k = d$ and $\mathbf{x}^{(k)}$ is given by (3), where $\rho_k = r_d^{(k)}$. Next we compute $r_i^{(k)}$ for $i = d_k + 1, \ldots, d_{2d_k}$ and, in case that $d_k > d_{k-1}$, also $r_i^{(k-1)}$ for $i = 2d_{k-1} + 1, \ldots, d_k + d_{k-1}$. We form the matrix $R^{(k)}$ according to (5), the vector $\mathbf{s}^{(k)} = [\,r_{d_{k-1}+i-1}^{(k-1)}\,]_{i=1}^{d_k+1}$, and solve the triangular Toeplitz system $R^{(k)}\mathbf{c} = \mathbf{s}^{(k)}$. From its solution $\mathbf{c}$ we form the symmetric vector

$$\mathbf{p}^{(k)} = \frac{\rho_k}{\rho_{k-1}}\begin{bmatrix}\mathbf{c}\\ *\end{bmatrix} \in \mathbb{F}_+^{\,2d_k+1}\,,$$

where the asterisk denotes the symmetric extension. Then $\mathbf{p}^{(k)}$ is monic and

$$\mathbf{u}^{(k+1)} = M_{2d_k+1}(\mathbf{u}^{(k)})\,\mathbf{p}^{(k)} - \begin{bmatrix}\mathbf{0}_{d_{k-1}+d_k}\\ \mathbf{u}^{(k-1)}\\ \mathbf{0}_{d_{k-1}+d_k}\end{bmatrix} \in \mathbb{F}_+^{\,n_k+2}. \tag{6}$$

Relation (6) describes one step of the recursive algorithm to find $\mathbf{u}^{(\ell)}$ and $\mathbf{u}^{(\ell+1)}$. The solutions $\mathbf{x}^{(\ell)} = \mathbf{x}_n$ and $\mathbf{x}^{(\ell+1)} = \mathbf{x}_{n+2}$ that are involved in the inversion formula in Subsection 2.2 are given now by (3). In the special case $d_k = 1$ for all k this algorithm is closely related to the algorithms described in [35] and [13]. Relation (6) can also be described in polynomial language as follows.

Theorem 3.1 The polynomials $\mathbf{u}^{(k)}(t)$ satisfy the recursion

$$\mathbf{u}^{(k+1)}(t) = \mathbf{p}^{(k)}(t)\,\mathbf{u}^{(k)}(t) - q^{(k)}\,t^{d_{k-1}+d_k}\,\mathbf{u}^{(k-1)}(t),$$

the polynomials $\mathbf{x}^{(k)}(t)$ the recursion

$$\mathbf{x}^{(k+1)}(t) = t^{d_{k+1}-d_k}\mathbf{p}^{(k)}(t)\,\mathbf{x}^{(k)}(t) - q^{(k)}\,t^{d_{k+1}+d_k}\,\mathbf{x}^{(k-1)}(t),$$

where $k = 1, \ldots, \ell$ and $q^{(k)} = \dfrac{\rho_k}{\rho_{k-1}}$.

An analogue recursion holds for the solution of the equations $T_j\mathbf{x}_j^- = \mathbf{e}_1^-$, where j runs over all numbers for which $n - j$ is even and the restriction of T_j to the

subspace of skewsymmetric polynomials is invertible. The split Levinson algorithm for the centrosymmetric T+H matrix $C_n = T_n^+ P_n^+ + T_n^- P_n^-$ is now a straightforward generalization, since T_n^+ and T_n^- are both symmetric Toeplitz matrices. To compute the solutions of (2) one has to run the split Levinson algorithm described above for T_n^+ and its skewsymmetric counterpart for T_n^-.

3.2. Skewsymmetric Toeplitz Matrices

In the style of the previous subsection let $n_1 < \ \ldots \ < n_\ell = n$ be the integers $j \in \{\, 1, 2, \ldots, n \,\}$ for which T_j is nonsingular, $d_k = \frac{1}{2}\,(n_k - n_{k-1})$. But now $\mathbf{u}^{(k)}$ is the vector spanning the kernel of T_{n_k+1} with last component equal to 1 and $\mathbf{x}^{(k)}$ is a solution of

$$T_{n_k+1}\mathbf{x}^{(k)} = \mathbf{e}_{n_k+1} - \mathbf{e}_1.$$

The residuals $r_j^{(k)}$ and $s_j^{(k)}$ of $\mathbf{u}^{(k)}$ and $\mathbf{x}^{(k)}$ are defined by

$$r_j^{(k)} = [\, a_{j+n_k} \ \cdots \ a_j \,]\, \mathbf{u}^{(k)}, \quad s_j^{(k)} = [\, a_{j+n_k} \ \cdots \ a_j \,]\, \mathbf{x}^{(k)}\,, \tag{7}$$

respectively, for $j = 0, \ldots, n - n_k$. Clearly, $r_0^{(k)} = 0$ and $s_0^{(k)} = 1$.

Our aim is to find $\mathbf{u} = \mathbf{u}^{(\ell)}$ and $\mathbf{x} = \mathbf{x}^{(\ell)}$. As it was shown in [27], $\mathbf{x}^{(k)}$ is of the form

$$\mathbf{x}^{(k)} = \frac{1}{r_{d_k}^{(k-1)}} \begin{bmatrix} \mathbf{0}_{d_k} \\ \mathbf{u}^{(k-1)} \\ \mathbf{0}_{d_k} \end{bmatrix}$$

and

$$s_j^{(k)} = \frac{1}{r_{d_k}^{(k-1)}}\ r_{j+d_k}^{(k-1)}\ .$$

That means it is sufficient to compute the residuals $r_j^{(k)}$ and to construct the vectors $\mathbf{u}^{(k)}$.

For initialization we set $n_0 = 0$ and $\mathbf{u}^{(0)} = 1$. Then $r_j^{(0)} = a_j$. If $a_1 = \cdots = a_{d-1} = 0$ and $a_d \neq 0$, then $n_1 = 2d$. The vector $\mathbf{u}^{(1)}$ is the normalized solution of the homogeneous system $T_{2d+1}\mathbf{v} = 0$. Let us show how this solution can be found. We form the matrix

$$\widetilde{R}^{(0)} = \begin{bmatrix} a_d & \cdots & a_{2d} \\ & \ddots & \vdots \\ 0 & & a_d \end{bmatrix}.$$

Let $\mathbf{c}$ be the solution of the triangular Toeplitz system $(\widetilde{R}^{(0)})^T\mathbf{c} = \mathbf{e}_1$ and $\mathbf{v} =$

$\begin{bmatrix} \mathbf{c} \\ \mathbf{c}' \end{bmatrix} \in \mathbb{F}_+^{2d+1}$ its symmetric extension. Then $T_{2d+1}\mathbf{v} = \mathbf{0}$. Hence $\mathbf{u}^{(1)} = \frac{1}{c}\,\mathbf{v}$, where c is the first component of $\mathbf{c}$.

We assume now that n_{k-1}, n_k, $\mathbf{u}^{(k-1)}$ and $\mathbf{u}^{(k)}$ are given. We also need some of the values $r_j^{(k-1)}$ $(j = 1, \ldots, 2d_k)$ that are computed in the previous step. Now n_{k+1} and $\mathbf{u}^{(k+1)}$ are computed as follows. If $r_1^{(k)} = \cdots = r_{d-1}^{(k)} = 0$ and $r_d^{(k)} \neq 0$, then $d_{k+1} = d$, i.e. $n_{k+1} = n_k + 2d$. We compute the numbers $r_{d_{k+1}+1}^{(k)}, \ldots, r_{2d_{k+1}}^{(k)}$ and form the matrix $\widetilde{R}^{(k)}$ as

$$\widetilde{R}^{(k)} = \begin{bmatrix} r_{d_{k+1}}^{(k)} & \cdots & r_{2d_{k+1}}^{(k)} \\ & \ddots & \vdots \\ 0 & & r_{d_{k+1}}^{(k)} \end{bmatrix}.$$

If $d_{k+1} > d_k$, then it will be necessary to compute also the numbers $r_j^{(k-1)}$ for $j = 2d_k + 1, \ldots, d_k + d_{k+1}$ to form the vector $\mathbf{r}'^{(k-1)} = (r_j^{(k-1)})_{j=d_k}^{d_k+d_{k+1}}$.

Let $\mathbf{c}^{(k)}$ be the solution of the triangular Toeplitz system $(\widetilde{R}^{(k)})^T\mathbf{c}^{(k)} = \mathbf{r}'^{(k-1)}$, $q^{(k)} = \frac{1}{c}$, where c is the first component of $\mathbf{c}^{(k)}$, and $\mathbf{p}^{(k)} = q^{(k)} \begin{bmatrix} \mathbf{c}^{(k)} \\ \mathbf{c}'^{(k)} \end{bmatrix} \in \mathbb{F}_+^{2d_{k+1}+1}$ be the symmetric extension of $q^{(k)}\mathbf{c}^{(k)}$. Then

$$\mathbf{u}^{(k+1)} = M_{2d_{k+1}+1}(\mathbf{u}^{(k)})\mathbf{p}^{(k)} - q^{(k)} \begin{bmatrix} \mathbf{0}_{d_k+d_{k+1}} \\ \mathbf{u}^{(k-1)} \\ \mathbf{0}_{d_k+d_{k+1}} \end{bmatrix}.$$

In polynomial language the recursion can be written as follows.

Theorem 3.2 The polynomials $\mathbf{u}^{(k)}(t)$ satisfy the three-term recursion

$$\mathbf{u}^{(k+1)}(t) = \mathbf{p}^{(k)}(t)\mathbf{u}^{(k)}(t) - t^{d_k+d_{k+1}}\, q^{(k)}\mathbf{u}^{(k-1)}(t).$$

Example. Consider the skewsymmetric Toeplitz matrix $T_6 = [\, a_{i-j} \,]_{i,j=1}^6$, with $(\, a_k \,)_{k=1}^5 = (1,\ 2,\ 3,\ 5,\ 6)$. Since we need also an extension of T_6 we set $a_6 = 0$. The standard setting for initialization is $n_0 = 0, \mathbf{u}^{(0)} = 1$ and $r_j^{(0)} = a_j$. Since $r_1^{(0)} = 1 \neq 0$ we have $d_1 = 1$ and $n_1 = 2$. We obtain $\mathbf{x}^{(1)} = [\, 0\ 1\ 0 \,]^T$ and $\mathbf{u}^{(1)} = [\, 1,\ -2,\ 1 \,]^T$. With $\mathbf{u}^{(0)}$ and $\mathbf{u}^{(1)}$ we can start the recursion.

We compute the residuals as $r_1^{(1)} = 0$, $r_2^{(1)} = 1$. Thus $d_2 = 2$, $n_2 = n_1 + 2d_1 = 6$, and $\mathbf{x}^{(2)} = [\, 0,\ 0,\ 1,\ -2,\ 1,\ 0,\ 0 \,]^T$. In order to form the matrix $\widetilde{R}^{(1)}$ we find that $r_3^{(1)} = -1$ and $r_4^{(1)} = -7$, and in order to form the vector $\mathbf{r}'^{(0)}$ we observe that

$r_2^{(0)} = a_2 = 2$, $r_3^{(0)} = a_3 = 3$. The solution of the system $(\widetilde{R}^{(1)})^T \mathbf{c}^{(1)} = \mathbf{r}'^{(0)}$ is $\mathbf{c}^{(1)} = [\,1,\,3,\,13\,]^T$. Hence $\mathbf{p}^{(1)} = [\,1,\,3,\,13,\,3,\,1\,]^T$, which gives

$$\mathbf{u}^{(2)} = [\,1,\,1,\,8,\,-21,\,8,\,1,\,1\,]^T.$$

The inverse of T_6 is now given by Theorem 2.4 with $\mathbf{x} = \mathbf{x}^{(2)}$ and $\mathbf{u} = \mathbf{u}^{(2)}$. A check shows that this really gives the inverse matrix.

3.3. Hermitian Toeplitz Matrices

In this section we have $\mathbb{F} = \mathbb{C}$. The algorithms described in the previous sections cannot be generalized to hermitian Toeplitz matrices $T_n = [\,a_{i-j}\,]_{i,j=1}^n$, $a_{-i} = \overline{a}_i$. The reason for this is that T_n is not the direct sum of T_n^+ and T_n^-. Nevertheless an algorithm exists that generalize, in principle, the algorithms in [33] and [34]. In contrast to the previous cases, where the step size was even, the step size in this algorithm will be always odd.

Besides T_n we consider its leading principal submatrices T_j and a nonsingular hermitian Toeplitz extension T_{n+1}. We extend the definition of $\mathcal{C}_j$ given in Section 2.4 for $j = n, n+1$ to all j and denote by κ_j the dimension (over the reals) of $\mathcal{C}_j$. Let $n_1 = 1$ and $n_2 < \cdots < n_\ell = n+1$ all integers $j > 1$ for which $\kappa_j = \kappa_{j-1} = 1$. Let $\mathbf{w}^{(k)}$ $(k = 1, \ldots, \ell)$ be a vector spanning the subspace $\mathcal{C}_{n_k}$. Each of these vectors is unique up to a real factor. The following observation is important for the design of our algorithm (see [30]).

Proposition 3.3 The difference $n_{k+1} - n_k$ is always odd.

We set $d_k = \frac{1}{2}(n_{k+1} - n_k + 1)$, i.e. $n_{k+1} = n_k + 2d_k - 1$. Suppose that $\mathbf{w}^{(k-1)} \in \mathcal{C}_{n_{k-1}}$ and $\mathbf{w}^{(k)} \in \mathcal{C}_{n_k}$ are given. We show how to find n_{k+1} and $\mathbf{w}^{(k+1)}$. Let the "residuals" r_{jk} be defined by

$$r_{jk} = [\,a_{n_k+j-1}\ \ldots\ a_{j-1}\,]\,\mathbf{w}^{(k)} \quad (j = 1, 2 \ldots)\,.$$

The recursion step starts with computing the residuals r_{jk} for $j = 1, 2, \ldots$ until a nonzero one appears. Assume that $r_{1k} = \cdots = r_{d-1,k} = 0$ and $r_{dk} \neq 0$. We will see that $d = d_k$. Then we compute the residuals $r_{d+1,k}, \ldots, r_{2d-1,k}$ and form the lower triangular $d \times d$ Toeplitz matrix

$$R_k = \begin{bmatrix} r_{dk} & & \\ \vdots & \ddots & \\ r_{2d-1,k} & \cdots & r_{dk} \end{bmatrix}. \tag{8}$$

If $d > d_{k-1}$, then we also compute the residuals $r_{j,k-1}$ for $j = 2d_{k-1}, \ldots, d_{k-1} + d - 1$ in order form the vector $\mathbf{r}_{k-1} = (r_{d_{k-1}+j-1,k-1})_{j=1}^{d}$. The other components of this vector were computed in the previous step.

We have

$$\partial T_{n_k+2d-1} M_{2d}(\mathbf{w}^{(k)}) = \begin{bmatrix} O & R_k^* \\ O & O \\ R_k & O \end{bmatrix} ,$$

where $R_k^* = \overline{R}_k^T$. From this representation we can see that if $d = 1$, i.e. $r_{1k} \neq 0$, then $\kappa_{n_k+1} = 0$. Hence $d_k = d = 1$. Furthermore, we see that if $d > 1$, then $\kappa_{n_k+j} > 0$ for $j = 1, \ldots, 2d-3$ but $\kappa_{n_k+2d-2} = 0$. The vector $\begin{bmatrix} \mathbf{0} \\ \mathbf{w}_k \\ \mathbf{0} \end{bmatrix}$ spans $\mathcal{C}_{n_k+2d-2}$. According to Proposition 3.3 we have also $\kappa_{n_k+2d-1} = 0$. Hence $n_{k+1} = n_k + 2d - 1$, i.e. $d = d_{k+1}$.

Hereafter, for a vector $\mathbf{u} \in \mathbb{C}^d$ we denote by $\mathbf{u}^{\#}$ the vector $\mathbf{u}^{\#} = J_d \overline{\mathbf{u}}$.

Let $\mathbf{c}_k$ be the solution of the triangular Toeplitz system

$$R_k \mathbf{c}_k = \mathbf{r}_{k-1} ,$$

and $\mathbf{p}^{(k)} = \begin{bmatrix} \mathbf{c}_k \\ \mathbf{c}_k^{\#} \end{bmatrix}$. Then we have

$$\partial T_{n_k+2d-1} \left(M_{2d}(\mathbf{w}_k)\mathbf{p}^{(k)} - \begin{bmatrix} \mathbf{0} \\ \mathbf{w}_{k-1} \\ \mathbf{0} \end{bmatrix} \right) = \begin{bmatrix} R_k^* \mathbf{c}_k^{\#} - \mathbf{r}_{k-1}^{\#} \\ R_k \mathbf{c}_k - \mathbf{r}_{k-1} \end{bmatrix} = \mathbf{0} .$$

That means that the vector

$$\mathbf{w} = M_{2d}(\mathbf{w}_k)\mathbf{p}^{(k)} - \begin{bmatrix} \mathbf{0} \\ \mathbf{w}_{k-1} \\ \mathbf{0} \end{bmatrix}$$

belongs to the kernel of ∂T_{n_k+2d-1} and satisfies $\mathbf{w}(1) = 0$. Hence $\mathbf{w}_{k+1} = \mathbf{w}$. In polynomial language this can be expressed as follows.

Theorem 3.4 For $k = 2, \ldots, \ell - 1$, the polynomials $\mathbf{w}_k(t)$ satisfy the three-term recursion

$$\mathbf{w}^{(k+1)}(t) = \mathbf{p}^{(k)}(t)\mathbf{w}^{(k)}(t) - t^{d_k + d_{k-1} - 1}\mathbf{w}^{(k-1)}(t) .$$

To complete the algorithm we have to find $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$. We set $n_1 = 1$ and $\mathbf{w}^{(1)}(t) = \mathrm{i}(t-1)$. The integer n_2 and the vector $\mathbf{w}^{(2)}$ are found in the following

way. As in the case $k > 1$ we compute r_{j1} until the result is nonzero. Suppose that $r_{11} = \cdots = r_{d-1,1} = 0$, and $r_{d1} \neq 0$. Then we compute $r_{d1}, \ldots, r_{2d-1,1}$ and observe that

$$\partial T_{2d} M_{2d}(\mathbf{w}^{(1)}) = \begin{bmatrix} 0 & & 0 & \overline{r}_{d1} & \ldots & \overline{r}_{2d-1,1} \\ & & & & \ddots & \vdots \\ r_{d1} & & & & & \overline{r}_{d1} \\ \vdots & \ddots & & & & \\ r_{2d-1,1} & \ldots & r_{d1} & 0 & & 0 \end{bmatrix}. \tag{9}$$

Note that the middle row of the matrix on the right-hand side has, in contrast to the case $k > 1$, nonzero elements at the left and right ends, which makes the construction different to the case $k > 1$. But as in the case $k > 1$ we can conclude from (9) that $d_1 = d$. We form the $d \times d$ lower triangular Toeplitz matrix R_1 by (8) for $k = 1$, find the solution of $R_1\mathbf{c} = \mathrm{i}\mathbf{e}_1$ and set $\mathbf{p}^{(1)} = \begin{bmatrix} \mathbf{c} \\ \mathbf{c}^{\#} \end{bmatrix}$. Then we have

$$\partial T_{2d} M_{2d}(\mathbf{w}^{(1)})\mathbf{p}^{(1)} = \begin{bmatrix} \mathbf{0} \\ R_1\mathbf{c} \end{bmatrix} + \begin{bmatrix} R_1^*\mathbf{c}^{\#} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathrm{i} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathrm{i} \\ \mathbf{0} \end{bmatrix} = \mathbf{0}\,.$$

Hence $d_1 = d$ and $\mathbf{w}^{(2)} = M_{2d}(\mathbf{w}^{(1)})\mathbf{p}^{(1)}$ or, in polynomial language, $\mathbf{w}^{(2)}(t) = \mathbf{p}^{(1)}(t)\mathbf{w}^{(1)}(t)$.

The algorithm described above computes vectors $\mathbf{w}^{(\ell)} \in \mathbb{C}^{n+2}$ and $\mathbf{w}^{(\ell-1)} \in \mathbb{C}^{n_{\ell-1}+1}$. We show how from these vectors the data in the inversion formula (1) can be computed. For the inversion formula we need a nontrivial vector in $\mathcal{C}_n$. If $n_{\ell-1} = n$, then $\mathbf{w}^{(\ell-1)}$ is the desired vector. If $n_{\ell-1} < n$, then we consider the vector $\begin{bmatrix} \mathbf{0} \\ \mathbf{w}^{(\ell-1)} \\ \mathbf{0} \end{bmatrix}$, where the zero vectors have length $d_{\ell-1} - 1$. The corresponding polynomials have to be divided by $\mathrm{i}(t-1)$ in order to get real multiples of $\mathbf{q}_n$ and $\mathbf{q}_{n+1}$. It remains to divide the two vectors by real numbers that are obtained by applying a test functional.

Example 1. Let $T_4 = [\,a_{i-j}\,]_{i,j=1}^4$ with $(a_i)_{i=0}^3 = (\,1,\,1,\,\mathrm{i},\,0)$. Then $r_{11} = 0$ and $r_{21} = 1 + \mathrm{i}$. Hence $d_1 = 2$ and $n_2 = 4$. We find that

$$R_1 = \begin{bmatrix} 1+\mathrm{i} & 0 \\ -1 & 1+\mathrm{i} \end{bmatrix} \quad \text{and} \quad \mathbf{c}_1 = \frac{1}{2}\begin{bmatrix} 1+\mathrm{i} \\ 1 \end{bmatrix}.$$

Applying the recursion formula we obtain

$$\mathbf{w}^{(2)} = \frac{1}{2}[\,1-\mathrm{i},\ -1,\ 0,\ -1,\ 1+\mathrm{i}\,]^T\,.$$

Example 2. Let $T_4 = [\, a_{i-j} \,]_{i,j=1}^{4}$ with $(a_i)_{i=0}^{3} = (\, 1, \, -1, \, \mathrm{i}, \, 0)$. The matrix T_2 is singular but, nevertheless, $d_1 = 1$, since $r_{11} = 2\mathrm{i} \neq 0$. Clearly, $\mathbf{w}_2 = \frac{1}{2}[\, -\mathrm{i}, \;\; 0, \;\; \mathrm{i}\,]^T$. Applying the recursion formula we obtain $d_2 = 1$,

$$\mathbf{w}^{(3)} = [\, 1 - \mathrm{i}, \;\; -1, \;\; -1, \;\; 1 + \mathrm{i}\,]^T$$

and $d_3 = 1$,

$$\mathbf{w}^{(4)} = \frac{1}{4}\,[\, 2, \;\; -1 - \mathrm{i}, \;\; -2, \;\; -1 + \mathrm{i}, \;\; 2\,]^T .$$

4. SPLIT SCHUR ALGORITHMS

The split Levinson algorithms include inner product calculations. To avoid these one can precompute the residuals using a Schur-type algorithm. The Schur-type algorithms follow immediately from the Levinson recursions. As we show in the next section, the Schur-type algorithm produces a factorization of the matrix. This factorization can be used to solve a system of equation without using an inversion formula.

4.1. Symmetric Toeplitz and Centrosymmetric T+H Matrices

We consider the residuals $r_i^{(k)}$ defined by (4) for $i = 1, \ldots, n - n_{k-1} + 1$. From Theorem 3.1 we obtain immediately the recursion

$$r_i^{(k+1)} = \sum_{j=1}^{2d_k+1} p_j^{(k)} r_{i+d_{k+1}-d_k+j-1}^{(k)} - q^{(k)} r_{i+d_{k+1}+d_k}^{(k-1)},$$

where $p_j^{(k)}$ are the coefficients of $\mathbf{p}^{(k)}(t)$. Introducing the polynomials $\mathbf{r}^{(k)}(t) = \sum_{i=1}^{n-n_{k-1}+1} r_i^{(k)} t^{i-1}$ this can be written in the following polynomial form.

Theorem 4.1 The polynomials $\mathbf{r}^{(k)}(t)$ satisfy the recursion

$$\mathbf{r}^{(k+1)}(t) = P_{n-n_k}(\mathbf{p}^{(k)}(t^{-1})\mathbf{r}^{(k)}(t)t^{-d_{k+1}+d_k} - q^{(k)}\,\mathbf{r}^{(k-1)}(t)t^{-d_k-d_{k+1}}).$$

Here and in all what follows we denote by P_j the projection that cuts off all powers t^i for $i \geq j$ and $i < 0$. Note that according to the construction in the recursion of Theorem 4.1 no negative powers of t appear.

The theorem above provides a split Schur algorithm for computing the residuals. It can replace the inner product calculations in the Levinson algorithm of Theorem

3.1 for the computation of the vectors $\mathbf{x}_n$ and $\mathbf{x}_{n+2}$, which are involved in the inversion formula. This gives a slightly higher complexity in sequential computing but might be convenient in parallel processing. As we will see in Section 5 Theorem 4.1 also provides, in principle, a factorization of the symmetric part T_n^+ of T_n, which can be used to solve linear systems via factorization and back substitution, which is an advisable method if T_n is ill-conditioned. The initializations for these algorithms are obtained from the initializations of the corresponding vectors in the split Levinson algorithms presented in Section 3.1. We can use the recursion of Theorem 4.1 and a similar recursion for the residuals of skewsymmetric solutions to obtain a split Schur algorithm for centrosymmetric T+H matrices.

4.2. Skewsymmetric Toeplitz Matrices

We consider again the full residual vectors $\mathbf{r}^{(k)} = (r_j^{(k)})_{j=1}^{n-n_k}$ and the corresponding polynomials $\mathbf{r}^{(k)}(t)$, where $r_j^{(k)}$ are defined in (7). By the definition of the integer d_{k+1}, $\widetilde{\mathbf{r}}^{(k)}(t) = t^{-d_{k+1}+1}\,\mathbf{r}^{(k)}(t)$ is a polynomial. The monic, symmetric polynomial $\mathbf{p}^{(k)}(t)$ and $q^{(k)} \in \mathbb{F}$ have been constructed in such way that the polynomial

$$\widetilde{\mathbf{r}}^{(k)}(t)\mathbf{p}^{(k)}(t) - q^{(k)}\widetilde{\mathbf{r}}^{(k-1)}(t)$$

has a zero of order $d_{k+1}+1$ at $t=0$. According to Theorem 3.2 the remainder will give us $\mathbf{r}^{(k+1)}(t)$ and the following recursion formula for the residuals is immediately deduced.

Theorem 4.2 The polynomials $\mathbf{r}^{(k)}(t)$ satisfy the recursion

$$\mathbf{r}^{(k+1)}(t) = P_{n-n_{k+1}}\left(t^{-2d_{k+1}}\,\mathbf{p}^{(k)}(t)\mathbf{r}^{(k)}(t) - t^{-d_{k+1}-d_k}\,q^{(k)}\,\mathbf{r}^{(k-1)}(t)\right).$$

To write this recursion in matrix form we introduce the matrix $Q^{(k)}$ by

$$Q^{(k)} = \left[\, r^{(k)}_{2d_{k+1}+i-j+1} \,\right]_{i=1\;\;j=1}^{\mu_k\;\;2d_{k+1}+1},$$

where $\mu_k = n - n_{k+1} = n - n_k - 2d_{k+1}$. Now we have

$$\mathbf{r}^{(k+1)} = Q^{(k)}\,\mathbf{p}^{(k)} - q^{(k)}\check{\mathbf{r}}^{(k-1)},$$

where $\check{\mathbf{r}}^{(k-1)} = [\, r^{(k-1)}_{d_k+d_{k+1}+i} \,]_{i=1}^{\mu_k}$. The recursion starts with $\check{\mathbf{r}}^{(-1)} = 0$, $\mathbf{r}^{(0)} = [\, a_j \,]_{j=1}^{n}$, $\mathbf{p}^{(0)} = \mathbf{u}^{(1)}$, and

$$Q^{(0)} = [a_{n_1+i-j+1}]_{i=1\;\;j=1}^{n-n_1\;\;n_1+1}.$$

The vector $\mathbf{u}^{(1)}$ will be computed as described in the initialization of the Levinson recursion of Subsection 3.2 Theorem 3.2 can be combined with Theorem 4.2 to compute the parameters in the inversion formula of Theorem 2.4.

4.3. Hermitian Toeplitz Matrices

Now we present the Schur-type counterpart of the split Levinson recursion in Theorem 3.4. We introduce the $(2n+2-k)\times k$ Toeplitz matrix

$$\widehat{T}_k = \begin{bmatrix} \overline{a}_{n+1-k} & \dots & \overline{a}_n \\ \vdots & & \vdots \\ \overline{a}_1 & \dots & \overline{a}_k \\ & T_k & \\ a_k & \dots & a_1 \\ \vdots & & \vdots \\ a_n & \dots & a_{n+1-k} \end{bmatrix} \tag{10}$$

and observe that

$$\widehat{T}_{n_k+1}\mathbf{w}_k = \begin{bmatrix} \mathbf{r}_k^{\#} \\ \mathbf{0} \\ \mathbf{r}_k \end{bmatrix}$$

for vectors $\mathbf{r}_k \in \mathbb{C}^{n-n_k+1}$, $(k=1,\dots,\ell-1)$. In order to transform the Levinson recursion in Theorem 3.4 into Schur recursions for the residual vectors $\mathbf{r}_k$ we apply the following lemma.

Lemma 4.3 Let $\widehat{T}_m\mathbf{w} = \mathbf{b} \in \mathbb{C}^{2n+2-m}$ and $\widehat{T}_{m+r}M_{r+1}(\mathbf{w})\mathbf{c} = \widetilde{\mathbf{b}} \in \mathbb{C}^{2n+2-m-r}$, $\mathbf{c}\in\mathbb{C}^{r+1}$. Then

$$\widetilde{\mathbf{b}}(t) = P_{2n+2-m-r}t^{-r}\mathbf{c}(t)\mathbf{b}(t)\,.$$

Proof. Let $\mathbf{w}^{(j)} \in \mathbb{C}^{m+r}$ $(j=0,\dots,r)$ be defined by $\mathbf{w}^{(j)}(t) = t^j\mathbf{w}(t)$ and $\mathbf{b}^{(j)} = \widehat{T}_{m+r}\mathbf{w}^{(j)}$. Then it is immediately checked that $\mathbf{b}^{(j)}(t) = P_{2n+2-m-r}t^{j-r}\mathbf{b}^{(0)}(t)$. The rest follows by linear combination. ∎

Indeed, with the help of this lemma we conclude from Theorem 3.4 the following result.

Theorem 4.4 The polynomials of the residual vectors $\mathbf{r}_k(t)$ satisfy the three-term recursion

$$\mathbf{r}_{k+1}(t) = P_{n-n_{k+1}+1}(\mathbf{p}_k(t)\mathbf{r}_k(t)t^{-2d_k+1} - t^{-d_k-d_{k-1}+1}\mathbf{r}_{k-1}(t))\,.$$

The recursion starts with $\mathbf{r}_0(t) = 0$ and $\mathbf{r}_1(t) = \mathrm{i}\sum_{i=1}^{n}(a_{i-1} - a_i)t^{i-1}$. For Example 1 considered in Subsection 3.3 the recursion for the residuals makes no sense, because we have only one step. We check Example 2. We have in this case

$$\mathbf{r}_1 = [\,2\mathrm{i},\ \ 1-\mathrm{i},\ \ -1\,]^T \quad \text{and} \quad \mathbf{r}_2 = \frac{1}{2}[\,1+\mathrm{i},\ \ -\mathrm{i}\,]^T.$$

The recursion of Theorem 4.4 gives $\mathbf{r}_3 = 2$, which can be verified directly.

5. BLOCK ZW-FACTORIZATION

We show that the algorithm described in the previous section can be used to compute factorization of the matrix. First we recall some concepts.

5.1. General ZW-Factorization

A matrix $A = [\,a_{ij}\,]_{i,j=1}^{n}$ is called *W-matrix* if $a_{ij} = 0$ for all (i,j) for which $i > j$ and $i+j > n+1$ or $i < j$ and $i+j \le n$. The matrix A will be called *unit W-matrix* if in addition $a_{ii} = 1$ for $i = 1,\ldots,n$ and $a_{i,n+1-i} = 0$ for $i \ne \frac{n+1}{2}$. The transpose of a W-matrix is called a *Z-matrix*. A matrix which is both a Z- and a W-matrix will be called *X-matrix*. The names arise from the shapes of the set of all possible positions for nonzero entries, which are as follows:

$$W = \begin{bmatrix} \bullet & & & & & \bullet \\ \bullet & \circ & & & \circ & \bullet \\ \bullet & \circ & \circ & \circ & \circ & \bullet \\ \bullet & \circ & \bullet & \bullet & \circ & \bullet \\ \bullet & \bullet & & & \bullet & \bullet \\ \bullet & & & & & \bullet \end{bmatrix}, \quad Z = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ & \circ & \circ & \circ & \bullet & \\ & & \circ & \bullet & & \\ & & \bullet & \circ & & \\ & \bullet & \circ & \circ & \circ & \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix}, \quad X = \begin{bmatrix} \bullet & & & & & \bullet \\ & \bullet & & & \bullet & \\ & & \bullet & \bullet & & \\ & & \bullet & \bullet & & \\ & \bullet & & & \bullet & \\ \bullet & & & & & \bullet \end{bmatrix}.$$

For sake of simplicity of notation we assume that n is even, $n = 2m$. The case of odd n is similar. A unit W- or Z-matrix A is obviously nonsingular and a system $A\mathbf{f} = \mathbf{b}$ can be solved by back substitution with $\frac{n^2}{2}$ additions and $\frac{n^2}{2}$ multiplications. A representation $A = ZXW$ of a nonsingular matrix A in which Z is a Z-matrix, W is a W-matrix, and X an X-matrix is called *ZW-factorization*. Analogously WZ-factorization is defined. A necessary and sufficient condition for a matrix $A = [\,a_{jk}\,]_{j,k=1}^{n}$ to admit a ZW-factorization is that all central submatrices $[\,a_{jk}\,]_{j,k=l}^{n+1-l}$ are nonsingular for all natural numbers $l = 1,\ldots,[\,\frac{n}{2}\,]$. A matrix with this property will be called *centro-nonsingular*. Under the same condition A^{-1} admits a WZ-factorization. Among all ZW-factorizations of A there is a unique one in which the factors are unit. Symmetry properties of the matrix are inherited in the factors of the unit ZW-factorization. If A is symmetric or skewsymmetric, then $W = Z^T$,

and X is symmetric or skewsymmetric, respectively. If A is centrosymmetric, then all factors Z, X and W are also centrosymmetric, if A is centro-skewsymmetric, then Z and W are centrosymmetric and X is centro-skewsymmetric (in particular, anti-diagonal). All this follows from the uniqueness of the unit ZW-factorization.

If the matrix is not centro-nonsingular, then one might look for a block ZW-factorization. We show that for some cases the matrix classes under consideration such a factorization exists and can be evaluated with the help of the algorithms described above. However, the block factorization for centrosymmetric T+H matrices we present below will not be a generalization of the unit ZW-factorization but of a modification of it which will be described next. We introduce the $n \times n$ X-matrix

$$\Theta_n = \begin{bmatrix} -1 & & & & & 1 \\ & \ddots & & & \iddots & \\ & & -1 & 1 & & \\ & & 1 & 1 & & \\ & \iddots & & & \ddots & \\ 1 & & & & & 1 \end{bmatrix}.$$

Obviously, $\Theta_n^{-1} = \frac{1}{2}\,\Theta_n$. If Z_0 is an $n \times n$ centrosymmetric Z-matrix, then the matrix $Z_1 = Z_0\Theta_n$ has the property $J_nZ_1 = Z_0J_n\Theta_n$. That means that the first m columns of Z_1 are skewsymmetric, whereas the last m columns are symmetric. Let us call a matrix with this property *column-symmetric*. If moreover the X-matrix built from the diagonal and antidiagonal of Z_1 is equal to Θ_n, then Z_1 will be referred to as *unit*. The unit ZW-factorization $A = Z_0X_0Z_0^T$ of a centrosymmetric, symmetric matrix A can be transformed into a ZW-factorization $A = ZXZ^T$ in which Z is unit column-symmetric. We will call this factorization *unit column-symmetric ZW-factorization*. Since the product

$$\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$

is a diagonal matrix, the X-factor in the column-symmetric ZW-factorization is actually a diagonal matrix. Thus, provided that A is centro-nonsingular, A admits a factorization $A = ZDZ^T$ in which Z is unit column-symmetric and D is diagonal.

5.2. Centrosymmetric T+H Matrices

We show now that any nonsingular centrosymmetric T+H matrix $C_n = T_n^+P_n^+ + T_n^-P_n^-$ has a *block* unit column-symmetric ZW-factorization. This is a representation $A = ZDZ^T$ in which D is a block diagonal matrix and Z is a column-symmetric Z-matrix. If the diagonal blocks of D have size $m_k^\pm \times m_k^\pm$, then the corresponding

diagonal blocks of Z are assumed to be $\pm I_{m_k^{\pm}}$. In this case we call the Z-matrix *block unit* column-symmetric. Clearly, such a factorization is, if it exists, unique. It can be expected that the block sizes m_k are equal to the numbers d_k that were introduced in Section 3.1. Let Σ_d denote the $(2d-1)\times d$ matrix

$$\Sigma_d = \left.\begin{bmatrix} & & 1 & & \\ & 1 & & 1 & \\ \iddots & & & & \ddots \\ 1 & & & & 1 \end{bmatrix}\right\} d\,.$$

The $n_k\times(2d_k-1)$ matrix $M_{2d_k-1}(\mathbf{u}^{(k)})\Sigma_{d_k}^T$ has symmetric or skewsymmetric columns, depending whether we have the + or – case, and has the form

$$\begin{bmatrix} \pm J_{d_k} U^{(k)} \\ * \\ U^{(k)} \end{bmatrix},$$

where $U^{(k)}$ is the (nonsingular) upper triangular Toeplitz matrix

$$U^{(k)} = \begin{bmatrix} u_0^{(k)} & \dots & u_{d_k-1}^{(k)} \\ & \ddots & \vdots \\ & & u_0^{(k)} \end{bmatrix} \tag{11}$$

and $u_j^{(k)}$ $(k=0,\dots,d_k-1)$ are the first components of $\mathbf{u}^{(k)} \in \mathbb{F}^{\,n_{k-1}+2}$ possibly extended by zeros in case where $n_{k-1}+2 < d_k$. We evaluate, for the + and – cases, the matrices

$$M_{2d_k-1}(\mathbf{u}^{(k)})\Sigma_{d_k}^T (U^{(k)})^{-1}.$$

These matrices will be extended to $n\times d_k$ matrices by adding symmetrically zeros at the top and the bottom. The resulting matrices will be denoted by $W_{\pm}^{(k)}$. Clearly, $W_{\pm}^{(k)}$ has again symmetric or skewsymmetric columns. From now on we indicate by a subscript or superscript the + or – case.

We form the block matrix

$$W = [\, W_-^{(\ell_-)} J_{d_{\ell_-}^-}\,,\ \dots\ W_-^{(1)} J_{d_1^-}\,,\ W_+^{(1)}\,,\ \dots\,,\ W_+^{(\ell_+)}\,], \tag{12}$$

which is a block unit column-symmetric W-matrix. Next we evaluate $C_n W$. We have

$$T_n^{\pm} W_{\pm}^{(k)} = \begin{bmatrix} \pm J_{\nu_k+d_k^{\pm}} R_{\pm}^{(k)} \\ O \\ R_{\pm}^{(k)} \end{bmatrix} (U_{\pm}^{(k)})^{-1}\,,$$

where

$$R_{\pm}^{(k)} = \begin{bmatrix} & & r_0^{\pm(k)} \\ & \iddots & \vdots \\ r_0^{\pm(k)} & \dots & r_{d_k^{\pm}-1}^{\pm(k)} \\ r_1^{\pm(k)} & \dots & r_{d_k^{\pm}}^{\pm(k)} \\ \vdots & & \vdots \\ r_{\nu_k^{\pm}}^{\pm(k)} & \dots & r_{\nu_k^{\pm}+d_k^{\pm}-1}^{\pm(k)} \end{bmatrix},$$

$\nu_k^{\pm} = \frac{1}{2}(n - n_k^{\pm})$. We introduce matrices $V_{\pm}^{(k)}$ by

$$V_{\pm}^{(k)} = \begin{bmatrix} r_0^{\pm(k)} & \dots & r_{d_k^{\pm}-1}^{\pm(k)} \\ & \ddots & \vdots \\ & & r_0^{\pm(k)} \end{bmatrix}$$

and the matrices $Z_{\pm}^{(k)}$ by

$$Z_{\pm}^{(k)} = T_n^{\pm} W_{\pm}^{(k)} U_{\pm}^{(k)} (V_{\pm}^{(k)})^{-1} J_{d_k^{\pm}} = \begin{bmatrix} \pm J_{\nu_k^{\pm}+d_k^{\pm}} R_{\pm}^{(k)} \\ O \\ R_{\pm}^{(k)} \end{bmatrix} (V_{\pm}^{(k)})^{-1} J_{d_k^{\pm}}.$$

We arrange the matrices $Z_{\pm}^{(k)}$ to the $n \times n$ matrix

$$Z = [\, Z_-^{(\ell_-)} J_{d_{\ell_-}^-} \,,\, \dots \, Z_-^{(1)} J_{d_1^-} \,,\, Z_+^{(1)} \,,\, \dots \,,\, Z_+^{(\ell_+)} \,].$$

This matrix is a block unit column-symmetric Z-matrix.

We have now the relation

$$C_n W = ZD\,, \tag{13}$$

where $D = \operatorname{diag}(D_-^{(\ell_-)}, \dots, D_-^{(1)}, D_+^{(1)}, \dots, D_+^{(\ell_+)})$ and

$$D_+^{(k)} = J_{d_k^+} V_+^{(k)} (U_+^{(k)})^{-1}, \quad D_-^{(k)} = V_-^{(k)} (U_-^{(k)})^{-1} J_{d_k^-}.$$

Note that the matrices $D_+^{(k)}$ are lower triangular and the matrices $D_-^{(k)}$ are upper triangular Hankel matrices. Since the inverse of a W-matrix is a W-matrix again we have a block ZW-factorization $C_n = ZDW^{-1}$. The matrix $2\,W^{-1}$ is a block unit row-symmetric W-matrix. Taking the uniqueness of unit ZW-factorization into account we conclude that $Z^T = 2\,W^{-1}$. Summing up, we obtain the following result.

Theorem 5.1 Any nonsingular centrosymmetric T+H matrix C_n admits a representation $C_n = ZDZ^T$ in which Z is a block unit column-symmetric Z-matrix and

D is a block diagonal matrix the diagonal blocks of which are triangular Hankel matrices.

In order to find the block ZW-factorization of C_n one has to run the split Schur algorithm from Section 4.1. This gives the factor Z. In order to evaluate the block diagonal factor one has to find the first $d_k^{\pm}$ components of the vectors $\mathbf{u}_{\pm}^{(k)}$. This can be done running partly the split Levinson algorithm described in Section 3.1. Unfortunately the numbers d_k are not known a priori, so some updating might be necessary during the procedure.

5.3. Symmetric Toeplitz Matrices

Symmetric Toeplitz matrices are special centrosymmetric T+H matrices, so Theorem 5.1 holds. However, there are some specific relations between the left and the right parts of the matrices Z and D, which are not fully understood yet. The following can be shown. The matrix D is of the form

$$D = \operatorname{diag}\left(J_{\mu_p} K_p^- J_{\mu_p}\,,\ \ldots\ ,\ J_{\mu_1} K_1^- J_{\mu_1}\,,\ K_1^+,\ \ldots\ ,\ K_p^+ \right)$$

where the following cases are possible:

1. The matrices K_j^+ and K_j^- are both $\mu_j \times \mu_j$ lower triangular Hankel matrices.
2. K_j^+ is a $\mu_j \times \mu_j$ lower triangular Hankel matrix and K_j^- is of the form

$$\begin{bmatrix} * & 0 & 0 \\ \mathbf{0} & K' & \mathbf{0} \\ 0 & 0 & * \end{bmatrix}, \tag{14}$$

   where K' is a $(\mu_j - 2) \times (\mu_j - 2)$ lower triangular Hankel matrix.
3. K_j^- is a $\mu_j \times \mu_j$ lower triangular Hankel matrix and K_j^+ is of the form (14).

5.4. Skewsymmetric Toeplitz Matrices

The construction of the block ZW-factorization of general skewsymmetric Toeplitz matrices is, to some extend, similar to that of centrosymmetric T+H matrices. However, the result has some different features. In the skewsymmetric case, the Z-factor will not be column-symmetric but centrosymmetric, the middle factor will block anti-diagonal and skewsymmetric rather than block diagonal and symmetric, and the blocks will be triangular Toeplitz rather than triangular Hankel.

We introduce the $n \times d_k$ matrices $\widetilde{W}_{\pm}^{(k)}$ by

$$\widetilde{W}_{-}^{(k)} = \begin{bmatrix} O_{\nu_k \times d_k} \\ M_{d_k}(\mathbf{u}^{(k-1)}) \\ O_{d_k \times d_k} \\ O_{\nu_k \times d_k} \end{bmatrix}, \quad \widetilde{W}_{+}^{(k)} = \begin{bmatrix} O_{\nu_k \times d_k} \\ O_{d_k \times d_k} \\ M_{d_k}(\mathbf{u}^{(k-1)}) \\ O_{\nu_k \times d_k} \end{bmatrix},$$

where $\nu_k = \frac{1}{2}(n - n_k)$, and form the matrix

$$\widetilde{W} = [\, \widetilde{W}_{-}^{(\ell)}, \, \cdots \, , \widetilde{W}_{-}^{(1)}, \, \widetilde{W}_{+}^{(1)}, \, \cdots \, , \widetilde{W}_{+}^{(\ell)} \,] \, . \tag{15}$$

Then $\widetilde{W}$ is a centrosymmetric W-matrix. Furthermore, we set

$$\widetilde{Z}_{+}^{(k)} = T_n \widetilde{W}_{-}^{(k)}, \quad \widetilde{Z}_{-}^{(k)} = -T_n \widetilde{W}_{+}^{(k)},$$

and form the matrix

$$\widetilde{Z} = [\, \widetilde{Z}_{-}^{(\ell)} \, \cdots \, \widetilde{Z}_{-}^{(1)} \, \widetilde{Z}_{+}^{(1)} \, \cdots \, \widetilde{Z}_{+}^{(\ell)} \,]. \tag{16}$$

Then $\widetilde{Z}$ is a centrosymmetric Z-matrix and

$$T_n \widetilde{W} = \widetilde{Z} \widetilde{K},$$

where $\widetilde{K}$ is the skewsymmetric block antidiagonal matrix

$$\widetilde{K} = \begin{bmatrix} 0 & & & & & -I_{d_\ell} \\ & & & & \iddots & \\ & & & -I_{d_1} & & \\ & & I_{d_1} & & & \\ & \iddots & & & & \\ I_{d_\ell} & & & & & 0 \end{bmatrix}. \tag{17}$$

We represent now $\widetilde{W}$ and $\widetilde{Z}$ in the form $\widetilde{W} = W D_1$ and $\widetilde{Z} = Z D_2$, where W and Z are block unit and D_1 and D_2 are block diagonal with triangular Toeplitz diagonal blocks and centrosymmetric. From this we obtain a unit block ZW-factorization

$$T = Z D_2 \widetilde{K} D_1^{-1} W^{-1}.$$

The matrix $K = D_2 \widetilde{K} D_1^{-1}$ is now skewsymmetric and block anti-diagonal with triangular Toeplitz anti-diagonal blocks. Taking the uniqueness of such a factorization into account we arrive at the following result.

Theorem 5.2 Any nonsingular skewsymmetric Toeplitz matrix T_n admits a representation $T_n = ZKZ^T$ in which Z is a block unit centrosymmetric Z-matrix and K is a skewsymmetric block anti-diagonal matrix the anti-diagonal blocks of which are triangular Toeplitz matrices.

5.5. Hermitian Toeplitz Matrices

ZW-factorization of centrononsingular hermitian Toeplitz matrices is studied in [29]. It is an open problem how to generalize this to the case where the matrix has singular central submatrices.

ACKNOWLEDGEMENT

This work was supported by Research Grant SM05/02 of Kuwait University.

REFERENCES

[1] Ammar, G. and Gader, P., A variant of the Gohberg-Semencul formula involving circulant matrices, *SIAM J. Matrix Analysis Appl.*, 12, 3 (1991), 534–540.

[2] Al-Rashidi, A. Fast algorithms for skewsymmetric Toeplitz matrices with any rank profile, Master Thesis, Kuwait University, 2003.

[3] Bistritz, Y., Lev-Ari, H., and Kailath, T. Immitance-type three-term Schur and Levinson recursions for quasi-Toeplitz complex Hermitian matrice, *SIAM J. Matrix. Analysis Appl.*, 12, 3 (1991), 497–520.

[4] Delsarte, P. and Genin, Y., The split Levinson algorithm, *IEEE Transactions on Acoustics Speech, and Signal Processing* ASSP-34 (1986), 470–477.

[5] Delsarte, P. and Genin, Y., On the splitting of classical algorithms in linear prediction theory, *IEEE Transactions on Acoustics Speech, and Signal Processing* ASSP- 35 (1987), 645–653.

[6] Delsarte, P., Genin, Y. and Kamp, Y. A generalization of the Levinson algorithm for Hermitian Toeplitz matrices with any rank profile, *IEEE Transactions on Acoustics Speech, and Signal Processing* ASSP-33 (1985), 964–971.

[7] Demeure, C. J. Bowtie factors of Toeplitz matrices by means of split algorithms, *IEEE Transactions on Acoustics Speech, and Signal Processing*, ASSP-37, 10 (1989), 1601–1603.

[8] Evans, D.J. and Hatzopoulos, M., A parallel linear systems solver, *Internat. J. Comput. Math.*, 7, 3 (1979), 227–238.

[9] Gohberg, I. and Olshevsky, V., Circulants, displacements and decompositions of matrices, *Integral Equations Operator Theory*, 15, 5 (1992), 730–743.

[10] Heinig, G. Inversion of Toeplitz and Hankel matrices with singular sections, *Wiss. Zeitschr. d. TH Karl-Marx-Stadt*, 25, 3 (1983), 326–333.

[11] Heinig, G. Formulas and algorithms for block Hankel matrix inversion and partial realization. In: *Progress in Systems and Control*, 5, Birkhäuser 1990, 79-90.

[12] Heinig, G. Matrix representations of Bezoutians, *Linear Algebra Appl.*, 223-224 (1995), 337–354.

[13] Heinig, G. Chebyshev-Hankel matrices and the splitting approach for centrosymmetric Toeplitz-plus-Hankel matrices, *Linear Algebra Appl.*, 327 (2001), 181–196.

[14] Heinig, G. Inversion of Toeplitz-plus-Hankel matrices with any rank profile, In: V. Olshevsky (Ed.), *Fast Algorithms for Structured Matrices*, AMS-Series *Contemporary Mathematics*, vol. 323 (2003), 75–90 .

[15] Heinig, G. and Jankowski, P., Parallel and superfast algorithms for Hankel systems of equations. *Numerische Mathematik* 58 (1990), 109–127.

[16] Heinig, G., Jankowski, P. and Rost, K., Fast inversion algorithms of Toeplitz-plus-Hankel matrices, *Numerische Mathematik*, **52** (1988), 665–682.

[17] Heinig, G. and Rost, K., *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhäuser Verlag, Basel, Boston, Stuttgart, 1984.

[18] Heinig, G. and Rost, K., DFT representations of Toeplitz-plus-Hankel Bezoutians with application to fast matrix-vector multiplication, *Linear Algebra Appl.*, 284 (1998), 157–175.

[19] Heinig, G. K. Rost, Hartley transform representations of symmetric Toeplitz matrix inverses with application to fast matrix-vector multiplication. *SIAM J. Matrix Analysis Appl.* 22, 1 (2000), 86–105.

[20] Heinig, G. and Rost, K., Hartley transform representations of inverses of real Toeplitz-plus-Hankel matrices, *Numerical Functional Analysis and Optimization*, 21 (2000), 175–189.

[21] Heinig, G. and Rost, K., Efficient inversion formulas for Toeplitz-plus-Hankel matrices using trigonometric transformations, In: V. Olshevsky (Ed.), *Structured Matrices in Mathematics, Computer Science, and Engineering*, AMS-Series *Contemporary Mathematics*. vol. 2 (2001), 247–264.

[22] Heinig, G. and Rost, K., Centro-symmetric and centro-skewsymmetric Toeplitz matrices and Bezoutians, *Linear Algebra Appl.* 343–344 (2002), 195–203.

[23] Heinig, G. and Rost, K., Fast algorithms for skewsymmetric Toeplitz matrices, *Operator Theory: Advances and Applications*, Birkhäuser Verlag, Basel, Boston, Berlin, 135 (2002), 193–208.

[24] Heinig, G. and Rost, K., Centrosymmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices and Bezoutians, *Linear Algebra Appl.*, 366 (2003), 257–281.

[25] Heinig, G. and Rost, K., Fast algorithms for centro-symmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices, *Numerical Algorithms* 33 (2003), 305–317.

[26] Heinig, G. and Rost, K., New fast algorithms for Toeplitz-plus-Hankel matrices, *SIAM J. Matrix Analysis Appl.*, 25, 3 (2004), 842–857.

[27] Heinig, G. and Rost, K., Split algorithms for skewsymmetric Toeplitz matrices with arbitrary rank profile, *Theoretical Computer Science*, 315 (2004), 453–468.

[28] Heinig, and Rost, G. K., Split algorithms for symmetric Toeplitz matrices with arbitrary rank profile, *Numerical Linear Algebra with Applications*, to appear.

[29] Heinig, G. and Rost, K., Schur-type Algorithms for the solution of Hermitian Toeplitz systems via factorization, *Operator Theory: Advances and Applications*, to appear.

[30] Heinig, G. and Rost, K.,Split algorithms for hermitian Toeplitz matrices with arbitrary rank profile, *Linear Algebra Appl.*, to appear.

[31] Heinig, G. and Rost, K., Split algorithms for centrosymmetric Toeplitz-plus-Hankel matrices with arbitrary rank profile, *Linear Algebra Appl.*, in preparation.

[32] Kailath, T. and Sayed, A. H., *Fast reliable algorithms for matrices with structure*, SIAM, Philadelphia 1999.

[33] Krishna, B. and Krishna, H., Computationally efficient reduced polynomial based algorithms for hermitian Toeplitz matrices, *SIAM J. Appl. Math.* 49, 4 (1989), 1275–1282.

[34] Krishna, H. and Morgera, S. D., The Levinson recurrence and fast algorithms for solving Toeplitz systems of linear equations, *IEEE Transactions on Acoustics Speech, and Signal Processing* ASSP-35 (1987), 839–848.

[35] Melman, A. A two-step even-odd split Levinson algorithm for Toeplitz systems, *Linear Algebra Appl.*, 338 (2001), 219–237.

[36] Merchant, G. A. and Parks, T. W., Efficient solution of a Toeplitz-plus-Hankel coefficient matrix system of equations, *IEEE Transactions on Acoustics Speech, and Signal Processing* ASSP-30, 1 (1982), 40–44.

[37] Nersesjan, A. B. and Papoyan, A. A., Construction of the matrix inverse to the sum of Toeplitz and Hankel matrices (Russian), *Izv AN Arm. SSR, Matematika*, 8, 2 (1983), 150–160.

[38] Pal, D. and Kailath, T. Fast triangular factorization and inversion of Hermitian Toeplitz, and related matices with arbitrary rank profile, *SIAM J. Matrix Analysis Appl.*, 14, 4 (1993), 1016–1042.

[39] Pan, V. Y. *Structured Matrices and Polynomials*, Birkhäuser Verlag, Boston and Springer–Verlag, New York, 2001.

[40] Voyevodin, V. V. and Tyrtyshnikov, E. E., *Numerical Processes with Toeplitz Matrices* (in Russian), Nauka, Moscow 1987.

[41] Van Barel, M., Heinig, G. and Kravanja, P., A stabilized superfast solver for nonsymmetric Toeplitz systems, *SIAM J. Matrix Analysis Appl.*, 23, 2 (2001), 494–510.

PRODUCTIVITY OF OIL WELLS IN ARBITRARILY SHAPED RESERVOIRS

M. N. M. Ibrahim
School of Chemical Sciences, Universiti Sains Malaysia
11800 Minden, Pulau Pinang, Malaysia
email: mnm@usm.my

1. INTRODUCTION

A boundary element approach for predicting the productivity of oil wells arranged in complex configurations within irregularly shaped reservoirs were developed. The integral equations are written for boundary points as well as for the locations of the wells which are treated as point sources and sinks with specified pressures but unknown strengths. Using this approach, the solution to the resulting matrix gives the values of the nodal boundary pressure and their normal derivatives, as well as the unknown flow rates of all the wells.

2. PROBLEM FORMULATION

Consider a hypothetical two-dimensional homogeneous reservoir S having NSS sources and /or sinks located randomly within an arbitrarily shaped reservoir. The following assumptions were used in developing the theory: a) the reservoir is in steady-state flow with reservoir pressure above bubble points i.e. undersaturated condition; b) single phase fluid having small (and constant) compressibility and constant viscosity is flowing in the system; c) the reservoir has a uniform thickness and it has a finite boundary; and d) gravitational effects are negligible.

The differential equation describing the unknown functions i.e. pressure, at all points in the reservoir, is obtained by the introduction of Darcy's law into the continuity equation. By imposing the conditions and assumptions stated above, the differential equation describing the pressure distribution in the reservoir is [1,2]:

$$\frac{\partial^2 p}{\partial X^2} + \frac{\partial^2 p}{\partial Y^2} + \frac{\mu}{k} \sum_{m=1}^{NSS} q_m \, \delta(X - X_m, Y - Y_m) = 0, \tag{1}$$

where p is the pressure , μ is the dynamic viscosity of the fluid, k is the permeability, q_m is the flow rate of the m^{th} well per unit area (positive for injectors and negative for producers), δ is the Dirac delta function, X, Y are coordinates axes,

and X_m, Y_m are coordinates of the m^{th} source and/or sink, where m goes from 1 to *NSS*.

Equation (1) can be transformed into an integral equation by multiplying it with the free space Green's function and integrating it twice by parts. The free-space Green's function is also called the fundamental solution [1,2,3] and is given as:

$$G = \frac{1}{2\pi} \ln\left(\frac{1}{r}\right), \tag{2}$$

where r is the distance between a field point (X, Y) and a point of application of a unit charge (X_c, Y_c). After standard manipulation [1], equation (1) then becomes:

$$\alpha p(X_i, Y_i) = \frac{1}{2\pi} \sum_{j=1}^{N} \frac{\partial p}{\partial n_j} \int_{sj} \ln\left(\frac{1}{r_{i,j}}\right) ds -$$

$$\frac{1}{2\pi} \sum_{j=1}^{N} p_i \int_{sj} \frac{\partial}{\partial n} \left[\ln\left(\frac{1}{r_{i,j}}\right)\right] ds + \frac{1}{2\pi} \frac{\mu}{k} \sum_{m=1}^{NSS} q_m \ln\left(\frac{1}{r_{i,m}}\right), \tag{3}$$

where the boundary of the reservoir is divided into N constant elements with constant properties as shown in Figure 1. α is the included angle at the i^{th} pivot point. It is assigned a value of $\frac{1}{2}$ when the pivot point is on a smooth boundary (*i.e.* not on a corner), and a value of 1 when the pivot point is inside the problem domain. For simplicity, let

$$G_{i,j} = \frac{1}{2\pi} \int_{sj} \ln\left(\frac{1}{r_{i,j}}\right) ds \tag{4}$$

$$H_{i,j} = \frac{1}{2\pi} \int_{sj} \frac{\partial}{\partial n} \left[\ln\left(\frac{1}{r_{i,j}}\right)\right] ds \tag{5}$$

$$GSS_{i,m} = \frac{1}{2\pi} \ln\left(\frac{1}{r_{i,m}}\right), \tag{6}$$

where X_i, Y_i are coordiantes of any pivot point, $r_{i,j}$ is the distance between the pivot point and the j^{th} element where j runs from 1 to N, and $r_{i,m}$ is the distance between the pivot point and the m^{th} source and/or sink. Equation (3) now simplifies to :

$$\alpha p(X_i, Y_i) = \sum_{j=1}^{N} \frac{\partial p}{\partial n_j} G_{i,j} - \sum_{j=1}^{N} p_j H_{i,j} + \sum_{m=1}^{NSS} q_m GSS_{i,m} \tag{7}$$

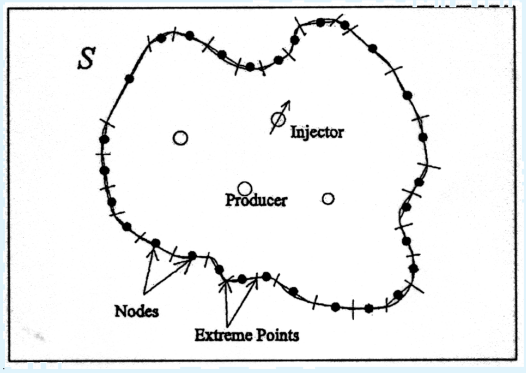


Figure 1: Reservoir having NSS sources and sinks where its boundary is divided into N segments or elements

The boundary of the reservoir S can be of the type S_p or $S_{dp/dn}$ or a combination of the two types. Over the S_p type boundary, the pressure p is specified as constant throughout the element while dp/dn is unknown. Over the $S_{dp/dn}$ type boundary, the dp/dn is prescribed as constant and the pressure p is unknown. Similarly, the sources and/or sinks can also have known and unknown rates. For the known flow rate well the well-bore pressure p_w is unknown and for the unknown flow rate well, the well-bore pressure is prescribed.

The idea is to apply Equation (7) at all the boundary nodes ($\alpha = \frac{1}{2}$), as well as at the entire source and/or sink locations ($\alpha = 1$). By doing so, a system of $N + NSS$ equations with $N + NSS$ unknowns can be obtained and simplified to matrix form as follows:

$$[HGGSS] \quad \vec{U} \; = \; \vec{A}, \tag{8}$$

where $[HGGSS]$ consists of the coefficients H, G and GSS. The vector $\vec{U}$ contains all the $N + NSS$ unknowns of $p, dp/dn, p_w$ and q and $\vec{A}$ is a vector containing all the known values.

3. VALIDATION

The flow rates obtained from Muskat's analytical equation [4] are compared with the BEM solutions for a circular battery of wells located at a radius $r = 50$ feet in a circular reservoir as of radius $R = 5,000$ feet as shown in Figure 2. The wells in the

battery are symmetric about the center. In order to have uniform pressures around the boundary of the reservoir, it was necessary to place the center of the battery at the reservoir center.

The ratio of the total production of the battery Q_n to the production of a single well Q_1 is plotted against the number of wells and compared with the Muskat's results as shown in Figure 3. The perfect match of the plots in Figure 3 clearly shows that the BEM solutions agree with the Muskat's analytical solutions.

Even though regular well patterns and boundary geometries are presented in these example applications, this was done simply to allow comparison with published analytical solutions. The method is equally applicable to non-pattern well clusters arbitrarily located in reservoirs with irregular boundary shapes.

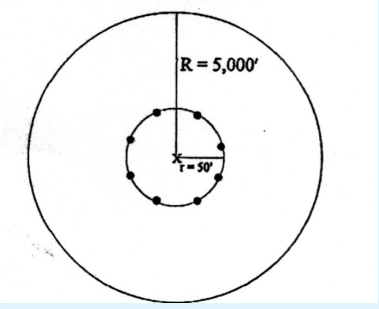


Figure 2: A circular battery of n wells at the center of a circular reservoir

4. CONCLUSIONS

The concept of formulating differential equations at source and/or sink points as well as at boundary node points was investigated and found to give excellent results. The formulation has the advantage of calculating the unknown source and/or sink rates directly as part of the matrix solution. Other potential uses include (i) the calculation of the production of individual wells within leases in a multiple lease reservoir and (ii) the identification of candidate wells in a field that may need work-over by comparing the predicted production rates with the actual field production rates.

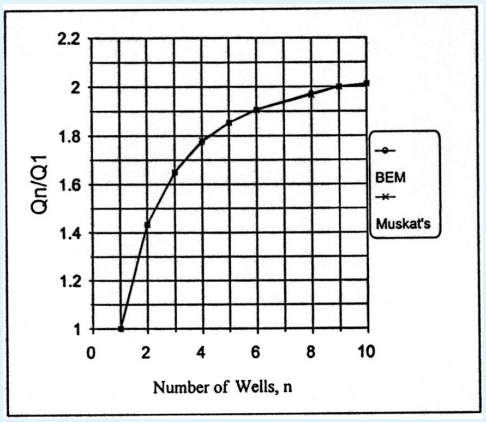


Figure 3: Comparison between Muskat's solution and the BEM solution

ACKNOWLEDGEMENTS

The author would like to express his appreciations to Universiti Sains Malaysia for the financial support of this project. The author also would like to acknowledge the Department of Mathematics & Computer Science at Kuwait University and Kuwait Foundation for the Advancement of Science for organizing this conference

REFERENCES

[1] Numbere, D. T. and Tiab, D., An improved streamline-generating technique that uses the boundary (Integral) element method, *SPE Reservoir Engineering*, (1988), 1061-68.

[2] Numbere, D. T., A general streamline modeling technique for homogenous and heterogeneous porous media with applications to stemaflood prediction, *Ph. D. dissertation,* University of Oklahoma, Norman, 1982.

[3] Brebbia, C. A., The boundary element method for engineers, John Wiley and Sons Inc., New York City, (1978), 46-103.

[4] Muskat, M., The flow of Homogeneous Fluids through Porous Media, McGraw-Hill Book Co., Inc., New York City, (1937), 507-514.

MATHEMATICAL MODELING AND SCIENTIFIC COMPUTING EXEMPLARY FOR CHEMICAL PROCESSES

M. Lehn and R. Scherer
Institute of Practical Mathematics, Universität Karlsruhe (TH),
D-76128 Karlsruhe, Germany
e-mail: lehn@math.uni-karlsruhe.de, scherer@math.uni-karlsruhe.de

1. INTRODUCTION

Various problems arising in Natural Sciences, Industry, and Economy are solved by mathematical modeling and scientific computing. Recently, several monographs were published on this topic (e.g., MacCluer [7], Shier and Wallenius [8]). Given such a problem with the relevant parameters, the basic laws or empiric experience are used to derive a mathematical description of the problem. Mathematical models may consist of equations representing systems of linear or nonlinear equations, ordinary or partial differential equations, or time series, Markov chains, etc. For the solution of mathematical models efficient numerical methods together with appropriate software have to be developed. Visualization allows to simulate the given problem. Finally, the numerical results have to be calibrated with real data and then the mathematical model has to be refined. The realization of these different steps requires the cooperation of engineers, mathematicians and computer scientists.

Interesting problems arise in chemical processes. This note is addressed to mathematical modeling and numerical treatment of chemical processes, especially the chemical vapor deposition, which is widely used for the production of advanced materials. To mention is the production of carbon fibre-reinforced carbon and carbon-carbon composites. Further, the production of catalytic converters, solar cells, or microelectronic devices involve such processes. The chemical vapor deposition of carbon on graphite by methane pyrolysis is considered. In an exemplary reactor the gas mixture flows into the reactor from the bottom to meet the heated substrate, where the deposition takes place. The reactor walls are cold to avoid chemical reactions there. The gas mixture leaves the reactor on top after being cooled down (Bammidipati et al. [1]).

In section two the mathematical model to describe the chemical vapor deposition process is derived. The Navier-Stokes equations to model the flow in the reactor are coupled with convection-diffusion-reaction equations to describe the concentra-

tions of the chemical species in the reactor. Chemical reactions in the gas-phase as well as on the surface are taken into account. The idea of separating the hyperbolic part from the parabolic part of the convection-diffusion equations leads to the concept of operator splitting (Karlsen et al. [4], [5]) that is introduced in section three. More precisely, the convection-diffusion-reaction equations are separated in four parts: convection, diffusion, gas-phase and surface reactions. The numerical splitting method to solve the hyperbolic and parabolic part of the equations (Crandall and Majda [3], LeVeque [6], Strang [9], Vreugdenhil and Koren [10]) are considered in section four. A test equation is used to discuss local discretization errors and numerical results obtained by splitting methods with some basic schemes.

2. THE MATHEMATICAL MODEL

To describe chemical vapor deposition mathematically the simple two dimensional geometry of figure 1 is used. In the reactor model there are at least four different processes, which have to be described. The defining equations are the Navier-Stokes equations to describe the flow in the reactor and a convection-diffusion-reaction system to describe the concentrations of the chemical species. The reaction terms in the convection-diffusion-reaction system are given by the gas-phase reactions, whereas the surface reactions appear as boundary conditions.

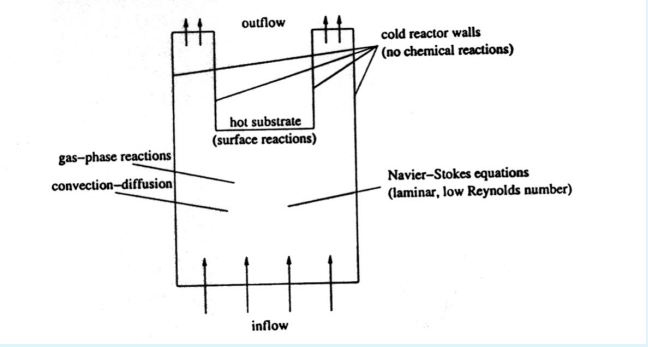


Figure 1: Reactor model

2.1.Navier-Stokes equations

The Navier-Stokes equations are used to model the flow in the reactor. They consist of the equations for the conservation of mass, momentum, and energy. Additionally, an equation of state is needed. Here we consider laminar flow in two space dimensions.

Conservation of mass

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \vec{u}) \qquad \text{with density } \rho \text{ and flow velocity } \vec{u} = (u, v)^T.$$

Conservation of momentum

$$\frac{\partial (\rho \vec{u})}{\partial t} = -\nabla \cdot (\rho \vec{u} \vec{u}^T) + \nabla \cdot [\mu(\nabla \vec{u} + (\nabla \vec{u})^T) - \frac{2}{3}\mu(\nabla \cdot \vec{u})I] - \nabla p + \rho \vec{g}$$

with viscosity μ, thermodynamical pressure p and gravity $\vec{g} = (g_1, g_2)^T$.

Conservation of energy

$$\rho c_p \frac{\partial T}{\partial t} = -\rho c_p (\vec{u} \cdot \nabla T) + \nabla \cdot (\lambda \nabla T) + \rho \dot{q_s} + \mu \Phi$$

with specific heat capacity c_p, temperature T, heat conductivity λ, external energy $\dot{q_s}$ and dissipativity Φ.

Equation of state

$$p = \rho \; r \; T \qquad \text{with gas constant } r.$$

2.2. Gas phase reactions and surface reactions

The chemical species mainly participating in the gas phase reactions and the surface reactions are identified and the reaction schemes are given (Birakayala and Evans [2]). E.g., for surface reactions the following results are possible:

1.	$C_2H_2 + 2C(S) \Rightarrow 2C(S) + 2C(D) + H_2$
2.	$C_6H_6 + 6C(S) \Rightarrow 6C(S) + 6C(D) + 3H_2$
3.	$C_2H_4 + 2C(S) \Rightarrow 2C(S) + 2C(D) + 2H_2$
4.	$H + CH(S) \Leftrightarrow C(S) + H_2$
5.	$H + C(S) \Leftrightarrow CH(S)$

Chemical reaction kinetics

A simplified reaction scheme is considered with the following reactions (CH_4 : methane gas, C : carbon, C_ν : carbon compound with ν atoms, C_∞ compact carbon):

gas phase reactions $\quad CH_4 \xrightarrow{k_{12}} C_1 \xrightarrow{k_{23}} C_2 \xrightarrow{k_{34}} C_{\geq 6},$

surface reactions $\quad C_1 \xrightarrow{k_{25}} C_\infty,\; C_2 \xrightarrow{k_{35}} C_\infty,\; C_{\geq 6} \xrightarrow{k_{45}} C_\infty,$

and the reaction constants for $T = 1398K$ and $p = 20kPa$.

Consider a simple model for gas phase reactions as well as for surface reactions. Let A and B be given substances with concentrations y_1 and y_2 which react to form the substances C and D with concentrations y_3 and y_4 with kinetic parameter k, i.e.,

$$A + B \ \stackrel{k}{\rightarrow} \ C + D.$$

The reacting amount per time of a substance is a multiple of the product $y_1 y_2$ with the multiplicative constant k. This reaction can be described by stiff ordinary differential equations

$$\frac{dy_1}{dt} = -k\ y_1 y_2,\ \frac{dy_2}{dt} = -k\ y_1 y_2,\ \frac{dy_3}{dt} = k\ y_1 y_2,\ \frac{dy_4}{dt} = k\ y_1 y_2$$

with suitable initial values. This simple model is only valid under certain assumptions such as constant temperature, constant volume, and no additional substances.

2.3. System of convection-diffusion-reaction equations

The concentrations of the chemical species are given by a system of convection-diffusion-reaction equations derived by the principle of conservation of mass.

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot c_i \vec{u} + d_i \Delta c_i + r_i^g, \quad i = 1, \ldots, K,$$

where the following abbreviations are used:

c_i: the concentration of the i-th chemical substance $c_i = c_i(x, y, t)$,
$\vec{u}$: the velocity (in direction of x and y) $\vec{u} = (u_1(x, y, t), u_2(x, y, t))^T$
d_i: the diffusion coefficient of the i-th substance (d_i constant),
r_i^g: the growth rate of the i-th substance by gas-phase reactions $r_i^g = r_i^g(x, y, t)$,
r_i^s: the growth rate of the i-th substance by surface reactions $r_i^s = r_i^s(x, y, t)$,
K: the number of substances ($K \approx 150 - 200$, test equation K small).

The growth rate of the $i - th$ chemical substance c_i by surface reactions appears as a boundary condition at the surface of the substrate:

$$\frac{\partial c_i}{\partial n} = r_i^s, \quad i = 1, \ldots, K.$$

In the convection-diffusion-reaction system the four different processes are coupled together. In each time-step the field of velocity, described by the Navier-Stokes equation, the gas phase and surface reactions, described by two systems of stiff ordinary differential equations, and then the solution of the convection-diffusion-reaction equation have to be computed, which is a very complex and difficult problem.

The treatment of the Navier-Stokes equations is a separate topic, therefore in this paper the velocity field of the flow is taken from measurements done by chemical engineers. Thus the treatment of the flow is decoupled from the other processes. We concentrate on the convection-diffusion equations and the gas-phase reactions.

3. SPLITTING METHODS

With some additional simplifications such as constant flow velocities and a single chemical substance the general convection-diffusion-reaction system from section 2.3 reduces to the convection-diffusion-reaction *model equation*

$$c_t = -(uc_x + vc_y) + d(c_{xx} + c_{yy}) + r$$

for the concentration $c = c(x, y, t)$ (constant velocity $\vec{u} = (u, v)^T$, constant diffusion d, reaction $r = r(x, y, t)$).

It is very difficult to construct a method convenient and effective for all three parts in the equation. Originally operator splitting was used to reduce the dimension from 2D to 1D. Now the idea of operator splitting is to separate the convection part from the diffusion part and also to separate the reaction part of the equation (LeVeque [6]). Then it is possible to use methods especially designed for solving the given type of equation. In the following sections the model equation is considered for the function $u = u(x, y, t)$ instead of $c = c(x, y, t)$.

3.1. Dimensional splitting

To introduce the idea of operator splitting consider the 2D convection equation

$$u_t + au_x + bu_y = 0 \tag{1}$$

with initial value

$$u(x, y, 0) = u_0(x, y)$$

and exact solution

$$u(x, y, t) = u_0(x - at, y - bt)$$

in $0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le t \le T$. The 2D equation is splitted into two 1D equations, using the splitting stepsize τ, to determine the solution of the two problems at time τ (one step):

$$v_t + av_x = 0,\ v(x, y, 0) = u_0(x, y) \quad \text{resp.} \quad w_t + bw_y = 0,\ w(x, y, 0) = v(x, y, \tau) \tag{2}$$

with the exact solution

$$v(x, y, \tau) = v(x - a\tau, y, 0) = u_0(x - a\tau, y)$$

and

$$w(x,y,\tau) = w(x,y-b\tau,0) = v(x,y-b\tau,\tau) = u_0(x-a\tau,y-b\tau) = u(x,y,\tau),$$

respectively. In this simple case, w is the exact solution of the 2D convection equation, i.e., there is no splitting error. Usually there arises a splitting error depending on the splitting stepsize τ.

3.2. General splitting

Consider the general equation

$$u_t + (\mathcal{A}+\mathcal{B})u = 0,\ u(x,y,0) = u_0(x,y)$$

in some region of $\mathbb{R}^2$ and $0 \le t \le T$, where $\mathcal{A}$ and $\mathcal{B}$ are suitable differential operators. The solution is given using the corresponding solution operator by

$$u(x,y,t) = S_t u_0(x,y).$$

Splitting of the previous equation into the corresponding parts of $\mathcal{A}$ and $\mathcal{B}$ with the splitting stepsize τ yields

$$\begin{aligned} v_t &= \mathcal{A}v, \qquad v(x,y,0) = u_0(x,y), \\ w_t &= \mathcal{B}w, \qquad w(x,y,0) = v(x,y,\tau), \end{aligned}$$

where the solutions at time τ (one step) are denoted by

$$v(x,y,\tau) = \mathcal{S}_\tau^{\mathcal{A}} v(x,y,0) \ \text{ and } \ w(x,y,\tau) = \mathcal{S}_\tau^{\mathcal{B}} w(x,y,0),$$

respectively. The splitting solution u^* for the original problem is achieved by

$$u^*(x,y,\tau) = \mathcal{S}_\tau^* u_0(x,y) \ \text{ and } \ u^*(x,y,n\tau) = (\mathcal{S}_\tau^*)^n\, u_0(x,y),\ n = 1,\dots,N,\ N\tau = T,$$

where $\mathcal{S}_\tau^*$ is constructed using $\mathcal{S}_\tau^{\mathcal{A}}$ and $\mathcal{S}_\tau^{\mathcal{B}}$.

The method is called *Godunov* resp. *Strang splitting*, if it holds

$$\mathcal{S}_\tau^* = \mathcal{S}_\tau^{\mathcal{B}}\mathcal{S}_\tau^{\mathcal{A}} \ \text{ resp. } \ \mathcal{S}_\tau^* = \mathcal{S}_{\tau/2}^{\mathcal{A}}\mathcal{S}_\tau^{\mathcal{B}}\mathcal{S}_{\tau/2}^{\mathcal{A}}.$$

Considering a linear system of equations $u_t + Au_x + Bu_y = 0$ with initial value $u(x,y,0) = u_0(x,y)$, the splitting error of one step $u(x,y,\tau) - u^*(x,y,\tau)$ of the Godunov resp. Strang splitting satisfies $\mathcal{O}((\tau)^2)$ resp. $\mathcal{O}((\tau)^3)$ for $\tau \to 0$. Godunov

splitting implies order one, Strang splitting order two. The splitting error depends on the commutator $\mathcal{AB} - \mathcal{BA}$ and should be studied in detail.

3.3. Special problems

For the m-dimensional convection-diffusion equation

$$u_t + \sum_{i=1}^{m} (f_i(u))_{x_i} = \varepsilon \sum_{i=1}^{m} u_{x_i x_i}, \quad u(x,0) = u_0(x), \quad x = (x_1, \ldots, x_m),$$

u scalar, operator splitting can be introduced analogously. The constant $\varepsilon > 0$ regulates the dominance of the convection. The convection equation

$$v_t + \sum_{i=1}^{m} (f_i(v))_{x_i} = 0$$

with inital value $v(x,0) = v_0(x)$ and exact solution $v(x,t) = S_t^{(K)} v_0(x)$ and the diffusion equation

$$w_t = \varepsilon \sum_{i=1}^{m} w_{x_i x_i}$$

with initial value $w(x,0) = w_0(x)$ and exact solution $w(x,t) = S_t^{(D)} w_0(x)$ are solved separately to get the solution of the convection-diffusion equation $u(x,t)$ as

$$u^*(x,y,n\tau) = (S_\tau^{(D)} S_\tau^{(K)})^n u_0(x), \quad n = 1, \ldots, N, \ N\tau = T.$$

A result on convergence of the splitting solution $u^*(x,y,n\tau)$ is known (Karlsen and Risebro [4]): Let $u_0 \in L_\infty \cap B.V.$ and $f(u)$ Lipschitz continuous, then $(S_\tau^{(D)} S_\tau^{(K)})^n u_0(x)$ converges for $\tau \to 0$ to the solution of the given initial value problem.

Problems of convection-reaction or diffusion-reaction or convection-diffusion-reaction type can also be considered, where the aspect of stiffness appears in the reaction part.

4. NUMERICAL SPLITTING METHODS

The idea of splitting can also be applied to the numerical solution of differential equations. There exist different methods for convection and for diffusion equations and they can complement each other.

4.1. Numerical methods

Replace the operators $S_\tau^{\mathcal{A}}$ and $S_\tau^{\mathcal{B}}$ by the operators $\mathcal{H}_\tau^{\mathcal{A}}$ and $\mathcal{H}_\tau^{\mathcal{B}}$ corresponding to numerical methods and define

$$\mathcal{H}_\tau^* = \mathcal{H}_\tau^{\mathcal{A}}\mathcal{H}_\tau^{\mathcal{B}} \quad \text{resp.} \quad \mathcal{H}_\tau^* = \mathcal{H}_{\tau/2}^{\mathcal{A}}\mathcal{H}_\tau^{\mathcal{B}}\mathcal{H}_{\tau/2}^{\mathcal{A}},$$

then the splitting method reads

$$U^{n+1} = \mathcal{H}_\tau^* U^n = (\mathcal{H}_\tau^*)^n\, U^0,\ n = 1, \ldots, N,$$

where U^n is the vector consisting of the numerical solution in each grid point of the space and at time $n\tau$. Using Strang splitting and applying numerical methods $\mathcal{H}_\tau^{\mathcal{A}}$ and $\mathcal{H}_\tau^{\mathcal{B}}$ of consistency order two leads to a splitting method of order two for a linear system. This result still holds for smooth solutions of the nonlinear system of conservation laws

$$u_t + (f(u))_x + (g(u))_y = 0$$

with initial value $u(x, y, 0) = u_0(x, y)$ (Strang [9]).

The numerical methods use a grid size in the different space coordinates and the time stepsize Δt. Usually the time step Δt is equivalent to the splitting stepsize τ, but it could be possible that τ is a multiple of Δt. This depends on the splitting error, the local error of the numerical methods, and of the propagation of the different errors. It is very important that the numerical methods used in the different parts harmonize with each other, then error compensations are possible and good results can be expected.

4.2. Error estimates

Splitting methods deliver approximations to the true solution and error bounds are needed to judge on the accuracy of the method. In this section a simple example is discussed, which is also used for numerical experiments in section 3.

The linear 1D convection-diffusion equation

$$u_t + a u_x = D u_{xx}, \quad x \in \mathbb{R},\ 0 \le t \le T$$

is taken as a test equation for the numerical methods (Vreugdenhil and Koren [10]). With the initial values $u(x, 0) = u_0(x) = \sin \pi \frac{x-\alpha}{\beta-\alpha}$ if $x \in [\alpha, \beta]$ and $u_0(x) = 0$ if $x \notin [\alpha, \beta]$ the exact solution is given by

$$u(x, t) = \frac{1}{\sqrt{4\pi D t}} \int_{-\infty}^{\infty} \exp\left(\frac{-\xi^2}{4Dt}\right)\ u_0(x - at - \xi)\, d\xi.$$

Operator splitting does not introduce an additional error since the convection equation

$$u_t + au_x = 0, \quad u(x,0) = u_0(x)$$

has the exact solution $u^*(x,\tau) = u_0(x - a\tau)$ and the diffusion equation

$$u_t = Du_{xx}, \quad u(x,0) = u^*(x,\tau)$$

has the exact solution $u^{**}(x,\tau) = u(x,\tau)$.

Different finite difference methods for the convection and diffusion parts are discussed. Numerical methods suggested for the convection equation are the first order upwind method and the second order Beam-Warming method. The diffusion equation is solved using the explicit centered difference scheme of order one as well as the implicit second order Crank-Nicolson method (LeVeque [6]).

Let $u(x_j, t_{n+1})$ denote the exact solution of the convection-diffusion equation and $u_{j,n+1}$ the numerical solution each at the grid point x_j and at time t_{n+1}. The numerical approximation is obtained with one step of the splitting method $\mathcal{H}_\tau^*$ using the corresponding numerical schemes. Concerning the local discretization it is assumed that the values $u_{j,n}$ are exact. The local discretization error for the pairs upwind method/centered difference scheme and Beam-Warming/Crank-Nicolson satiesfies

$$u(x_j, t_{n+1}) - u_{j,n+1} = \frac{1}{2}\left(a(a-\nu)u_{xx} + D^2 u_{xxx}\right)|_{(x_j,t_n)}\ (\Delta t)^2 + \mathcal{O}((\Delta t)^3)$$

and

$$u(x_j, t_{n+1}) - u_{j,n+1} = \frac{1}{12}\left(a(3\nu a - 2\nu^2 - a^2)u_{xxx} - D\nu^2 u_{xxxx} - D^3 u_{xxxxxx}\right)|_{(x_j,t_n)}\ (\Delta t)^3 + \mathcal{O}((\Delta t)^4),$$

respectively, where $\nu = \frac{\Delta x}{\Delta t}$ and $\Delta t \to 0$.

In both cases the leading error terms are small considering the constants multiplied with the space derivatives of u. First, for a grid ratio ν chosen close to the convection velocity a the term $a(a-\nu)$ as well as $3\nu a - 2\nu^2 - a^2$ is close to zero. Further, for typical diffusion coefficients D, it holds that $D \ll 1$ (*e.g.*, $D = 0.002$), so the remaining constants also turn out to be of small size.

4.3. Numerical experiments

Very satisfactory results are achieved by the combination of the Beam-Warming method (upwind difference quotients, second order) for the convection equation

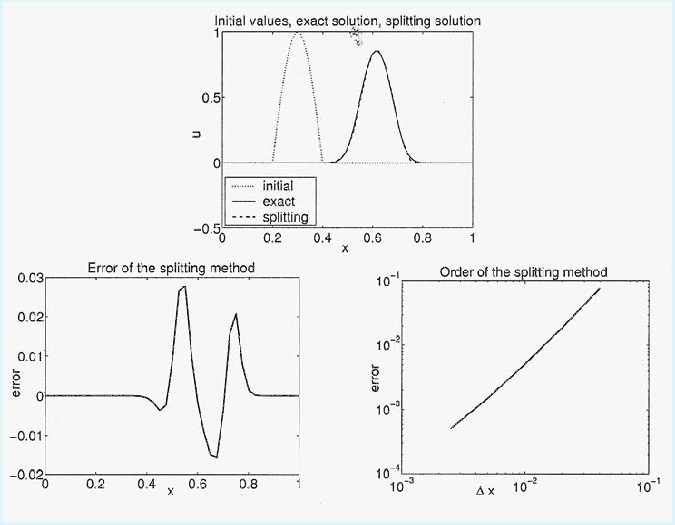


Figure 2: Numerical solution of the convection-diffusion equation

and the Crank-Nicolson method (implicit centered differences, second order) for the diffusion equation.

Figure 2 shows the obtained results. The parameters in the linear 1D convection diffusion-equation are $\alpha = 0.2,\ \beta = 0.4,\ a = 1,\ D = 0.002,\ x \in [0, 1]$. The first plot shows the numerical solution of the convection-diffusion equation produced by the splitting method using the Beam-Warming method for the convection part and the Crank-Nicolson method for the diffusion part. Together with the numerical solution, the initial values as well as the exact solution is shown. The numerical solution was computed up to $T = 0.3$ with the space discretization $\Delta x = 0.025$ and time-steps $\tau = \Delta t = 0.9\Delta x$ satisfying the CFL condition for the convection equation. The second plot shows the error of the numerical solution compared to the exact solution. In the third plot using logarithmic scales the second order of convergence is shown. Computations for different stepsizes Δx have been performed to achieve this result.

Future plans involve the generalization of the test equations in various ways. The modeling of chemical reactions will introduce source terms to the equations. The 1D case should be extended to the 2D case where real data from chemical vapor deposition processes can be used. Concerning the numerical methods more stress will be put on the interaction between the convective and the diffusive part of the splitting methods.

REFERENCES

[1] Bammidipati, S., et al., Chemical Vapor Deposition of Carbon on Graphite by Methane Pyrolysis, *AIChE Journal,* 42 (1996), 3123-3132.

[2] Birakyala, N. and Evans, A. A reduced reaction mechanism for carbon CVD/CVI processes, *Carbon,* 40 (2002), 675-683.

[3] Crandall, M. and Majda, A., The Method of Fractional Steps for Conservation Laws, *Numer. Math.,* 34 (1980), 285-314.

[4] Karlsen, K.H. and Risebro N.H., An Operator Splitting Method for Nonlinear Convection-Diffusion Equations, *Numer. Math.,* 77 (1997), 365-382.

[5] Karlsen, K.H., et al., Operator Splitting Methods for Systems of Convection-Diffusion Equations: Nonlinear Error Mechanisms and Correction Strategies, *J. Comp. Phys.,* 173 (2001), 636-663.

[6] LeVeque, R.J., *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, 2002.

[7] MacCluer C.R., *Industrial mathematics: Modeling in Industry, Science, and Government*, Prentice Hall, 2000.

[8] Shier, D.R. and Wallenius, K.T., *Applied Mathematical Modeling: A Multidisciplinary Approach*, Chapman & Hall, London, 2000.

[9] Strang, G., On the Construction and Comparison of Difference Schemes, *SIAM J. Numer. Anal.,* 5 (1968), 506-517.

[10] Vreugdenhil, C.B. and Koren, B., *Numerical Methods for Advection-Diffusion Problems*, Vieweg, Braunschweig, 1993.

WEIGHTED (0,2)-INTERPOLATION WITH ADDITIONAL INTERPOLATORY CONDITION

M. Lénárd
Department of Mathematics & Computer Science
Kuwait University, P. O. Box 5969, Safat 13060, Kuwait
e-mail: lenard@mcc.sci.kuniv.edu.kw

1. INTRODUCTION

P. Turán initiated the study of the (0,2)-interpolation in order to get an approximate solution to the differential equation

$$y'' + f \cdot y = 0.$$

In 1961 Balázs [2] introduced a generalization of this problem, the *weighted (0,2)-interpolation problem*: Let the system of distinct nodes

$$x_1, x_2, \ldots, x_n \in (a, b) \tag{1}$$

be given in the finite (or infinite) open (or closed) interval (a, b) and let $w \in C^2(a, b)$ be a given function, called *weight function*. Find a polynomial R_n of minimal degree satisfying the conditions

$$R_n(x_k) = y_k; \quad (wR_n)''(x_k) = y_k'', \quad (k = 1, \ldots, n; n \in \mathbb{N}) \tag{2}$$

where y_k, y_k'' are arbitrary given real numbers.

If for any choice of the values y_k, y_k'', there exists a unique polynomial R_n of degree less than $2n$, which fulfils the equations (2), then the problem is called *regular* on the nodes (1) with the weight function w. The questions are how to choose the nodal points and the weight function so that the problem is regular, and in the regular case find a simple explicit form of R_n in order to prove convergence.

Balázs [2] investigated the above problem on the interval $[-1, 1]$, when the nodes are the zeros of the ultraspherical polynomial $P_n^{(\alpha)}$ $(\alpha > -1)$, and the weight function is $w(x) = (1 - x^2)^{(\alpha+1)/2}$. He showed, that in this case the problem is not regular, there does not exist a polynomial of degree $\leq 2n - 1$ satisfying the requirements (2). He proved, that if n is even, then under the condition

$$R_n(0) = \sum_{k=1}^{n} y_k l_k^2(0) \tag{3}$$

the problem is regular, but if n is odd, the uniqueness fails. ($l_k(x)$ represent the fundamental polynomials of Lagrange interpolation corresponding to the nodal points x_k, He gave the explicit form of the interpolational polynomial and proved convergence theorem.

Several authors investigated the weighted (0,2)-interpolation with the additional Balázs-type condition (3) on the roots of the classical orthogonal polynomials (Joó [6], Joó and Szili [7], Prasad [9], [10], [11], [12], Szili [16], [17]). For special choice of the nodes, Bajpai [1], Eneduanya [5], and Balázs [3] substituted the additional condition (3) with interpolatory type conditions. For more results on (0,2)-interpolation we refer to the survey paper of Szili [18].

In order to replace the additional Balázs-type condition (3) with interpolatory conditions, in [8] we have studied the weighted (0,2)-interpolation problem with two additional interpolatory conditions in a unified way with respect to the existence, uniqueness and representation (explicit formulae). In this paper we study the weighted (0,2)-interpolation problem with one interpolatory condition:

The problem: On the finite or infinite interval $[a, b]$ let x_i, $(i = 0, \dots, n; n \in \mathbb{N})$ be distinct points, the *nodal points of interpolation*, and let $w \in C^2(a, b)$ be a given function, called *weight function*. Find a polynomial R_n of minimal degree satisfying the weighted (0,2)-interpolational conditions

$$R_n(x_i) = y_i, \quad (wR_n)''(x_i) = y_i'' \qquad (i = 1, \dots, n-1),$$

with the additional interpolatory condition

$$R_n(x_0) = y_0,$$

or

$$R_n'(x_0) = y_0',$$

or

$$R_n'(x_1) = y_1',$$

where y_i, y_0', y_1' and y_i'' are arbitrary real numbers.

As the number of conditions is $2n - 1$, the problem is regular, if for any choice of the values y_i, y_i'', y_0' and y_1' there exists a unique polynomial Q_n of degree at most $\leq 2n - 2$. We formulate sufficient conditions on the nodal points and the weight function w, for the problem to be regular. In the regular cases we find simple explicit forms for R_n. Finally, applying the theorems, we present some results on the zeros of the classical orthogonal polynomials.

2. PRELIMINARIES

Let $[a, b]$ be a finite or infinite interval, and let

$$x_0, x_1, \dots, x_n \in [a, b] \tag{4}$$

be distinct real numbers, the nodal points of the interpolation. Let $p_{n-1}(x)$ be a polynomial of degree $n-1$, for which

$$p_{n-1}(x_i) = 0 \qquad (i = 1, \dots, n-1), \tag{5}$$

and

$$l_j(x) = \frac{p_{n-1}(x)}{p'_{n-1}(x_j)(x - x_j)} \qquad (j = 1, \dots, n-1) \tag{6}$$

are basis polynomials of Lagrange interpolation, and hence

$$l_j(x_i) = \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \qquad (i, j = 1, \dots, n-1), \tag{7}$$

and

$$l'_j(x_j) = \frac{p''_{n-1}(x_j)}{2p'_{n-1}(x_j)} \qquad (j = 1, \dots, n-1). \tag{8}$$

Furthermore, let us introduce the notations

$$r(x) = (x - x_0)^{\varepsilon_1}(x - x_n)^{\varepsilon_2},$$

$$q(x) = (x - x_0)^{\delta_1}(x - x_n)^{\delta_2},$$

where $\varepsilon_i, \delta_i \in \{0, 1, 2, \dots\}$ and $\varepsilon_i \geq \delta_i$ for $i = 1, 2$.

We recall the general explicit formulae of the fundamental polynomials of first and second kind for the weighted (0,2)-interpolation ([8]):

Lemma 1. If on the system of nodes (4) the weight function w satisfies the conditions

$$w(x_i) \neq 0, \quad (qwp_{n-1})''(x_i) = 0 \qquad (i = 1, \dots, n-1), \tag{9}$$

then for $k = 1, \dots, n-1$ the polynomials

$$\begin{aligned} A_k(x) = {} & \frac{r(x)}{r(x_k)}\, l_k^2(x) + \frac{q(x)p_{n-1}(x)}{r(x_k)p'_{n-1}(x_k)} \\ & \times \Big\{ c_k + \int_{x_0}^{x} \Big[\frac{l'_k(x_k)l_k(t) - l'_k(t)}{t - x_k} \cdot \frac{r(t)}{q(t)} + a_k l_k(t) + b_k p_{n-1}(t) \Big] dt \Big\}, \end{aligned} \tag{10}$$

where b_k, c_k are arbitrary constants,

$$a_k = -\frac{(rw)''(x_k)}{2(qw)(x_k)} - 2l_k'(x_k)\Big(\frac{r}{q}\Big)'(x_k), \tag{11}$$

satisfy the weighted (0,2)-interpolational conditions

$$A_k(x_i) = \delta_{i,k}, \qquad (wA_k)''(x_i) = 0 \qquad (i = 1, \dots, n-1). \tag{12}$$

Furthermore, for $k = 1, \dots, n-1$ the polynomials

$$B_k(x) = \frac{q(x)p_{n-1}(x)}{2q(x_k)w(x_k)p_{n-1}'(x_k)}\Big\{\tilde{a}_k + \int_{x_0}^{x}\Big[l_k(t) + \tilde{b}_k p_{n-1}(t)\Big]dt\Big\} \tag{13}$$

satisfy the weighted (0,2)-interpolational conditions

$$B_k(x_i) = 0, \qquad (wB_k)''(x_i) = \delta_{i,k} \qquad (i = 1, \dots, n-1) \tag{14}$$

with $\tilde{a}_k, \tilde{b}_k$ arbitrary constants.

3. RESULTS

Theorem 1. Let $\{x_i\}_{i=0}^{n-1}$ $(n \geq 2)$ be a set of distinct nodes in $[a, b]$, let $w \in C^2(a, b)$ be a weight function, and $p_{n-1}(x) = c(x - x_1)\dots(x - x_{n-1})$.
If

$$w(x_i) \neq 0, \qquad \big(wp_{n-1}\big)''(x_i) = 0 \qquad (i = 1, \dots, n-1), \tag{15}$$

then for arbitrary real numbers y_i, y_i'' and y_0 there exists a unique polynomial R_n of degree at most $2n-2$, which fulfils weighted (0,2)-interpolational conditions at $x_1, \dots, x_{n-1}$

$$R_n(x_i) = y_i, \qquad (wR_n)''(x_i) = y_i'' \qquad (i = 1, \dots, n-1), \tag{16}$$

with the additional condition at x_0

$$R_n(x_0) = y_0. \tag{17}$$

Proof. We apply Lemma 1 with $r(x) = (x - x_0)$ and $q(x) = 1$. In order to get the minimal degree $2n-2$ for A_k, let $b_k = 0$ in (10). From the condition $A_k(x_0) = 0$ we get $c_k = 0$, hence we have for $k = 1, \dots, n-1$

$$\begin{aligned} A_k(x) \;=\; & \frac{x - x_0}{x_k - x_0}\, l_k^2(x) + \frac{p_{n-1}(x)}{(x_k - x_0)p_{n-1}'(x_k)} \\ & \int_{x_0}^{x}\Big[\frac{l_k'(x_k)l_k(t) - l_k'(t)}{t - x_k}(t - x_0) + a_k l_k(t)\Big]dt, \end{aligned} \tag{18}$$

where

$$a_k = -\frac{\big((x-x_0)w\big)''(x_k)}{2w(x_k)} - 2l_k'(x_k). \tag{19}$$

Furthermore, let

$$A_0(x) = \frac{p_{n-1}(x)}{p_{n-1}(x_0)}, \tag{20}$$

and for $k = 1, \dots, n-1$

$$B_k(x) = \frac{p_{n-1}(x)}{2w(x_k)p_{n-1}'(x_k)} \int_{x_0}^{x} l_k(t)dt. \tag{21}$$

As the polynomials A_k, B_k $(k = 1, \dots, n-1)$, and A_0, defined by (18)–(21) are the basis polynomials of the interpolational problem, the polynomial

$$R_n(x) = \sum_{k=0}^{n-1} y_k \; A_k(x) + \sum_{k=1}^{n-1} y_k'' \; B_k(x) \tag{22}$$

is of degree at most $2n-2$ and fulfils the conditions (16) and (17).

For the proof of the uniqueness we study the homogeneous problem: Find a polynomial $\bar{R}_n$ of degree at most $2n-2$ such that $\bar{R}_n(x_i) = 0$, $(w\bar{R}_n)''(x_i) = 0$ $(i = 1, \dots, n-1)$, and $\bar{R}_n(x_0) = 0$. From these conditions it is obvious, that

$$\bar{R}_n(x) = p_{n-1}(x)\bar{g}_{n-1}(x),$$

where $\bar{g}_{n-1}$ is a polynomial of degree at most $n-1$. As for $i = 1, \dots, n-1$

$$(w\bar{R}_n)''(x_i) = 2w(x_i)p_{n-1}'(x_i)\bar{g}_{n-1}'(x_i) = 0,$$

and $w(x_i) \neq 0$, $p_{n-1}'(x_i) \neq 0$, therefore with a constant $\bar{c}$ we get $\bar{g}_{n-1}(x) \equiv \bar{c}$. Finally, from the condition $\bar{R}_n(x_0) = 0$ we obtain $\bar{c} = 0$, that is $\bar{R}_n(x) \equiv 0$, which completes the proof. □

Theorem 2. Let $\{x_i\}_{i=0}^{n-1}$ $(n \geq 2)$ be a set of distinct nodes in $[a, b]$, let $w \in C^2(a, b)$ be a weight function, and $p_{n-1}(x) = c(x - x_1) \dots (x - x_{n-1})$.
If

$$w(x_i) \neq 0, \qquad \big(wp_{n-1}\big)''(x_i) = 0 \qquad (i = 1, \dots, n-1), \tag{23}$$

and

$$p_{n-1}'(x_0) \neq 0, \tag{24}$$

then for arbitrary real numbers y_i, y_i'' and y_0' there exists a unique polynomial R_n of degree at most $2n-2$, which fulfils weighted (0,2)-interpolational conditions at $x_1, \ldots, x_{n-1}$

$$R_n(x_i) = y_i, \qquad (wR_n)''(x_i) = y_i'' \qquad (i = 1, \ldots, n-1), \tag{25}$$

with the additional condition at x_0

$$R_n'(x_0) = y_0'. \tag{26}$$

Proof. We apply Lemma 1 with $r(x) = (x - x_0)$ and $q(x) = 1$. In order to get the minimal degree $2n-2$ for A_k, let $b_k = 0$ in (10). From the equation $A_k'(x_0) = 0$ we get c_k, hence we have for $k = 1, \ldots, n-1$

$$\begin{aligned} A_k(x) &= \frac{x - x_0}{x_k - x_0}\, l_k^2(x) + \frac{p_{n-1}(x)}{(x_k - x_0)p_{n-1}'(x_k)} \\ &\quad \Big\{ c_k + \int_{x_0}^{x} \Big[\frac{l_k'(x_k)l_k(t) - l_k'(t)}{t - x_k}(t - x_0) + a_k l_k(t) \Big] dt \Big\}, \end{aligned} \tag{27}$$

where

$$a_k = -\frac{\big((x - x_0)w\big)''(x_k)}{2w(x_k)} - 2l_k'(x_k), \tag{28}$$

and

$$c_k = -\frac{p_{n-1}(x_0)l_k(x_0)}{p_{n-1}'(x_0)} \Big(\frac{1}{x_0 - x_k} + a_k \Big). \tag{29}$$

Furthermore, let

$$C_0(x) = \frac{p_{n-1}(x)}{p_{n-1}'(x_0)}, \tag{30}$$

and for $k = 1, \ldots, n-1$

$$B_k(x) = \frac{p_{n-1}(x)}{2w(x_k)p_{n-1}'(x_k)} \Big\{ -\frac{p_{n-1}(x_0)l_k(x_0)}{p_{n-1}'(x_0)} + \int_{x_0}^{x} l_k(t)dt \Big\}. \tag{31}$$

As the polynomials A_k, B_k $(k = 1, \ldots, n-1)$, and C_0, defined by (27)–(31), are the basis polynomials of the interpolational problem, the polynomial

$$R_n(x) = \sum_{k=1}^{n-1} y_k\, A_k(x) + \sum_{k=1}^{n-1} y_k''\, B_k(x) + y_0'\, C_0(x) \tag{32}$$

is of degree at most $2n-2$ and fulfils the conditions (25) and (26).

The uniqueness can be proved in a similar way as in Theorem 1, which completes the proof. □

Theorem 3. Let $\{x_i\}_{i=0}^{n-1}$ $(n \geq 2)$ be a set of distinct nodes in $[a,b]$, let $w \in C^2(a,b)$ be a weight function, and $p_{n-1}(x) = c(x-x_1)\dots(x-x_{n-1})$.
If

$$\begin{aligned} \big((x-x_0)wp_{n-1}\big)''(x_i) =& 0, \qquad (i=1,\dots,n-1),\\ w(x_i) \neq& 0, \qquad (i=0,1,\dots,n-1), \end{aligned} \tag{33}$$

then for arbitrary real numbers y_i, y_i'' and y_0', there exists a unique polynomial R_n of degree at most $2n$, which fulfils weighted (0,2)-interpolational conditions at $x_0, x_1, \dots, x_{n-1}$ with the additional condition

$$R_n'(x_0) = y_0'.$$

Proof. We apply Lemma 1 with $r(x) = (x-x_0)^2$ and $q(x) = (x-x_0)$. From the condition $A_k'(x_0) = 0$ we obtain $c_k = 0$, hence we have for $k = 1, \dots, n-1$

$$\begin{aligned} A_k(x) =& \frac{(x-x_0)^2}{(x_k-x_0)^2}\, l_k^2(x) + \frac{(x-x_0)p_{n-1}(x)}{(x_k-x_0)^2 p_{n-1}'(x_k)}\\ &\times \int_{x_0}^{x} \Big[\frac{l_k'(x_k)l_k(t) - l_k'(t)}{t-x_k}(t-x_0) + a_k l_k(t) + b_k p_{n-1}(t)\Big]dt, \end{aligned} \tag{34}$$

where

$$a_k = -\frac{\big((x-x_0)^2 w\big)''(x_k)}{2(x_k-x_0)w(x_k)} - 2l_k'(x_k). \tag{35}$$

It is clear, that $A_k(x_i) = \delta_{k,i}$ for $i = 0, 1, \dots, n-1$, $A_k'(x_0) = 0$ and by Lemma 1 also $(wA_k)''(x_i) = 0$ for $i = 1, \dots, n-1$. From $(wA_k)''(x_0) = 0$ we obtain the constant

$$b_k = \frac{1}{(x_k-x_0)p_{n-1}'(x_k)}\Big(\frac{1}{x_0-x_k} + a_k\Big). \tag{36}$$

The polynomial A_0, which fulfils the conditions

$$A_0(x_i) = \delta_{0,i}\,, \quad (wA_0)''(x_i) = 0 \quad (i=0,1,\dots,n-1), \quad A_0'(x_0) = 0,$$

will be determined in the form

$$A_0(x) = \frac{p_{n-1}^2(x)}{p_{n-1}^2(x_0)} + (x-x_0)p_{n-1}(x)g_0(x), \tag{37}$$

where g_0 is a polynomial of degree at most n. Applying the conditions $(wA_0)''(x_i) = 0$ for $i = 1, \dots, n-1$, we get

$$g_0'(x_i) = -\frac{1}{p_{n-1}^2(x_0)} \cdot \frac{p_{n-1}'(x_i)}{x_i - x_0} \qquad (i = 1, \dots, n-1),$$

which inspires us to define the polynomial g_0', as it follows

$$g_0'(x) = -\frac{1}{p_{n-1}^3(x_0)} \cdot \frac{p_{n-1}'(x)p_{n-1}(x_0) - p_{n-1}'(x_0)p_{n-1}(x)}{x - x_0} + c_0 p_{n-1}(x),$$

Thus

$$g_0(x) = a_0 - \frac{1}{p_{n-1}^3(x_0)} \int_{x_0}^{x} \frac{p_{n-1}'(t)p_{n-1}(x_0) - p_{n-1}'(x_0)p_{n-1}(t)}{t - x_0} dt + c_0 \int_{x_0}^{x} p_{n-1}(t)dt, \tag{38}$$

where

$$a_0 = -\frac{2p_{n-1}'(x_0)}{p_{n-1}^2(x_0)}, \qquad c_0 = \frac{-w''(x_0)}{2w(x_0)p_{n-1}^2(x_0)} \tag{39}$$

are determined from the conditions $A_0'(x_0) = 0$ and $(wA_0)''(x_0) = 0$.

For the fundamental polynomials of second kind we apply Lemma 1 with the conditions $B_k'(x_0) = 0$ and $(wB_k)''(x_k)$, and hence we obtain for $k = 1, \ldots, n-1$

$$B_k(x) = \frac{(x - x_0)p_{n-1}(x)}{2(x_k - x_0)p_{n-1}'(x_k)w(x_k)} \Big\{ \int_{x_0}^{x} l_k(t)dt + \frac{1}{(x_k - x_0)p_{n-1}'(x_k)} \int_{x_0}^{x} p_{n-1}(t)dt \Big\}. \tag{40}$$

It is clear, that $B_k(x_i) = 0$ for $i = 0, 1, \ldots, n-1$, $B_k'(x_0) = 0$, $(wB_k)''(x_0) = 0$, and by Lemma 1 also $(wB_k)''(x_i) = \delta_{k,i}$ for $i = 1, \ldots, n-1$.

It is easy to verify that the polynomial

$$B_0(x) = \frac{(x - x_0)p_{n-1}(x)}{2w(x_0)p_{n-1}^2(x_0)} \int_{x_0}^{x} p_{n-1}(t)dt \tag{41}$$

fulfils the conditions

$$B_0(x_i) = 0\,, \quad (wB_0)''(x_i) = \delta_{0,i} \quad (i = 0, 1, \ldots, n-1), \quad B_k'(x_0) = 0.$$

The polynomial C_0, which fulfils the conditions

$$C_0(x_i) = 0\,, \quad (wC_0)''(x_i) = 0 \quad (i = 0, 1, \ldots, n-1), \quad C'(x_0) = 1,$$

will be determined in the form

$$C_0(x) = (x - x_0)p_{n-1}(x)\tilde{g}_0(x),$$

where $\tilde{g}_0$ is a polynomial of degree at most n. From the equations

$$(wB_0)''(x_i) = 2(x_i - x_0)w(x_i)p_{n-1}'(x_i)\tilde{g}_0'(x_i) = 0 \qquad (i = 1, \ldots, n-1),$$

we get

$$\tilde{g}_0'(x_i) = 0 \qquad (i = 1, \dots, n-1),$$

and so

$$\tilde{g}_0(x) = \tilde{c} \int_{x_0}^{x} p_{n-1}(t) + \tilde{d},$$

where the constants $\tilde{c}$ and $\tilde{d}$ are defined by the conditions

$$C_0'(x_0) = 1, \qquad (wC_0)''(x_0) = 0.$$

Finally

$$C_0(x) = (x - x_0)p_{n-1}(x)\Big\{\frac{1}{p_{n-1}(x_0)} - \frac{(wp_{n-1})'(x_0)}{w(x_0)p_{n-1}^3(x_0)} \int_{x_0}^{x} p_{n-1}(t)dt\Big\}. \tag{42}$$

As the polynomials A_k and B_k $(k = 0, 1, \dots, n-1)$ and C_0, defined by (34)–(42), are the basis polynomials of the interpolational problem, the polynomial

$$R_n(x) = \sum_{k=0}^{n-1} y_k \; A_k(x) + \sum_{k=0}^{n-1} y_k'' \; B_k(x) + C_0(x)y_0' \tag{43}$$

is of degree at most $2n$ and fulfils the interpolational conditions.

For the proof of the uniqueness we study the homogeneous problem: Find a polynomial $\bar{R}_n$ of degree at most $2n-1$ such that $\bar{R}_n(x_i) = 0$, $(w\bar{R}_n)''(x_i) = 0$ $(i = 0, 1, \dots, n-1)$, and $\bar{R}_n'(x_0) = 0$. From these conditions it is obvious, that

$$\bar{R}_n(x) = (x - x_0)^2 p_{n-1}(x) g_{n-1}(x),$$

where g_{n-1} is a polynomial of degree at most $n-1$. As for $i = 1, \dots, n-1$

$$(w\bar{R}_n)''(x_i) = 2(x_i - x_0)w(x_i)p_{n-1}'(x_i)\big((x - x_0)g_{n-1}\big)'(x_i) = 0,$$

and $x_i \neq x_0$, $w(x_i) \neq 0$, $p_{n-1}'(x_i) \neq 0$, therefore with a constant $\bar{c}$ we get

$$(x - x_0)g_{n-1}(x) = \bar{c} \int_{x_0}^{x} p_{n-1}(t)dt.$$

Finally, from the condition

$$(w\bar{R}_n)''(x_0) = 2\bar{c}\, w(x_0)p_{n-1}^2(x_0) = 0,$$

on using $w(x_0) \neq 0$ and $p_{n-1}(x_0) \neq 0$ we obtain $\bar{c} = 0$, that is $\bar{R}_n(x) \equiv 0$, which completes the proof. □

4. RESULTS ON THE ZEROS OF THE CLASSICAL ORTHOGONAL POLYNOMIALS

Lemma 2. The classical orthogonal polynomials (Jacobi-, Laguerre-, and Hermite polynomials) satisfy the homogeneous differential equation

$$(wy)'' + f \cdot (wy) = 0 \tag{44}$$

with an appropriate weight function w and function f.

Proof. We refer to [15] ((4.24.1), (5.1.2), (5.5.2)):

For the Jacobi polynomials $y = P_n^{(\alpha,\beta)}$, $(\alpha, \beta > -1)$,

$$w(x) = (1-x)^{\frac{\alpha+1}{2}}(1+x)^{\frac{\beta+1}{2}},$$

and

$$f(x) = \frac{1}{4}\frac{1-\alpha^2}{(1-x)^2} + \frac{1}{4}\frac{1-\beta^2}{(1+x)^2} + \frac{2n(n+\alpha+\beta+1)+(\alpha+1)(\beta+1)}{2(1-x^2)}.$$

For the Laguerre polynomials $y = L_n^{(\alpha)}$, $(\alpha > -1)$,

$$w(x) = e^{-\frac{x}{2}}x^{\frac{\alpha+1}{2}},$$

and

$$f(x) = \frac{2n+\alpha+1}{2x} + \frac{1-\alpha^2}{4x^2} - \frac{1}{4}.$$

For the Hermite polynomials $y = H_n$,

$$w(x) = e^{-\frac{x^2}{2}}, \qquad f(x) = 2n + 1 - x^2.$$

□

4.1. Hermite polynomials

COROLLARY 1. (S. Datta and P. Mathur [4]) If the weight function is $w(x) = e^{-x^2/2}$, then under the condition

$$R_n(0) = y_0, \qquad \text{for even } n,$$

or

$$R_n'(0) = y_0', \qquad \text{for odd } n,$$

the weighted (0,2)-interpolation is regular on the zeros of the Hermite polynomial H_n of degree n.

Proof. As $H_n(0) = 0$ for odd n, $H_n(0) \neq 0$ for even n, $H_n'(x) = 2nH_{n-1}(x)$ (cf. (5.5.10) in [15]), and $w(0) \neq 0$, thus by Lemma 2 we can apply Theorem 1 and 3 with $x_0 = 0$, $p_n(x) = H_n(x)$, and with $x_0 = 0$, $p_{n-1}(x) = H_n(x)/x$, respectively, which completes the proof. □

COROLLARY 2. If the weight function is $w(x) = e^{-x^2/2}$, x_0 is the smallest or largest zero of H_{n+1}, then under the condition

$$R_n(x_0) = y_0,$$

or

$$R_n'(x_0) = y_0',$$

the weighted (0,2)-interpolation is regular on the zeros of the Hermite polynomial H_n of degree n.

Proof. The zeros of H_n and H_{n+1} form an interscaled system, so $H_n(x_0) \neq 0$ and $H_n'(x_0) \neq 0$, therefore by Lemma 2 we can apply Theorem 1 and 2, which completes the proof. □

4.2. Laguerre polynomials

COROLLARY 3. If the weight function is $w(x) = e^{-x/2}x^{(\alpha+1)/2}$, then under the condition

$$R_n(0) = y_0,$$

or

$$R_n'(0) = y_0',$$

the weighted (0,2)-interpolation is regular on the zeros of the Laguerre polynomial $L_n^{(\alpha)}$ $(\alpha > -1)$ of degree n.

Proof. As $L_n^{(\alpha)}(0) \neq 0$, $L_n^{(\alpha)'}(0) \neq 0$, so by Lemma 2 we can apply Theorem 1 and 2 with $x_0 = 0$ and $p_n(x) = L_n^{(\alpha)}(x)$, which completes the proof. □

COROLLARY 4. If the weight function is $w(x) = e^{-x/2}$, then under the condition

$$R_n'(0) = y_0',$$

the weighted (0,2)-interpolation is regular on the zeros of $L_n^{(-1)}(x) = -\frac{1}{n} x L_{n-1}^{(1)}(x)$.

Proof. By Lemma 2 we can apply Theorem 3 with $x_0 = 0$, $p_{n-1}(x) = L_{n-1}^{(1)}(x)$, and we obtain the statement. □

4.3. Jacobi polynomials

In a similar way one can prove

COROLLARY 5. If the weight function is $w(x) = (1-x^2)^{(\alpha+1)/2}$, then under the additional condition

$$R_n'(0) = y_0', \qquad \text{for odd } n,$$

or

$$R_n(0) = y_0, \qquad \text{for even } n,$$

the weighted (0,2)-interpolation is regular on the zeros of the ultraspherical polynomial $P_n^{(\alpha)}$ $(\alpha > -1)$ of degree n.

COROLLARY 6. If the weight function is $w(x) = (1-x)^{(\alpha+1)/2}(1+x)^{(\beta+1)/2}$, then under the additional condition

$$R_n'(1) = y_0',$$

or

$$R_n(1) = y_0,$$

the weighted (0,2)-interpolation is regular on the zeros of the Jacobi polynomial $P_n^{(\alpha,\beta)}$ $(\alpha, \beta > -1)$ of degree n.

COROLLARY 7. If the weight function is $w(x) = (1+x)^{(\beta+1)/2}$, then under the additional condition

$$R_n'(1) = y_0',$$

the weighted (0,2)-interpolation is regular on the zeros of $P_n^{(-1,\beta)}(x) = \frac{n+\beta}{2n}(x-1)P_{n-1}^{(-1,\beta)}(x)$ $(\beta > -1)$.

Remark. On using $P_n^{(\alpha,\beta)}(-x) = (-1)^n P_n^{(\beta,\alpha)}(x)$ ([15] (4.1.3)) we obtain similar results, if we substitute x with $-x$ and $x_0 = -1$.

5. ACKNOWLEDGEMENT

The research was supported by the grant SM 01/03, given by the Research Administration of Kuwait University.

REFERENCES

[1] Bajpai, P., Weighted (0,2)-interpolation on the extended Tchebycheff nodes of second kind, *Acta Math. Hung.*, 63 (2) (1994), 167–181.

[2] Balázs, J., Weighted (0,2)-interpolation on the zeros of ultraspherical polynomials, (in Hungarian), *MTA III.oszt. Közl.*, 11 (1961), 305–338.

[3] Balázs, J., Modified weighted (0,2) interpolation, Govil N. K. (ed.) et al., *Approximation Theory. In memory of A. K. Varma.* New York, NY: Marcel Dekker. Pure Appl. Math., 212, (1998), 61-85.

[4] Datta, S. and Mathur, P., On weighted (0,2)-interpolation on infinite interval $(-\infty, \infty)$, *Annales Univ. Sci. Budapest., Sect. Math.*, 42 (1999), 45–57.

[5] Eneduanya, S. A. N., The weighted (0,2) interpolation I, *Demonstracio Mathematica*, 18 (1) (1985), 9–13.

[6] Joó, I., On weighted $(0, 2)$ interpolation, *Annales Univ. Sci. Budapest., Sect. Math.*, 38 (1995), 185–222.

[7] Joó, I. and Szili, L., On weighted $(0, 2)$-interpolation on the roots of Jacobi polynomials, *Acta Math. Hung.*, 66 (1-2) (1995), 25–50.

[8] Lénárd, M., Weighted (0,2)-interpolation with interpolatory boundary conditions, (submitted to *Annales Univ. Sci. Budapest., Sect. Comp.*)

[9] Prasad, J., Some interpolatory polynomials on Hermite abscissas, *Math. Japonica*, 12 (1967), 73–80.

[10] Prasad, J., Balázs-type interpolation on Laguerre abscissas, *Math. Japonica*, 13 (1967), 47–53.

[11] Prasad, J., On the weighted $(0, 2)$-interpolation, *SIAM J. Numer. Anal.*, 7 (1970), 428–446.

[12] Prasad, J., On the representation of functions by interpolatory polynomials, *Mathematica (Cluj)*, 15 (1973), 289–305.

[13] Prasad, J., Verma, A., An analogue problem of J. Balázs and P. Turán, III, *Math. Japonica*, 14 (1969), 85–99.

[14] Surányi, J. and Turán, P., Notes on Interpolation I, On some interpolatorical properties of the ultraspherical polynomials, *Acta Math. Acad. Sci. Hung.*, 6 (1955), 66–79.

[15] Szegő, G., *Orthogonal polynomials*, Amer. Math. Soc. Coll. Publ., New York, 1959. (Fourth Edition in 1975)

[16] Szili, L., Weighted $(0, 2)$-interpolation on the roots of Hermite polynomials, *Annales Univ. Sci. Budapest., Sect. Math.*, 27 (1985), 153–166.

[17] Szili, L., Weighted $(0, 2)$-interpolation on the roots of the classical orthogonal polynomials, *Bull. Allahabad Math. Soc.*, 8–9 (1993–94), 111–120.

[18] Szili, L., A survey on (0,2) interpolation, *Annales Univ. Sci. Budapest., Sect. Comp.*, 16 (1996), 377–390.

GROUPS WITH PRESCRIBED ORDERS OF ELEMENTS

V. D. Mazurov
Institute of Mathematics, Novosibirsk, 630090, Russia;
e-mail: mazurov@math.nsc.ru

Abstract

The paper is devoted to latest results of author and his colleagues concerning the structure of groups with various finiteness conditions, which contain elements of small orders.

1. GROUPS WITH SMALL SPECTRUM

For a group G, define the *spectrum* of G as the set $\omega(G)$ of element orders of G. This set consists of some natural numbers and, possibly, the symbol ∞.

It is obvious that a group with $\omega(G) = \{1, 2\}$ is elementary Abelian. F.Levi and B.L. van der Waerden [14] proved that if $\omega(G) = \{1, 3\}$ then G is nilpotent of class at most 3. B.H.Neumann [22] described groups with $\omega(G) = \{1, 2, 3\}$. I.N.Sanov [25] and M.Hall [10], respectively, stated that a group G with $\omega(G) \subseteq \{1, 2, 3, 4\}$ and $\omega(G) \subseteq \{1, 2, 3, 6\}$ is locally finite. Nothing is known about the locally finiteness of groups of exponent 5, but the following results are proved for groups with small spectrum which contains the number 5.

Theorem 1.1. *Let G be a group.*

1 [24, 9]. *If $\omega(G) = \{1, 2, 5\}$ then G contains an Abelian normal Sylow subgroup.*

2 [9]. *If $\omega(G) = \{1, 3, 5\}$ then one of the following holds:*

(i) *$G = FT$ where F is a normal 5-subgroup which is nilpotent of class at most 2 and $|T| = 3$;*

(ii) G contains a normal 3-subgroup T which is nilpotent of class at most 3 and G/T is a 5-group.

3 [34]. *If $\omega(G) = \{1, 2, 3, 5\}$ then G is isomorphic to the alternating group A_5.*

4 [9]. *If* $\omega(G) = \{1, 2, 4, 5\}$ *then one of the following holds:*

(i) $G = TD$ *where* T *is a non-trivial elementary Abelian* 2*-group and* D *is a non-Abelian group of order* 10*;*

(ii) $G = FT$ *where* F *is an elementary Abelian normal* 5*-subgroup and* T *is isomorphic to a subgroup of quaternion group of order* 8*.*

(iii) G *contains a normal* 2*-subgroup* T *which is nilpotent of class at most* 6 *and* G/T *is a* 5*-group.*

5 [17]. *If* $\omega(G) = \{1, 2, 3, 4, 5\}$*. Then one of the following holds:*

(i) $G \simeq A_6$;

(ii) $G = VC$ *where* V *is a non-trivial elementary Abelian normal* 2*-subgroup of* G *and* $C \simeq A_5$*. More precise,* V *is the direct product of minimal normal subgroups of* G *of order* 16 *and every of those is the natural two-dimensional module for* $SL_2(4) \simeq C$*.*

Questions.

1. Is a group G locally finite if $2 \in \omega(G) \subseteq \{1, 2, 3, 4, 5\}$? In view of Theorem 1.1. this question is equivalent to the following: Is a 5-group of exponent 5 acting as a group of fixed-point-free automorphisms on a non-trivial elementary Abelian group necessarily cyclic?

2. Is a group G locally finite if $\omega(G)$ coincides with $\{1, 2, 3, 4, 6\}$ or $\{1, 2, 3, 4, 5, 6\}$?

Since A_5 is isomorphic to the first member of the infinite family of finite simple groups $L_2(2^n) = SL(2, 2^n), n \geq 2$, the part 2 of Theorem 1.1 is a particular case of the following characterization of $L_2(2^n)$.

Theorem 1.2 [34]. *Let* G *be a group. If* $\omega(G) = \omega(L_2(2^n))$ *for some* $n \geq 2$ *then* G *is isomorphic to* $L_2(2^n)$*.*

Under assumption that G is finite, this theorem was proved in [3]. In its turn, Theorem 1.2 follows from

Theorem 1.3. *Let G be a periodic group such that $\omega = \omega(G)$ satisfies the following conditions:*

a) $2, 3 \in \omega$;

b) if $n \in \omega$ *and* $n \neq 2$ *then* n *is odd;*

c) there exists a prime $p > 3$ *such that* $p \in \omega, 3p \notin \omega$.

Then there exists a locally finite field P of characteristic 2 such that $G \simeq L_2(P) = PSL_2(P)$, projective special linear group of dimension 2 over P.

The crucial point in the proof of this theorem is that the centralizer of every involution in a group G satisfying the conditions of this theorem is elementary Abelian, so that we can use Theorem 2.1 below. Notice that, under conditions of Theorem 1.3, the cases 2.1 and 2.2 of Theorem 2.1 are not presented by Corollary 2 of Theorem 4.2.

Theorem 1.3 can be used to point out examples of infinite groups which are recognizable up to isomorphism by their spectra.

Theorem 1.4 [16]. *Let P be a field which is the union of an ascending series of finite fields of orders $2^{m_i}, m_i > 1, i = 1, 2, ...$, and let $L = PGL_2(P)$. If there exists a natural s such that 2^s does not divide m_i for every $i = 1, 2, ...$ then L is recognizable by $\omega(L)$, i.e. every group G with $\omega(G) = \omega(L)$ is isomorphic to L. In all other cases, there exist infinitely many pair-wise non-isomorphic groups G such that $\omega(G) = \omega(L)$.*

This theorem answers affirmatively a question by H.Deng and W.Shi [5] on the existence of a infinite group G which is recognizable by $\omega(G)$. Notice that every group G which is recognizable by $\omega(G)$ should be periodic because, for a group G containing an element of infinite order and arbitrary torsion-free group X, $\omega(G) = \omega(G \times X)$.

2. GROUPS WITH ABELIAN CENTRALIZERS OF INVOLUTIONS

An involution t of a group G is said to be *a finite involution* (in G) if, for every $g \in G$, the order of the commutator $[t, g] = tt^g$ is finite. This definition is equivalent to the condition that the order of ti is finite for every involution $i \in G$. It is obvious that in a periodic group every involution is finite.

Theorem 2.1 [16]. *Let G be a group containing a finite involution t. Suppose that the centralizer of every involution in G is an Abelian 2-group.*

1. *If the centralizer $C_G(t)$ of t in G contains an involution other than t then one of the following statements holds:*

1.1. *$C_G(t)$ is normal in G.*

1.2. *$C_G(t)$ is elementary Abelian.*

2. *If $C_G(t)$ is elementary Abelian then one of the following statements holds:*

2.1. *$G = A < t >$ where A is an Abelian periodic subgroup containing no involutions and $a^t = a^{-1}$ for every $a \in A$.*

2.2. *G is an extension of an Abelian 2-group by a group containing no involutions.*

2.3. *There exists a locally finite field P of characteristic two such that G is isomorphic to $PGL_2(P)$.*

For the finite groups, this theorem is a particular case of a result by M.Suzuki [28]. Finite case of part 2 was proved firstly by R.Brauer, M.Suzuki and G.E.Wall [4] (see also [12, Theorem XI.2.7]) with an aid of the character theory of finite groups. Later, D.Goldschmidt [7] found an elementary proof in which the finiteness of G was used also substantially.

The condition of the existence of a finite involution cannot be weakened as shows an example of free product $PGL_2(P) * X$ where P is an arbitrary field of characteristic 2 and X is an arbitrary torsion free group.

On the other hand, it is easy to notice that, under conditions of Theorem 2.1, groups of form 2.1 and 2.3 are locally finite, but there exist non-locally finite groups of form 2.2. For example, the natural semi-direct product of the additive group of an arbitrary field P of characteristic 2 and the multiplicative group of P acting on P by the multiplication satisfies the conditions of Theorem 2.1.

To prove Theorem 2.1, suppose that $T = C_G(t)$ is not normal in G and contains more than one involution. Then one can show that T is a Sylow 2-subgroup of G and $N_G(T) = TK$ where K is a periodical Abelian group which contains no involutions.

Moreover, G acts by right multiplications on the set of right cosets of $N = N_G(T)$ as a triply transitive group, and one can apply

Theorem 2.2 [16]. *Let G be a triply transitive group in which the stabilizer of some two points is periodic and contains no involutions. If the stabilizer of some three points is trivial then there exists a locally finite field P of characteristic 2 such that G is similar to $PGL_2(P)$ in its natural action on projective line $P \cup \{\infty\}$.*

Theorem 2.2 generalizes the well known Zassenhaus theorem [30] on finite sharply triply transitive groups of odd degree (see also [12, Theorem XI.2.1]).

Questions. 1. Is it true that a simple group with Abelian centralizers of involutions which contains a finite involution is isomorphic to $PGL_2(P)$ for some locally finite field P of characteristic two?

2. Let G be a group with a finite involution T such that $C_G(t)$ is a locally cyclic 2-group. Is it true that $G = NC_G(t)$ for some normal Abelian subgroup N?

3. QUADRATIC AUTOMORPHISMS OF ABELIAN GROUPS

An automorphism x of an Abelian group V is said to be a *quadratic automorphism* if there exist natural numbers a, b such that $vx^2 + avx + bv = 0$ for all $v \in V$, in other words, $x^2 + ax + b = 0$ in the endomorphism ring of G. A pair (a, b) is said to be *type* of quadratic automorphism x.

The most important examples of quadratic automorphisms are generators of cyclic groups of fixed-point-free automorphisms of orders 3, 4, 6 for which, respectively, $x^2 + x + 1 = 0$, $x^2 + 1 = 0$ and $x^2 - x + 1 = 0$.

Let F be a field, $a, b \in \mathbb{Z}, \lambda \in F, \lambda \neq 0$, $A(a, b) = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix}, B(a, b, \lambda) = \begin{pmatrix} -a & \lambda^{-1}b \\ \lambda & 0 \end{pmatrix}$. If $A(a, b)$ and $B(c, d, \lambda)$ are non-degenerated then denote by $L(a, b; c, d; \lambda, F)$ a subgroup H of $GL_2(F)$ generated by $A(a, b)$ and $B(c, d, \lambda)$ and call H the group *of type* $(a, b; c, d)$ *over* F *corresponding to* λ.

Theorem 3.1 [31, 15]. *Let V be an Abelian group, G is a group generated by two quadratic automorphisms α, β of V of types $(a, b), (c, d)$, respectively, such that orders of $\alpha, \beta, \alpha\beta$ are finite. Then G is an extension of a finite nilpotent group N by a subgroup of the direct product of a finite Abelian group A and a finite number*

of groups H_i of type $(a, b; c, d)$ over an extensions of a prime field $P_i, i = 1, ..., s$, by a n-th root of unity where n is equal to order of $\alpha\beta$, and the following conditions holds:

1. *If V is torsion-free then $N = 1$ and $H_i = L(a, b; c, d; \lambda, \mathbb{Q}(\lambda_i)), i = 1, ..., s$, where λ_i is some n-th root of unity in complex number field $\mathbb{C}$.*

2. *If the exponent m of V is finite then G is a finite group, and every prime divisor of $|N|$ and the characteristic of $P_i, i = 1, ..., s$, divide m.*

Corollary. *If G is a group generated by two quadratic automorphisms α, β of an Abelian group of a finite exponent m such that orders of $\alpha, \beta, \alpha\beta$ are finite then G is a finite group and every composition factor of G is isomorphic either to alternating group A_5, or to projective special linear group $L_2(q)$ of degree 2 over a field of order q whose characteristic divides m.*

4. GROUPS GENERATED BY FIXED-POINT-FREE AUTOMORPHISMS OF SMALL ORDERS

An action of a non-trivial group G on an (additive) non-zero group V is said to be *free,* if $vg \neq v$ for $1 \neq g \in G, 0 \neq v \in V$. In other words, G acts freely on V if G is a group of fixed-point-free automorphisms of V.

This section is devoted to generalizations of the following well-known result which is a particular case in the classification (given by H.Zassenhaus [30]) of finite groups acting freely on an Abelian group.

Theorem 4.1. *Let G be a finite group acting freely on an Abelian group. If G is generated by elements of prime order p then either G is cyclic, or $p = 5$ and G is isomorphic to $SL_2(5)$, or $p = 3$ and G is isomorphic to one of the groups $SL_2(3), SL_2(5)$.*

A proof of the following theorem is based on Theorem 3.1.

Theorem 4.2 [32, 20]. *Suppose that a group G acting freely on a non-zero Abelian group is generated by a non-empty normal set X of elements of order 3. If one of the following conditions holds:*

a) the order of $x^{-1}y$ is finite for every two elements $x, y \in X$,

b) the order of xy is finite for every two elements $x, y \in X$,

then G is a finite group which is isomorphic to either a cyclic group of order 3, or $SL_2(3)$, or $SL_2(5)$. In particular, if G is periodic then G is finite.

Corollary 1. *Let x be an element of order 3 in a group G which acts freely on a non-trivial Abelian group. If, for every $g \in G$, the order of the commutator $[x, g]$ is finite, then x belongs to a finite normal subgroup of G.*

Corollary 2. *Periodic group acting freely on an Abelian group and containing an element of order 3 contains a central element of order 2 or 3.*

A proof of Theorem 4.2, together with Theorem 3.1, uses the following two results.

Proposition 1. *Let x, y be elements of order 3 in $SL_2(\mathbb{C})$ such that the order n of xy is finite. If one of the following conditions holds:*

a) the order of $[x, y] = x^{-1}x^y$ is finite,

b) the order of xx^y is finite and non-equal to 3,

then either n is equal to one of the numbers $1, 2, 3, 4, 6, 10$ and $\langle x, y\rangle$ is finite, or the case b) holds, $n = 14$ and the order of xx^y is equal to 7.

Proposition 2. *Let x, y be an order 3 automorphisms of an Abelian group V such that the order of xy is finite and $G = \langle x, y\rangle$ acts freely on V. If one of the following conditions holds:*

a) the order of $[x, y] = x^{-1}x^y$ is finite,

b) the order of xx^y is finite,

then either n is equal to one of the numbers $1, 2, 3, 4, 6, 10$ and $\langle x, y\rangle$ is finite, or the case b) holds, $n = 14$ and the order of xx^y is equal to 7.

There exists an example which shows that, in Proposition 2, the alone condition of finiteness of order of xy cannot imply the finiteness of G. More exact, for every natural n not equal to 1, 2, 3, 4, 6, 10, there exists a subgroup $G = G(n)$ of $SL_2(\mathbb{C})$ such that:

a) G is generated by two elements of order 3 whose product is of order n,

b) G is infinite,

c) G acts freely on $\mathbb{C}^2$.

Moreover, there exists an infinite group which acts freely on an Abelian group and is generated by three elements of order 3 every pair of which generates a finite group.

Theorem 4.3 [33]. *Let G be a group of automorphisms of a non-trivial Abelian group A, $x \in G$ is an element of prime order p and, for every $g \in G$, the subgroup $\langle x, x^g \rangle$ is finite and acts freely on A. Then x is contained in a finite normal subgroup of G.*

More exactly, either $H = \langle x^G \rangle$ is cyclic, or one of the following conditions holds:

a) $p = 5$ and $H \simeq SL_2(5)$;

b) $p = 3$ and H is isomorphic to one of the groups $SL_2(3), SL_2(5)$.

Let S be a set of arbitrary (not necessarily finite) cardinality n and A_n the (locally finite) alternating group on S.

Theorem 4.4 [19]. *Let G be a group acting faithfully on an Abelian group V. Suppose that G is generated by a conjugacy class X of elements of order 3 such that every two non-commuting members of X generate a finite subgroup which acts freely on V. Then either G is Abelian, or G contains a central subgroup C of order 2 such that $G/C \simeq A_n$ for some cardinality n.*

On the other hand, for every cardinality $n \geq 3$ (finite or infinite) there exists an extension G of a group of order 2 by A_n such that every two elements of order 3 from pre-images in G of 3-cycles in A_n either commute, or generate a group which acts freely on V.

The first part of this result is a direct consequence of the following

Theorem 4.5 [19]. *Suppose that a group G contains a normal subset X of elements of order 3 such that, for every non-commuting $x, y \in X$, the subgroup $\langle x, y \rangle$ is isomorphic to A_4, or to A_5. Then $\langle x^G \rangle$ is locally finite and is isomorphic to*

the (discrete) direct product of groups every of which is isomorphic either to A_n for some cardinality n or to an extension of an elementary Abelian 2-group by a group of order 3.

5. SELFCENTRALIZING SUBGROUPS OF ORDER 3.

W.Feit and J.G.Thompson [6] obtained a description of finite groups containing a subgroup T of order 3 which coincides with its centralizer. We extend this results on arbitrary groups with the condition that T generates with every its conjugate a finite subgroup.

Theorem 5.1 [18]. *Suppose that a group G contains a subgroup T of order 3 such that $C_G(T) = T$. If, for every $g \in G$, the subgroup $\langle T, T^g \rangle$ is finite, then one of the following holds:*

1. *$G = NN_G(T)$ for a periodic nilpotent normal subgroup N of class 2 and NT is a Frobenius group with core N and complement T.*

2. *$G = NA$ where A is isomorphic to $A_5 \simeq SL_2(4)$ and N is a normal elementary Abelian 2-subgroup such that N is the direct product of subgroups of order 16 normal in G and isomorphic to the natural $SL_2(4)$-module of dimension 2 over a field of order 4.*

3. *G is isomorphic to $L_2(7)$.*

In particular, G is locally finite.

Our arguments are based on the using of the coset enumeration algorithm for a proof of finiteness of some finitely presented groups. All necessary calculations are made in GAP [26].

Let t be a generator of T and $X = t^G$. By the mentioned Feit-Thompson theorem, a group generated by any two members of X can be described up to defining relations. This gives a possibility to calculate orders of various groups generated by specially chosen triples of members of X and then to prove the finiteness of a group generated by a finite number of members of X.

The following example shows that the finiteness condition for subgroups $\langle T, T^g \rangle$ cannot be omitted.

Example. Let G be the periodic product of odd exponent $n \geq 665$ (see [1]) of two subgroups G_1, G_2 of order 3. Then G is an infinite simple group of exponent $3n$, $C_G(G_1) = G_1$, and G is not a locally finite group.

6. FROBENIUS GROUPS

Let G be a transitive permutation group on a (possibly, infinite) set Ω such that the stabilizer $H = G_\alpha$ of a point $\alpha \in \Omega$ is non-trivial but the stabilizer of every two distinct points is trivial. In particular, H is *detached* in G (this term was introduced by Yu.M. Gorchakov in [8]), that is H is a proper subgroup of G and $H \bigcap H^g = 1$ for every $g \notin H$. In V.P.Shunkov's notation [27], this means that (G, H) is a *Frobenius pair*. By famous result of G.Frobenius, if Ω is finite then H has a normal complement F in G consisting of trivial element and all elements in G which fix no points in Ω, and F is a regular subgroup, i.e. F is transitive and $F_\alpha = 1$. In this situation, the set of subgroups consisting of F and all $G_\beta, \beta \in \Omega$ form a *partition* of G, i.e. the set of proper subgroups having pair-wise trivial intersections and covering G. For infinite group G, the set $F = (G \setminus \bigcup\limits_{\alpha \in \Omega} G_\alpha) \cup \{1\}$ is not necessarily a subgroup (for example, if G is a free group of rank 2 and H is a maximal Abelian subgroup of its commutator subgroup, see [13]), and if F is a regular subgroup we call G a *Frobenius group*. Note, that this definition differs from the definition of Frobenius group by P.Neumann and P.Rowly [23] (our Frobenius group is a particular case of their split Frobenius group).

If G is a Frobenius group then

(a) F is a non-trivial proper normal subgroup of G, $G = FH$ and $F \cap H = 1$;

(b) $H \cap H^g = 1$ for every $g \in G \setminus H$;

(c) $F \setminus \{1\} = G \setminus (\bigcup_{g \in G} H^g) = \bigcap_{g \in G} (G \setminus H^g)$.

On the other hand, if F and H are subgroups of an arbitrary group G which satisfy the conditions (a)-(c) then one can easy to see that G acts by the right multiplication on the set Ω of all cosets $Hg, g \in G$, as a permutation Frobenius group and, after Shunkov [27], we call such G an (abstract) Frobenius group. The subgroup F is said to be the Frobenius core and H a Frobenius complement (in respect to the decomposition (a).)

For example, the semidirect product of the additive group of arbitrary skew-field

F and a non-trivial subgroup H of its multiplicative group such that H acts on F by multiplication from the right is a Frobenius group.

For finite Frobenius groups, the structure of the core F and a complement H is well-studied: F is nilpotent by Thompson theorem [29] and H is either soluble or contains a normal subgroup $N \simeq SL_2(5)$ such that H/N is metacyclic. Moreover, if H contains no elements of order 3 or 4 then H is super-soluble [30]. Thus elements of order 3 and 4 play a special role in the structure of Frobenius groups. For infinite group the situation differs radically. For example, every group can be embedded in the core of some Frobenius group, and every right orderable group is isomorphic to a complement of some Frobenius group [2]. The state changes, if a Frobenius complement contains elements of order 2 or 3.

Let G be a Frobenius group with core F and complement H. Then H *acts freely* on F by conjugation, i.e. $f^h = f$ for $f \in F, h \in H$ only if $f = 1$ or $h = 1$. Furthermore, H acts *co-freely*, i.e., for every non-trivial $h \in H$, the map $\phi_h : F \longrightarrow F$ with $\phi_h(f) = f^h f^{-1}$ is onto. On the other hand, if H is an automorphism group of a group F acting freely and co-freely then the natural semi-direct product FH is a Frobenius group with core F and complement H.

Suppose that a is an element of order 3 or 4 in H. Then it easy to prove that $b^{a^2} = b^{-1}(b^{-1})^a = b^{-1}b^{-a}$ or, respectively, $b^{a^2} = b^{-1}$ for every $b \in F$. Hence, a is a quadratic automorphism in the sense of the following definition.

An automorphism a of a group X is a *quadratic automorphism* if there exist integers $m = m(a), n = n(a)$ such that, for every $x \in X$, $x^{a^2} = x^n(x^m)^a = x^n x^{ma}$. If G is a Frobenius group we say that $g \in G$ is quadratic, if g induces in the core F by conjugation a quadratic automorphism.

Theorem 6.1 [35]. *A Frobenius group generated by two quadratic elements of finite order is finite and its core is Abelian.*

Corollary. *Let G be a Frobenius group generated by two elements of orders at most 4. Then G is finite and the core of G is Abelian.*

For a proof of Theorem 6.1, we study a co-free action of a cyclic group C on an Abelian group V and prove that C is finite if V is generated by a finite number of orbits of C. This fact is used also in a proof of the following

Theorem 6.2 [35]. *Suppose that a non-trivial group H acts co-freely on a*

soluble group F. If, for every nontrivial $h \in H$, F contains a finite subset $M = M(h)$ such that $F = \langle mh^i | m \in M, i \in \mathbb{Z} \rangle$ then F and H are finite.

Corollary 1. *A Frobenius group with finitely generated soluble core is finite.*

Corollary 2. *Let G be a Frobenius group with core F and complement H. Suppose that H is generated by elements of order 3 and the order of the product of any two elements of order 3 in H is finite. Then H is finite. If, under such conditions, G is finite generated then it is finite.*

REFERENCES

[1] Adian, S. I. Classifications of periodic words and their application in group theory. Burnside groups, Proc. Workshop, Bielefeld 1977, Lect. Notes Math. 806, 1-40 (1980).

[2] Bludov, V.V., On Frobenius groups. Siberian Math. J. 38, No.6 (1997), 1054-1056; translation from Siberian Mat. Zh. 38, No.6 (1997), 1219-1221.

[3] Brandl, R., Shi, W. J., The characterization of $PSL(2, q)$ by its element orders, J.Algebra, 163, No.1 (1994), 109-114.

[4] Brauer, R., Suzuki, M. and Wall, G.E. A characterization of the one-dimensional unimodular projective groups over finite fields. Illinois J. Math., 2, No.3 (1958), 718-742.

[5] Deng, H., Shi, W., The characterization of Ree groups ${}^2F_4(q)$ by their element orders. J.Algebra, 217, No.1 (1999), 180-187.

[6] Feit, W., Thompson, J. G., Finite groups which contain a self-centralizing subgroup of order 3 (English) Nagoya Math. J. 21, 185-197 (1962).

[7] Goldshmidt, D., Elements of order two in finite groups. Delta (Waukesha), 4 (1974/75), 45-58.

[8] Yu.M.Gorchakov. On infinite Frobenius groups. Sov. Math., Dokl. 4 (1963), 1397-1399.

[9] Gupta, N. D. and Mazurov, V. D., On groups with small orders of elements. Bull. Austral. Math. Soc., 60 (1999), 197-205.

[10] M.Hall Jr, Solution of the Burnside problem for exponent six. Illinois J. Math. 2 (1958), 764–786.

[11] Huppert, B., Endliche Gruppen I. Springer Verlag, 1979.

[12] Huppert, B. and Blackburn, N., Finite groups III. Springer Verlag, 1982.

[13] Kegel, O.H., Lectures on locally finite groups. Oxford, 1969.

[14] Levi, F. and Van der Waerden, B.L., Über eine besondere Klasse von Gruppen. Abh. Math. Semin. Hamburg Univ. 9 (1932), 154-158.

[15] Makarenko, E. N. Quadratic automorphisms of Abelian groups (to appear).

[16] Mazurov, V. D. Infinite groups with Abelian centralizers of involutions. Algebra and Logic, 39, No.1 (2000), 42-49.

[17] Mazurov, V. D., On groups of exponent 60 with prescribed orders of elements. Algebra and Logic, 39, No.3 (2000), 189-198.

[18] Mazurov, V. D., Groups containing a selfcentralizing subgroup of order 3. Algebra and Logic, 42, No.1 (2003), 29-36.

[19] Mazurov, V. D., A characterization of infinite alternating groups (to appear).

[20] Mazurov, V.D. and Churkin., V. A., About a group that acts freely on an Abelian group. Siberian Math. J., 42, No.4 (2001), 748-750.

[21] Mazurov, V. D. and Churkin, V. A., On a free action of a group on an Abelian group. Siberian Math. J., 43, No.3 (2002), 480-486.

[22] Neumann, B. H., Groups whose elements have bounded orders. J. London Math. Soc., 12 (1937), 195-198.

[23] Neumann, B.H. and Rowley, P.J., Free actions of Abelian groups on groups. Geometry and cohomology in group theory (Durham, 1994). London Math. Soc. Lecture Note Ser., 252, Cambridge Univ. Press, Cambridge, 1998, 291–295.

[24] Newman., M. F. Groups of exponent dividing seventy. Math. Scientist., 4 (1979), 149-157.

[25] Sanov, I. N. Solution of Burnside's problem for exponent 4 (Russian). Leningrad State Univ. Annals [Uchenye Zapiski] Math.Ser. 10 (1940), 166-170.

[26] Schönert, M., et al, Groups, Algorithms and Programming, Lehrstuhl D für Mathematik, RWTH Aachen, 1993.

[27] Shunkov, V. P. A nonsimplicity criterion for groups. Algebra Logic 14, No.5 (1975), 355-372 (1976).

[28] Suzuki, M. On characterizations of linear groups. I,II. Trans. Amer. Math. Soc., 92 (1959), 191-219.

[29] Thompson, J. G., Normal p-complements for finite groups. Math. Z., 72 (1959/1960), 332–354.

[30] Zassenhaus, H. Kennzeichnung endlicher linearen Gruppen als Permutationsgruppen. Abhandl. math. Semin. Univ. Hamburg, 11 (1936), 17-40.

[31] Zhurtov, A.Kh. On quadratic automorphisms of Abelian groups (in Russian). Algebra i logika, 39, No.3 (2000), 320-328.

[32] Zhurtov, A. Kh. On regular automorphisms of order 3 and Frobenius pairs (in Russian). Sibirskij Mat. Zhurn., 41, No.2 (2000), 329-338.

[33] Zhurtov, A. Kh. On a group acting freely on an Abelian group (in Russian). Sibirskij Mat. Zhurn., 43, No.2 (2002), 343-346.

[34] Zhurtov, A. Kh. and Mazurov, V. D., A recognition of simple groups $L_2(2^m)$ in the class of all groups. Siberian Math. J., 40, No.1 (1999), 42-44.

[35] Zhurtov, A. Kh. and Mazurov, V. D., On Frobenius groups generated by quadratic elements. Algebra and Logic, 42, N 3 (2003), 153-164.

INTEGRABLE NONLOCAL BURGERS EQUATION

P. Miškinis[1,2]
[1]NORDITA, Nordic Institute of Theoretical Physics,
Blegdamsvej 17, DK-2100 Cøpenhagen Ø, Denmark

[2]Department of Physics, Faculty of Fundamental Sciences,
Vilnius Gediminas Technical University, Saulėtekio Ave 11, LT-2040 Vilnius,
Lithuania
email:paulius.miskinis@fm.vtu.lt

Abstract

A nonlocal generalization of the one-dimensional Burgers equation is suggested. The explicit form of a particular analytical solution, existence of the travelling wave solution, interaction of nonlocal perturbation, asymptotic behaviour of solutions as well as the symmetries and conservation laws of the equation are considered. Its relation with the Burgers equation of integer order and the universal character of the Reynolds number are discussed.

1. INTRODUCTION

The Burgers equation initially had been proposed as a simple nonlinear partial differential equation in studies on turbulence: Burgers [1], Whitham [2]. This equation can be viewed as a simplified version of the Navier–Stokes equation and related to the heat equation via the Hopf–Cole transformation: Hopf [3], Cole [4]. Presently, the number of applications of the Burgers equation is immense (see, for instance: Gurbatov *et al.* [5], Woyczynski [6], Smaoui and Belgacem [7] and references below).

In the case when the properties of a system in a certain point of configuration or phase space depend not only on the properties of this system at this point, but also on the properties of at least one point of the environment, we deal with nonlocal phenomena. From the mathematical point of view, such phenomena are usually described by integro-differential equations: Agarwal and O'Regan [8]. Over the last few years more attention has been given to a special part of theory of integro-differential equations, the so-called fractional calculus: Miller and Ross [9], Oldham and Spanier [10], Samko, Kilbas and Marichev [11]. This approach is applied not only in the theory of fractals, but also for description of electrical, biological and diffusion phenomena. The latter topic, as follows from the growing number of publications, receives the bulk of attention: Podlubny [12].

Thus the nonlocal Burgers equation can arise in continuum mechanics as a result of the cumulative (memory) effects and be defined by considering the fractional powers of corresponding differential operators.

2. THE NONLOCAL BURGERS EQUATION AND ITS SOLUTIONS

Below, we consider some properties of the nonlocal Burgers equation (NBE), such as a nonlinear and nonlocal generalization of diffusion equation, based on a fractional generalization of the Hopf and Cole transformation:

$$\phi_t + \frac{1}{2}\,{}_aD_x^p\left({}_aD_x^{1-p}\phi\right)^2 - \alpha\phi_{xx} = 0\,, \tag{1}$$

in which $\phi(x,t)$, $\phi_0(x) \in \mathbb{R}\,, -\infty < x < +\infty\,;\ t \geq 0$ and the parameter $\alpha > 0$. ${}_aD_x^p$ is the fractional derivative in the sense of Riemann–Liouville [11]:

$${}_aD_x^p\phi(x,t) = \frac{1}{\Gamma(1-p)}\,\frac{d}{dx}\int_a^x \frac{\phi(\xi,t)}{(x-\xi)^p}\,d\xi\,, \tag{2}$$

in which $0 < p < 1$, and a is the parameter of nonlocality. From the physical point of view, we may consider this spatial fractional derivative as a Fourier transformation of the fractional power of the wave number k.

2.1. The Reynolds number and solutions of the nonlocal Burgers equation

The Reynolds number is a very convenient dimensionless quantity which is used in the nonlinear Burgers equation. For NBE (1) we may introduce a dimensionless generalization of the Reynolds number

$$\mathsf{Re} \sim \frac{{}_aD_x^p\left({}_aD_x^{1-p}\phi\right)^2}{\alpha\phi_{xx}} \sim \frac{\phi x^p}{\alpha}\,. \tag{3}$$

Note that in the case when $\mathsf{Re} \ll 1$, the influence of the nonlinear-nonlocal term is minimal and NBE (1) turns into an ordinary diffusion equation. This implies a deep correlation between the diffusion equation and the NBE.

Suppose the solution of the diffusion equation $w(x,t)$ is known. Then, due to a useful relation between the Burgers equation and diffusion equation at $p = 1$, we can express the solution of NBE (1) for any fractional or integer value of the parameter p:

$$\phi(x,t) = -2\alpha\,{}_aD_x^p \log w(x,t)\,. \tag{4}$$

2.1.1. An example

In the case when the solution of the diffusion equation is

$$w(x,t) = \exp\left(-\frac{cx}{2\alpha} + \frac{c^2 t}{4\alpha} - b\right), \tag{5}$$

the solution of NBE (1) at $a = 0$ is

$$\phi(x,t) = cx - \frac{c^2}{2}t + 2\alpha b\,, \qquad p = 0\,, \tag{6}$$

$$\phi(x,t) = c\,, \qquad p = 1\,, \tag{7}$$

$$\phi(x,t) = \frac{c}{\Gamma(2-p)}x^{1-p} + \frac{x^{-p}}{\Gamma(1-p)}\left(2\alpha b - \frac{c^2}{2}t\right), \qquad 0 < p < 1\,. \tag{8}$$

Note that solution (8) transforms from solution (6) into (7) when the parameter p runs from $p = 0$ to $p = 1$.

2.2 Evolution of the initial conditions

Let the initial conditions for NBE (1) and for the diffusion equation be related by the expression

$$w_0 = e^{-\frac{1}{2\alpha}\,{}_aD_y^{-p}\phi_0}\,. \tag{9}$$

This allows us to express the solution $\phi(x,t)$ of the NBE through the initial condition $\phi_0(x)$ and the general form of the solution of the diffusion equation

$$\phi(x,t) = -2\alpha\,{}_aD_x^p \log\frac{1}{\sqrt{4\pi\alpha t}}\int_{-\infty}^{+\infty} e^{-\frac{|x-y|^2}{4\alpha t} - \frac{1}{2\alpha}\,{}_aD_y^{-p}\phi_0(y)}dy\,. \tag{10}$$

As follows from NBE (1), depending on the values of the parameter p we deal with not one, but with an infinite number or a whole hierarchy of integro-differential equations. One of the most important properties of NBE (1) is interrelation between nonlinearity and nonlocality: for the fractional meaning of the parameter p we have the nonlinear-nonlocal and for the integer positive p only a nonlinear generalization of the Burgers equation. In this hierarchy, due to the substitution $\phi(x,t) \to {}_aD_x^{q-p}\phi(x,t)$, the low order equations turn into the higher order ones, but in the inverse direction this transformation is multivalued.

2.3 The travelling wave solution

Now consider a special type of the travelling wave solution when $\phi(x,t) = \phi(x-ut) \equiv \phi(\xi)$, where $\xi \equiv x - ut$. Then the NBE takes the form

$$\frac{1}{2}{}_aD_\xi^p \left[{}_aD_\xi^{1-p}\phi\right]^2 = \alpha\phi'' + u\phi' \,. \tag{11}$$

The travelling wave solution of (11), according to the above mentioned property, for $p > 1$ is

$$\phi(\xi) = {}_aD_\xi^{p-1}\psi(\xi) \tag{12}$$

where $\psi(\xi)$ is

$$\psi(\xi) = \psi_1 + \frac{\psi_2 - \psi_1}{1 + \exp\left(\frac{\psi_2-\psi_1}{2\alpha}\xi\right)} \,, \quad p = 1 \,, \tag{13}$$

with the asymptotics $\psi(\xi \to +\infty) = \psi_1 \,,\ \psi(\xi \to -\infty) = \psi_2 \,, \quad \psi_2 > \psi_1$. This means that NBE (1) can describe transition from one asymptotic state $\psi(-\infty) = \psi_2\xi^{1-p}$ to the other – $\psi(+\infty) = \psi_1\xi^{1-p}$, which takes place in the region $\Delta x = 2\alpha/(\psi_2 - \psi_1)$ in the presence of nonlinearity and nonlocality.

2.4 Interaction of nonlinear and nonlocal perturbations

The relation (4) between the solutions of NBE and the diffusion equation allows to consider an interaction of nonlocal and nonlinear perturbations. Two or more perturbations moving with a different velocity can overtake each other or flow together into a new intensive perturbation. The NBE also describes the interaction process of two or more moving nonlocal perturbations. The principle of superposition is not valid for the nonlinear NBE, but it is valid for the linear diffusion equation. The fractional Hopf–Cole transformation (4) interrelates the solutions of nonlocal and nonlinear NBE and of linear diffusion equation. Thus, if $w_i(x,t)$ are the solutions of diffusion equation, then $\phi(x,t) = -2\alpha\,{}_aD_x^p(\log\sum w_i)$ are the solutions of NBE.

For instance, for two solutions of the diffusion equation in the form (5) we obtain a nonlocal and nonlinear interaction of these perturbations $-\phi(x,t)/2\alpha =$

$$\log\left(w_1 + w_2\right) \,, \quad p = 0 \,, \tag{14}$$

$$\frac{1}{\Gamma(-p)}\int_a^x \frac{\log\left[w_1(\xi,t) + w_2(\xi,t)\right]}{(x-\xi)^{(p+1)}}\,d\xi \,, \ 0 < p < 1 \,, \tag{15}$$

$$\frac{c_1w_1 + c_2w_2}{w_1 + w_2} \,, \quad p = 1 \,. \tag{16}$$

3. THE CONSERVATION LAWS

In the case of the usual BE, for $x \in E^1, \forall\, t > 0, \phi(\pm, t) = \phi_x(\pm\infty, t) = 0$, we have a conservation value of

$$I^{(1)} = \int_{-\infty}^{+\infty} \phi(x,t)\, dx = inv\,, \tag{17}$$

since

$$\frac{\partial I^{(1)}}{\partial t} = \int_{-\infty}^{+\infty} \left[\alpha\phi_x - \frac{1}{2}\phi^2\right]_x dx = \left(\alpha\phi_x - \frac{1}{2}\phi^2\right)\Big|_{-\infty}^{+\infty} = 0\,. \tag{18}$$

In the applications, this conservation law is called the "mass" conservation law, because $\phi(x,t)$ can be a one-dimensional density or gradient of any physical, chemical or biological magnitude.

In the case of NBE (1), if $\forall\, t > 0, \phi_x(\pm\infty, t) = \phi_{xx}(\pm\infty, t) = 0$, we again deal with the conservation value:

$$I^{(0)} = \phi(+\infty, t) - \phi(-\infty, t) = inv\,, \tag{19}$$

since by applying the derivative ∂_x to the evolutionary equation (1) followed by integration we obtain

$$\frac{\partial I^{(0)}}{\partial t} = \frac{\partial}{\partial t}\int_{-\infty}^{+\infty} \phi_x\, dx = \left(\alpha\phi_{xx} - \frac{1}{2}\phi_x^2\right)\Big|_{-\infty}^{+\infty} = 0\,. \tag{20}$$

This conservation value shows that the difference in asymptotic values for any time moment remains unchanged. If, for instance, we deal with the evolution of the potentials, the conservation value $I^{(0)}$ (19) shows that the difference of potentials for $x \to \pm\infty$ does not change.

In the case of NBE (1), if $\forall\, t > 0, {}_aD_x^{2-p}\phi(\pm\infty, t) = {}_aD_x^{1-p}\phi(\pm\infty, t) = 0$, again we deal with a conservation value:

$$I^{(p)} = \int_{-\infty}^{+\infty} {}_aD_x^{1-p}\phi(x,t)\, dx = inv\,, \tag{21}$$

as

$$\frac{\partial I^{(p)}}{\partial t} = \int_{-\infty}^{+\infty} \left[\alpha_a D_x^{2-p}\phi - \frac{1}{2}({}_aD_x^{1-p}\phi)^2\right]_x dx = \left[\alpha_a D_x^{2-p}\phi - \frac{1}{2}({}_aD_x^{1-p}\phi)^2\right]\Big|_{-\infty}^{+\infty} = 0\,. \tag{22}$$

Even this simple example highlights two important properties of the nonlocal conservation law (21): it interrelates the conservation values of two different dynamical systems, which can be of different mathematical nature (*e.g.*, in our case these values are integral and discrete).

Note that in the "common" case of the nonlocal BE

$$\phi_t + \phi\phi_x - \alpha_a D_x^{2-p}\phi = 0\,, \tag{23}$$

an analogous conservation integral exists at other asymptotic values:

$$\phi(\pm\infty, t) = {}_aD_x^{1-p}\phi(\pm\infty, t) = 0\,. \tag{24}$$

At this point, it is not difficult to characterize the "mass" conservation law of the nonlinear nonlocal evolution equation as

$$\phi_t + \frac{1}{2}{}_aD_x^p\left({}_aD_x^{1-p}\phi\right)^2 - \alpha_a D_x^{2-q}\phi = 0\,, \tag{25}$$

The magnitude $I^{(p,q)}$ is the invariant of the evolution equation (25)

$$I^{(p,q)} = \int_{-\infty}^{+\infty} {}_aD_x^{1-p}\phi(x,t)\,dx = inv\,, \tag{26}$$

for ${}_aD_x^{1-p}\phi(\pm\infty, t) = {}_aD_x^{2-(p+q)}\phi(\pm\infty, t) = 0$. However, in this case the integrability is sacrificed: the fractional generalization of the Hopf–Cole transformation does not exist.

The "energy" of travelling excitation

$$K = \frac{1}{2}\int_{-\infty}^{+\infty}\left({}_aD_x^{1-p}\phi\right)^2\,dx\,, \tag{27}$$

if $\forall t > 0, {}_aD_x^{2-p}\phi(\pm\infty, t) = {}_aD_x^{1-p}\phi(\pm\infty, t) = 0$, as in the case of BE, is not in variable, but it is constantly decreasing:

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\int_{-\infty}^{+\infty}\left({}_aD_x^{1-p}\phi\right)^2\,dx &= -\frac{1}{3}\left({}_aD_x^{1-p}\phi\right)^3\Big|_{-\infty}^{+\infty} + \\ +\alpha\left({}_aD_x^{1-p}\phi\right)\left({}_aD_x^{2-p}\phi\right)\Big|_{-\infty}^{+\infty} - \alpha\int_{-\infty}^{+\infty}\left({}_aD_x^{2-p}\phi\right)\cdot\left({}_aD_x^{2-p}\phi\right)\,dx &= \\ = -\alpha\int_{-\infty}^{+\infty}\left({}_aD_x^{2-p}\phi\right)^2\,dx < 0\,.\end{aligned} \tag{28}$$

Like in the case of the nonlocal "mass" conservation law, the "energy" K (27) links the energy of the travelling excitation in the case of BE ($p = 1$) and the one-dimensional density of energy in the case of NDE ($p = 0$).

3.1 The asymptotic form of the solutions

The momentum conservation law predetermines the asymptotic form of the NBE solution. To arrive to such a result, let us consider some estimates. From the general form of NBE solution (11) it follows that the limit $t \to +\infty$ corresponds to a rather low value of the parameter α. At a low α, to calculate the values of the corresponding integrals we can apply the saddle point approximation.

The critical point y_0 can be determined from the equation

$$\frac{y_0 - x}{t} + {}_aD_y^{1-p}\phi_0(y_0) = 0\,. \tag{29}$$

Then the asymptotic expression of the NBE solution acquires the form

$$\phi(x,t) = {}_aD_x^p\left[\frac{(x-y_0)^2}{2t} + C\right] \sim \frac{(x-a)^{2-p}}{t\Gamma(3-p)} + C_1(x-a)^{-p}\,. \tag{30}$$

For $x \to +\infty$ and $p > 2$ the solution $\phi(x,t) \to 0$, and for $p < 2$ the solution $\phi(x,t) \to (x-a)^{2-p}/[t\Gamma(3-p)]$. Thus we obtain a power-deformed perturbation of the usual solution of the Burgers equation . Note here that these estimates are valid not only in the environment of the meaning $p = 1$, but also for any $p \in \mathbb{R}$.

We have to show the region of the validity of solution (30). In both its limit cases the integral in the expression of the momentum conservation law diverges. Therefore, for $x > x_0$ the solution $\phi(x) \equiv 0$. To determine the value x_0, we insert the asymptotic form of solution (30) in the expression of the momentum conservation law. This means that $x_0^2/2t \sim I$. Thus the maximum meaning of the solution

$$\phi_{max}(x,t) \sim \frac{I^{1-\frac{p}{2}}}{(2t)^{\frac{p}{2}}} \quad \text{and} \quad x_0 \sim \sqrt{2It}\,. \tag{31}$$

3.2 The width and spectrum of the nonlocal shock wave

The Reynolds number or its powers express various regimes of the physical processes. As an example, consider the determination of the shock wave width, i.e. the width of the region Δx where smoothing of the single perturbation near the border of the wave turnover takes place. The influence of the nonlinear-nonlocal

term ${}_aD_x^p\,({}_aD_x^{1-p}\phi)^2$ and of the dissipative term $\alpha\phi_{xx}$ on the interval Δx becomes equal. Therefore,

$$\frac{\phi(\Delta x)^p}{\alpha} \sim 1\,, \tag{32}$$

and using the universal expressions of the Reynolds number

$$\mathsf{Re} = 2I/\alpha\,, \tag{33}$$

which is valid for any integer or fractional value of parameter p and expression (31) we have

$$\frac{\Delta x}{x_0} \sim \frac{1}{\mathsf{Re}^{1/p}}\,. \tag{34}$$

In other words, the relative width of the nonlocal shock wave front is the inverse power of $\mathsf{Re}^{1/p}$.

The coefficients of the Fourier transformation of the asymptotic solution of the nonlocal shock wave propagation decrease as $1/k^{(3-p)}$ under an increase of the wave number k, therefore the spectrum of ϕ_k does not cut off effectively at any meaning of k. The turnover perturbation cutoff is effective under condition that $k\Delta x \gg 1$. The effective width of the space spectrum

$$k_0 \sim \frac{1}{\Delta x}\,. \tag{35}$$

The wave number Δk corresponding to the wave length x_0 is $\Delta k = 2\pi/x_0$. Therefore the dimensionless width of the wave perturbation

$$\frac{k_0}{\Delta k} \sim \frac{x_0}{\Delta x} = \mathsf{Re}^{1/p}\,. \tag{36}$$

Equation (36) actually determines a certain number of single perturbation degrees of freedom, i.e. the number of the space modes that format the wave packet of a single perturbation periodically repeated in the space direction.

Such interpretation of the Reynolds number as a number of the perturbation degrees of freedom may look somewhat artificial. Nevertheless, this interpretation is universal for a wide class of systems and not only for dissipative ones and, as we see now, can be extended to nonlocal systems too.

3.3 The symmetries

Note here a property that allows us to get new solutions of the NBE. Let $v(x,t)$ be a known solution of NBE (1), and $u(x,t)$ is the solution of the linear equation

$$u_t + \left({}_aD_x^{1-p}v\right)\,u_x - \alpha u_{xx} = 0. \tag{37}$$

Then

$$\phi(x,t) = -2\alpha_a D_x^p \log u + v \tag{38}$$

is a new solution of NBE (1). At an integer meaning of p we have a local equation (37), in particular, for $p = 1$ we obtain a new solution of the Burgers equation .

An important and not yet clear problem is the nonlocal and nonclassical symmetries of NBE. Some general aspects have already been presented in Abraham-Shrauner and Guo [13]. Promising seems an attempt of computer symmetry analysis, as was done for nonlinear heat equation in Clarcson and Mansfield [14]. Symmetries in the discrete Burgers equation have been studied in Hernandez, Lavi and Winternitz [15].

Note here that the existence of a transformation T expressed by relation (4) allows to solve the problem of symmetry group of NBE (1). Let G_1 be a group of symmetry of the diffusion equation, then $G = TG_1T^{-1}$ is a symmetry group of NBE (1).

4. THE SUPERSYMMETRIC NONLINEAR NONLOCAL DIFFUSION EVOLUTION EQUATION

We shall show that the NNDE has a supersymmtric generalization. Let the superfield $\chi = \theta_a D_x^{1-p}\phi + \psi$ unite two fields of different properties: the "bosonic" field $\phi(x,t)$ and its spinor superpartner $\psi(x,t)$; θ is the constant Majorana spinor. The transformations of the fields ϕ, ψ, are nonlocal because of the fractional derivatives ${}_aD_x^p f(x)$.

$$\delta_\eta \psi = \eta_a D_x^{1-p}\phi\,, \qquad \delta_{\eta a} D_x^{1-p}\phi = \eta\psi_x\,. \tag{39}$$

However, the commutator of the two transformations (39) are a spatial translation:

$$\left[\delta_\eta, \delta_\xi\right] = 2\xi\eta\partial_x\,. \tag{40}$$

The supersymmetric equation

$$\chi_t = (\chi_x + \frac{1}{2}\chi\mathcal{D}\chi)_x\,, \tag{41}$$

(here $\mathcal{D} = \theta\partial_x + \partial_\theta$ is a supersymmetric derivative) is a system of two evolutionary equations,

$$\psi_t = \psi_{xx} + \frac{1}{2}\left({}_aD_x^{1-p}\phi\cdot\psi\right)_x\,, \qquad \phi_t = \phi_{xx} + \frac{1}{2}{}_aD_x^p\left[\left({}_aD_x^{1-p}\phi\right)^2 - \psi\psi_x\right]\,, \tag{42}$$

which is invariant with respect to the supertransformations (39). In the general case, the system (42) is a system of two nonlinear nonlocal evolution equations, which becomes local when

$$p = 0: \qquad \psi_t = \psi_{xx} + \frac{1}{2}\left(\phi_x \psi\right)_x , \qquad \phi_t = \phi_{xx} + \frac{1}{2}\left(\phi_x^2 - \psi\psi_x\right) ; \tag{43}$$

$$p = 1: \qquad \psi_t = \psi_{xx} + \frac{1}{2}\left(\phi\psi\right)_x , \qquad \phi_t = \phi_{xx} + \frac{1}{2}\left(\phi^2 - \psi\psi_x\right)_x . \tag{44}$$

The supersymmetric equation (41) and the corresponding system of equations (42) unites two fields of different nature, and only one of them is nonlocal.

Here to the point is one general note related to the application of the nonlocal systems. Suppose that a dynamic system is characterized by two interacting fields, one of which for instance, the "fermionic" field $\psi(x,t)$, can be measured in the course of experiment, whereas the other, "bosonic" field $\phi(x,t)$ is assessed only phenomenologically. Such assessment in the class of local evolution equations results in a qualitatively erroneous mathematical model of a dynamic system.

5. CONCLUSIONS AND DISCUSSION

It is important to note that the influence of nonlocality can be arbitrarily large. Therefore we do not describe nonlocality by an additional term in the Burgers equation.

Remind here that the classical Burgers equation belongs to a unique group of the three completely integrable second-order PDEs. I suggest that the NBE also belongs to a unique group of the completely integrable nonlocal PDEs of fractional order.

The fractional diffusion process is related to the non-Gaussian statistics which leads to slow diffusion correlators $\langle(\Delta x)^2\rangle \propto Dt^\gamma$ with $\gamma \neq 1$, and D is a generalized diffusion coefficient of the dimension L^2/T^γ. In our case, the NBE is related to the so-called Lévi statistics: Shlesinger, Zaslavsky and Frisch [16]; at the same time the initial Burgers equation as well as the diffusion equation are related to the usual Gauss statistics.

From our point of view, the NBE has at least two important topics:

i) the influence of nonlocality is not assumed to be insignificant;

ii) the relation of the NBE with the usual diffusion equation allows a lot of analytical solutions of the NBE.

Besides, despite the nonlocality in the proposed nonlinear and nonlocal NBE,

• space-localized solutions are possible;

• nonlocal perturbations of a system described by the NBE can interact;

• the Reynolds number is a universal dimensionless parameter both for the local and nonlocal Burgers equation;

• there are analogies of the momentum conservation law and dissipation of kinetic energy .

In some fields of physics we need the vectorial form of the Burgers equation, e.g. in astrophysics to describe the large-scale structure of the Universe: Shandarin and Zeldovich [17], Miskinis [18]. In such cases the vectorial NBE can be proposed:

$$\boldsymbol{\phi}_t + \frac{1}{2}\,{}_{\alpha}\boldsymbol{D}^{p}_{x}\left({}_{\alpha}\boldsymbol{D}^{1-p}_{x}\boldsymbol{\phi}\right)^2 - \alpha\boldsymbol{\nabla}(\boldsymbol{\nabla}\boldsymbol{\phi}) = 0\,, \tag{45}$$

where $\boldsymbol{\phi} = (\phi^1, \ldots, \phi^n) \in \mathbb{R}^n$, ${}_{\alpha}\boldsymbol{D}^{p}_{x}$ is the fractional generalization of the gradient operator $\boldsymbol{\nabla}$.

Note also that for $\alpha = 0$ from NBE (1) follows the fractional generalization of the Riemann equation, which also has numerous applications.

The proposed NBE, because of its general character, allows a wide range of applications. Actually we may try to introduce the nonlocal generalization in almost all fields where the BE is applied. These are the nonlocal effects in shock wave propagation in acoustics, the effective model of the process of nonlinear heat distribution in the environment in the presence of heat sources and sinks, the Kardar–Parisi–Zhang (KPZ) equation in the crystal growth phenomena in (1+1)-dimensions: Kardar, Parisi and Zhang [19], the nonlinear dynamics of moving line: Hwa and Kardar [20], galaxy formation: Shandarin and Zeldovich [17], Miškinis [18], behaviour of the magnetic flux line in superconductor: Hwa [21], and spin glasses: Fisher and Huse [22], as well as numerous examples of the application of the usual Burgers equation presented in the above mentioned monographs Gurbatov *et al.* [5], Woyczynski [6].

5. ACKNOWLEDGEMENTS

The author would like to express his most cordial thanks for financial support and the possibility to partake at the International Conference on Mathematics and its Applications (ICMA 2004).

REFERENCES

[1] Burgers, J.M. *The nonlinear diffusion equation* Dordrecht, Reidel, 1974.

[2] Whitham, G.B. *Lienear and nonlinear waves* Wiley-Int. Publ., New York, 1974.

[3] Hopf, E., Comm. Pure Appl. Math., 3, 201 (1950).

[4] Cole, J.D., Q. Appl. Math., 9, 225 (1951).

[5] Gurbatov, S.N., Malakhov, A.N. and Saichev, A.I., *Nonlinera Random Waves and Turbulence in Non-dissipative Media: Waves, Rays, Particles* Mancester Univ. Press, 1991.

[6] Woyczynski, W.A., *Burgers-KPZ Turbulence* Springer, New York, 1998.

[7] Smaoui, N. and Belgacem, F. Journal of Applied Mathematics and Stochastic Analysis. 15, 1 (2002) 57-75.

[8] Agarwal, R.P. and O'Regan, D., *Integral and Integrodifferential Equations Theory, Methods and Applications* G & B Science Pub., Amsterdam, 2000).

[9] Oldham, K.B. and Spanier, J., *The fractional calculus* Acad. Press, New York and London, 1974.

[10] Miller, K.S. and Ross, B., *An Introduction to the Fractional Calculus and Fractional Differential Equations,* John Wiley, New York, 1996.

[11] Samko, S.G., Kilbas A.A. and Marichev, O.I. *Fractional integrals and derivatives. Theory and applications* (Gordon and Breach, Amsterdam, 1993).

[12] Podlubny, I. *Fractional differential Equations, Ser. Math. Science and Engineering* Acad. Press, San Diego, 1999.

[13] Abraham-Shrauner, B. and Guo, A.: *in Modern Goup Analysis: Advanced Analytical and Computational Methods in Mathematical Physics*, Kluver Acad Pb, Dordrecht, 1993, 1–5.

[14] Clarcson, P.A. and Mansfield, E.L. *Advanced Analytical and Computational Methods in Mathematical Physics* (Kluver Acad PbCo, Dordrecht, 155–171, 1993).

[15] Hernandez, R.H., Lavi D. and Winternitz, P. *in SIDE III - Symmetry and integrabilities of Differential equations* (AMS, Providence, 1999).

[16] Shlesinger, M.F., Zaslavsky, G.M. and Frisch, U. *Lévi flights and related topics in Physics*, Lecture Notes in Physics, Vol. 450 Springer-Verlag, Berlin, 1995.

[17] Shandarin, S.F. and Zeldovich, Ya.B. Rev. Mod. Phys., 61, 185 (1989).

[18] Miškinis, P. The Astrophysical Journal, 543, L95 (2000).

[19] Kardar, M., Parisi, G. and Zhang, Z. Phys. Rev. Lett., 56, 889 (1986).

[20] Hwa, T. and Kardar, M. Phys. Rev. A, 45, 7002 (1992).

[21] Hwa, T. Phys. Rev. Lett., 69, 1552 (1992).

[22] Fisher, D.S. and Huse, D.A. Phys. Rev. B, 43, 10728 (1991).

UNIVALENCY OF A CONVOLUTION OPERATOR CONCERNING HYPERGEOMETRIC FUNCTIONS

S. Naik and S. Ponnusamy
Department of Mathematics,
Indian Institute of Technology, IIT-Madras, Chennai- 600 036, India.
email: samy@iitm.ac.in

Abstract

Main objective of this article is to discuss the univalency of the function defined by the integral operator of the form

$$\Lambda_f(z) = \int_0^1 t^{b-1}(1-t)^{c-a-b}\varphi(1-t)\frac{f(tz)}{t}\,dt$$

under suitable restrictions on the parameters a, b, c and the functions $f(z)$ and $\phi(t)$.

1. INTRODUCTION

Let $\mathcal{A}$ denote the class of all analytic functions f in the unit disk $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ with the normalization $f(0) = 0 = f'(0) - 1$. For $\beta < 1$, let

$$\mathcal{P}_1(\beta) = \{f \in \mathcal{A} : \mathrm{Re}\,\{e^{i\eta}(f'(z) - \beta)\} > 0, \quad z \in \Delta\},$$

and

$$\mathcal{P}(\beta) = \{f \in \mathcal{A} : \mathrm{Re}\,\{e^{i\eta}(\frac{f(z)}{z} - \beta)\} > 0, \quad z \in \Delta\}$$

for some $\eta \in \mathbb{R}$. For $0 \le \beta < 1$, functions in $\mathcal{P}_1(\beta)$ are known to be univalent in Δ whereas functions in $\mathcal{P}(\beta)$ are not necessarily univalent in Δ (see [2]). For $f \in \mathcal{A}$, the object of our study concerns the integral transform

$$V_\lambda(f)(z) = \int_0^1 \lambda(t)\frac{f(tz)}{t}\,dt.$$

Here $\lambda(t)$ is a real valued nonnegative weight function normalized so that $\int_0^1 \lambda(t)\,dt = 1$. This operator contains some of the well known operators such as Libera, Bernardi, Carlson-Shaffer and Komatu as its special cases (see [8]). This operator has been studied by a number of authors for various choices of $\lambda(t)$ which includes these special cases [1, 3, 6, 7, 8, 9]. The most interesting choice concerns the integral representation for $F(a, b; c; z)$ and hence for $H_{a,b;c}(f(z)) := zF(a, b; c; z) * f(z)$, where

$f \in \mathcal{A}$ and $*$ denotes the usual convolution/Hadamard product between two power series represented by two functions. We recall that [1]

$$H_{a,b;c}(f(z)) := [H_{a,b;c}(f)](z) = \int_0^1 \lambda(t)\frac{f(tz)}{t}\,dt \tag{1}$$

where

$$\lambda(t) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c-a-b+1)}t^{b-1}(1-t)^{c-a-b}F\left(\begin{array}{c} c-a, \quad 1-a \\ c-a-b+1 \end{array};1-t\right).$$

In [1], the authors studied this integral in detail and, in particular, concerning its convexity, and starlikeness properties. The so called hypergeometric integral operators were studied earlier by Love, Kalla etc. e.g. see V. Kiryakova, Generalized Fractional Calculus and Applications, Longman - J. Wiley, New York, 1994. We introduce an auxiliary function

$$\varphi(1-t) = 1 + \sum_{m=1}^{\infty} b_m(1-t)^m$$

and consider $P_{a,b,c}(f)(z) := P(f(z))$ defined by

$$P_{a,b,c}(f)(z) := C\int_0^1 (1-t)^{c-a-b}t^{b-1}\varphi(1-t)\frac{f(tz)}{t}\,dt \tag{2}$$

where C is a normalized constant so that

$$C\int_0^1 \lambda(t)\,dt = 1, \tag{3}$$

where

$$\lambda(t) = t^{b-1}(1-t)^{c-a-b}\varphi(1-t).$$

We are interested in obtaining conditions so that $P_{a,b,c}(f)(z) \in \mathcal{P}_1(\beta')$, whenever $f \in \mathcal{P}(\beta)$.

2. DISCUSSION FOR OUR MAIN RESULTS

To proof our results we need the following lemma.

Lemma 1 [10] If f, g are analytic in unit the disc Δ and ϕ, ψ are convex (need not be normalized) functions in Δ such that $f \prec \phi$, $g \prec \psi$, then $f * g \prec \phi * \psi$.

In Lemma 1, $\prec$ denotes subordination (see [2]).

Let $P_{a,b,c}(f)(z) := P(f(z))$ be given by (2) and $f(z) = z + \sum_{n=2}^{\infty} a_n z^n$. Then, with $a_1 = 1$, we have

$$\begin{aligned} P'(f(z)) &= C\int_0^1 t^{b-2}(1-t)^{c-a-b}\varphi(1-t)\sum_{n=1}^{\infty} n a_n t^n z^{n-1}\,dt \\ &= C\int_0^1 t^{b-2}(1-t)^{c-a-b}\varphi(1-t)\sum_{n=1}^{\infty} n t^n z^{n-1} * \sum_{n=1}^{\infty} a_n z^{n-1}\,dt \\ &= \frac{f(z)}{z} * M(z) \end{aligned} \tag{4}$$

where

$$\begin{aligned} M(z) &= C\int_0^1 t^{b-2}(1-t)^{c-a-b}\varphi(1-t)\sum_{n=0}^{\infty}(n+1)t^{n+1}z^n\,dt \\ &= C\sum_{n=0}^{\infty}(n+1)z^n\int_0^1 t^{n+b-1}(1-t)^{c-a-b}\sum_{m=0}^{\infty} b_m(1-t)^m\,dt \quad (b_0=1) \\ &= C\sum_{n=0}^{\infty}(n+1)z^n\sum_{m=0}^{\infty} b_m\int_0^1 t^{n+b-1}(1-t)^{m+c-a-b}\,dt \\ &= C\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(n+1)z^n b_m\frac{\Gamma(n+b)\Gamma(m+c-a-b+1)}{\Gamma(n+m+c-a+1)} \\ &= C\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(n+b)z^n b_m\frac{\Gamma(n+b)\Gamma(m+c-a-b+1)}{\Gamma(n+m+c-a+1)} \\ &\quad +(1-b)C\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} z^n b_m\frac{\Gamma(n+b)\Gamma(m+c-a-b+1)}{\Gamma(n+m+c-a+1)}. \end{aligned}$$

After a simple calculation we see that first double summation equals

$$\int_0^1 t^b(1-t)^{c-a-b-1}\frac{1}{1-tz}\sum_{m=0}^{\infty} b_m(1-t)^m(m+c-a-b)\,dt$$

whereas second double summation equals

$$\int_0^1 t^{b-1}(1-t)^{c-a-b}\frac{1}{1-tz}\sum_{m=0}^{\infty} b_m(1-t)^m\,dt.$$

In view of the above observations, we have

$$\begin{aligned} M(z) = C\int_0^1 t^{b-1}(1-t)^{c-a-b-1}\frac{1}{1-tz}\sum_{m=0}^{\infty}[b_m(1-t)^m \quad \times \\ \{(m+c-a-b)t-(b-1)(1-t)\}]dt \end{aligned} \tag{5}$$

which after a simple calculation, gives the representation

$$M(z) = C \int_0^1 t^{b-1}(1-t)^{c-a-b-1} \frac{1}{1-tz} \left[\sum_{m=0}^{\infty} R_m (1-t)^{m+1} + (c-a-b) \right] dt$$

where

$$R(m) = b_{m+1}(m+c+1-a-b) - b_m(m+c-a-1).$$

The above discussion gives the following result.

Theorem 2. Let $c+1-a-b > 0$ and $0 < b \leq 2$. Suppose $\varphi(t) = \sum_{m=0}^{\infty} b_m t^m$ with $b_0 = 1$, and $\{b_m\}_{m\geq 1}$ is an increasing sequence of non-negative real numbers. Suppose that $f(z) \in \mathcal{P}(\beta)$. Then $P_{a,b,c}(f)$ defined by (2) is in $\mathcal{P}_1(\gamma)$, where $\gamma = 1 - 2(1-\beta)(1-\beta')$ with

$$\beta' = C \int_0^1 (1-t)^{c-a-b} t^{b-2} (1+t)^{-2} \varphi(1-t)\, dt.$$

where C is given by (3).

Proof. . Let $c+1-a-b > 0$ and $0 < b \leq 2$. Assume that $b_0 = 1$, and $\{b_m\}_{m\geq 1}$ is an increasing sequence of non-negative real numbers. Then $R(m)$ defined above gives

$$R(m) \geq b_m(m+c+1-a-b) - b_m(m+c-a-1) = b_m(2-b)$$

which is nonnegative if $0 < b \leq 2$. Thus, we have $\operatorname{Re} M(z) > M(-1)$ and the estimate here is clearly sharp. Finally, we assume that $f(z) \in \mathcal{P}(\beta)$. Now, we choose

$$\phi(z) = 1 + 2(1-\beta)\frac{z}{1-z} \quad \text{and} \quad \psi(z) = 1 + 2(1-\beta')\frac{z}{1-z}.$$

Both $\phi(z)$ and $\psi(z)$ are known to be convex in Δ. Further,

$$(\phi * \psi)(z) = 1 + 4(1-\beta)(1-\beta')\frac{z}{1-z} = 1 + 2(1-\gamma)\frac{z}{1-z}$$

and the desired conclusion follows from Lemma 1 and (4). □

Similarly, the following result can be proved.

Theorem 3. Let $a > 0$, $0 < b \leq 1$, and $c > a+1$. Suppose $\varphi(t) = \sum_{m=0}^{\infty} b_m t^m$ with $b_0 = 1$, and $\{b_m\}_{m\geq 1}$ is a sequence of non-negative real numbers. Suppose that $f(z) \in \mathcal{P}(\beta)$. Then P defined by (2) is in $\mathcal{P}_1(\gamma)$ where $\gamma = 1 - 2(1-\beta)(1-\beta')$ with

$$\beta' = C \int_0^1 (1-t)^{c-a-b} t^{b-2} (1+t)^{-2} \varphi(1-t)\, dt.$$

where C is given by (3).

Proof. By hypothesis, we have $c - a - b > 0$. The desired conclusion follows easily from (3) and (5), since the square bracketed term in (5) is non-negative for all $0 \leq t \leq 1$.

REFERENCES

[1] Balasubramanian, R., Ponnusamy, S. and Vuorinen, M., On hypergeometric functions and function spaces, *J. Computational and Applied Maths.* 139 (2) (2002), 299-322.

[2] Duren, P.L. Univalent Functions, Springer-Verlag, Berlin-New York, 1983.

[3] Fournier, R. and St. Ruscheweyh, On two extremal problems related to univalent functions, *Rocky Mountain J. of Math.* 24 (2) (1994), 529-538.

[4] Goodman, A. W. Univalent Functions, Vol. II. Mariner Publishing Co., Inc., Tampa, FL, 1983.

[5] Miller, S. S. and Mocanu, P.T., Differential subordinations: Theory and Applications, Marcel Dekker, Inc. New York Basel, No. 225, 2000.

[6] Ponnusamy, S. Differential subordination concerning starlike functions, *Proc. Ind. Acad. Sci.* (*Math. Sci.*) (2) 104 (1994), 397-411.

[7] Ponnusamy, S. Inclusion theorems for convolution product of second order polylogarithms and functions with the derivative in a halfplane, *Rocky Mountain J. Math.* 28 (2)(1998), 695-733 (Also Reports of the Department of Mathematics, Preprint 92, 1995, University of Helsinki, Finland).

[8] Ponnusamy, S. and Rønning, F., Duality for Hadamard products applied to certain integral transforms, *Complex Variables: Theory and Appl.* 32 (1997), 263-287.

[9] Ponnusamy, S. and Sabapathy, S., Polylogarithms in the theory of univalent functions, *Results in Mathematics.* 30 (1996), 136-150.

[10] St. Ruscheweyh and Stankiewicz, J., Subordination and convex univalent functions, *Bull. Pol. Acad. Sci. Math.,* 33 (1985), 499-502.

ON UNIFIED ELLIPTIC-TYPE INTEGRALS

M. A. Pathan
Department of Mathematics,
Aligarh Muslim University, Aligarh-202002, India
email: mapathan@postmark.net

Abstract

Various generalizations of certain families of elliptic-type integrals are studied in a number of earlier works on the subject due to their importance for possible applications in certain problems arising in physics and nuclear technology. The object of this paper is to present certain theorems on generating functions and to show how easily these theorems can be associated with the families of elliptic-type integrals in a unified and generalized form. Many single and double integrals of Euler-type which are seemingly relevant to the present investigation, are also considered here.

Keywords and Phrases: Elliptic-type integrals, Euler-type integrals, Lauricella functions, and Appell's functions.
AMS Subject Classification: Primary 33C55, Secondary 33E05.

1. INTRODUCTION

Let

$$(k) \;=\; \int_0^{\pi/2} \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}} \quad (k^2<1) \tag{1}$$

$$E(k) \;=\; \int_0^{\pi/2} \sqrt{1-k^2\sin^2\theta}\, d\theta \quad (k^2<1) \tag{2}$$

and

$$u(k) \;=\; \int_0^{\pi} \frac{d\phi'}{\sqrt{1-k^2\sin^2\phi'}} \quad (k^2<1) \tag{3}$$

where $K(k)$ and $E(k)$ denote the **complete** elliptic integrals of the first and second kinds, respectively and $u(k)$ denotes the **incomplete** elliptic integrals of the first kind.

The integral

$$\Omega_j = \int_0^{\pi} (1-k^2\cos\theta)^{-j-1/2} d\theta \tag{4}$$

where $j = 0, 1, 2, \ldots;\ 0 \leq k < 1$ is called the Epstein-Hubbell integral. (4) has been studied by many workers during the last three decades due to its importance in certain problems arising in radiation physics and nuclear technology [see, eg. 3,4,6,13,19].

We now describe the various generalizations of (4) studied earlier by a number of authors, notably by Kalla [7,8], Kalla, Conde and Hubbell [10], Kalla and Al-Saqabi [9], Siddiqi [19], Srivastava and Siddiqi [22], Kalla and Tuan [12], Al-Zamel et.al [2], Saxena et.al [16] and Saxena and Pathan [18].

The integral

$$R_\mu(k, \alpha, \gamma) \ = \ \int_0^\pi \frac{\cos^{2\alpha-1}(\theta/2) \sin^{2\gamma-2\alpha-1}(\theta/2) d\theta}{(l - k^2 \cos\theta)^{\mu+1/2}}, \tag{5}$$

where $0 \leq k < 1,\ Re(\gamma) > Re(\alpha) > 0,\ Re(\mu) > -1/2$, which is a generalization of (4) was discussed by Kalla et.al [9,11] and Glasser and Kalla [5].

An interesting generalization of the elliptic-type integral (4) follows from Al-Saqabi [1]

$$B_\mu(k, m, \nu) \ = \ \int_0^\pi \frac{\cos^{2m}(\theta) \sin^{2\nu}(\theta) d\theta}{(1 - k^2 \cos\theta)^{\mu+1/2}}, \tag{6}$$

where $0 \leq k < 1$; $m \in N_0$, $\mu \in C$; $\mathrm{Re}(\mu) > -1/2$.

Siddiqi [19] studied yet another generalization of (4) in the form

$$\bar{\lambda}_\nu(\alpha, k) \ = \ \int_0^\pi \frac{\exp[\alpha \sin^2(\theta/2)] d\theta}{(1 - k^2 \cos\theta)^{\nu+1/2}}, \tag{7}$$

where $0 \leq k < 1$, $\alpha, \nu \in R$.

An interesting unification and extension of the families of elliptic-type integrals (4) to (7) was given by Siddiqi and Srivastava [20] in the form

$$\bar{\lambda}^{(\alpha,\beta)}_{(\lambda,\mu)}(\rho, k) = \int_0^\pi \frac{\cos^{2\alpha-1}(\theta/2) \sin^{2\beta-1}(\theta/2)(1 - \rho \sin^2(\theta/2) d\theta}{(1 - k^2 \cos\theta)^{\mu+1/2}}, \tag{8}$$

where $0 \leq k < 1$, $\mathrm{Re}(\alpha) > 0$, $\mathrm{Re}(\beta) > 0$; $\lambda, \mu \in C$, $\mid \rho \mid < 1$.

Kalla and Tuan [12] extended (8) by means of the following integral and derived its asymptotic expansion.

$$\bar{\lambda}^{(\alpha,\beta)}_{,(\lambda,\gamma,\mu)}(\rho,\delta,k) = \int_0^{\pi} \frac{\cos^{2\alpha-1}(\theta/2)\sin^{2\beta-1}(\theta/2)[1+\delta\cos^2(\theta/2)]^{-\gamma}[1-\rho\sin^2(\theta/2)]^{-\lambda}d\theta}{(1-k^2\cos\theta)^{\mu+1/2}}, \tag{9}$$

where $0 \leq k < 1$, $\mathrm{Re}(\alpha) > 0$, $\mathrm{Re}(\beta) > 0$; $\lambda, \mu, \gamma \in C$, either $|\rho|.|\delta| < 1$ or ρ(or $\delta) \in C$ whenever $\lambda = -m$(or $\gamma = -m$), $m \in N_0$.

Al-Zamel et.al [2] discussed a generalized family of elliptic-type integrals in the form

$$\begin{aligned} Z^{(\alpha,\beta)}_{(\gamma)}(k) &= Z^{(\alpha,\beta)}_{(\gamma_1,\ldots,\gamma_n)}(k_1,\ldots,k_n) \\ &= \int_0^{\pi} \cos^{2\alpha-1}(\theta/2)\sin^{2\beta-1}(\theta/2)\prod_{j=1}^{n}(1-k_j^2\cos\theta)^{-\gamma_j}d\theta, \end{aligned} \tag{10}$$

where $\mathrm{Re}(\alpha) > 0$, $\mathrm{Re}(\beta) > 0$, $|k_j| < 1$, $\gamma_j \in C$, $j = 1,\ldots,n$.

Saxena et.al [16] introduced yet another unification and extension of (9) in the form

$$\begin{aligned} \Omega^{(\alpha,\beta)}_{(\sigma_1,\ldots,\sigma_{n-2};\delta,\mu)}(\rho_1,\ldots,\rho_{n-2},\delta:k) &= \int_0^{\pi} \cos^{2\alpha-1}(\theta/2)\sin^{2\beta-1}(\theta/2) \\ \times \prod_{j=1}^{n-2}[1-\rho_j\sin^2(\theta/2)]^{-\sigma_j}[1+\delta\cos^2(\theta/2)]^{-\gamma}(1-k^2\cos\theta)^{-\mu-1/2}d\theta, \end{aligned} \tag{11}$$

where $0 \leq k < 1$, $\mathrm{Re}(\alpha) > 0$, $\mathrm{Re}(\beta) > 0$; $\sigma_j (j = 1,\ldots,n-2)$. $\gamma, \mu \in C$; $\max\left\{|\rho_j|, \left|\frac{\delta}{1+\delta}\right|, \left|\frac{2k^2}{k^2-1}\right|\right\} < 1$.

For $n = 3$, (11) reduces to (9).

In a recent paper, Saxena and Pathan [18] introduced a new generalized family of the elliptic-type integrals in the following general form, which includes, both the generalisations given by (10) and (11).

$$\Omega = \Omega^{(\alpha,\beta)}_{(\sigma_1,\ldots,\sigma_m,\gamma;\tau_1,\ldots,\tau_n)}(\rho_1,\ldots,\rho_m,\delta;\lambda_1,\ldots,\lambda_n) = \int_0^{\pi} \cos^{2\alpha-1}(\theta/2)\sin^{2\beta-1}(\theta/2)$$

$$\times \prod_{j=1}^{m}[1-\rho_j \sin^2(\theta/2)]^{-\sigma_j}[1+\delta\cos^2(\theta/2)]^{-\gamma}\prod_{j=1}^{n}[1-\lambda_j^2\cos\theta]^{-\tau_j}d\theta, \qquad (12)$$

where $\min(\mathrm{Re}(\alpha),\ \mathrm{Re}(\beta)) > 0;\ |\lambda_j| < 1; \sigma_i, \gamma, \tau_j \in C;$
$\max\left\{|\rho_i|,\ \left|\frac{2\lambda_j^2}{\lambda_j^2-1}\right|,\ \left|\frac{\delta}{1+\delta}\right|\right\} < 1 \quad (i=1,\ldots,m; j=1,\ldots,n)$

Upon a closer examination of (12), it is obvious that Ω is contained in the following Euler-type integral

$$\int_0^1 t^{\beta-1}(1-t)^{\alpha-1}\prod_{j=1}^{m}(1-\rho_j t)^{-\sigma_j}(1-At)^{-\gamma}\prod_{j=1}^{n}(1-B_j t)^{-\tau_j}\,dt \qquad (13)$$

where $A = \dfrac{\delta}{(1+\delta)}$, and $B_j = \dfrac{2\lambda_j}{(\lambda_j^2-1)}$

In this article, we will study a new family of the elliptic-type integrals which generalizes the integral given by Ω (mentioned above). In Section 2, we will give three theorems associated with 2 and 4 variables generating functions. Our strategy to obtain these theorems is an extension of the idea given in Mohammed [14] and Srivastava and Yeh [23]. We apply these theorems in Section 3 to obtain explicit representations of elliptic-type integrals given in Section 1 and their generalizations. Moreover, we obtain a number of single and double Euler-type integrals for various choices of generating functions involved in Theorems 1 to 3.

2. THEOREMS

Let a two variable generating function $F(x,t)$ possesses a formal (not necessarily convergent for $t \neq 0$) power series expansion in t such that

$$F(x,t) \ = \ \sum_{n=0}^{\infty} C_n f_n(x) t^n \qquad (14)$$

where each member of the generalized set $\{f_n(x)\}_{n=0}^{\infty}$ is independent of x and t. Also, let $(\lambda)_n$ denote the Pocchammer symbol defined by

$$(\lambda)_n \ = \ \frac{\Gamma(\lambda+n)}{\Gamma(\lambda)} \ = \ \begin{cases} \lambda(\lambda+1)\cdots(\lambda+n-1), & n \in N = \{1,2,3,\cdots\} \\ 1, & n=0, \quad \lambda \neq 0 \end{cases}$$

Motivated essentially by earlier results of Saran [15] and Mohammad [14; p.262], Srivastava and Yeh [23] gave two theorems on bilinear and bilateral generating functions associated with Lauricella functions $F_A^{(r)}$ and $F_C^{(r)}$ in r-variables [15]. Making use of the integral representation of Lauricella function $F_D^{(r)}$ of r-variables, given by

$$F_D^{(r)}[\alpha, \beta_1, \cdots, \beta_r;\ \gamma;\ x_1, \cdots, x_r]$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma-\alpha)} \int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1} \prod_{i=1}^{r}(1-ux_i)^{-\beta_i}\, du \tag{15}$$

where $\mathrm{Re}(\gamma) > \mathrm{Re}(\alpha) > 0$ and $F_D^{(r)}$ is defined by [15], [21].

$$F_D^{(r)}[a, b_1, \cdots, b_r;\ c;\ x_1, \cdots, x_r]$$

$$= \sum_{m_1, \cdots, m_r=0}^{\infty} \frac{(a)_{m_1+\cdots+m_r}(b_1)_{m_1}\cdots(b_r)_{m_r}}{(c)_r\ m_1!\cdots m_r!}\ x_1^{m_1}\cdots x_r^{m_r}, \tag{16}$$

it is not difficult to prove the following theorem.

Theorem 1. Let the generating function $F(x,t)$ and the Lauricella function $F_D^{(r)}$ be given by (14) and (16), respectively. Then

$$\int_0^1 \frac{u^{\alpha-1}}{(1-u)^{\alpha-\gamma+1}} \prod_{i=1}^{r}(1-ux_i)^{-\beta_i} F(x, tu(1-u))\, du$$

$$= \frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)} \sum_{n=0}^{\infty} \frac{(\alpha)_n(\gamma-\alpha)_n}{(\gamma)_{2n}}\ c_n t^n f_n(x) F_D^{(r)}[\alpha+n, \beta_1, \cdots, \beta_r; \gamma+2n; x_1, \cdots, x_r] \tag{17}$$

provided the each side of (17) exists.

Due to importance of Theorem 1 in various generalizations of Elliptic-type integrals, we state an obvious modified version of this theorem which provides an interesting generalization of Elliptic-type integrals (12) and (13) of Saxena and Pathan [18].

Theorem 2. Let the generating function $F(x,t)$ and the Lauricella function $F_D^{(r)}$ be given by (14) and (16), respectively. Then

$$\int_0^1 u^{\beta-1}(1-u)^{\alpha-1} \prod_{j=1}^{m}(1-\rho_j u)^{-\sigma_j} \left(1 - \frac{\delta}{(1+\delta)}\, u\right)^{-\gamma} \prod_{j=1}^{n}\left(1 - \frac{2\lambda_j^2 u}{\lambda_j^2 - 1}\right)^{-\tau_j}$$

$$F(x, tu(1-u))du \;=\; B(\alpha,\beta)\sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n}{(\alpha+\beta)_{2n}}\,c_n t^n f_n(x)$$

$$F_D^{(m+n+1)}\left[\beta+n,\sigma_1,\cdots,\sigma_m;\gamma,\tau_1,\cdots,\tau_n;\alpha+\beta+2n;\rho_1,\cdots,\rho_m,\frac{\delta}{1+\delta},\frac{2\lambda_1^2}{\lambda_1^2-1},\cdots\frac{2\lambda_n^2}{\lambda_n^2-1}\right] \tag{18}$$

provided that each side of (18) exists.

Next, we consider a generating function $F(x, y; t, T)$ (not necessarily, convergent for $t \neq 0,\ T \neq 0$) in t and T such that

$$F(x,y;t,T) \;=\; \sum_{m,n=0}^{\infty} c_{m,n}\; f_{m,n}(x,y)t^m T^n \tag{19}$$

where each member of the generated set $\{f_{m,n}(x,y)\}_{m,n=0}^{\infty}$ is independent of t and T and the coefficient set $\{c_{m,n}\}_{m,n=0}^{\infty}$ may contain the parameters of the set $\{f_{m,n}(x,y)\}_{m,n=0}^{\infty}$ but is independent of x, y, t and T.

In our attempt to generalize the above theorems, we are easily led to the following extension of the Eulerian integral.

Theorem 3. Let the generating function $F(x, y; t, T)$ be given by (19). Then

$$\begin{aligned}&\sum_{m,n=0}^{\infty}\frac{(\lambda)_{\rho m}(\mu-\lambda)_{\sigma m}(k)_{\rho n}(\nu-k)_{\sigma n}t^m T^n}{(\mu)_{(\rho+\sigma)m}(\nu)_{(\rho+\sigma)n}}\,c_{m,n}f_{m,n}(x,y)\\ =\;&\frac{\Gamma(\mu)\Gamma(\nu)}{\Gamma(\lambda)\Gamma(\mu-\lambda)\Gamma(k)\Gamma(\nu-k)}\\ &\int_0^1\int_0^1\frac{u^{\lambda-1}v^{k-1}}{(1-u)^{\lambda-\mu+1}(1-v)^{k-\nu+1}}\\ &F(x,y;tu^{\rho}(1-u)^{\sigma},Tv^{\rho}(1-v)^{\sigma})\,du\,dv\end{aligned} \tag{20}$$

provided that $\mathrm{Re}(\mu) > \mathrm{Re}(\lambda) > 0$, $\mathrm{Re}(\nu) > \mathrm{Re}(k) > 0$, $\rho \geq 0,\ \sigma \geq 0,\ \rho + \sigma > 0$ and each side of (19) exists.

Proofs of theorems. To prove Theorem 1, we replace $F(x, t)$ by its power series (14) in the integral of (17). On changing the order of integration and summation, which is permissible due to the uniform convergence of the series involved and evaluating the resulting integral using (15), we arrive at the result (17).

The proofs of Theorems 2 and 3 are similar to that of Theorem 1.

3. APPLICATIONS

In view of the importance and usefulness of Theorem 1 to 3, we first mention some interesting applications of Theorem 1.

(i) Consider the generating function

$$F(x,t) \;=\; (1-xt)^{-\lambda} \;=\; \sum_{n=0}^{\infty}(\lambda)_n\, x^n\, \frac{t^n}{n!} \tag{21}$$

and apply Theorem 1 to get

$$\int_0^1 \frac{u^{\alpha-1}}{(1-u)^{\alpha-\gamma+1}} \prod_{i=1}^{r}(1-ux_i)^{-\beta_i}(1-xtu(1-u))^{-\lambda}\, du$$

$$= \frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)} \sum_{n=0}^{\infty} \frac{(\alpha)_n(\gamma-\alpha)_n(\lambda)_n}{(\gamma)_{2n}\, n!} x^n t^n\; F_D^{(r)}[\alpha+n, \beta_1, \cdots, \beta_r; \gamma+2n; x_1, \cdots, x_r] \tag{22}$$

$$\mathrm{Re}(\gamma) > \mathrm{Re}(\alpha) > 0$$

Setting $u = \cos^2\theta/2$ and using $\cos\theta = 2\cos^2\theta/2 - 1$, (22) give us the following representation which is a generalization of the elliptic-type integral of Kalla et.al [9,11] and Glasser and Kalla [5]

$$\int_0^{\pi} \cos^{2\alpha-1}\theta/2 \sin^{2\gamma-2\alpha-1}\theta/2 \prod_{i=1}^{r}\left(1-\frac{x_i}{2-x_i}\cos\theta\right)^{-\beta_i}\left(1-\frac{xt}{4}\sin^2\theta\right)^{-\lambda} d\theta$$

$$= \frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)} \prod_{i=1}^{r}\left(1-\frac{x_i}{2}\right)^{\beta_i} \sum_{n=0}^{\infty} \frac{(\alpha)_n(\gamma-\alpha)_n)(\lambda)_n}{(\gamma)_{2n}\, n!}\, x^n t^n$$

$$F_D^{(r)}[\alpha+n, \beta_1, \cdots, \beta_r; \gamma+2n; x_1, \cdots, x_r] \tag{23}$$

$$\mathrm{Re}(\gamma) \;>\; \mathrm{Re}(\alpha) \;>\; 0,\; \left|\frac{x_i}{2-x_i}\right| < 1,\; i = 1, \cdots, r$$

If we set $r = 3$ and $x_3 = 1$ in (16) and apply

$$F_D^{(3)}(a, b_1, b_2, b_3; c; x_1, x_2, 1) \;=\; \frac{\Gamma(c)\Gamma(c-a-b_3)}{\Gamma(c-a)\Gamma(c-b_3)}\, F_1(a, b_1, b_2; c-b_3; x_1, x_2) \tag{24}$$

where F_1 is Appell's function [21], we get

$$\int_0^1 \frac{u^{\alpha-1}(1-ux_1)^{-\beta_1}(1-ux_2)^{-\beta_2}}{(1-u)^{\alpha-\gamma+1}(1-xtu(1-u))^{\lambda}}\,du$$
$$= \frac{\Gamma(\gamma-\alpha)\Gamma(\alpha)}{\Gamma(\gamma)}\sum_{n=0}^{\infty}\frac{(\alpha)_n(\lambda)_n(\gamma-\alpha)_n}{n!\,(\gamma)_{2n}}x^n t^n\, F_1(\alpha+n,\beta_1,\beta_2;\gamma+2n;x_1,x_2)$$
$$\mathrm{Re}(\gamma-\alpha)>0, \max(|x_1|,\ |x_2|)<1 \tag{25}$$

Furthermore, since

$$F_D^{(3)}(a,b_1,b_2,b_3;c_3;x,1,1) = \frac{\Gamma(c)\Gamma(c-a-b_2-b_3)}{\Gamma(c-a)\Gamma(c-b_2-b_3)}\,{}_2F_1\left(\begin{matrix}a, & b\\ & \\ c-b_2-b_3 & \end{matrix};x\right) \tag{26}$$

a special case of (22) when $r=3$ and $x_2=x_3=1$ yields

$$\int_0^1 \frac{u^{\alpha-1}(1-ux_1)^{-\beta_1}}{(1-u)^{\alpha-\beta+1}(1-xtu(1-u))^{\lambda}}\,du$$
$$= \frac{\Gamma(\alpha)\Gamma(\beta-\alpha)}{\Gamma(\beta)}\sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta-\alpha)_n(\lambda)_n}{n!\,(\beta)_{2n}}x^n t^n\,{}_2F_1\left(\begin{matrix}\alpha+n,\beta_1 & \\ & \\ \beta+2n & \end{matrix};x_1\right), \tag{27}$$
$$\mathrm{Re}(\gamma-\alpha)>0,\ |x|<1$$

In view of the familiar transformation [21, p.55(15)], for $r=3$ and $x_1=x_2=1$, (25) yields

$$\int_0^1 \frac{u^{\alpha-1}(1-xtu(1-u))^{-\lambda}}{(1-u)^{-\beta+1}}\,du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\,{}_3F_2\left(\begin{matrix}\alpha, & \lambda, & \beta\\ & & \\ \dfrac{\beta+\alpha}{2}, & \dfrac{\beta+\alpha+1}{2} & \end{matrix};\frac{xt}{4}\right), \tag{28}$$

(ii) Consider the generating relation [21]

$$F(x,t) = (1-X_1t)^{-\alpha_1}(1-X_2t)^{-\alpha_2} = \sum g_n^{\alpha_1,\alpha_2}(X_1,X_2)t^n \tag{29}$$

where $g_n^{\alpha_1,\alpha_2}(X_1,X_2)$ is the Lagrange polynomial defined by

$$g_n^{\alpha,\beta}(x,y) = \sum_{n=0}^{\infty}\frac{(\alpha)_r(\beta)_{n-r}}{r!\,(n-r)!}\,x^r y^{n-r} \tag{30}$$

and apply Theorem 1 to get

$$\int_0^1 \frac{u^{\alpha-1}}{(1-u)^{\alpha-\gamma+1}} \prod_{i=1}^{r}(1-ux_i)^{-\beta_i} \prod_{j=1}^{2}(1-X_j tu(1-u))^{-\alpha_j}\, du$$

$$= \frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)} \sum_{n=0}^{\infty} \frac{(\alpha)_n(\gamma-\alpha)_n}{(\gamma)_{2n}}$$

$$g_n^{\alpha_1,\alpha_2}(X_1,X_2)t^n F_D^{(r)}(\alpha+n,\beta_1,\cdots,\beta_r;\gamma+2n;x_1,\cdots,x_r) \qquad (31)$$

(iii) Consider the generating function

$$F(x,t) \;=\; e^{-xt} \;=\; \sum_{n=0}^{\infty} \frac{(-1)^n\, x^n t^n}{n!}$$

and apply Theorem 1 to get

$$I \;=\; \int_0^1 \frac{e^{-xt(u(1-u))}}{(1-u)^{\alpha-\gamma+1}}\, u^{\alpha-1} \prod_{i=1}^{r}(1-ux_i)^{-\beta_i}\, du$$

$$= \frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)} \sum_{n=0}^{\infty} \frac{(\alpha)_n(\gamma-\alpha)_n\,(-x)^n t^n}{n!\,(\gamma)_{2n}}$$

$$F_D^{(r)}(\alpha+n,\beta_1,\cdots,\beta_r;\gamma+2n;x_1,\cdots,x_r) \qquad (32)$$

or, equivalently

$$I \;=\; \int_0^{\pi} \frac{e^{-(xt/4)\sin\theta}\sin^{2\alpha-1}\theta/2}{\cos^{2\alpha-2\gamma+1}\theta/2} \prod_{i=1}^{r}(1-x_i\sin^2\theta/2)^{-\beta_i}\, d\theta \qquad (33)$$

Making use of Theorem 2 and generating function (21), it is not difficult to prove that

$$\int_0^1 u^{\beta-1}(1-u)^{\alpha-1} \prod_{j=1}^{m}(1-\rho_j u)^{-\sigma_j} \left(1-\frac{\delta}{1+\delta}u\right)^{-\gamma}$$

$$\prod_{j=1}^{n}(1-A_j u)^{-\tau_j}(1-xtu(1-u))^{-\lambda}\, du$$

$$= B(\alpha,\beta) \sum_{n=0}^{\infty} \frac{(\alpha)_n(\beta)_n(\lambda)_n}{(\alpha+\beta)_{2n}\, n!}\, t^n\, x^n$$

$$\times F_D^{(m+n+1)}\left[\beta+n,\sigma_1,\cdots,\sigma_m;\gamma,\tau_1,\cdots,\tau_n;\alpha+\beta+2n;\right.$$

$$\left.\rho_1,\cdots,\rho_m,\frac{\delta}{1+\delta},A_1,\cdots,A_n\right] \qquad (34)$$

where $A_j = \dfrac{2\lambda_j}{\lambda_j^2 - 1}$, $j = 1, \cdots, n$

For $\lambda \to 0$ and $u = \sin^2\theta/2$, (33) yields an explicit representation of a generalized family of the elliptic-type integrals given recently by Saxena and Pathan [18] in the form

$$\begin{aligned}
& \Omega^{(\alpha,\beta)}_{(\sigma_1,\cdots,\sigma_n,\gamma;\tau_1,\ldots,\tau_n)}(\rho_1,\cdots,\rho_m,\delta;\lambda_1,\cdots,\lambda_n) \\
= & \int_0^{\pi} \cos^{2\alpha-1}\theta/2\ \sin^{2\beta-1}\theta/2 \prod_{j=1}^{m}(1-\rho_j\sin^2\theta/2)^{-\sigma_j}\ (1+\delta\cos^2\theta/2)^{-\gamma} \\
& \prod_{j=1}^{n}(1-\lambda_j^2\cos\theta)^{-\tau_j}\ d\theta \\
= & B(\alpha,\beta)(1+\delta)^{-\gamma}\prod_{j=1}^{n}(1-\lambda_j^2)^{-\tau_j} \\
& \times F_D^{(m+n+1)}\left[\beta,\sigma_1,\cdots,\sigma_m,\gamma,\tau_1,\cdots,\tau_n;\alpha+\beta;\rho_1,\cdots,\right. \\
& \left.\rho_m,\frac{\delta}{1+\delta},\frac{2\lambda_1^2}{\lambda_1^2-1},\cdots,\frac{2\lambda_n^2}{\lambda_n^2-1}\right], \qquad (35)
\end{aligned}$$

where $\min(\mathrm{Re}(\alpha),\mathrm{Re}(\beta)) > 0$, $|\lambda_j| < 1$; $\sigma_i, \gamma, \tau_j \notin C$, $\max\left\{|\rho_i|\left|\dfrac{2\lambda_1^2}{\lambda_1^2-1}\right|, \left|\dfrac{\delta}{1+\delta}\right|\right\} < 1$ $(i = 1,\cdots,m;\ j = 1,\cdots,n)$.

(iv) With a view to obtaining numerous families on double integral representations of Euler-type as an application of Theorem 3, we first observe that

$$(1-z_1-z_2)^{-a} = \sum_{m,n=0}^{\infty}\frac{(a)_{m+n}z_1^m\ z_2^n}{m!\ n!}$$

yields easily the generating function

$$\begin{aligned}
& \sum_{m,n=0}^{\infty}\frac{(b)_m(c)_n}{m!\ n!}\ F_2(a,-m,-n,b,c;x,y)t^mT^n \\
= & (1-t)^{-b}(1-T)^{-c}\left(1+\frac{xt}{1-t}+\frac{yT}{1-T}\right)^{-a} = F(x,y;t,T) \qquad (36)
\end{aligned}$$

where F_2 is Appell's function of second kind [21].

Now, upon using this last result (35) in Theorem 3, we get

$$\int_0^1\int_0^1 \frac{u^{\lambda-1}v^{k-1}(1-tu^\rho(1-u)^\sigma)^{-d}(1-Tv^\rho(1-v)^\sigma)^{-d'}}{(1-u)^{\lambda-\mu+1}(1-v)^{k-\nu+1}} F_2(a,d,d',b,c;-X,-Y)\,dudv$$

$$= \frac{\Gamma(\lambda)\Gamma(k)\Gamma(\mu-\lambda)\Gamma(\nu-k)}{\Gamma(\lambda)\Gamma(\nu)} \sum_{m,n=0}^{\infty} \theta \dot{F}_2(a,-m,-n,b,c;x,y)t^mT^n \qquad (37)$$

where $X = \dfrac{xtu^\rho(1-u)^\sigma}{1-tu^\rho(1-u)^\sigma}$, $Y = \dfrac{yTv^\rho(1-v)^\sigma}{1-Tv^\rho(1-v)^\sigma}$,

$$\theta = \frac{(\lambda)_{\rho m}(\mu-\lambda)_{\sigma m}(k)_{\rho n}(v-k)_{\sigma n}}{(\mu)_{(\rho+\sigma)m}(\nu)_{(\rho+\sigma)n}} \frac{(d)_m(d')_n}{m!\,n!}$$

$\mathrm{Re}(\mu) > \mathrm{Re}(\lambda) > 0$, $\mathrm{Re}(\nu) > \mathrm{Re}(k) > 0$, $\rho \geq 0$, $\sigma \geq 0$ and $\rho + \sigma > 0$.

Furthermore, put $b = c = a$ in (36) and use [20, p.305(108)]

$$F_2(\alpha,\beta,\beta',\alpha,\alpha;x,y) = (1-x)^{-\beta}(1-y)^{-\beta'}{}_2F_1\left(\begin{matrix}\beta,\beta'\\ \\ \alpha\end{matrix}\ ;\ \frac{xy}{(1-x)(1-y)}\right),$$

to get

$$\int_0^1\int_0^1 \frac{u^{\lambda-1}v^{k-1}(1-tu^\rho(1-u)^\sigma)^{-d}(1-Tv^\rho(1-v)^\sigma)^{-d'}}{(1-u)^{\lambda-\mu+1}(1-v)^{k-\nu+1}} (1+X)^{-d}(1+Y)^{-d'}$$

$$\times {}_2F_1\left(\begin{matrix}d,d'\\ \\ a\end{matrix}\ ;\ \frac{XY}{(1+X)(1+Y)}\right) dudv$$

$$= \frac{\Gamma(\lambda)\Gamma(\mu-\lambda)\Gamma(k)\Gamma(\nu-k)}{\Gamma(\lambda)\Gamma(\nu)}$$

$$\sum_{m,n=0}^{\infty} \theta(1-x)^m(1-y)^n{}_2F_1\left(\begin{matrix}-m,-n\\ \\ a\end{matrix}\ ;\ \frac{xy}{(1-x)(1-y)}\right) t^mT^n \qquad (38)$$

where θ, X, Y are defined in (36).

REFERENCES

[1] Al-Saqabi, B. N., *A generalization of elliptic-type integrals*, Hadronic J., 10 (1987), 331.

[2] Al-Zamel, A., Tuan, V. K. and Kalla, S. L., *Generalized Elliptic-type integrals and Asymptotic formulas*, App. Math. Comput., 114 (2000), 13-25.

[3] Berger, M. J. and Lamkin, J. C., *Simple calculations of Gamma ray penetration into shelters*, contribution. J. Res. NBS, 60(1958), 109.

[4] Bjorkberg, J. and Kristensson, G., *Electromagnetic scattering by a perfectly conducting elliptic disk*, Canad. J. Phys., 65 (1987), 723.

[5] Glasser, M. L. and Kalla, S. L., *Recursion relations for a class of generalized elliptic-type integrals*, Rev. Tec. Ing. Univ. Zulia, 12 (1989), 47.

[6] Hubbell, J. H., Bach, R. I. and Herbold, R. J., *Radiation field from a circular disc source*, J. Res. NBS, 65 (1961), 254-264.

[7] Kalla, S. L., *Results on generalized elliptic-type integrals Mathematical Structures*, computational Mathematics-Mathematical Modelling (Ed. El. Sendov) Bulg. Acad. Special Vol. (164), 216-219.

[8] Kalla, S. L., *The Hubell rectangular source integral and its generalizations*, Radiat. Phys. Chem., 41 (1993), 775-781.

[9] Kalla, S. L. and Al-Saqabi, B., *On a generalized elliptic-type integral*, Rev. Bra. Fis., 16 (1986), 145-156.

[10] Kalla, S. L., Conde, S. and Hubbell, J. H., *Some results on generalized elliptic-type integrals*, Appl. Anal., 22 (1986), 273-286.

[11] Kalla, S. L., Leubner, C. and Hubbell, J. H., *Further results on generalized elliptic-type integrals*, Appl. Anal., 25 (1987), 269-274.

[12] Kalla, S. L., Tuan, Vu Kim, *Asymptotic formulas for generalized Elliptic-type integrals*, Computers Math. Appl., 32 (1996), 49-55.

[13] Kliuga, P. and Khanna, S. M., *Dose rate to the inner ear during Mosebauer experiments*, Phys. Med. Biol., 28 (1983), 359-366.

[14] Mohammed, Ch. W., *Bilinear and bilateral generating functions of generalized polynomials*, J. Austral. Math. Soc. Sec. B 39 (1997), 257-270.

[15] Saran, S., *Theorems on bilinear generating functions*, Indian J. Pure Appl. Math. 3 (1972), 12-20.

[16] Saxena, R. K. and Kalla, S. L., *A new method for evaluating Epstein-Hubbell generalized Elliptic-type integral*, Int. J. Appl. Math. 2 (2000), 732-742.

[17] Saxena, R. K., Kalla, S. L. and Hubbell, J. H., *Asymptotic expansion of a unified Elliptic-type integral*, Math. Balkanica 15 (2001).

[18] Saxena, R. K. and Pathan, M. A., *Asymptotic formulas for unified elliptic-type integrals*, Demonstratio Mathematica 36(3) 2003, 581-590.

[19] Siddiqi, R. N., *On a class of generalized elliptic-type integral*, Rev. Brasileira Fis., 19 (1989), 137-147.

[20] Srivastava, H. M. and Karlsson P. W., *Multiple Gaussian hypergeometric series*, Ellis Horwood Ltd., Chichester, 1985.

[21] Srivastava, H. M. and Manocha, H. L., *A treatise on generating functions*, Ellis Horwood Ltd., Chichester, 1984.

[22] Srivastava, H. M. and Siddiqi, R. N., *A unified presentation of certain families of elliptic-type integrals related to radiation field problems*, Radiat. Phys. Chem., 46 (1995), 303-315.

[23] Srivastava, H. M. and Yeh, Yeong Nan, *Certain theorems on bilinear and bilateral generating functions*, ANZIAM J. 43 (2002), 567-574.

MODELING OF AN UNRELIABLE PRODUCTION MERGING PROCESS WITH INTERMEDIATE STORAGE TANK

M. Savsar
College of Engineering & Petroleum,
Department of Industrial & Management Systems Engineering,
Kuwait University P.O. Box 5969 Safat 13060, Kuwait.
email: Mehmet@kuc01.kuniv.edu.kw

Abstract

This paper consider a two-stage unreliable continuous production process with merging configuration. A stochastic model, which consists of twelve differential equations, is developed and solved to study the reliability and productivity of the production system under random failure and repair rates.

1. INTRODUCTION

In most of the continuous processes, including chemical and petrochemical industries, a storage tank, or an array of tanks, is provided between the production stages to decouple the stages and to reduce the effects of variation in one stage over the others. Without intermediate storage, random equipment failures and variable operation times significantly reduce the process output rate and line efficiency. Since providing a large storage tank is costly, it is important to be able to determine the exact effect of a given tank size on production output rate. Different aspects of this problem have been considered in several previous studies (1-5), with emphasis being on discrete parts manufacturing systems. In this paper, mathematical modeling of a continuous process with two unreliable stages is considered, where the first stage consists of two parallel machines and the second stage has one machine as shown in figure 1. The system is modeled using continuous Markov processes and the state of the system is described by set of differential equations, which are then solved to determine the reliability and productivity of the system under different operational conditions.

2. THE STOCHASTIC MODEL

In order to develop a stochastic model to describe the state of the system at any time and to analyze its performance measures, the following notations are introduced:

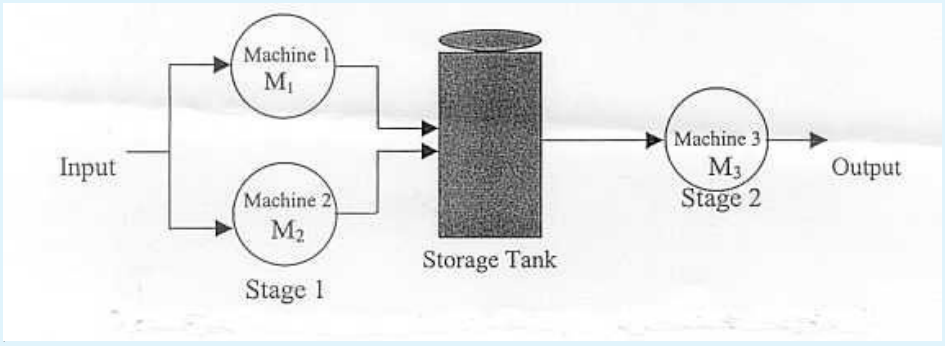


Figure 1: A Two-stage process with a decoupling intermediate storage tank.

1. $M_i, i = 1, 2$ denotes two identical machines at stage 1 with failure, repair and production rates denoted by λ_1, μ_1, and q_1 respectively.

2. M_3 is a machine at stage 2 with parameters λ_2, μ_2, and q_2. It is assumed that $q_2 = 2q_1$.

3. The storage tank has a finite capacity of z.

4. Each stage has its own repair crew; i.e., repairs start without waiting.

5. Machine failures and repairs are random quantities with exponentially distributed time to failure and time to repairs.

6. Failure rate of M_3 reduces to $\lambda_2' = \lambda_2/2$ when its operation rate reduces to q_1 from q_2.

7. The system operates until one machine fails. If M_1 (or M_2) fails, the operation continues until one of the following events occurs:

i) The tank level is reduced to zero by machine M_3;

ii) machine M_3 fails; or

iii) machine M_2 (or M_1) fails. If the tank level reduces to zero before the repair of failed machine is completed, the second stage (M_3) slows down and operates at rate q_1 instead of its normal rate q_2. If both machines M_1 and M_2 fail, the second stage continues operation until the tank is empty, at which time; M_3 is forced down due to unavailability of incoming flow. The failure of M_3 will

force M_1 and/or M_2 down, i.e., blocked, when the tank reaches its maximum level z. In any case, a forced down machine will not fail.

State of the system is described by the following variables: $S_{ijx}(t,x)$ =State of the system at time t with tank level $x; t > 0,\ 0 \leq x \leq z; i$ and j machines operating at the first and the second stages respectively ($i = 0, 1, 2$ and $j = 0, 1$).

x = Tank level, $0 \leq x \leq z$ at a time t.

z = Maximum tank capacity.

$f_{ijx}(t,x)$ = Probability distribution function of state $S_{ijx}(t,x)$ with $i = 0, 1, 2;$ $j = 0, 1;\ 0 < x < z$.

$P_{ijx}(t)$ = Marginal probability of state S_{ijx} for the states in which x is variable.

For those states in which tank level varies, i.e., $0 < x < z$, the system changes its state with respect to tank level x as well as the time t. There are six such states, namely, $S_{00x(}t,x), S_{01x}(t,x), S_{11x}(t,x), S_{21x}(t,x), S_{20x}(t,x)$, and $S_{10x}(t,x)$.

For the states in which storage tank is either empty or full, the system changes its state with respect to time t only and there are also six such states, which are: $S_{010}(t,0), S_{110}(t,0),\ S_{210}(t,0), S_{10z}(t,z), S_{20z}(t,z), \text{and} S_{21z}(t,z)$. Thus, the system operates within twelve states. Note that the probability of two machines failing at the same time, while the tank is full or empty, is negligible since such an event has infinitely small probability. The marginal probabilities for the first six states are given by the equation , $P_{ijx}(t) = \int_0^z f_{ijx}(t,x)dx,\ i = 0, 1, 2;\ j = 0, 1;\ 0 < x < z$.

The equivalent probabilities for the last six states do not depend on x, an thus x is fixed at 0 or z. Operation of the system is governed by the following set of differential equations, which describe the twelve system states. Here f_{ijx} is used to denote $f_{ijx}(t,x)$ for simplification.

$$\frac{\partial f_{00x}}{\partial t} = -(\mu_1 + \mu_2) f_{00x} + \lambda_2 f_{01x} + \lambda_1 f_{10x} \tag{1}$$

$$\frac{\partial f_{21x}}{\partial t} = -(2\lambda_1 + \lambda_2) f_{21x} + \mu_1 f_{11x} + \mu_2 f_{20x} \tag{2}$$

$$\frac{\partial f_{01x}}{\partial t} - q_2 \frac{\partial f_{01x}}{\partial x} = \mu_2 f_{00x} - (\mu_1 + \lambda_2) f_{01x} + \lambda_1 f_{11x} \tag{3}$$

$$\frac{\partial f_{11x}}{\partial t} - q_1 \frac{\partial f_{11x}}{\partial x} = 2\lambda_1 f_{21x} + \mu_1 f_{01x} - (\lambda_1 + \lambda_2 + \mu_1) f_{11x} + \mu_2 f_{10x} \tag{4}$$

$$\frac{\partial f_{10x}}{\partial t} - q_1 \frac{\partial f_{10x}}{\partial x} = \mu_1 f_{00x} + \lambda_2 f_{11x} - (\mu_1 + \mu_2 + \lambda_1) f_{10x} + 2\lambda_1 f_{20x} \tag{5}$$

$$\frac{\partial f_{20x}}{\partial t} - 2q_1 \frac{\partial f_{20x}}{\partial x} = \lambda_2 f_{21x} + \mu_1 f_{10x} - (\mu_2 + 2\lambda_1) f_{20x} \tag{6}$$

$$\frac{\partial p_{010}}{\partial t} = -\mu_1 p_{010} + \lambda_1 p_{110} + q_2 f_{01x}(0) \tag{7}$$

$$\frac{\partial p_{110}}{\partial t} = -\mu_1 p_{010} - (\lambda_1 + \mu_1 + \lambda_2') p_{110} + 2\lambda_1 p_{210} + q_1 f_{11x}(0) \tag{8}$$

$$\frac{\partial p_{10z}}{\partial t} = -(\mu_1 + \mu_2) p_{10z} + q_1 f_{10x}(z) \tag{9}$$

$$\frac{\partial p_{20z}}{\partial t} = \mu_1 p_{10z} - \mu_2 p_{20z} + \lambda_2 p_{21z} + 2q_1 f_{20x}(z) \tag{10}$$

$$\frac{\partial p_{210}}{\partial t} = \mu_1 p_{110} - (2\lambda_1 + \lambda_2) p_{210} \tag{11}$$

$$\frac{\partial p_{21z}}{\partial t} = -\mu_2 p_{20z} - (2\lambda_1 + \lambda_2) p_{21z} \tag{12}$$

3. THE SOLUTION

In order to solve the above system of equations, boundary conditions must be specified. For the system under consideration, there are four boundary conditions caused by the flows from states $S_{110}, S_{210}, S_{10z}$, and S_{21z} to the states $S_{10x}, S_{20x}, S_{11x}$, and S_{11x} respectively. These conditions are stated as:

$$\lambda_2' p_{110} = q_1 f_{10x}(0);\ \lambda_2 p_{210} = 2q_1 f_{20x}(0);$$

$$\mu_2 p_{10z} = q_1 f_{11x}(z);\ 2\lambda_1 p_{21z} = q_1 f_{11z}(z) \tag{13}$$

Equations (1) to (6), which are decoupled from equations (7) to (12), can be represented in matrix notations as follows:

$$[\dot{F}]_t = [q_1][\dot{F}]_x = [A][F] \tag{14}$$

where,

$$[\dot{F}]_t = \frac{\partial}{\partial t}\begin{vmatrix} F_1 \\ F_2 \end{vmatrix}, [\dot{F}]_x = \frac{\partial}{\partial x}\begin{vmatrix} F_1 \\ F_2 \end{vmatrix}, [F] = \begin{vmatrix} F_1 \\ F_2 \end{vmatrix}, [F_1] = \begin{vmatrix} f_{00x}(t,x) \\ f_{21x}(t,x) \end{vmatrix}, [F_2] = \begin{vmatrix} f_{01x}(t,x) \\ f_{11x}(t,x) \\ f_{10x}(t,x) \\ f_{20x}(t,x) \end{vmatrix}$$

$$[q_1] = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -q_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -q_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2q_1 \end{vmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & q_{01} \end{bmatrix}$$

$$[A] = \begin{vmatrix} -(\mu_1+\mu_2) & 0 & \lambda_2 & 0 & \lambda_1 & 0 \\ 0 & -(2\lambda_2+\lambda_2) & 0 & \mu_1 & 0 & \mu_2 \\ \mu_2 & 0 & -(\mu_1+\lambda_2) & \lambda_1 & 0 & 0 \\ 0 & 2\lambda_2 & \mu_1 & -(\mu_1+\lambda_1+\lambda_2) & \mu_2 & 0 \\ \mu_1 & 0 & \lambda_2 & -(\mu_1+\mu_2+\lambda_1) & 2\lambda_1 & 0 \\ 0 & \lambda_2 & 0 & 0 & \mu_1 & -(\mu_2+\lambda_1) \end{vmatrix}$$

$$= \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$$

Where A_1 is 2×2; A_2 is 2×4; A_3 is 4×2; and A_4 is 4×4 sub matrix in A. At steady state, the partial derivatives with respect to time t approach zero, i.e., $[\dot{F}]_t = 0$ and $[F_1]$ & $[F_2]$ become functions of x only. The new system of differential equations is written as:

$$[A_1][F_1] + [A_2][F_2] = 0 \tag{15}$$

$$[A_3][F_1] + [A_4][F_2] = [q_{01}][\dot{F}_2]_x \tag{16}$$

Substituting $[F_1]$ from equation (15) into (16), we get

$$[\dot{F}_2]_x = [\Omega][F_2] \tag{17}$$

where, $[\Omega] = [q_{01}]^{-1}\{[A_4] - [A_3][A_1]^{-1}[A_2]\}$. This is a system of homogenous differential equations, which has a general solution as

$$[F_2] = [S][e^{kx}][c] = [\Psi_2][c] \tag{18}$$

Where, $[S]$ is a 4×4 matrix containing the eigenvectors of matrix $[\Omega]$ and $[e^{kx}]$ is a 4×4 diagonal matrix with $e^{k_i x}$ in the i^{th} diagonal; k_i is the i^{th} eigenvalue of $[\Omega]$ and $[C] = (c_1, c_2, c_3, c_4)^T$ is the set of constant coefficients to be determined by the initial conditions. By substituting $[F_2]$ from (18) into (15), $[F_1]$ is determined as follows:

$$[F_1] = -[A]^{-1}[A_2][S][e^{kx}][C] = [\Psi_1][C] \tag{19}$$

Similarly, equations (7) to (12), which constitute a linear system, are represented by:

$$[\dot{P}]_t = [B][P]_t + [q_{II}][F_0], \tag{20}$$

where

$$[\dot{P}]_t = \frac{\partial}{\partial t} \quad |P| \quad [P] = \begin{vmatrix} P_3 \\ P_4 \end{vmatrix} \qquad [P_3] = \begin{vmatrix} p_{010}(t) \\ p_{110}(t) \\ p_{10z}(t) \\ p_{20z}(t) \end{vmatrix} \qquad [p_4] = \begin{vmatrix} p_{210}(t) \\ p_{21z}(t) \end{vmatrix}$$

$$[B] = \begin{vmatrix} -\mu_1 & \lambda_1 & 0 & 0 & 0 & 0 \\ \mu_1 & -(\lambda_1 + \mu_1 + \lambda_2') & 0 & 0 & 2\lambda_1 & 0 \\ 0 & 0 & -(\mu_1 + \mu_2) & 0 & 0 & 0 \\ 0 & 0 & \mu_1 & -\mu_2 & 0 & \lambda_2 \\ 0 & \mu_1 & 0 & 0 & -(2\lambda_1 + \lambda_2) & 0 \\ 0 & 0 & 0 & \mu_2 & 0 & -(2\lambda_1 + \lambda_2) \end{vmatrix} = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$$

$$[q_{II}] = \begin{vmatrix} q_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & q_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2q_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix} = \begin{bmatrix} q_{02} & 0 \\ 0 & 0 \end{bmatrix} \qquad [F_0] \begin{vmatrix} f_{01x}(0) \\ f_{11x}(0) \\ f_{10x}(z) \\ f_{20x}(z) \\ 0 \\ 0 \end{vmatrix} = \begin{vmatrix} F_{01} \\ 0 \end{vmatrix}$$

Where, B_1 is 4×4; B_2 is 4×2; B_3 is 2×4; and B_4 is 2×2 sub matrix. For the steady state solution, $[\dot{P}]_t = 0$ and therefore, $[B_1][P_3] + [B_2][P_4] + [q_{02}][F_{01}] = 0$ and $[B_3][P_3] + [B_4][P_4] = 0$, which are solved to obtain

$$[P_3] = [D][C] \tag{21}$$

$$[P_4] = [H][C] \tag{22}$$

where, $[D] = -\{[B_1] - [B_2][B_4]^{-1}[B_3]\}^{-1}[q_{02}][R]$ $[H] = -[B_4]^{-1}[B_3][D]$

$$[R] = \begin{vmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31}e^{k_1 z} & S_{32}e^{k_2 z} & S_{33}e^{k_3 z} & S_{34}e^{k_4 z} \\ S_{41}e^{k_1 z} & S_{42}e^{k_2 z} & S_{43}e^{k_3 z} & S_{44}e^{k_4 z} \end{vmatrix}$$

Where, s_{ij} are the elements of matrix $[S]$. The constant coefficients, $[C]$ are determined by using the boundary conditions, the normalizing condition, and some matrix manipulations. Matrix $[C]$ is then substituted into equations (18), (19), (21) and (22) to obtain $[F_2], [F_1], [P_3]$, and $[P_4]$. Finally $[P_1]$ and $[P_2]$ are obtained by integrating $[F_1]$ and $[F_2]$ respectively. The solution set $[P_1], [P_2], [P_3]$, and $[P_4]$ are the steady sate probabilities of twelve system states. Combining all these into a single vector, one obtains the solution vector $[P_s] = [p_{00x}, p_{21x}, p_{01x}, p_{11x}, p_{10x}, p_{20x}, p_{010}, p_{110}, p_{10z}, p_{20z}, p_{210}, p_{21z}]$.

4. LINE PERFORMANCE MEASURES

Two performance measures, reliability and productivity, can be calculated from the solutions for the system state probabilities. Line reliability can be defined as full reliability or partial reliability. At full reliability, all the equipment operate at full rate. Fill reliability is determined by $R_{\text{full}} = p_{21x} + p_{210} + p_{21z}$. At partial reliability, some equipment operate at full rate and some at partial rate with no imposed stoppages. Partial reliability is determined from $R_{\text{partial}} = p_{01x} + p_{11x} + p_{10x} + p_{20x} + p_{110}$.

The production rate of the line is determined by determining the proportion of time that the last stage is producing at its full rate give by $\beta_1 = \sum_{i=0}^{2} p_{i1x} + p_{210} + p_{21z}$ and at its reduced rate by $\beta_2 = p_{110}$. Thus, line production rate is obtained from $\beta_l = q_2\beta_1 + q_1\beta_2 = q_1(2\beta_1 + \beta_2)$, while line efficiency is obtained from $E = \beta_l/q_2 = (2\beta_1 + \beta_2)q_1/q_2$.

These formulas can be used to evaluate line performance under different line operational characteristics. Computational results show that line efficiency increases with respect to increasing storage tank capacity. For example, for failure rates of equipmen given as $\lambda_1 = \lambda_2 = 0.2$ failures per time unit; repair rates given as $\mu_1 = \mu_2 = 2$ repairs per time unit, then the line efficiency approaches to about 88% with a tank capacity of $z = 20$ units. If the repairs rates are doubled, then the

line efficiency approaches to 95% for the same tank capacity of $z = 20$ units. While the line efficiency is about 78.6% at a tank capacity of $z = 0$ units, it exceeds 85% for $z = 10$ units for the same repair rates of $\mu_1 = \mu_2 = 2$ repairs per time unit. Similarly, while the line efficiency is about 88% for $z = 0$, it approaches to almost 95% for $z = 10$, for repair rates of $\mu_1 = \mu_2 = 4$ repairs per time unit. Similar results are obtained when the failure rates are changes. After the storage tank capacity reaches a certain value, no more improvement is possible in the line efficiency for a given set of failure and repair parameters.

The expected storage level can be determined using the state probabilities and the expectation formula as follows. In only nine states the storage level is above zero. In the remaining three states the storage level x is equal to zero.

$$E(x) = \int_0^z x[f_{00x}(x)+f_{01x}(x)+f_{10x}(x)+f_{11x}(x)+f_{20x}(x)+f_{21x}(x)]dx+z[p_{10z}+p_{20z}+p_{21z}]$$

5. CONCLUSION

In this paper, a stochastic model is presented for an unreliable two-stage continuous process. The model consists of twelve differential equation which describe operation of the system with unreliable equipment and a storage tank. The proposed model can be used to evaluate system efficiency as a function of storage tank capacity, failure rates, repair rates, and production rates of each stage (or machine). Expected storage tank level can also be determined at steady state for a given maximum tank capacity (z). The solution is in a closed form, with final equations of the production output rate given as a function of storage capacity and other line parameters. The model can be used to study efficiency of a given line and to determine the optimum storage tank capacity, which results in maximum line efficiency or through put under different operational parameters. It is not possible, with the given set of equations, to determine the optimum capacity in closed form, which is done by differentiation of a given nonlinear function. However, expected storage level can be determined under given operational conditions using the state probabilities and the expectation formula as stated in the previous section.

In order to determine the optimum storage capacity, one must try several capacity levels using the model and find out the capacity that results in best efficiency or production output rate. This is a trivial computational exercise using the mathematical model presented here. The problem presented in this paper becomes more complicated in case of longer production processes or several stages with parallel machines. Analytical solutions are very difficult. On must try to approximate solu-

tion approach or use the simulation modeling approach to study such processes in which number of stages exceeds three.

REFERENCES

[1] Hillier, F.S. and So, K.C., "The effect of the coefficient of variation of operation times on the allocation of storage space in production line system", IIE Transactions, 23, (1993), 198-206.

[2] Heavey, C. and Papadopoulos, H.T., "The throughput rate of multi-station unreliable production lines", *European Journal of Operational Research,* 68, (1993), 69-89

[3] Powell, S.G. "Buffer allocation in unbalanced three-station serial lines", *International Journal of Production Research,* 32, (1994), 2201, 2217.

[4] Cheng, D.W. "On the design of a tandem queue with blocking: Modeling, analysis, and gradient estimation", *Naval Research Logistic,* 41, (1994), 759-770

[5] Papadopoulos, H.T. and Heavey, C., "Queuing Theory in manufacturing systems analysis and design: A classification of models for production and transfer lines", it European Journal of Operational Research, 92, (1996), 1-27.

A MATHEMATICAL ANALYSIS FOR AN AGGREGATION MODEL OF PHYTOPLANKTON

N. El Saadi, M. Adioui and O. Arino
GEODES-IRD 32, avenue Henri Varagnat
F-93143 Bondy cedex, France.

Abstract

Our main goal in this paper is the development and analysis of an aggregation model of phytoplankton. The model consists of an integro-differential advection-diffusion equation, with convolution term. The Cauchy problem is well posed in a suitable function space. Existence, uniqueness and positivity of solutions are investigated.

1. INTRODUCTION

The role of aggregates in marine food webs and vertical transport processes is now well recognized (Alldredge and Silver,1988; Fowler and Knauer, 1986). Recent attention has been devoted to modeling studies of the mechanisms by which aggregates form and the dynamics governing their formation (review in Jackson and Lochmann, 1993).
Coagulation theory has more recently been applied to describe phytoplankton aggregation (Jackson, 1990; Hill, 1992; Riebesel and Gladrow, 1992). It requires that primary particles collide by some physical process and stick together upon collision. Brownian motion, differences in sinking velocity between particles, and fluid shear may all cause primary particles to collide.
However, studies of marine aggregates at small-scales have emphasized biological mechanisms for their formation. That is, although planktonic organisms can be thought of as particles, the richness of biological responses makes the nature of their interactions more complex than the simple physical ones described by coagulation theory. Some planktonic species (algae, bacteria, dinoflagellates that are motile species of phytoplankton) have chemosensory abilities (Fitt, 1985; Spero, 1985; Spero and Morée, 1981): they can sense the chemical field generated by the presence of other particles. The dinoflagellates and more generally algae are known to leak organic matter into solution (Mague et al., 1980). Bell and Mitchell (1972) noted that this leakage creates a zone around individual cells, the "phycosphere" where extracellular products exist in enhanced concentrations compared to the surrounding concentration. The released products such as amino-acids and sugar attract algae or bacteria that are present in a suitable neighborhood.

The remaining part of this paper is organized as follows. In section 2, we describe our model in detail. In section 3, we show that the Cauchy problem associated to the model is well posed in a suitable function space. Indeed, we investigate existence, uniqueness and positivity of solutions. We also prove that the solution satisfies the principle of conservation of mass for all positive time. Finally, we conclude with a brief discussion in section 4.

2. MODEL DERIVATION

In this paper, we perform the mathematical analysis of a non-local advection model for the motion of large plankton populations. This model is the continuum limit of an infinite system of planktonic particles subject to random dispersal modeled as Brownian motions and mutual interactions allowing the particles motions some dependence [23]. The model deals with temporal and spatial changes in the phytoplankton population density. To describe interactions between planktonic particles, we propose two hypotheses : a) a non uniform concentration fields around organisms, b) organisms considered (plankton particles) having chemosensory abilities and hence some "knowledge" of their neighbors' whereabouts and modifying their motion accordingly. So, aggregation here is due to "social" forces induced by interactions of each cell with others in the population which belong to a suitable neighborhood. As each particle has a limited knowledge of the spatial distribution of its neighbors (Berg and Purcell, 1977; Jackson, 1987, 1989), the particles are subject to their interaction aggregate within a range $R = r_1 - r_0 (0 < r_0 < r_1)$.

The model is given by the following partial differential equation:

$$\frac{\partial}{\partial t}u(x,t) = d\frac{\partial^2}{\partial x^2}u(x,t) - \frac{\partial}{\partial x}\left(u(x,t)\Phi(x)\left[G * u^0(.,t)\right](x)\right), \text{in}\Omega \times (0,\infty), \quad (1)$$

where $\Omega =]0, L[$ is a bounded domain with smooth boundary $\partial\Omega$ in $\mathbb{R}$, x is a one dimensional coordinate, t is time. $u(x,t)$ represents the proportion density function of phytoplankton at time t. That is $u(x,t)dA$ is the expected proportion of organisms in the sample area dA surrounding the point x at time t. Here, $\mathbb{R}$ represents the vertical axis oriented downward from the surface to the seabed. The point 0 is at the surface of water and L is the limit of the euphotic zone (the upper layers of oceans and lakes). Generally, phytoplankton particles can survive and multiply only in the "euphotic zone", that is why we restrict our model to Ω. d is a coefficient of diffusion and G is the attractive force given by:

$$G(x) = \begin{cases} |x|^2 - (r_0 + r_1)|x| + r_0 r_1 & \text{if} r_0 < x < r_1 \\ -|x|^2 + (r_0 + r_1)|x| - r_0 r_1 & \text{if} -r_1 < x < -r_0 \\ 0 & \text{otherwise.} \end{cases}$$

Typically, terms in the advective velocity have the form of convolution (Mogilner and Edelstein Keshet, 1996), i.e.,

$$\left[G * u^0(.,t)\right](x) = \int_{\mathbb{R}} G(x-x')\, u^0(x',t)dx',$$

where

$$u^0(x) = \begin{cases} u(x) & 0 < x < L \\ 0 & x \leq 0 \text{ or } x \geq L. \end{cases}$$

This form describes the velocity induced at the site x by the net effects of all individuals at various sites $x\prime$. The kernel $G(x-x')$ associates a strength of interaction per unit density as a function of the distance $x - x'$ between any two particles. Boundary conditions are imposed at the surface and at L :

$$\frac{\partial}{\partial x}u(x,t) = 0, \text{on} \partial\Omega \times \mathbb{R}^+ \tag{2}$$

and the initial condition is

$$u(x,0) = u_0(x) \geq 0, \text{in}\Omega. \tag{3}$$

We also assume that

$$u_0(x) \geq 0 \text{ and } \int_0^L u_0(x)dx = 1. \tag{4}$$

and

$$\Phi \in H_0^1(\Omega), \sup \Phi \subset [\delta, L-\delta], \delta \text{sufficientlysmall}, \tag{5}$$

where $\sup \Phi$ denotes the support of function Φ.

The normalization condition (4) is connected with the fact that $u(x,t)$ is a proportion and that the mass must be conserved. Namely, that

$$u(x,t) \geq 0 \text{ and } \int_0^L u(x,t)dx = 1, \forall t.$$

will be shown to hold when assumptions in (4) are made. We will not consider growth terms, and focus exclusively on non-linear and non-local transport properties of the population. The model describes motion of phytoplankton particles alive during their lifetimes.

3. EXISTENCE, UNIQUENESS AND POSITIVITY

3.1 Abstract formulation and well-posedness

The system (1)-(3) will be studied via the theory of operator semigroups [24]. For this purpose, we write (1)-(3) as an abstract Cauchy problem. We obtain a quasilinear problem with nonlinearities in the first order term:

$$\begin{cases} \dfrac{d}{dt}u(t) &= Au(t) - B\left[u(t)g_{(\Phi,G)}\left(u(t)\right)\right] \\ u(0) &= u_0. \end{cases} \tag{6}$$

in which $u(t)$ is used for $u(.,t)$. The operator $A : \mathcal{D}(A) \subset X := L^2(\Omega) \to X$ is defined by

$$\begin{aligned} Aw &= d\frac{d^2w}{dx^2}, \\ D(A) &= \left\{w \in H^2(\Omega) : w'_{|\partial\Omega} = 0\right\}, \end{aligned} \tag{7}$$

and the operator $B : \mathcal{D}(B) \subset X \to X$ by

$$\begin{aligned} Bw &= \frac{d}{dx}w, \\ \mathcal{D}(B) &= H^1(\Omega). \end{aligned} \tag{8}$$

$H^1(\Omega)$ and $H^2(\Omega)$ denote usual Sobolev functions spaces. We will denote by $\langle , \rangle$ and $\|.\|$, respectively, the scalar product and the norm in X. The operator A commutes with B and they are related by the following formula

$$\langle Bu, dBu\rangle = -\langle u, Au\rangle, \forall u \in \mathcal{D}(A). \tag{9}$$

We endow $D(B)$ with the graph norm $\|x\|_B = \|Bx\|$ for $x \in D(B)$.
The main existence result will be derived using successive approximations in a space of continuous functions from some suitable interval $[0, t_0]$ (where $t_0 > 0$ will be chosen later on) into $\mathcal{D}(B)$. On occasion, we will use the notation $Y = C([0, t_0], \mathcal{D}(B))$. The operator $g_{(\Phi,G)}$ is defined as follows:

$$g_{(\Phi,G)}(\varphi)(x) = \Phi(x)\left[G * \varphi\right](x) = \Phi(x) \int_{\mathbb{R}} G(x-y)\,\varphi(y)\,dy.$$

By straightforward consequence of standard calculations, we can establish that $g_{(\Phi,G)} : \mathcal{D}(B) \to \mathcal{D}(B)$, continuously so there exists a constant δ, so that

$$\left|g_{(\Phi,G)}(\varphi)\right|_{\mathcal{D}(B)} \leq \delta\,|\varphi|_{\mathcal{D}(B)}, \forall \varphi \in \mathcal{D}(B).$$

Note also that $G * u^0$ is uniformly bounded. Hence, $g_{(\Phi,G)}u$ is uniformly bounded. As a result of Hölder's inequality, we get

$$\left|g_{(\Phi,G)}u\right|_\infty \leq \sqrt{L}\,|G|_\infty\,|\Phi|_\infty\,\|u\|\,,\ \forall u \in D(B). \tag{10}$$

On the other hand, we have

$$\left|Bg_{(\Phi,G)}u\right|_\infty \leq \sqrt{L}\,|G|_\infty \max(|\Phi|_\infty\,,|B\Phi|_\infty)\,|u|_{\mathcal{D}(B)}\,,\ \forall u \in D(B). \tag{11}$$

We now return back to the Cauchy problem (6)

Proposition 1 The operator A defined by (7) is the generator of an analytic semigroup of contractions in X, $(T(t))_{t\geq 0}$, compact for $t > 0$. The restrictions $T(t)_{/\mathcal{D}(B)}$ send $\mathcal{D}(B)$ into itself and are uniformly bounded in $\mathcal{D}(B)$(that is, there exists $C_1 \geq 0$, such that, $\left|T(t)_{/\mathcal{D}(B)}\right|_{D(B)} \leq C_1$, for $t \geq 0$).

The solving of the problem (6) involves two steps: first, one deals with local existence; next, a noncontinuation principle will be established (Theorem 5) which will ensure solutions exist on as long a time interval as desired.
To prove local existence for problem (6), we write it in integral form by using the variation of constants formula

$$u(t) = T(t)u_0\ -\int_0^t T(t-s)B\left[u(s)g_{(\Phi,G)}\left(u(s)\right)\right]ds. \tag{12}$$

We remind that a solution of (12) is called a mild solution of the differential equation(6) (see [24]).

3.2 Local existence of solutions

This subsection is concerned with local existence of solutions for problem (6). For this purpose, we start by establishing some useful estimates.

Lemma 2
1) There exists a constant M, such that, for all u, $v \in \mathcal{D}\left(B\right)$, we have

$$\left\|B\left[ug_{(\Phi,G)}(u)\right] - B\left[vg_{(\Phi,G)}(v)\right]\right\| \leq M\max(|u|_{\mathcal{D}(B)}\,,|v|_{\mathcal{D}(B)})\,|u-v|_{\mathcal{D}(B)}\,.$$

2) There exists a positive constant Q, such that, for all $u \in \mathcal{D}\left(B\right)$, it holds that

$$\left\|B\left[ug_{(\Phi,G)}(u)\right]\right\| \leq Q\,|u|_{\mathcal{D}(B)}\,\|u\|\,.$$

3) There exists a positive constant C, such that, for all $u \in X$, it holds that

$$\|BT(t)u\| \leq \frac{C}{\sqrt{t}}\,\|u\|\,,\ \forall t > 0. \tag{13}$$

We can now state the main theorem of this subsection:

Theorem 3 *For every $R > 0$, there exists $t_0 > 0$, $t_0 = t_0(R)$, such that, for each $u_0 \in \mathcal{B}_{\mathcal{D}(B)}(R)$, (i.e., the ball of radius R centered at 0 of $\mathcal{D}(B)$), the Cauchy problem (6) has a unique mild solution u defined on the interval $[0, t_0]$. Moreover, the map $u_0 \to u$ is Lipschitz continuous from $\mathcal{B}_{\mathcal{D}(B)}(R)$ into Y. Finally, ${}_{\Omega}u(x,t)dx =_{\Omega} u_0(x)dx$, for all $t \geq 0$.*

3.3 Global existence

Global existence (i.e., the fact that the solutions are defined on the whole of $t > 0$) is established for positive solutions. For that, we will show, the boundedness of the solution $u(t)$ in the $\mathcal{D}(B)$ norm. This property, together with theorem 5, implies that $t_{\max} = \infty$. Prior to this, we will prove that (CP) preserves positiveness, which will be needed in the a priori estimates of the solutions. Our first result in this direction is the following theorem.

Theorem 4 *Equation (1)-(3) preserves positiveness, that is: $u_0 \geq 0$ implies that $u(x,t) \geq 0$ for all $t \geq 0$.*

The next two results are crucial in continuation of solutions.

Theorem 5 *For every initial data $u_0 \in \mathcal{D}(B)$, the abstract Cauchy problem (6) has a unique mild solution on a maximal interval of existence $[0, t_{\max}[$.*
If $t_{\max} < \infty$ then

$$\lim_{t \to t_{\max}} \sup |u(t)|_{\mathcal{D}(B)} = \infty.$$

Proposition 6 *There exists a function $K : \mathbf{R}^+ \to]0, +\infty]$, non increasing, such that, if $u(.,t)$ is a solution of (CP) with $u_0 \in \mathcal{D}(B)$ and $u_0 \geq 0$, then it holds that $|u(t)|_{\mathcal{D}(B)} \leq K_1(\|u_0\|)\,|u_0|_{\mathcal{D}(B)}$, for all $t \in [0, K(\|u_0\|)]$, where $K_1(x) = 2\,[C_1 + \exp(L_1 K(x))]$.*

3.4 Regularity

The following result describes the regularity of a mild solution of (6).

Theorem 7 *For every $u_0 \in \mathcal{D}(A)$, the mild solution of equation (6) is a classical solution, i.e. u is continuous on $[0, \infty)$, continuously differentiable on $(0, \infty)$, $u(t) \in D(A)$ for $t \in (0, \infty)$ and (6) is satisfied on $[0, \infty)$.*

4. CONCLUSION

Our main result in this paper is the development and analysis of an aggregation model of phytoplankton. Here, the clustering phenomenon is a consequence of social behavior and is due to the nonlinear interactions between particles. The model describes the evolution of the mean-field spatial density of phytoplankton population on the vertical water column by a deterministic nonlinear partial differential equation of the advection-diffusion type. We have proved that the Cauchy problem associated to this model is well posed in $\mathcal{D}(B)$. Solutions are fixed points of strict contractions and initial values in $\mathcal{D}(A)$ yield classical solutions. We have also proved the conservation of mass of phytoplankton for all positive time.

ACKNOWLEDGEMENT

The authors would like to thank the unknown referee for his comments, which allowed to improve the original manuscript.

REFERENCES

[1] Adioui, M., Arino, O., Smith, W.V. and Treuil, J.P., A mathematical analysis of a fish school model.J.Differential Equations 188 (2003)406 − 446.

[2] Alldredge, A. L. and Silver, M.W., Characteristics, dynamics and significance of marine snow. Progress in oceanography 20 (1983)41 − 82. Springer-Verlag, New York, 119.

[3] Bell, W. and Mitchell, R., Chemotactic and growth responses of marine bacteria to algal extracellular products. Biol.Bull.143 (1972)265 − 277.

[4] Dam, H.G. and Drapeau, D.T., Coagulation efficiency, organic- matter glues and the dynamics of particles during a phytoplankton bloom in a mesocosm study. Deep-Sea Res 42 (1995)111 − 123.

[5] Drapeau, D.T., Dam, H.G. and Grenier, G., An improved flocculator design for use in particle aggregation experiments. Limnol. Oceanogr. 39 (1994)723 − 729.

[6] Fittm, W. K. Chemosensory responses of the symbiotic dinoflagellate Symbiodinium microadriatica (Dinophycae). J.Phycol. 21 (1985)62 − 67.

[7] Fowler, S.W. and Knauer, G.A., Role of large particles in the transport of elements and organic compounds through the ocean water column. Prog. Oceanogr. 16(1986)147 – 194.

[8] Friedman, A., Partial Differential Equations of Parabolic Type, Holt, Reinhart, and Winston Inc., New York, (1969).

[9] Hill, P. S. Reconciling aggregation theory with observed vertical fluxes following phytoplankton blooms. J.Geophys Res. 97 (1992)2295 – 2308.

[10] Jackson, G. A. Simulating chemosensory responses of marine microorganisms. Limnol. Oceanogr 32 (6) (1987)1253 – 1266.

[11] Jackson, G. A. Simulation of bacterial attraction and adhesion to falling particles in an aquatic environment. Limnol. Oceanogr 34 (3)(1989)514 – 530.

[12] Jackson, G. A. A model of the formation of marine algal flocks by physical coagulation processes .Deep-Sea Research. 37 (1990)1197 – 1211.

[13] Jackson, G. A. and Lochmann, S. E., Modeling coagulation of algae in marine ecosystems. In: Environmental analytical and physical chemistry series:Environmental particles ,volume 2, J.Buffle and H.P.van Leeuwen, editors Lewis Publishers, Boca Raton, FL, USA, (1995)387 – 184.

[14] Jackson, G. A. Comparing observed changes in particle-size spectra with those predicted using coagulation theory. Deep-Sea Research.*I*. 42 (1990)159 – 1211.

[15] Kiørboe, T., Lundsgaard, C., Olesen, M. and Hansen,J. L. S., Aggregation and sedimentation processes during a spring phytoplankton bloom: A field experiment to test coagulation theory. Journal of Marine Research. 52(1994)297−323.

[16] Kiørboe, T., Hansen, J. L. S., Alldredge, A. L. and Jackson, G. A., et al., Sedimentation of phytoplankton during a diatom bloom: Rates and mechanisms. J.Mar.Res. 54(1996)1123 – 1148.

[17] Kiørboe, T., Tiselius, P., Mitchell-Innes, B., Hansen, J.L.S., Visser, A. W. and Mari, X., Intensive aggregate formation with low vertical flux during an upwelling-induced diatom bloom. Limnol. Oceanogr. 43(1998)104 – 116.

[18] Kiørboe, T. Formation and fate of marine snow:small-scale processes with large-scale implications. Sci.Mar.655(Suppl.2) (2001)57 – 71.

[19] Macnab, R. M. and Han, D. P., Asynchronous switching of flagellar motors on a single bacterial cell. Cell 32, (1984)109 – 117.

[20] Mague, T. H., Friberg, E. D., Hughes, J. and Morris, I., Extracellular release of carbon by marine phytoplankton; a physiological approach. Limnol. Oceanogr. 25(1980)262 − 279.

[21] McCave, I. N. Size spectra and aggregation of suspended particles in the deep ocean. Deep-Sea Research.31 (1984)329 − 352.

[22] Mogilner, A. and Keshet, E., A non_local model for a swarm, J. Math. Biol, 38 (1999) 534 − 570.

[23] Morale et al., D. An interacting particle system modelling aggregation behavior: from individuals to populations, J. Math. Biolgy, 2004. In press.

[24] Pazy, A. Semigroups of Linear Operators and Applications to Partial Differential Equations. Applied Mathematical Sciences, Vol. 119, Springer, New York, (1983).

[25] Riebesel, U. and Wolf-Gladrow, D. A., The relationship between physical aggregation of phytoplankton and particle flux: a numerical model. Deep-Sea Research. 39(1992)1085 − 1102.

[26] Spero, H. J., and Morée, M., Phagotrophic feeding and its importance to the life cycle of the holozoic dinoflagellate, Gymnodinium fungiforme. J.Phycol. 17(1981)43 − 51.

[27] Spero, H. J. Chemosensory capabilities in the phagotrophic dinoflagellate Gymnodinium fungiforme. J.Phycol 21(1985)181-184.

[28] Stewart, R. C. and Dahlquist, F. W., Molecular components of bacterial chemotaxis. Chem. Rev 87(1987)997-1025.

ITERATIVE METHODS FOR SOME NONLINEAR MATRIX EQUATIONS

S. M. El-Sayed[1,2]
[1]On leave from Department of Mathematics, Faculty of Science,
Benha university, Benha 13518, Egypt.
[2]Department of Mathematics, Scientific Departments,
Education College for Girls, Al-Montazah, Buraydah, Al-Qassim,
Kingdom of Saudi Arabia.
E-mail: ms4elsayed@yahoo.com

Abstract

The present paper treats iterative methods for a class of nonlinear matrix equations of the form $X \pm A^{\star}\mathcal{F}(X)A = Q$, where $\mathcal{F}(X)$ maps positive definite matrices either into positive definite matrices or into negative definite matrices, and satisfies some monotonicity property. Here the matrix A is arbitrary and Q is a positive definite matrix. The available iterative methods for solving the nonlinear matrix equations are based on basic fixed point, inversion free variant of the basic fixed point, applications of Newton's method and on cyclic reduction method. The nonlinear matrix equations have a solution if and only if a related iterative algorithm converges to a positive definite solution (PDS) under some conditions on the given matrix A.

Keywords: Nonlinear matrix equation; iterative methods; basic fixed point; inversion free variant of the basic fixed point; Newton's method; cyclic reduction method; convergence of a sequence; positive definite matrix.

Abbreviated Title: The matrix equation $X \pm A^{\star}\mathcal{F}(X)A = Q$.

1991 Mathematics Subject Classification: Primary 15A24, 65F10; Secondary 65H10, 93B40.

INTRODUCTION

Let $\mathcal{P}(n)$ denote the set of $n \times n$ positive semidefinite matrices. We consider the following class of nonlinear matrix equations

$$X \pm A^{\star}\mathcal{F}(X)A = Q, \tag{1}$$

where $\mathcal{F}(\cdot) : \mathcal{P}(n) \to \mathcal{P}(n)$ is either monotone (meaning that $0 \leq X \leq Y$ implies

that $\mathcal{F}(X) \leq \mathcal{F}(Y)$) or anti-monotone (meaning that $0 \leq X \leq Y$ implies that $\mathcal{F}(X) \geq \mathcal{F}(Y)$). In particular, we shall be interested in the case where $\mathcal{F}(X)$ is generated by a function from $[0,\infty)$ to $[0,\infty)$ which is either operator monotone or operator anti-monotone. For example, $\mathcal{F}(x) = x^r$ is operator monotone for $0 < r \leq 1$, while $\mathcal{F}(x) = x^{-1}$ is operator anti-monotone (see, e.g., [9], where a thorough study of operator monotone functions is presented). This type of nonlinear matrix equation (5),(for example when $\mathcal{F}(X) = X^{-1}$) often arises in the analysis of ladder networks, dynamic programming, control theory, stochastic filtering, statistics, see [1] and the references contained therein. Also in many mathematical and physical applications, we must solve a system of linear equations

$$M\ x = f, \tag{2}$$

where the positive definite matrix M arises from the finite difference approximation to an elliptic partial differential equation. As an example, let

$$M = \begin{pmatrix} I & A \\ A^* & I \end{pmatrix}.$$

We consider the matrix $M = \tilde{M} + diag[I - X, 0]$ where

$$\tilde{M} = \begin{pmatrix} X & A \\ A^* & I \end{pmatrix}.$$

We can decompose the matrix $\tilde{M}$ via two ways. First

$$\begin{pmatrix} X & A \\ A^* & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ A^*X^{-1} & I \end{pmatrix} \begin{pmatrix} X & A \\ 0 & X \end{pmatrix}. \tag{3}$$

In order that the decomposition (3) exists the matrix X must be a solution of the matrix equation $X + A^*X^{-1}A = I$.

Second

$$\begin{pmatrix} X & A \\ A^* & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ A^*X^{-1} & I \end{pmatrix} \begin{pmatrix} X & A \\ 0 & \sqrt{X} \end{pmatrix}. \tag{4}$$

In order that the decomposition (4) exists the matrix X must be a solution of the matrix equation $Y + A^*Y^{-2}A = I,\ Y = \sqrt{X}$. We can see the matrix equation (5) (when $\mathcal{F}(X) = X^{-1}$) in another point of view as a particular case from the discrete-time algebraic Riccati matrix equation

$$0 = Q + F^\star XF - X - (F^\star XB + A^\star)(R + B^\star XB)^{-1}(B^\star XF + A),$$

where Q is a positive definite matrix, see [6, 9]. The above equation can reduced to the equation (1) with $\mathcal{F}(X) = X^{-1}$, by substituting $F = 0$, $B = I$, and $R = 0$. Several authors [1-5,7-13] have considered such a nonlinear matrix equation (5) and its special cases.

The matrix equation

$$X + A^{\star}X^{-1}A = I \tag{5}$$

has been studied recently by several authors [1-4,11-13]. Anderson, Morley and Trapp [1] discussed the existence of a positive solution to the matrix equation (5) with right hand side an arbitrary matrix, while Engwerda, Ran and Rijkeboer [2] established necessary and sufficient conditions for the existence of a positive definite solution of same matrix equation as in [1]. They discussed both the real and complex case and proposed recursive algorithms to compute the largest and smallest solution of the equation. Engwerda [3] proved the existence of a positive definite solution of the real matrix equation (5) and also found an algorithm to calculate the solution. El-Sayed et al. [5, 7-11] obtained necessary and sufficient conditions for the existence of a positive definite solution of matrix equations with several forms instant of X^{-1} in (5). Zhan and Xie [12] proposed several numerical algorithms for finding solutions for (5). In [13], Zhan proposed an algorithm called inversion free variant of the basic fixed point iteration, that avoids matrix inversion.

$$\begin{aligned} Take \qquad & X_0 = Y_0 = I, \\ X_{n+1} &= I - A^{\star}Y_nA, \\ Y_{n+1} &= Y_n\left(2I - X_nY_n\right), \qquad n = 0, 1, 2, \cdots . \end{aligned} \tag{6}$$

Guo and Lancaster [4] modified Zhan's algorithm (6) to find the maximal positive definite solutions of Eq. (5) as follows:

$$\begin{aligned} Take \qquad & X_0 = Y_0 = I, \\ Y_{n+1} &= Y_n\left(2I - X_nY_n\right), \\ X_{n+1} &= I - A^{\star}Y_{n+1}A, , \qquad n = 0, 1, 2, \cdots . \end{aligned} \tag{7}$$

They presented a deeper discussion of convergence of the inversion free variant of the basic fixed point iteration method for Eq. (5) than was done in [13].

Our goal of this paper is to discuss the matrix equation (5) with a new inversion free variant of the basic fixed point iteration method.

$$\begin{aligned} Take \qquad & X_0 = Y_0 = I \\ Y_{n+1} &= \left(I - X_n\right)Y_n + I \\ X_{n+1} &= I - A^{\star}Y_{n+1}A, \qquad n = 0, 1, 2, \cdots . \end{aligned} \tag{8}$$

The suggested algorithm also avoids matrix inversion. Furthermore the algorithm requires only three matrix multiplications per step, whereas Zhan's algorithm (6) and Guo et al. algorithm (7) require four matrix multiplications per step. We use the algorithm to obtain numerically the maximal solution of Eq. (5) under some additional conditions. We obtain the rate of convergence for the sequence generated by our algorithm. Some numerical examples are given to show the behavior of the considered algorithm.

The paper is organized as follows. In section 2, under some conditions on matrix A we obtain the rate of convergence of the iterative sequence of approximate solutions. Section 3 illustrates the performance of the method with some numerical examples. Section 4 contains the conclusion and Remarks drawn from the results.

We will start with some notations, which are used throughout the rest of the paper. The notation $A > 0$ ($A \geq 0$) means that A is positive definite (semidefinite). $A^\star$ denotes the complex conjugate transpose of A, and I is the identity matrix. The notation $A > B$ ($A \geq B$) indicates that $A - B$ is positive definite (semidefinite). We denote by $\rho(A)$ the spectral radius of A (i.e. $\max_{\lambda_i} |\lambda_i|$, where λ_i are the eigenvalues of A). In the following we denote by $\|.\|$ the spectral norm (i.e. $\| A \| = \sqrt{\rho(AA^\star)}$) unless otherwise noted. We denote by X_+ the maximal solution.

1. CONDITIONS FOR THE EXISTENCE OF SOLUTIONS

In this section, we introduce an inversion free variant of the basic fixed point iteration method to avoid the computation of the matrix inverse in every iteration. We will obtain the conditions for existence of the solutions of Eq. (5).

We will prove that the sequence $\{X_n\}$ is monotone decreasing and converges to the maximal solution X_+.

Theorem 1.1 *If Eq. (5) has a positive definite solution and the two sequences $\{X_n\}$ and $\{Y_n\}$ are determined by Algorithm (8), then $\{X_n\}$ is monotone decreasing and converges to the maximal solution X_+. If the matrix A is nonsingular and $X_n > 0$ for every n, then (5) has a positive definite solution.*

Proof. First, we will prove that $I = X_0 \geq X_1 \geq X_2 \geq \cdots \geq X_n \geq X_+$ and $I = Y_0 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_n \leq X_+^{-1}$.
Since X_+ is a solution of (5), i.e.

$$X_+ = I - A^\star X_+^{-1} A,$$

then $X_0 = I \geq X_+$. Also

$$X_1 = I - A^\star A \geq I - A^\star X_+^{-1} A = X_+,$$

i.e. $X_0 \geq X_1 \geq X_+$. For

$$X_2 = I - A^\star Y_2 A = I - A^\star A - A^{\star 2} A^2 = X_1 - A^{\star 2} A^2$$

this implies $X_2 \leq X_1$, so that $X_0 \geq X_1 \geq X_2$.
For the sequence $\{Y_n\}$ we have $Y_0 = Y_1 = I$ and since $X_+^{-1} \geq I$, then $Y_0 = Y_1 \leq X_+^{-1}$

$$Y_0 = Y_1 = I \leq Y_2 = (I - X_1)\, Y_1 + I = A^\star A + I,$$

i.e. $Y_0 = Y_1 \leq Y_2$.

We have also

$$Y_2 = (I - X_1)\, Y_1 + I \leq (I - X_+)\, X_+^{-1} + I = X_+^{-1},$$

i.e. $Y_1 \leq Y_2 \leq X_+^{-1}$. Concerning $\{X_2\}$, we get

$$X_2 = I - A^\star Y_2 A \geq I - A^\star X_+^{-1} A = X_+.$$

i.e. $X_0 \geq X_1 \geq X_2 \geq X_+$.

That is means that the inequalities are true for $n = 0, 1, 2$. So, assume that the above inequalities are true for $n = k$, i.e.

$$I = X_0 \geq X_1 \geq X_2 \geq \cdots \geq X_k \geq X_+$$

and

$$I = Y_0 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_k \leq X_+^{-1}.$$

Now we will prove the inequality for $n = k + 1$. We have

$$Y_{k+1} = (I - X_k)\, Y_k + I \geq (I - X_{k-1})\, Y_{k-1} + I = Y_k.$$

We have also

$$\begin{aligned} Y_{k+1} &= (I - X_k)\, Y_k + I \leq (I - X_+)\, X_+^{-1} + I \\ &= X_+^{-1}, \end{aligned} \tag{9}$$

i.e. $Y_k \leq Y_{k+1} \leq X_+^{-1}$. Concerning the sequence $\{X_n\}$, we have

$$X_k - X_{k+1} = A^\star\, (Y_{k+1} - Y_k)\, A,$$

since $Y_{k+1} \geq Y_k$, hence $X_k \geq X_{k+1}$. Therefore,

$$X_{k+1} = I - A^\star Y_{k+1} A \geq I - A^\star X_+^{-1} A = X_+.$$

i.e. $X_k \geq X_{k+1} \geq X_+$.

This completes the proof of the inequality for $n = k+1$. Therefore, $I = X_0 \geq X_1 \geq X_2 \geq \cdots \geq X_n \geq X_+$ and $I = Y_0 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_n \leq X_+^{-1}$ are true for all n, and $\lim_{n\to\infty} X_n$ and $\lim_{n\to\infty} Y_n$ exist. By taking limits in the equations of (8) leads to $Y = X^{-1}$ and $X = I - A^\star X^{-1} A$. Moreover, as each $X_n \geq X_+$ then $X = X_+$, see [13].

If matrix A is nonsingular and $X_n > 0$ for every n. Hence the above proof of the monotonicity of $\{Y_n\}$ remains valid (monotone increasing). It follows that sequence $\{X_n\}$ is monotone decreasing and bounded from below by the zero matrix. So $\lim_{n\to\infty} X_n = X$ exists. Since A is nonsingular $Y_{n+1} = A^{-\star}\left(I - X_{n+1}\right) A^{-1}$. Thus $\lim_{n\to\infty} Y_n = Y$ exist. As $Y_0 = I$ and $\{Y_n\}$ is monotone increasing, $Y \geq I$. Taking limit in the Algorithm (8) implies

$$\begin{aligned} Y &= (I - X)\, Y + I \\ X &= I - A^\star Y A. \end{aligned} \tag{10}$$

Since $Y \geq I$, $X = Y^{-1} > 0$, and hence $X = I - A^\star X^{-1} A$. So Eq. (5) has a positive definite solution. ∎

Lemma 1.2 *Assume that Eq. (5) has a positive definite solution and* $\|A\| < 1/2$, *then the sequence* $\{Y_n\}$ *satisfies* $\|Y_n A\| < 1$ *for every* $n = 0, 1, \cdots$.

Proof. Since $Y_0 = Y_1 = I$, it is clear that $\|Y_0 A\| = \|Y_1 A\| < \frac{1}{2} < 1$. For Y_2 we have $Y_2 = (I - X_1)\, Y_1 + I = A^\star A + I$, thus $\|Y_2 A\| = \|A^\star A^2 + A\| \leq \|A^\star A^2\| + \|A\| < \frac{5}{8} < 1$. This means that the inequality is holds for $n = 0, 1, 2$. So, assume that the inequality is satisfied for $n = k$, i.e. $\|Y_k A\| < 1$. Now we will prove inequality when $n = k+1$.

$$\begin{aligned} Y_{k+1} A &= \left[(I - X_k)\, Y_k + I\right] A, \\ &= \left[(I - (I - A^\star Y_k A))\, Y_k + I\right] A, \\ &= A^\star Y_k A Y_k A + A. \end{aligned} \tag{11}$$

Then we get

$$\begin{aligned} \|Y_{k+1} A\| &\leq \|A^\star Y_k A Y_k A\| + \|A\|, \\ &\leq \|A^\star\| \|Y_k A\|^2 + \|A\|, \\ &\leq \|A^\star\| + \|A\| < 1. \end{aligned} \tag{12}$$

This completes the proof of the lemma. ■

We now establish the following result to obtain the rate of convergence for Algorithm (8).

Theorem 1.3 *If Eq. (5) has a positive definite solution and* $\|A\| < \frac{1}{2}$*, then the sequence* $\{X_n\}$ *satisfies*

$$\|Y_{n+1} - X_+^{-1}\| \leq \|AX_+^{-1}\|\|Y_n - X_+^{-1}\|, \tag{13}$$

and

$$\|X_{n+1} - X_+\| \leq \|A\|^2\|Y_n - X_+^{-1}\|, \tag{14}$$

for all n large enough. If the matrix A is nonsingular, we also have

$$\|X_{n+1} - X_+\| \leq \|X_+^{-1}A\|\|X_n - X_+\|. \tag{15}$$

Proof.

$$\begin{aligned} Y_{n+1} &= (I - X_n)\, Y_n + I, \qquad (16)\\ &= A^\star Y_n A Y_n + I,\\ &= A^\star \left(Y_n + X_+^{-1} - X_+^{-1}\right) A Y_n + I,\\ &= A^\star \left(Y_n - X_+^{-1}\right) A Y_n + A^\star X_+^{-1} A Y_n + Y_n - Y_n + I,\\ &= A^\star \left(Y_n - X_+^{-1}\right) A Y_n - \left(I - A^\star X_+^{-1} A\right) Y_n + Y_n + I,\\ &= A^\star \left(Y_n - X_+^{-1}\right) A Y_n - X_+ Y_n + Y_n + I. \end{aligned}$$

Then we get

$$\begin{aligned} X_+^{-1} - Y_{n+1} &= X_+^{-1} + A^\star \left(X_+^{-1} - Y_n\right) A Y_n + X_+ Y_n - Y_n - I, \qquad (17)\\ &= (I - X_+)\left(X_+^{-1} - Y_n\right) + A^\star \left(Y_n - X_+^{-1}\right) A Y_n,\\ &= A^\star X_+^{-1} A \left(X_+^{-1} - Y_n\right) + A^\star \left(Y_n - X_+^{-1}\right) A Y_n, \end{aligned}$$

i.e. we have

$$\begin{aligned} \|X_+^{-1} - Y_{n+1}\| &\leq \|A^\star X_+^{-1} A\|\|X_+^{-1} - Y_n\| + \|A^\star\|\|AY_n\|\|X_+^{-1} - Y_n\|, \qquad (18)\\ &\leq \left(\|X_+^{-1}A\| + \|AY_n\|\right)\|A^\star\|\|X_+^{-1} - Y_n\|. \end{aligned}$$

Since $\lim_{\to\infty} Y_n = X_+^{-1}$, then

$$\begin{aligned} \|Y_{n+1} - X_+^{-1}\| &\leq 2\|A^\star\|\|AX_+^{-1}\|\|Y_n - X_+^{-1}\|, \qquad (19)\\ &\leq \|AX_+^{-1}\|\|Y_n - X_+^{-1}\|. \end{aligned}$$

Thus the inequality (13) is true. The second inequality (14) can be proved to hold directly from the following equality.

$$X_{n+1} - X_+ = A^\star \left(X_+^{-1} - Y_{n+1}\right) A.$$

We now prove the last inequality (15). We have from Eq. (17) the following:

$$\begin{aligned} X_+^{-1} - Y_{n+1} &= A^\star X_+^{-1} A \left(X_+^{-1} - Y_n\right) + A^\star \left(Y_k - X_+^{-1}\right) A Y_n, \\ &= A^\star X_+^{-1} A A^{-\star} \left(A^\star X_+^{-1} A - A^\star Y_n A\right) A^{-1} + \left(A^\star Y_k A - A^\star X_+^{-1} A\right) Y_n, \\ &= A^\star X_+^{-1} A A^{-\star} \left(X_n - X_+\right) A^{-1} + \left(X_n - X_+\right) Y_n. \end{aligned} \tag{20}$$

Therefore,

$$\begin{aligned} X_{n+1} - X_+ &= A^\star \left(X_+^{-1} - Y_{n+1}\right) A, \\ &= \left(A^\star\right)^2 X_+^{-1} A A^{-\star} \left(X_n - X_+\right) + A^\star \left(X_n - X_+\right) Y_n A. \end{aligned} \tag{21}$$

Taking norm for the above equation, we get

$$\begin{aligned} \|X_{n+1} - X_+\| &\leq \|A^\star\|^2 \|X_+^{-1} A\| \|A^{-\star}\| \|X_n - X_+\| \\ &+ \|A^\star\| \|Y_n A\| \|X_n - X_+\|, \\ &\leq \left(\|X_+^{-1} A\| + \|Y_n A\|\right) \|A^\star\| \|X_n - X_+\|. \end{aligned} \tag{22}$$

Since $\lim_{\to\infty} Y_n = X_+^{-1}$, then

$$\begin{aligned} \|X_{n+1} - X_+\| &\leq 2\|A^\star\| \|X_+^{-1} A\| \|X_n - X_+\|, \\ &\leq \|X_+^{-1} A\| \|X_n - X_+^{-1}\|. \end{aligned} \tag{23}$$

Thus the inequality (15) holds. ∎

We note that from Algorithm (8) $I - X_n Y_n = Y_{n+1} - Y_n \to 0$, as $n \to \infty$. Thus one stopping criterion may be $\|I - X_n Y_n\| < \epsilon$, for small $\epsilon > 0$. The effect of the stopping criterion can be seen from the following Theorem.

Theorem 1.4 *If Eq. (5) has a solution and after n iterative steps of Algorithm (8), we have* $\|I - X_n Y_n\| < \epsilon$, *thus*

$$\|X_n + A^\star X_n^{-1} A - I\| \leq \epsilon \|A\|^2 \|X_+^{-1}\|.$$

Proof. Since,

$$\begin{aligned} X_n + A^\star X_n^{-1} A - I &= X_n - X_{n+1} + A^\star \left(X_n^{-1} - Y_{n+1}\right) A \\ &= A^\star \left(Y_{n+1} - Y_n\right) A + A^\star \left(X_n^{-1} - Y_{n+1}\right) A \\ &= A^\star \left(Y_{n+1} - X_n^{-1} + X_n^{-1} - Y_n\right) A + A^\star \left(X_n^{-1} - Y_{n+1}\right) A \\ &= A^\star X_n^{-1} \left(I - X_n Y_n\right) A \end{aligned} \tag{24}$$

Taking norm on both sides,

$$\begin{aligned} \|X_n + A^\star X_n^{-1} A - I\| &\leq \|A\|^2 \|X_+^{-1}\| \|I - X_n Y_n\| \\ &\leq \epsilon \|A\|^2 \|X_+^{-1}\| \end{aligned} \tag{25}$$

■

2. NUMERICAL EXPERIMENTS

In this section, numerical experiments are given to display the flexibility of the new inversion free variant of the basic fixed point iteration methods. The maximal solution are computed for some different matrices A with different orders. We will compare the suggested Algorithm (8)with Algorithm (6) and Algorithm (7). The numerical experiments were carried out on an IBM-PC Pentium IV 2000 MHz computer. Double precision is used in the following calculations. The machine precision is approximately 1.11022×10^{-16}. For the following examples, we use the practical stopping criterion $\|X + A^T X^{-1} A - I\| < 10^{-16}$.

Example 2.1 *Consider Eq. (5) with normal matrix*

$$A = \frac{1}{32} \begin{pmatrix} 0.2 & -0.1 & -0.5 & 0.1 \\ -0.1 & 0.6 & -0.5 & 0.7 \\ -0.5 & -0.5 & 0.1 & 0.8 \\ 0.1 & 0.7 & 0.8 & 0.5 \end{pmatrix}.$$

For this matrix the spectral norm is $\|A\| = 0.0412375$. The exact maximal solution can be found according to the formula

$$X_+ = \frac{1}{2}\left[I + (I - 4A^\star A)^{1/2}\right],$$

which is valid for any normal matrix A with $\|A\| \leq 1/2$ (see [12]). Therefore the exact maximal solution is

$$X_+ = \begin{pmatrix} 0.999697 & -0.234558\ 10^{-3} & 0.195301\ 10^{-4} & 0.391194\ 10^{-3} \\ -0.234558\ 10^{-3} & 0.998915 & -0.254492\ 10^{-3} & -0.352352\ 10^{-3} \\ 0.195301\ 10^{-4} & -0.254492\ 10^{-3} & 0.998876 & -0.784171\ 10^{-4} \\ 0.391194\ 10^{-3} & -0.352352\ 10^{-3} & -0.784171\ 10^{-4} & 0.99864 \end{pmatrix}.$$

Algorithm (6) needs 9 iterations to find the above maximal solution, Algorithm (7) needs 5 iterations and the suggested algorithm needs 5 iterations as Algorithm (7) but the number of operations is less than Algorithm (7).

Example 2.2 We consider Eq. (5) with nonnormal matrix

$$A = \frac{1}{100}\begin{pmatrix} 0.2 & -0.1 & 0.4 \\ 0.7 & 0.6 & -0.5 \\ 0.4 & 0.8 & 0.6 \end{pmatrix}.$$

For this matrix the spectral norm is $\|A\| = 0.00796591$. We will obtain the maximal solution X_+ (with first fifteen digits) by any iterative algorithm. Therefore the maximal solution is

$$X_+ = \begin{pmatrix} 0.999931 & -0.72008\ 10^{-4} & 0.299962\ 10^{-5} \\ -0.72008\ 10^{-4} & 0.999899 & -0.140023\ 10^{-4} \\ 0.299962\ 10^{-5} & -0.140023\ 10^{-4} & 0.999923 \end{pmatrix}.$$

Algorithm (6) needs 5 iterations to find the maximal solution, Algorithm (7) needs 3 iterations and the suggested algorithm needs 3 iterations as Algorithm (7) but the number of operations is less than Algorithm (7).

3. CONCLUSIONS AND REMARKS

In this paper we considered the nonlinear matrix equations (5). We suggested a new inversion free variant of the basic fixed point iteration method. We achieved the conditions for the existence of a positive definite solution. We discussed an iterative algorithm from which a solution can always be calculated numerically whenever the equation is solvable. Moreover, Two numerical examples are given to show the accuracy of the suggested algorithm. We observe that our suggested algorithm also avoid matrix inversion and involves only matrix-matrix multiplication. Furthermore the algorithm requires only three matrix multiplications per step, whereas Algorithm (6) and Algorithm (7) require four matrix multiplications per step.

ACKNOWLEDGEMENT

The author wishes to thank the referee for his careful reading of the manuscript and his fruitful comments and suggestions.

REFERENCES

[1] Anderson, W. N. Jr, Morley, T. D. and Trapp, G. E., Positive solutions to $X = A - BX^{-1}B^{\star}$, *Linear Algebra Appl.,* 134 (1990), 53-62.

[2] Engwerda, J. C., Ran, A. C. M. and Rijkeboer, A. L., Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^{\star}X^{-1}A = Q$, *Linear Algebra Appl.,* 186 (1993), 255-275.

[3] Engwerda, J. C., On the existence of the positive definite solution of the matrix equation $X + A^{T}X^{-1}A = I$, *Linear Algebra Appl.,* 194 (1993), 91-108.

[4] Guo, C.-H. and Lancaster, P., Iterative solution of two matrix equations, *The Math.Comput.,* 68 (1999), 1589-1603.

[5] Ivanov, I. G. and El-Sayed, Salah M., Properties of positive definite solutions of the equation $X + A^{\star}X^{-2}A = I$, *Linear Algebra Appl.,* 279 (1998), 303-316.

[6] Lancaster, P. and Rodman, L., *Algebraic Riccati Equations,* Oxford Science Publ., 1995.

[7] El-Sayed, Salah M. and Ramadan, M. A., On the existence of a positive definite solution of the matrix equation $X - A^{\star}\sqrt[2^m]{X^{-1}}A = I$, *Intern. J. Computer Math.,* 76 (2001), 331-338.

[8] El-Sayed, Salah M., Two iterations processes for computing positive definite solutions of the matrix equation $X - A^{\star}X^{-n}A = I$, *Comput. & Math. Appl.,* 41 (2001), 579-588.

[9] El-Sayed, Salah M. and Ran, A. C. M., On an iteration methods for solving a class of nonlinear matrix equations, *SIAM J. Matrix Anal. Appl.,* 23 (2001), 632-645.

[10] El-Sayed, Salah M. and El-Alem, M., Some properties for the existence of a positive definite solution of matrix equation $X + A^{\star}X^{-2^m}A = I$, *Appl. Math. Comput.,* 128 (2002), 99-108.

[11] El-Sayed, Salah M., Two sided iteration methods for computing positive definite solutions of a nonlinear matrix equation, *J. Aus. Math. Soc., Series B,* 44 (2003), 145-152.

[12] Zhan, X. and Xie, J., On the matrix equation $X+A^{\star}X^{-1}A = I$, *Linear Algebra Appl.,* 247 (1996), 337-345.

[13] Zhan, X., Computing the extremal positive definite solutions of a matrix equation, *Siam J. Sci. Comput.,* 17 (1996), 1167-1174.

FRACTIONAL CALCULUS AND INTERMEDIATE PHYSICAL PROCESSES

A. M. A. El-Sayed
Faculty of Science, Alexandria University, Alexandria, Egypt
e.mail: amasayed@hotmail.com and amasayed@maktoob.com

Abstract

One of the main applications of the fractional calculus, integration and differentiation of arbitrary (fractional) orders, is modeling of the intermediate physical processes. Here we survey our results on two of these processes, the diffusion-wave process and the convection-diffusion process.

1. INTRODUCTION

Let $J = [0, T]$, f be integrable on J and g be (at least) absolutely continuous on J.

Definition 2.1. The (Riemann-Liouville) fractional order integral of f of order $\beta \geq 0$ is defined by (see [11], [12] and [15]- [17])

$$I^{\beta}\ f(t)\ =\ \int_0^t \frac{(t-s)^{\beta-1}}{\Gamma(\beta)} f(s)\ ds = \int_0^t \phi(t-s)\ f(s)\ ds = f(t) * \phi_{\alpha}(t), \qquad (1)$$

where $\phi_{\beta}(t)\ =\ \frac{t^{\beta-1}}{\Gamma(\beta)}$, for $t\ >\ 0$, $\phi_{\beta}(t)\ =\ 0$, for $t\ \leq 0$, and ([10])

$$\lim_{\beta \to 0} I^{\beta}\ f(t)\ =\ \lim_{\beta \to 0}\ f(t) * \phi_{\beta}(t)\ =\ f(t) * \delta(t)\ =\ \int_0^t\ f(s)\ ds.$$

Definition 2.2. The (Caputo) fractional order derivative of g of order $\alpha \in (0, 1)$ is defined by ([1]- [5], [11] and [15]-[17])

$$D^{\alpha}\ g(t)\ =\ I^{1-\alpha}\ Dg(t), \quad D\ =\ \frac{d}{d\ t}, \qquad (2)$$

where

$$\lim_{\alpha \to 1}\ D^{\alpha} g(t)\ =\ \frac{dg(t)}{dt} \qquad \text{and} \qquad D^{\alpha} k = 0,\ k \neq 0 \text{ is constant.}$$

When $\alpha \in (n-1, n)$ and if $D^n g$ is integrable we have

$$D^{\alpha}\ g(t)\ =\ I^{n-\alpha}\ D^n g(t). \qquad (3)$$

Let X be a Banach space (with norm $\|.\|$) and $L(J, X)$ be the class of integrable functions defined on J with values in X. Let A be a closed linear operator with dense

domain $D(A) \subset X$. Consider the two Cauchy problems, the abstract fractional order diffusion problem,

$$D^{\gamma}u(t) = Au(t), \quad t \in (0,T], \text{ with } \quad u(0) = u_o, \quad \gamma \in (0,1] \tag{4}$$

and the abstract fractional order wave problem

$$D^{\beta}u(t) = Au(t), \quad t \in (0,T], \quad \text{with } \quad u(0) = u_o \quad u_t(0) = 0, \quad \beta \in [1,2). \tag{5}$$

The author (see [2]) proved, under certain conditions, that the problem (4) has a unique solution $u_\gamma(t) \in L(J, D(A))$ and the problem (5) has a unique solution $u_\beta(t) \in L(J, D(A))$, these two solutions satisfy the continuation properties

$$\lim_{\gamma \to 1^-} u_\gamma(t) = \lim_{\beta \to 1^+} u_\beta(t) = u(t) = T(t)u_o, \tag{6}$$

$$\lim_{\gamma \to 1^-} D^{\gamma}u_\gamma(t) = \lim_{\beta \to 1^+} D^{\beta}u_\beta(t) = AT(t)u_o = du(t)/dt, \tag{7}$$

where $u(t) = T(t)u_o$ is the solution of the Cauchy problem

$$du(t)/dt = Au(t), \quad t \in (0,T], \quad \text{and } \quad u(0) = u_o \tag{8}$$

and $\{T(t), t \geq 0\}$ is the semigroup generated by the operator A.
Combining these results the abstract diffusion-wave problem have been defined (see[3]) as

$$D^{\alpha}u(t) = Au(t), \quad t \in (0,T], \text{ with } \quad u(0) = u_o, \quad u_t(0) = 0, \quad \alpha \in (0,2). \tag{9}$$

Here we have two remarks:
(1) The continuation properties (6) and (7) have been proved under the assumption that $u_t(0) = 0$.

(2) $\alpha \in (0,2)$ is being understood to consider either problem (4) whenever $\alpha \in (0,1]$ or problem (5) when $\alpha \in [1,2)$.

In this paper we formulate the more general and accurate model of the (abstract) diffusion-wave problem as

$$D^{\alpha}u(t) = \int_0^t h(t-s)Au(s)\, ds, \quad t > 0, \quad \alpha \in (0,1], \quad u(0) = u_o. \tag{10}$$

The existence and uniqueness of the solution $u_\alpha \in L(J, D(A))$ will be proved. The continuation as $\alpha \to 1$ will be studied. The special cases, fractional-order diffusion

problem, diffusion problem, fractional-order wave problem and wave problem, will be given. Finally, we consider the model of the convection-diffusion process ([7])

$$\left.\begin{array}{rl} \frac{\partial u(x,t)}{\partial t} & =_x W^{\gamma} u(x,t),\ \gamma \in (1,2),\ t \in (0,T),\ x \in (0,b) \\ u(x,0) & = u_o(x), \\ u(0,t) & = u_x(b,t) \ = \ 0, \end{array}\right\} \tag{11}$$

where ${}_xW^{\gamma}u(x,t)$ is the finite Weyl fractional order derivative (see [4]). The existence and uniqueness of the solution $u_\gamma(x,t) \in H^2(0,b) \cap C^1(0,T)$ will be proved, The continuation (of the problem) to the convection problem ($\gamma \to 1$) and to the diffusion problem ($\gamma \to 2$) will be studied.

2. EVOLUTIONARY INTEGRAL EQUATIONS

Let X be a Banach space and A be a closed linear operator with domain $D(A)$ dense in X and satisfying the assumption
(I) A generates an analytic semigroup $\{T(t), t \geq 0\}$ on X . In particular
$\Lambda_1 = \{\lambda \in C : |arg\lambda| < \pi/2 + \delta_1\},\ 0 < \delta_1 < \pi/2$ is contained in the resolvent set of A and $\|(\lambda I - A)^{-1}\| \leq M_1/|\lambda|, Re\lambda > 0$ on Λ_1 , for $M_1 > 0$.
Example Let the operator A be defined by

$$D(A) = \{u(x,t) \in C^2(-\infty,\infty),\ \lim_{x\to\pm\infty} u(x,t) = 0\},\ Au(x,t) = \frac{\partial^2}{\partial x^2}u(x,t), \tag{12}$$

then ([18]) A satisfies the condition (I).
The following results have been proved for the Cauchy problem (10)(see [6])

Theorem 2.1 Let $\alpha \in (0,1),\ u_o \in D(A^2)$, and $e^{-t}h(t) \in L((0,\infty),R)$. If the operator A satisfies the condition (I), then the Cauchy problem (10) has the unique solution $u_\alpha \in L(J, D(A)),\ Du_\alpha \in L(J, D(A))$. given by

$$u_\alpha(t) = u_o - \int_0^t e^s r_\alpha(s) u_o ds, \tag{13}$$

where the resolvent operator r_α satisfies the relations

$$r_\alpha(t)x = -h(t)Ax + r_\alpha(t) * h(t)Ax, \quad a.e., \quad t > 0 \tag{14}$$

and

$$r_\alpha(t)x = h(t)Ax + h(t) * Ar_\alpha(t)x, \quad a.e., \quad t > 0. \tag{15}$$

Theorem 2.2 If the solutions of the initial value problem (10) exists then

$$\lim_{\alpha\to 1^-} u_\alpha(t) = u_1(t) \ \text{ and if } \ Au_o \in D(A^2), \ \text{ then } \ \lim_{\alpha\to 1^-} D^\alpha u_\alpha(t) = \frac{du_1(t)}{dt}, \tag{16}$$

where $u_1(t)$ is the solution of the Cauchy problem

$$\frac{du(t)}{dt} \;=\; \int_0^t h(t-s)\; Au(t), \quad t>0, \quad \gamma\in(0,1], \quad u(0)=u_o.$$

The following corollary can easily proved.

Corollary 2.1 If the solution of the diffusion wave problem (10) exists then it depends continously on the initial data u_o.

3. DIFFUSION-WAVE EQUATION

Now let A be given by (12), then the problem (10) will be of the form.

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} \;=\; \int_0^t h(t-s)\,\frac{\partial^2 u(x,s)}{\partial x^2}\,ds, \quad t>0, \quad \alpha\in(0,1], \quad u(x,0)=u_o(x). \quad (17)$$

Now applying the results of [2], [5] and ([6]) we get the following.

3.1 Fractional-order diffusion process Let $\alpha\in(0,1)$, $\gamma\in(0,1]$ such that $\gamma-\alpha\in(0,1)$ and $h(t)=\phi_{\gamma-\alpha}(t) \;=\; t^{\gamma-\alpha-1}/\Gamma(\gamma-\alpha), \quad t>0.$
Then the problem (17) will be the problem of fractional-order diffusion process

$$\frac{\partial^\gamma u(x,t)}{\partial t^\gamma} \;=\; \frac{\partial^2 u(x,t)}{\partial x^2}, \quad t>0, \quad u(x,0)=u_o(x)$$

with the solution

$$u_\gamma(x,t)=u_o(x)-\int_0^t e^s r_\gamma(s)u_o(x)ds \;\to\; \frac{1}{2\sqrt{\pi t}}e^{\frac{-x^2}{4t}}*u_0(x) \quad \text{as } \gamma\to 1^-.$$

3.2 Diffusion process

Let $\alpha\in(0,1)$ and $h(t)=\phi_{1-\alpha}(t)=\frac{t^{-\alpha}}{\Gamma(1-\alpha)}, \quad t>0.$
Then the problem (17) will be the problem of diffusion process

$$\frac{\partial u(x,t)}{\partial t} \;=\; \frac{\partial^2 u(x,t)}{\partial x^2}, \quad u(x,0) \;=\; u_0(x)$$

with the solution

$$u(x,t) = \frac{1}{2\sqrt{\pi t}}e^{\frac{-x^2}{4t}}*u_0(x).$$

3.3 Fractional-order wave process

Let $\alpha\in(0,1]$, $\beta\in(1,2)$ such that $\beta-\alpha\in(0,1]$ and $h(t)=\phi_{\beta-\alpha}(t)$. Then we deduce that $u_t(x,0)=0$ and the problem (17) will be the problem of fractional-order wave process

$$\frac{\partial^\beta u(x,t)}{\partial t^\beta} \;=\; \frac{\partial^2 u(x,t)}{\partial x^2}, \quad t>0, \quad u(x,0)=u_o(x) \;\text{ and }\; u_t(x,0)=0$$

with the solution

$$u_\beta(x,t) = u_o(x) - \int_0^t e^s r_\beta(s) u_o(x) ds \ \rightarrow \ \frac{1}{2\sqrt{\pi t}} e^{\frac{-x^2}{4t}} * u_0(x) \quad \text{as } \beta \rightarrow 1^+.$$

3.4 Wave process

Let $\alpha \rightarrow 1^-$ and $h(t) = 1$. Then we deduce that $u_t(x,0) = 0$ and the problem (17) will be the problem of wave process

$$\frac{\partial^2 u(x,t)}{\partial t^2} \ = \ \frac{\partial^2 u(x,t)}{\partial x^2}, \quad t > 0, \ \ u(x,0) = u_o(x) \ \text{ and } \ u_t(x,0) = 0$$

with the solution (D'Alembert solution of the Wave Equation)

$$u(x,t) = \lim_{\alpha \rightarrow 1^-} u_\alpha(x,t) = u_1(x,t) = \frac{1}{2}(u_o(x-t) + u_o(x+t)).$$

Remark

In Sect. 3 we deduced that the initial value $u_t(x,0) = 0$. Recently in [9] we study a modification of our model in order that $u_t(x,0) \neq 0$.

4. CONVECTION-DIFFUSION PROCESS

Let $\beta \in (0,1]$. Consider now the mixed problem (see [7])

$$\frac{\partial \, u(x,t)}{\partial \, t} \ = \ {}_xW_b^{-\beta} \ \frac{\partial^2 \, u(x,t)}{\partial \, x^2} \ = \int_t^b \ \frac{(s-t)^{-\beta}}{\Gamma(\beta)} \frac{\partial^2 \, u(x,t)}{\partial \, x^2}, \ t \in (0,T) \ , \ x \in (0,b) \tag{18}$$

$$u(x,0) \ = \ u_o(x) \quad \text{and} \ \ u(0,t) \ = \ u_x(b,t) \ = \ 0. \tag{19}$$

Definition 4.1 By a solution of the problem (18) and (19), we mean a function $u(x,t) \in H^2(0,b) \cap C^1(0,T)$ (the space $H^2(a,b)$ is the Sobolev space (see [18])) which satisfies the problem.

Let the operator A be defined as

$$D(A) \ = \{u(x,t) : \ u(x,t) \in H^2(0,b) \cap C^1(0,T)), \ u(0,t) \ = \ u_x(b,t) \ = \ 0, \quad \forall \, t \geq 0\} \tag{20}$$

$$Au(x,t) \ = \ {}_xW_b^{-\beta} \ \ \frac{\partial^2 \, u(x,t)}{\partial x^2}. \tag{21}$$

The following results have been proved in ([7]).

Theorem 4.1 Let $u_o(x) \in D(A)$. The initial value problem (18) and (19) has the unique solution

$$u(x,t) \ = \ T_\beta(t) \ u_o(x) \ \in \ H^2(0,b) \ \cap \ C^1(0,T), \tag{22}$$

where $\{ T_\beta(t) \ , \ t > 0 \}$ is the semi-group generated by the operator A.

Theorem 4.2 (Continuation theorem) Let the solution of (18) and (19) exist, then

1. When $\beta \rightarrow 1$, the problem (18) and (19) will be the convection problem

$$\frac{\partial\, u(x,t)}{\partial\, t} = - \frac{\partial\, u(x,t)}{\partial\, x}, \ t \in (0,T) \ , \ x \in (0,b)$$

$$u(x,0) = u_o(x) \quad \text{and } u(b,t) = 0.$$

2. When $\beta \rightarrow 0$, then the problem (18) and (19) will be the diffusion problem

$$\frac{\partial\, u(x,t)}{\partial\, t} = \frac{\partial^2\, u(x,t)}{\partial\, x^2}, \ t \in (0,T) \ , \ x \in (0,b)$$

$$u(x,0) = u_o(x), \quad \text{and } u(0,t) = u_x(b,t) = 0.$$

4.1 Continuation of the solution

Recently in [8] we study (by using the Trotter -Kato Theorem) the continuation as $\beta \rightarrow 1^-$ and $\rightarrow 0$ of the semigroup $T_\beta(t)$, consequently the solution (22) to the solution of the convection and diffusion problems respectively.

REFERENCES

[1] Caputo, M. Linear model of dissipation whose Q is almost frequency independent-II, *Geophys. J. R. Astr. Soc.* Vol. 13, pp. 529-539 (1967).

[2] Fractional order evolution equations. *J. of Frac. Calculus* Vol. 7, pp. 89-100 (1995).

[3] El-Sayed, A. M. A. Fractional-order diffusion-wave equation. *Int. J. Theoretical Physics.* Vol. 35 (2) , pp. 311-322 (1996).

[4] El-Sayed, A. M. A. Finite Weyl fractional calculus and abstract fractional differential equations. *J.of Frac. Calculus* Vol. 9, pp. 59-68, May (1996).

[5] El-Sayed, A. M. A. Fractional-order evolutionary integral equations, *Apll. Math. and Comput.* Vol. 98, No. 2-3, pp. 139-146 (1999).

[6] El-Sayed, A. M. A. and M. A.E. Aly, Continuation theorem of fractional order evolutionary integral equations, *Korean J. of Comput. and Appl. Math.* Vol. 9, No. 2, pp. 525-533 (2002).

[7] El-Sayed, A. M. A.and F. M. Gaafar., Fractional calculus and some intermediate physical processes *Appl. Math and Comput.*, Vol. 144, pp. 117-126, (2003).

[8] El-Sayed, A. M. A. On the continuation properties of the convection-diffusion process (in progress).

[9] El-Sayed, A. M. A. and Aly, M. A.E., On the diffusion-wave process (in progress).

[10] Gelfand, I.M. and Shilov, G.E., *Generalized functions,* Vol. 1, Moscow (1958).

[11] Gorenflo, R. and Mainardi, F., *Fractional calculus: integral and differential equations of fractional order.* In: Fractals and Fractional Calculus in Continuum Mechanics (A. Carpinteri and F. Mainardi, Ed-s), Springer Verlag - Wien - New York, pp. 223-276 (1997).

[12] Miller, K. S. and Ross, B., *An Introduction to the Fractional Calculus and Fractional Differential Equations,* John Wiley & Sons. Inc., New York (1993).

[13] Grimmer, R.C. and Pritchard, A.J., Analytic resolvent operators for integral equations in Banach spaces, *J. Diff. Equat.* Vol. 50, pp. 744-759 (1993).

[14] Pruss, J. *Evolutionary Integral Equations and Applications.* Birkhauser Verlag Basel- Boston- Berlin (1993).

[15] Podlubny, I. and El-Sayed, A. M.A., On two definitions of fractional calculus, *Slovak Academy Of Sciences Institute Of Experimental Physics* U E F - 03 -96 ISBN 80-7099- 252-2 (1996).

[16] Podlubny, I. *Fractional Differential Equation.* Acad. Press, San Diego - New York - London, 1999.

[17] Samko, S.G., Kilbas, A.A. and Marichev O.I., *Integral and derivatives of the fractional orders and some of their applications.* Nauka i Tekhnika Minisk, (1987).

[18] Oden, J. T. *Applied Functional Analysis.* Prentice-Hall, Inc., New York (1979).

COSET ENUMERATION ALGORITHM FOR SYMMETRICALLY GENERATED GROUPS

M. Sayed
Department of Mathematics and Computer Science,
Faculty of Science, Kuwait University,
P.O. Box:5969 Safat 13060, State of Kuwait
email: msayed@mcs.sci.kuniv.edu.kw

Abstract.

We conduct a computerized search for groups generated by symmetric sets of involutions. A coset enumeration algorithm for groups presented in this way together with its computer implementation is described.

Keywords. Coset enumeration, symmetric presentation, involutory generators, progenitor.

AMS Subject Classification (2000): 20Bxx

1. INTRODUCTION

The Todd-Coxeter algorithm described in [14] remains a primary reference for coset enumeration programs. It may be viewed as a means of constructing permutation representations of finitely presented groups. A statement of the basic technique and the early study appear in [7, 8]. A detailed survey and comparison of different strategies are given in [2]. A contemporary work is described in [6, 9]. All the strategies and variants of the algorithm perform essentially the same calculations as the original Todd-Coxeter algorithm, merely choosing different orders in which to process the available information.

In this paper we describe a particular technique of single coset enumeration which gives the action of elements of a group generated by symmetric set of involutions on the cosets of a group of automorphisms of these generators. The algorithm may appear significantly different from the Todd-Coxeter algorithm, but can still be viewed as another variant of the algorithm, one which uses some additional group-theoretical information. The basis of the algorithm was described in [11]. It is to be noted that our algorithm is practical in the sense that it can be programmed readily on a computer [1] and results can be obtained in reasonable time.

2. INVOLUTORY SYMMETRIC GENERATORS OF GROUPS

Let G be a group and let $T = \{t_0, t_1, \ldots, t_{n-1}\}$ be a set of elements of order m in G. Making the definitions $T_i = \langle t_i \rangle$ and $\overline{T} = \{T_0, T_1, \ldots, T_{n-1}\}$ allows us to define $N = \mathcal{N}_G(\overline{T})$, the set normalizer in G of $\overline{T}$. We say that T is a *symmetric generating set* for G if the following two conditions hold:

(i) $G = \langle T \rangle$, and

(ii) N permutes $\overline{T}$ transitively.

We call N the *control subgroup*. Conditions (i) and (ii) imply that G is a homomorphic image of the *progenitor*

$$m^{*n} : N,$$

where m^{*n} represents a free product of n copies of the cyclic group C_m and N is a group of automorphisms of m^{*n} which permutes the n cyclic subgroups by conjugation. For further information about the symmetric generations of groups the reader is referred to [4, 5, 10].

Since in this paper we are only concerned with involutory symmetric generators we restrict our attention to the case $m = 2$ (while N will simply act by conjugation as permutations of the n involutory symmetric generators).

Theorem 2.1. *All non-abelian finite simple groups can arise as finite homomorphic images of progenitors of the form* $2^{*n} : N$.

Proof. Let H be a maximal subgroup of a finite simple group G. Suppose that $1 \neq \mathbf{t} \in G$, $\mathbf{t}^2 = 1$. Under the subgroup H, $\mathbf{t}^G$, the conjugacy class of $\mathbf{t}$ in G, splits into orbits as

$$\mathbf{t}^G = \mathcal{T}_1 \dot{\cup} \mathcal{T}_2 \dot{\cup} \ldots \dot{\cup} \mathcal{T}_r.$$

Without loss of generality, we may assume that $\mathcal{T}_1 = \{\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_{n-1}\}$ is not a subset of H. It is clear that

$$\mathcal{N}_G(\langle \mathcal{T}_1 \rangle) \geq \langle H, \mathcal{T}_1 \rangle = G,$$

since H is maximal in G and $\mathcal{T}_1$ is not a subset of H. Therefore,

$$1 \neq \langle \mathcal{T}_1 \rangle \lhd G,$$

and, since G is simple, we have

$$\langle \mathcal{T}_1 \rangle = G.$$

Moreover, if $\pi \in H$ and $\mathbf{t}_i^{\pi} = \mathbf{t}_i$ $(i = 0, 1, \ldots, n-1)$ then $\pi \in \mathcal{Z}(G)$ and so $\pi = 1$, *i.e.* H permutes the elements of $\mathcal{T}_1$ faithfully (and transitively). Now, let 2^{*n} denote a free product of n copies of the cyclic group C_2 with involutory generators $t_0, t_1, \ldots, t_{n-1}$ and let $N \cong H$ consist of all automorphisms of 2^{*n} which permute the t_i as H permutes the $\mathbf{t}_i$:

$$\pi^{-1}t_i\pi = t_i^\pi = t_{\pi(i)} \text{ for } \pi \in N.$$

Then, clearly G is a homomorphic image of $2^{*n} : N$, a split extension of 2^{*n} by the permutation automorphisms N. $\diamondsuit$

Since the progenitor is a semi-direct product (of $\langle T \rangle$ with N), it follows that in any homomorphic image G, we may use the equation:

$$t_i\pi = \pi t_i^\pi = \pi t_{\pi(i)},$$

or $i\pi = \pi i^\pi$ as we will more commonly write (see below), to gather the elements of N over to the left. Each element of the progenitor can be represented as πw, where $\pi \in N$ and w is a word in the symmetric generators. Indeed, this representation is unique provided w is simplified so those adjacent symmetric generators are distinct. Thus any additional relator by which we must factor the progenitor to obtain G must have the form

$$\pi w(t_0, t_1, \ldots, t_{n-1}),$$

where $\pi \in N$ and w is a word in T. Another consequence of this is that a relation of the form $(\pi t_i)^n = 1$ for some $\pi \in N$ in a permutation progenitor becomes:

$$\pi^n = t_i t_{\pi(i)} \ldots t_{\pi^{n-1}(i)}.$$

3. COSET ENUMERATION ALGORITHM

In this section we describe how a factor group

$$G \cong \frac{2^{*n}:N}{\pi_1 w_1, \pi_2 w_2, \ldots, \pi_s w_s},$$

may be identified. We need to establish the order of G by enumerating the cosets of a subgroup (of G) of known order. Naturally, we would like this subgroup to be the control subgroup N.

We will allow i to stand for the coset Nt_i, ij for the coset Nt_it_j and so on. The coset N is denoted by $*$. We will also let i stand for the symmetric generator t_i when there is no danger of confusion. Thus we write, for instance, $ij \sim k$ to mean $Nt_it_j = Nt_k$ and $ij = k$ to mean $t_it_j = t_k$.

We define the subgroups N^i, N^{ij}, N^{ijk}, ...(for i, j and k distinct) as follows:

$$N^i = \mathcal{C}_N(\langle t_i \rangle),$$
$$N^{ij} = \mathcal{C}_N(\langle t_i, t_j \rangle),$$
$$N^{ijk} = \mathcal{C}_N(\langle t_i, t_j, t_k \rangle),$$

or, more generally,

$$N^{i_1 i_2 \ldots i_m} = \mathcal{C}_N(\langle t_{i_1}, t_{i_2}, \ldots, t_{i_m} \rangle),$$

for $i_1, i_2, \ldots, i_m$ distinct.

It is sometimes useful to have the notation of a *length* of a coset. The fact that for $\pi \in N$ we have

$$Nw(t_i)\pi = N\pi^{-1}w(t_i)\pi = Nw(t_i^\pi) = Nw'(t_i)$$

shows that all N-cosets have a representative in $\langle T_0 \cup T_1 \cup \ldots \cup T_{n-1} \rangle$. We are in a position to define the length, $L(Nw) = L(w)$, of a coset Nw. Firstly, we have $L(N) = 0$. If Nw has length k and $t \in T$ then Nwt has length at most $k+1$ and has length precisely $k+1$ if it does not have (or has not been proved to have) length at most k. We specify that all cosets of length $k+1$ have the form Nwt where $L(Nw) = k$ and $t \in T$.

Following the Todd-Coxeter algorithm we maintain a set C of cosets and a table (for each additional relation of the progenitor) which can be considered as a function $f : C \times T \longrightarrow C$. We read $f(w, t_i) = w'$ as meaning that $Nwt_i = Nw'$ where w and w' are words in the symmetric generators of length k and at most $k+1$ respectively. It is also convenient to have some way of recording in the table when a coset w' has been proved to be the same (in G) as an earlier coset w so that references to w' can be diverted to w. We have to ensure that cosets which are distinct in G remains distinct in C. We also demand that if Nw is earlier in the table than Nw', then $L(Nw) \leq L(Nw')$.

We start with an empty table and progressively modify it, using the above function, until it represents the permutation action of T on the cosets of N (in G). We will know that we have completed the coset enumeration when the set of right cosets obtained is closed under right multiplication. We can say that if N is of finite index, closure must be reached after finite number of steps. The proof is very similar the proof of a theorem of Mendelson (see, for example, [13]) which states that the Todd-Coxeter method will succeed in finding the permutation representation of a group G provided that the index $|G : H|$ is finite, where H is a subgroup of G.

With the observation of this section, we are in a position to carry out simple coset enumeration. We consider the group

$$G \cong \frac{2^{*4}{:}S_4}{(0,1){=}t_0t_1t_0},$$

which means that the progenitor $2^{*4} : S_4$ quotiented out by the relation $(0, 1) = t_0t_1t_0$. Here the control subgroup $N \cong S_4$ acts on the 2^{*4} in its normal action on four points.

It is clear that $N = N(0,1) = Nt_0t_1t_0$, which we write as $* \sim 010$ in our notation. By postmultiplying both sides by t_0, we deduce that $Nt_0 = Nt_0t_1$, that is $01 \sim 0$. In general, we have $* \sim iji$ and thus that $ij \sim i$. The relation table for the coset enumeration of G over N is given below (see Table 1):

The symmetric generators t_0 acts on these five cosets by right multiplication as the transposition interchanging the coset $*$ with the coset 0. Since $(*,0)(*,1)(*,0) = (0,1)$ we see that our additional relation is satisfied by these transpositions, and we have a symmetric presentation of S_5.

4. IMPLEMENTATION

This section of the paper describes the implementation of the coset enumeration algorithm of section 3. A more detailed description is provided. The program is heavily Magma-dependent, but can be readily modified for other packages such as GAP.

Given a control subgroup N (of a group G) as permutations on n letters together with some relations, in which elements of N are written in terms of n (involutory) symmetric generators (of G), the program performs a coset enumeration for G over N. The program returns what is essentially a Cayley graph of the action of G on the cosets of N. Each element of G is represented by a permutation of N followed by a word in the symmetric generators. Indeed, the program allows the user readily to pass between the symmetric representation of an element of G and its action on the cosets of N. If the index $|G : N|$ is finite, the procedure does finish and succeed in finding the permutation representation of the group G.

The system contains a set of routines (procedures) such as *eqcoset*, *canon*, *main*, *x2per*, *t2per*, *sym2per* and *per2sym*. In the remainder of this paper, we give a detailed description and an outline of some difficulties which arise with the implementation.

4.1 Eqcoset Procedure

The same coset will often have many names and the purpose of the procedure is to find these coincidences by using the additional relations of the group G.

Given any two sequences (each represents a word in the symmetric generators):

$$e_1 = [a_1, a_2, \ldots, a_r] \text{ and } e_2 = [b_1, b_2, \ldots, b_r],\ a_i, b_i \in \{1, 2, \ldots, n\};$$

the procedure checks the equivalence ($e_2^p = e_1$, for some $p \in N$) between them. If

they are equivalent it determines the permutation p of N such that $e_2^p = e_1$.

We know that

$$N \geq N^{b_1} \geq N^{b_1 b_2} \geq \cdots \geq N^{b_1 b_2 \ldots b_r}.$$

Assume that

$$n_1 = |N : N^{b_1}| \text{ and } n_i = |N^{b_1 \ldots b_{i-1}} : N^{b_1 \ldots b_i}|,\ i \in \{2, 3, \ldots, r\}.$$

Consequently, there exist tarnseversals

$$\{\tau_1, \ldots, \tau_{n_1}\},\ \{\sigma_1, \ldots, \sigma_{n_2}\},\ \{\rho_1, \ldots, \rho_{n_3}\},\ \ldots \text{ , and } \{\phi_1, \ldots, \phi_{n_r}\}$$

such that

$$\begin{gathered} N = N^{b_1}\tau_1 \dot{\cup} N^{b_1}\tau_2 \dot{\cup} \ldots \dot{\cup} N^{b_1}\tau_{n_1}, \\ N^{b_1} = N^{b_1 b_2}\sigma_1 \dot{\cup} N^{b_1 b_2}\sigma_2 \dot{\cup} \ldots \dot{\cup} N^{b_1 b_2}\sigma_{n_2}, \\ N^{b_1 b_2} = N^{b_1 b_2 b_3}\rho_1 \dot{\cup} N^{b_1 b_2 b_3}\rho_2 \dot{\cup} \ldots \dot{\cup} N^{b_1 b_2 b_3}\rho_{n_3}, \\ \vdots \\ N^{b_1 \ldots b_{r-1}} = N^{b_1 \ldots b_r}\phi_1 \dot{\cup} N^{b_1 \ldots b_r}\phi_2 \dot{\cup} \ldots \dot{\cup} N^{b_1 \ldots b_r}\phi_{n_r}. \end{gathered}$$

Now, if $e_2^p = e_1$, $p \in N$, then we can find the permutation $p = \phi' \ldots \rho'\sigma'\tau'$, where
$\tau' \in \{\tau_1, \ldots, \tau_{n_1}\}$, $\sigma' \in \{\sigma_1, \ldots, \sigma_{n_2}\}$, $\rho' \in \{\rho_1, \ldots, \rho_{n_3}\}$, ... , and $\phi' \in \{\phi_1, \ldots, \phi_{n_r}\}$, as follows: Since $\sigma', \rho', \ldots, \phi'$ fix b_1, the equation $b_1^P = a_1$ can be reduced to $b_1^{\tau'} = a_1$. Also, the permutations $\rho', \ldots, \phi'$ fix b_2, so the equation $b_2^P = a_2$ can be reduced to $b_2^{\sigma'\tau'} = a_2$. Similarly we have $b_3^{\rho'\sigma'\tau'} = a_3$, ... , $b_r^{\phi' \cdots \rho'\sigma'\tau'} = a_r$. Thus, we can easily identify (in a recursive manner) the permutations $\tau', \sigma', \rho', \ldots, \phi'$ and consequently p.

eqcoset 1: Set p as the identity element of S_n.
eqcoset 2: For $i \in \{1, 2, \ldots, r\}$ do
If $i = 1$ then $trans = Transversal(N, N^{b_1})$,
else $trans = Transversal(N^{b_1 b_2 \ldots b_{i-1}}, N^{b_1 b_2 \ldots b_i})$.
If there exist a permutation $trans[j]$ such that $a_i^{(p^{-1} \cdot trans[j]^{-1})} = b_i$ then set $p =$ trans[j]$\cdot p$,
otherwise e_1 is not equivalent to e_2, leave the loop and return with a proper prompt.

4.2 Canon Function

The function takes such a word in the symmetric generators and reduces it to its shortest form, using the following recursive function:

canon: If any two adjacent elements of a given sequence are equal, delete them and

call the function again with a new reduced sequence, otherwise return.

4.3 Main Processing

We show how the program generates the cosets and give an efficient method for handling the collapses. We will allow the variable *level* to stand for the length of the coset corresponding to the last row in the coset tables.

Input and initialization. The control subgroup N is defined as a permutation group of degree n. Now Magma and other theoretical packages handle permutations of a high degree with immense ease. The two sequences

$$\pi = [\pi_1, \pi_2, \ldots, \pi_m] \text{ and } w = [w_1, w_2, \ldots, w_m],$$

where $\pi_i \in N$ and w_i are words in the symmetric generators, represent the left and the right hand side of the problem relations. Also, C and $CT[i]$, $i \in \{1, 2, \ldots, m\}$, are defined as sequences of sequences whose terms are the complete set of coset representative words and the coset tables respectively.

Reduction. Any word w in the symmetric generators is put by the procedure into its canonically shortest form. No other representations of group elements are used; words in the symmetric generators are simply shortened by application of the relations (and their conjugates under N). The relations

$$w_i = \pi_i,\ i \in \{1, 2, \ldots, m\},$$

where $w_i = t_{i_1} t_{i_2} \ldots t_{i_r}$ and $\pi_i \in N$, can be written as

$$t_{i_1} t_{i_2} \ldots t_{i_k} = \pi_i t_{i_r} t_{i_{r-1}} \ldots t_{i_{k+1}},$$

where k equals to $\frac{r}{2}$ or $\frac{r+1}{2}$ according to whether r is even or odd respectively.

The procedure checks if a part of any given word w in the symmetric generators of length equal to k is equivalent to $t_{i_1} t_{i_2} \ldots t_{i_k}$, the left hand side of one of the previous relations using the *eqcoset* procedure, if so, the procedure replaces this part by $t_{i_r} t_{i_{r-1}} \ldots t_{i_{k+1}}$ after permuting by p^{-1} (a permutation obtained from the *eqcoset* procedure) and moves the permutation $\pi^{p^{-1}}$ over to the left of the whole word.

If a new word w' of length less than the length of w is obtained; in this case the procedure replaces the coset represented by w by the new coset represented by w'. References to the coset represented by w can be diverted to the coset represented by w'. On the other hand, if a new word w' of length equal to the length of w and equivalent to w is obtained, we call that the coset represented by w' has been proved to be the same (in G) as the coset represented by w. It is also convenient to have

some way of recording, in a sequence, all the coincidences that were obtained during the reduction step.

reduce 1: For any sequence in the symmetric generators, set the pointer at the first letter.

reduce 2: Check if the first k elements starting from the pointer position are equivalent to the left hand side of the given relation, replace them by the right hand side of the same relation after permuting by p^{-1} and conjugate the preceding elements by the permutation $\pi_i^{p^{-1}}$.

reduce 3: Call the *canon* function, shift the pointer one position and go to 2.

Collapses. From time to time we pack the sequence of coset representative words, reclaiming the space that was occupied by the redundant elements and this might lead to the collapse of part or the entire coset diagram. The strategy here is to use the collapse procedure at the end of each *level* which leads to minimizing the storage needed. We use all the coincidences obtained during the reduction step together with the problem relations to reduce the coset tables. If the coincidences generate further coincidences, the process must be repeated.

collapse 1: Set $level1 = level$.

collapse 2: If new coincidences have been defined during $level1$ which is equal n, say, try to reduce again the cosets of length equal $n-1$ using these coincidences, together with the problem relations, otherwise stop.

collapse 3: Set $level1 = n-1$ and go to 2.

Termination. We call a coset table is closed if it has no cosets in our tables of length greater than *level*. In this case we call the set of right cosets obtained is closed under right multiplication. Since N is a finitely generated subgroup of countable index in a finitely presented group G, the point of termination will always be reached.

4.4 X2per Procedure

As mentioned earlier, each element of the group G can be represented by a permutation on n letters; n is the cardinality of the permutation group N, followed

by a word in the n involutory symmetric generators. Given a permutation $x \in N$, the procedure constructs a permutation, xp, say, which gives the action of x on the cosets of N in G.

x2per 1: Initialize xs as a sequence of integers of length equal to the number of the cosets of N.

x2per 2: For each i, j such that $length(C[i]) = length(C[j])$, if $(C[i])^x = C[j]$ then set $xs[i] = j$.

x2per 3: Convert the sequence of integers xs to xp, a permutation on the cosets of N.

4.5 T2per Procedure

This procedure gives the action of the symmetric generators on the cosets of N in G. In general t_i in its action on the cosets of N has the form:

$$(*, i)(j, ji) \dots (jk, jki) \dots (jkl, jkli) \dots \text{ , for } i, j, k, l \text{ distinct.}$$

In practice our symmetric presentation is given in terms of a set of generators of N together with one of the symmetric generators. So, if the action of one of the symmetric generators, t_i, on the cosets of N is known, we can obtain the action of the others on the cosets of N by permuting this symmetric generator by NN (the control subgroup N in its action on the cosets of N).

t2per 1: Initialize ts as a sequence of integers of length equal to the number of the cosets of N.

t2per 2: For each i, j such that $length(C[i]) = length(C[j]) + 1$, if $(C[i]\ cat\ 1) = C[j]$ then set $ts[i] = j$ and $ts[j] = i$.

t2per 3: Convert the sequence of integers ts into $tp[1]$, a permutation on the cosets of N.

t2per 4: Construct $tp[i],\ i = 2, 3, \dots, n$ by permuting $tp[1]$ by NN.

4.6 Sym2per Procedure

This procedure converts a symmetrically represented element πw, where $\pi \in N$ and $w = [i, j, k, \dots]$ a word in the symmetric generators, into a permutation acting

on the cosets of N in G. Then $p = x2per(\pi) \cdot tp[i] \cdot tp[j] \cdot tp[k] \ldots$, where p is a permutation acting on the cosets of N.

4.7 Per2sym Procedure

The procedure converts a permutation p (acting on the cosets of N) of G into its symmetric representation. The image of one under p gives the coset representative for Np as a word w in the symmetric generators. Multiplication of p by the symmetric generators in w, in reverse order, yields a permutation which can be identified with an element of N by its action on cosets of length one.

per2sym 1: Assume $p = \pi w$, obtain w as $w = C[1^p]$.

per2sym 2: Obtain π as a permutation on the cosets of N, using the equation $\pi = pw^{-1}$.

per2sym 3: Identify the action of π on the cosets of length one as
$\pi = [j \mid (1^{tp[i]})^\pi = (1^{tp[j]}), \forall\ i, j \in \{1, 2, \ldots, n\}]$.
Finally write π as a permutation of S_n.

5. EXAMPLES

We give here three illustrative examples of the use of the program. We will consider the progenitors $2^{*n} : N$, for $N \cong S_3$, S_4 and $L_2(5)$, for $n = 3, 4$ and 6 respectively.

Example 5.1. Consider the group:

$$G \cong \frac{2^{*3}{:}S_3}{(0,1){=}t_0t_1t_0t_1t_0,\ (0,2,1){=}t_0t_1t_2t_0t_1}.$$

The result of the coset enumeration of G over S_3 is shown in Table 1 and Table 2. Thus, $|G : N| \leq 10$, so $|G| \leq 60 = |L_2(4)|$, and the (relatively) easy task of finding generators for $L_2(4)$, see [3], satisfying the required relations completes the identification of G with $L_2(4)$.

Example 5.2. Consider the group:

$$G \cong \frac{2^{*4}{:}S_4}{(2,3){=}[t_0t_1]^2}.$$

0	1	0	
$*$	0	$01 \sim 0$	$*$
0	$*$	1	$10 \sim 1$
1	$10 \sim 1$	$*$	0
2	$20 \sim 2$	$21 \sim 2$	$20 \sim 2$
3	$30 \sim 3$	$31 \sim 3$	$30 \sim 3$

Table 1: The relation table of the enumeration of S_5 over S_4

0	1	0	1	0	
$*$	0	01	$010 \sim 01$	0	$*$
0	$*$	1	10	$101 \sim 10$	1
1	10	$101 \sim 10$	1	$*$	0
2	20	$201 \sim 02$	$020 \sim 02$	$021 \sim 20$	2
01	$010 \sim 01$	0	$*$	1	10
10	1	$*$	0	01	$010 \sim 01$
02	$020 \sim 02$	$021 \sim 20$	2	21	$210 \sim 12$
20	2	21	$210 \sim 12$	$121 \sim 12$	$120 \sim 21$
12	$120 \sim 21$	2	20	$201 \sim 02$	$020 \sim 02$
21	$210 \sim 12$	$121 \sim 12$	$120 \sim 21$	2	20

Table 2: The first relation table of the enumeration of $L_2(4)$ over S_3

0	1	2	0	1	
$*$	0	01	$012 \sim 10$	1	$*$
0	$*$	1	12	$120 \sim 21$	2
1	10	$101 \sim 10$	$102 \sim 01$	$010 \sim 01$	0
2	20	$201 \sim 02$	0	$*$	1
01	$010 \sim 01$	0	02	$020 \sim 02$	$021 \sim 02$
10	1	$*$	2	20	$201 \sim 02$
02	$020 \sim 02$	$021 \sim 20$	$202 \sim 20$	2	21
20	2	21	$212 \sim 21$	$210 \sim 12$	$121 \sim 12$
12	$120 \sim 21$	2	$*$	0	01
21	$210 \sim 12$	$121 \sim 12$	1	10	$101 \sim 10$

Table 3: The second relation table of the enumeration of $L_2(4)$ over S_3

The result of the coset enumeration of G over S_4 shown in Table 3 indicates that $|G : N| \leq 14$, so $|G| \leq 336 = |PGL_2(7)|$. In fact, if we make the correspondence between the N-cosets and the 7 points and the 7 lines of the projective plane of order 2, see [12], we can easily identify G with $PGL_2(7)$.

0	1	0	1	
$*$	0	$01 \sim 10$	1	$*$
0	$*$	1	$10 \sim 01$	0
1	$10 \sim 01$	0	$*$	1
2	20	$201 \sim 310$	31	3
3	30	$301 \sim 210$	21	2
10	1	$*$	0	$01 \sim 10$
20	2	21	$210 \sim 301$	30
30	3	31	$310 \sim 201$	20
21	$210 \sim 301$	30	3	31
31	$310 \sim 201$	20	2	21
32	320	$3201 \sim 32$	320	$3201 \sim 23$
210	21	2	20	$201 \sim 310$
310	31	3	30	$301 \sim 210$
320	32	$321 \sim 320$	32	$321 \sim 230$

Table 4: The relation table of the enumeration of $PGL_2(7)$ over S_4

Example 5.3. Consider the group:

$$G \cong \frac{2^{*6}{:}L_2(5)}{[(0,1,2,3,4)t_0]^4=1}.$$

Here the $L_2(5)$ acts on the 2^{*6} as the group of linear fractional transformations (of determinant 1) on the projective line of order 5 whose points may be labelled with the elements of $\mathbf{F}_5 \cup \{\infty\}$. The result of the coset enumeration of G over $L_2(5)$ is shown in Table 4. It is easy to recognize the group $G-$in this case the projective general linear group $PGL_2(11)$ of order $22 \times 60 = 1320-$and check that it does contain such a symmetric generating set.

0	1	2	3	
$*$	0	$01 \sim 32$	3	$*$
∞	$\infty 0 \sim 12$	$121 \sim 212$	$21 \sim \infty 3$	∞
0	$*$	1	$12 \sim 43$	4
1	10	$101 \sim 242$	$24 \sim 03$	0
2	$20 \sim 41$	4	$42 \sim 13$	1
3	$30 \sim 14$	$141 \sim 232$	23	2
4	$40 \sim 21$	2	$*$	3
$\infty 0$	∞	$\infty 1 \sim 23$	$232 \sim 323$	$32 \sim \infty 4$
$\infty 1 \sim 04$	$040 \sim 131$	$13 \sim 42$	4	$43 \sim \infty 0$
$\infty 2 \sim 10$	1	$*$	2	$23 \sim \infty 1$
$\infty 3 \sim 40$	4	$41 \sim 20$	$202 \sim 343$	$34 \sim \infty 2$
$\infty 4 \sim 01$	$010 \sim 101$	$10 \sim \infty 2$	∞	$\infty 3$
0∞	$0\infty 0 \sim 141$	$14 \sim 2\infty$	$2\infty 2 \sim 313$	$31 \sim 4\infty$
$1\infty \sim 03$	$030 \sim 121$	12	1	$13 \sim 0\infty$
$2\infty \sim 30$	3	$31 \sim 02$	0	$03 \sim 1\infty$
$3\infty \sim 20$	2	21	$212 \sim 303$	$30 \sim 2\infty$
$4\infty \sim 02$	$020 \sim 1\infty 1$	$1\infty \sim 24$	$242 \sim 3\infty 3$	3∞
$\infty 0\infty \sim 0\infty 0$	$0\infty \sim 13$	$131 \sim 2\infty 2$	$2\infty \sim 30$	$303 \sim \infty 4\infty$
$\infty 1\infty \sim 020$	$02 \sim 31$	3	32	$323 \sim \infty 0\infty$
$\infty 2\infty \sim 040$	$04 \sim \infty 1$	∞	$\infty 2 \sim 34$	$343 \sim \infty 1\infty$
$\infty 3\infty \sim 010$	01	0	$02 \sim 31$	$313 \sim \infty 2\infty$
$\infty 4\infty \sim 030$	$03 \sim 1\infty$	$1\infty 1 \sim 202$	$20 \sim 3\infty$	$3\infty 3 \sim \infty 3\infty$

Table 5: The relation table of the enumeration of $PGL_2(11)$ over $L_2(5)$

6. CONCLUDING REMARK

A number of techniques have been proposed to reduce the calculation during the reduction step. For example, we may write all the problem relations and their conjugates under N of the form $t_{i_1}t_{i_2}\dots t_{i_k} = \pi_i t_{i_r} t_{i_{r-1}} \dots t_{i_{k+1}}$, where k equals to $\frac{r}{2}$ or $\frac{r+1}{2}$ according to whether r is even or odd respectively. In the reduction step, if a string $t_{i_1}t_{i_2}\dots t_{i_k}$ appears, replace it by $\pi_i t_{i_r} t_{i_{r-1}} \dots t_{i_{k+1}}$ and move the permutation π_i over the preceding symmetric generators in the standard manner. We can also apply a consistency condition to our tables that $f(w, t_i) = w' \iff f(w', t_i) = w$, so that inverses have the behavior we expect.

REFERENCES

[1] Bosma, W. Cannon, J. and Playoust, C., *The MAGMA Algebra System I: The User Language*, J. Symb. Comp. 24 (1997), 235–265.

[2] Cannon, J. J., Dimino, L. A. , Havas, G. and Watson, J. M., *Implementation and Analysis of the Todd-Coxeter Algorithm*, Math. Comp. 27 (1973), 463–490.

[3] Conway, J. H., Curtis, R. T., Norton, S. P., Parker, R. A. and Wilson, R. A., *An Atlas of Finite Group*, Oxford UP (1985).

[4] Curtis, R. T. *Symmetric Presentation I: Introduction with Particular Reference to Mathieu Groups* $\mathbf{M}_{12}$ *and* $\mathbf{M}_{24}$, Proc. of the LMS Durham Conference on Graphs and Combinatorics (1990), 380–396.

[5] Curtis, R. T. and Hasan, Z., *Symmetric Representation of the Elements of the Janko Group* $\mathbf{J}_1$, J. Symb. Comp. 22 (1996), 201–214.

[6] Havas, G. *Coset Enumeration Strategies*, University of Queensland, key Center for Software Technology, Technical Report No. 200 (1991).

[7] Leech, J. *Coset Enumeration on Digital Computers*, Proc. Camb. Phill. Soc. Vol. 59 (1963), 257–267.

[8] Leech, J. *Coset Enumeration*, In Computational Group Theory, Ed Michael D. Atkinson, Academic Press, (1984), 3–18.

[9] Linton, S. A. *The Maximal Subgroups of the Sporadic Group* **Th**, **Fi**24 *and* **Fi**24′ *and Other Topics*, Ph.D. Thesis, Univ. of Cambridge (1989).

[10] Sayed, M. *Computational Methods in Symmetric Generation of Groups*, Ph.D. thesis, Univ. of Birmingham (1998).

[11] Sayed, M. *Coset Enumeration Algorithm for Symmetrically Presented Groups*, Alexandria Engineering Journal Vol. 40 No. 2 (2001), 319–323.

[12] Sayed, M. *Nested Symmetric Representation of Elements of the Suzuki Chain Groups*, Int. J. Math. Math. Sci. 62 (2003), 3931–3948.

[13] Suzuki, M. *Group Theory I*, Springer Verlag (1982).

[14] Todd, J. A. and Coxeter, S. M., *A practical Method for Enumerating Cosets of Finite Abstract Groups*, Proc. Edinburgh Math. Soc. 5 (1936), 26–34.

PYTHAGORAS NUMBERS OF FUNCTION FIELDS OF GENUS ZERO CURVES DEFINED OVER HEREDITARILY PYTHAGOREAN FIELDS

S. V. Tikhonov[1] and V. I. Yanchevskiĭ[2]
[1]Institute of Mathematics of the National Academy of Sciences of Belarus
ul. Surganova 11, 220072 Minsk, Belarus
e-mail: tsv@im.bas-net.by

[2]Institute of Mathematics of the National Academy of Sciences of Belarus
ul. Surganova 11, 220072 Minsk, Belarus
e-mail: yanch@im.bas-net.by

1. INTRODUCTION

Let F be a field of characteristic 0 and ΣF^2 the set of all sums of squares of elements of F. For $a \in \Sigma F^2$ the minimal $n \in \mathbb{N}$ such that

$$a = a_1^2 + \cdots + a_n^2, \quad a_i \in F,$$

is called the length of a. It is denoted by $l(a)$. If $-1 \in \Sigma F^2$, then the number $s(F) = l(-1)$ is called the level of F.

The number $p(F) = sup\{l(a) \,|a \in \Sigma F^2\}$ is called the Pythagoras number of F. If F is nonreal, then $s(F) \leq p(F) \leq s(F) + 1$. A formally real field (= real field) F is called pythagorean if $p(F) = 1$, hereditarily pythagorean (= h. p.) if any real algebraic extension of F is pythagorean. Note that any nonrial extension of a h. p. field contains $\sqrt{-1}$. It is also know that a field F is a h. p. field iff $p(F(x)) = 2$ where $F(x)$ is the rational function field over F in one variable. We refer the reader to [1] for properties of h. p. fields.

Let C be a conic defined over a h.p. field F. In this paper we compute the Pythagoras number of the function field $F(C)$ of C. This article is an extended version of the shot communication [4]. We will consider separately the cases of real and nonreal $F(C)$.

2. PRELIMINARIES

Below we fix the following notations and conventions. For an abelian group A, the kernel of the multiplication by 2 is denoted by ${}_2A$. For a field k, we denote its

algebraic closure by $\overline{k}$. If R is a commutative ring, R^* denotes the group of units in R, and R^{*2} denotes the subgroup of squares in R^*. If $s \in R^*$, then for the brevity the class sR^{*2} will be denoted by the same symbol s.

Br L denotes the Brauer group of a field L. For any finite dimensional L-central simple algebra $\mathcal{A}$, we use $[\mathcal{A}]$ to denote its class in Br L. For finite dimensional L-central simple algebras $\mathcal{A}$ and $\mathcal{B}$, we write $\mathcal{A} \sim \mathcal{B}$ if $[\mathcal{A}] = [\mathcal{B}]$ in Br L. We write $\mathcal{A} \sim 1$ if $[\mathcal{A}] = 0$. For $a, b \in L^*$, we denote by (a, b) the corresponding quaternion algebra over L. Note that $(a, b) \sim 1$ iff $b \in N_{L(\sqrt{a})/L}(L(\sqrt{a})^*)$. Br F/L denotes the relative Brauer group of the extension F/L.

Let k be an arbitrary field of characteristic zero, X a smooth projective variety over k, $k(X)$ its function field. The set of k-points of X is denoted by $X(k)$. If L/k is a field extension, we set $X_L = X \times_{Spec\, k} Spec\, L$.

For a discrete valuation v of $k(X)$ trivial on k with the residue field $k(v)$, there exists the homomorphism of ramification at v

$$\mathrm{ram}_v : \mathrm{Br}\, k(X) \longrightarrow \mathrm{Hom}_{cont}(G_v, \mathbb{Q}/\mathbb{Z}) = H^1(G_v, \mathbb{Q}/\mathbb{Z}),$$

where $G_v = \mathrm{Gal}(\overline{k}/k(v))$. The ramification map is described in [3, Ch. 10]. A central simple algebra $\mathcal{A}$ over K is said to be ramified at v if $\mathrm{ram}_v([\mathcal{A}]) \neq 0$; then v is called a ramification point. The subgroup $\cap_v \ker \mathrm{ram}_v$, where v runs over the set of valuations with the aforementioned properties, is called the unramified Brauer group of the field $k(X)$ and is denoted by $\mathrm{Br}_{nr}k(X)$.

Let $\mathcal{A}$ be a $k(X)$-algebra of exponent 2. Then $ram_v([\mathcal{A}]) \in {}_2H^1(G_v, \mathbb{Q}/\mathbb{Z}) \cong H^1(G_v, \mathbb{Z}/2) \cong k(v)^*/k(v)^{*2}$. Note that for a quaternion algebra $\mathcal{A} = (a, b)$, $ram_v([\mathcal{A}]) = (-1)^{v(a)v(b)}\overline{a^{v(b)}b^{v(a)}} \in k(v)^*/k(v)^{*2}$ by "tame symbol" formula.

Let C be a conic over k. For any point $P \in C$, there is the corresponding valuation v_P of $k(C)$. The residue field of v_P is $k(P)$. There is a natural inclusion of Br C in Br $k(C)$ where Br C denotes the Brauer group of a conic C. This inclusion identifies Br C with the unramified Brauer group $\mathrm{Br}_{nr}k(C)$. Bellow we will write Br C instead of $\mathrm{Br}_{nr}k(C)$ keeping in mind this identification. We need the following

PROPOSITION 1. *Let C be a curve over a h. p. field k and $f_i \in k(C)^*$, $i = 1, \ldots, n$. Then the $k(C)$-algebra $\mathcal{A} = (-1, \sum_{i=1}^n f_i^2)$ is unramified.*

Proof. Let $g = \sum_{i=1}^n f_i^2$. The algebra $\mathcal{A}$ can be ramified at zeros or nodes of g. Let P be a pole or a zero of g. Choosing numeration one can assume that for the valuation

v_P of $k(C)$ at the point P there are the following unequalities $v_P(f_1) \leq v_P(f_i)$, $i = 1, \ldots, n$. Then

$$\mathcal{A} \sim (-1, f_1^2(1 + f_2^2/f_1^2 + \cdots + f_r^2/f_1^2)) \sim (-1, 1 + f_2^2/f_1^2 + \cdots + f_r^2/f_1^2).$$

If $1 + (f_2^2/f_1^2)(P) + \cdots + (f_r^2/f_1^2)(P) \neq 0$, then $\mathcal{A}$ is unramified at P by "tame symbol" formula.

If $1 + (f_2^2/f_1^2)(P) + \cdots + (f_r^2/f_1^2)(P) = 0$, then $k(P)$ is nonreal. Hence -1 is a square in $k(P)$. Therefore $\mathrm{ram}_P([\mathcal{A}]) = (-1)^{v_P(g)}\overline{g^{v_P(-1)}} = 1$.

Thus $\mathcal{A}$ is unramified and hence $[\mathcal{A}] \in \mathrm{Br}\, C$. The proposition is proved.

3. THE CASE OF A REAL FIELD

We begin with the following

LEMMA 2. *Let C be a conic over a h. p. field k, $\mathcal{D} = (-1, u)$, where $u \in k^*$. Let also $k(C)$ be real. Assume that $C(k) = \emptyset$ and $\mathcal{D} \otimes k(C) \not\sim 1$. Then $\mathcal{A} \not\sim \mathcal{D} \otimes k(C)$ where $\mathcal{A}$ is as in Proposition 1.*

Proof. Without loss of generality we will assume that C is defined by an affine equation $y^2 = ax^2 + b$, $a, b \in k$. Since $C(k) = \emptyset$, then $a, b \notin k^{*2}$ and $-ab \notin k^{*2}$. Moreover, $-a \notin k^{*2}$ or $-b \notin k^{*2}$ since otherwise $k(C)$ is not real.

We will give the proof by breaking it into several cases.

Case 1. Assume $ub \notin k^{*2}$.

(i) Let $b \notin -k^{*2}$. There exists an ordering of k such that $u < 0$, $b > 0$. Indeed, $k(\sqrt{b})$ is real and $u \notin k(\sqrt{b})$. Then $L = k(\sqrt{b})(\sqrt{-u})$ is real and there is an ordering of L such that $b > 0$, $u < 0$. We can extend this ordering to $k(x)$ viewing x as an infinitely small element. Then $ax^2 + b > 0$ and $k(C)$ has an ordering such that $u < 0$. Consider a real closure $k(C)^R$ of $k(C)$ corresponding to the constructed ordering. Then

$$\mathcal{A} \otimes k(C)^R \sim 1 \text{ and } \mathcal{D} \otimes k(C)^R \sim (-1, u) \otimes k(C)^R \sim (-1, -1) \otimes k(C)^R \not\sim 1.$$

Thus $\mathcal{A} \not\sim \mathcal{D} \otimes k(C)$.

(ii) Let $b \in -k^{*2}$. If $ua \in k^{*2}$, then

$$\mathcal{D} \otimes k(C) \sim (-1, u) \otimes k(C) \sim (-1, a) \otimes k(C) \sim (a, b) \otimes k(C) \sim 1.$$

Since $\mathcal{D} \otimes k(C) \not\sim 1$, then $ua \notin k^{*2}$.

Then there exists an ordering of k such that $u < 0$, $a > 0$. We can extend this ordering to $k(x)$ viewing x as an infinitely big element. Then $ax^2 + b > 0$ and $\mathcal{A} \not\sim \mathcal{D} \otimes k(C)$, by an argument analogous to that above.

Case 2. Assume $ub \in k^{*2}$. If $a \in -k^{*2}$, then

$$\mathcal{D} \otimes k(C) \sim (-1, u) \otimes k(C) \sim (-1, b) \otimes k(C) \sim (a, b) \otimes k(C) \sim 1.$$

For $au \in k^{*2}$, one has

$$\mathcal{D} \otimes k(C) \sim (-1, u) \otimes k(C) \sim (u, u) \otimes k(C) \sim (a, b) \otimes k(C) \sim 1.$$

Since $\mathcal{D} \otimes k(C) \not\sim 1$, then $a \notin -k^{*2}$ and $au \notin k^{*2}$.

There exists an ordering of k such that $u < 0$, $a > 0$. Hence by arguments analogous to that above we conclude again that $\mathcal{A} \not\sim \mathcal{D} \otimes k(C)$. The lemma is proved.

Now we are in a position to formulate the main result of this section.

Theorem 3 *Let C be a conic defined over a h.p. field k such that $k(C)$ is real. Then $p(k(C)) = 2$.*

Proof. Without loss of generality we will assume that C is defined by an affine equation $y^2 = ax^2 + b$, $a, b \in k$. Moreover, we will assume that $C(k) = \emptyset$ since otherwise $k(C)$ is a rational function field in one variable over k and then $p(k(C)) = 2$. As in the proof of Lemma 2 we obtain that $a, b \notin k^{*2}$ and $-ab \notin k^{*2}$. Besides, $-a \notin k^{*2}$ or $-b \notin k^{*2}$ since otherwise $k(C)$ is not real. Let $g \in \Sigma k(C)^2$ and $\mathcal{A} = (-1, g)$.

Proposition 1 implies that $[\mathcal{A}] \in \mathrm{Br}\ C$. Since $C(k) = \emptyset$, then there is an exact sequence ([2])

$$0 \longrightarrow < [(a, b)] > \longrightarrow \mathrm{Br}\ k \xrightarrow{res} \mathrm{Br}\ C \longrightarrow 0,$$

where $< [(a, b)] >$ is a subgroup generated by $[(a, b)] \in \mathrm{Br}\ k$.

We will prove that $[\mathcal{A}] = 0$. Let $\mathcal{B}$ be a central simple k-algebra such that $[\mathcal{A}] = res([\mathcal{B}])$. Assume that $\mathcal{B} \otimes k(C) \not\sim 1$. If $\mathcal{B} \otimes k(\sqrt{-1}) \sim 1$, then $B \sim (-1, u)$, where $u \in k^*$, and by Lemma 2 $[\mathcal{A}] \neq res([\mathcal{B}])$.

Now we consider the case where $\mathcal{B} \otimes k(\sqrt{-1}) \not\sim 1$. For the conic C we have two possibilities:

(i) $C(k(\sqrt{-1})) \neq \emptyset$;

(ii) $C(k(\sqrt{-1})) = \emptyset$.

In the case (i) $k(\sqrt{-1})(C)$ is a rational function field in one variable over $k(\sqrt{-1})$. Hence $\mathcal{B} \otimes k(\sqrt{-1})(C) \not\sim 1$. Since $\mathcal{A} \otimes k(\sqrt{-1})(C) \sim 1$, then $\mathcal{A} \not\sim \mathcal{B} \otimes k(C)$.

In the case (ii) we use an exact sequence

$$0 \longrightarrow < [(a, b) \otimes k(\sqrt{-1})] > \longrightarrow \text{Br } k(\sqrt{-1}) \overset{res_{k(\sqrt{-1})}}{\longrightarrow} \text{Br } C_{k(\sqrt{-1})} \longrightarrow 0.$$

If $\mathcal{A} \sim \mathcal{B} \otimes k(C)$, then $\mathcal{B} \otimes k(\sqrt{-1})(C) \sim 1$. Since $\mathcal{B} \otimes k(\sqrt{-1}) \not\sim 1$, we obtain that $[\mathcal{B}] \in \ker(res_{k(\sqrt{-1})})$. Therefore $\mathcal{B} \otimes (a, b) \otimes k(\sqrt{-1}) \sim 1$ and $\mathcal{B} \sim (a, b) \otimes (-1, u)$ for some $u \in k^*$. Note that $(-1, u) \otimes k(C) \not\sim 1$ since otherwise $res([\mathcal{B}]) = 0$. Then $[\mathcal{A}] = res([\mathcal{B}]) = res([(-1, u)])$, i.e. $\mathcal{A} \sim (-1, u) \otimes k(C)$. But this contradicts to Lemma 2. Thus $\mathcal{A} \sim 1$ and hence g is a sum of two squares. The theorem is proved.

4. THE CASE OF A NONREAL FIELD

In the case of a nonreal $k(C)$ we have the following

Theorem 4 *Let C be a conic defined over a h.p. field k. Assume that $k(C)$ is nonreal. Then*

$$p(k(C)) = \begin{cases} 2 & \text{if } |\text{Br } k(\sqrt{-1})/k| = 2, \\ 3 & \text{if } |\text{Br } k(\sqrt{-1})/k| > 2. \end{cases}$$

Proof. Without loss of generality we will assume that C is defined by an affine equation $y^2 = ax^2 + b$, $a, b \in k$. Since $k(C)$ is nonreal, then $C(L) = \emptyset$ for any real algebraic extension L/k. Indeed, assume that there exists a point $P \in C(L)$ for some real extension L/k. The completion of $L(C)$ with respect to the valuation corresponding to P is $L((z))$ for some uniformizer z. Since $L(P) = L$ is real, then $L((z))$ is also real. Then $k(C)$ is real in view of the inclusions $k(C) \subset L(C) \subset L((z))$.

This remark implies that we can assume without loss of generality that an affine equation of the conic C is $y^2+x^2=-1$. To see this one can observe that a is negative at any ordering of k since otherwise C has a point $(\sqrt{a}:1:0)$ in a real extension. This implies that $-a$ is a sum of squares in k and therefore is a square since k is pythagorean. In a similar way we can obtain that $-b \in k^{*2}$. Then C is k-birationally equivalent to the conic with an affine equation $y^2+x^2=-1$.

Let $g \in \Sigma k(C)^2$ and $\mathcal{A}=(-1,g)$. Proposition 1 implies that $[\mathcal{A}] \in \mathrm{Br}\, C$. Note that $(-1,-1)\otimes k(C) \sim 1$. Since $C(k)=\emptyset$, then there is an exact sequence

$$0 \longrightarrow < [(-1,-1)] > \longrightarrow \mathrm{Br}\, k \xrightarrow{res} \mathrm{Br}\, C \longrightarrow 0.$$

Note that $[\mathcal{A}] \in res(\mathrm{Br}\, k(\sqrt{-1})/k)$. Indeed, assume that $[\mathcal{A}] \notin res(\mathrm{Br}\, k(\sqrt{-1})/k)$. Then $[\mathcal{A}]=res[\mathcal{B}]$ for some central simple k-algebra $\mathcal{B}$ such that $\mathcal{B}\otimes k(\sqrt{-1}) \not\sim 1$. Since $C(k(\sqrt{-1})) \neq \emptyset$, then $\mathcal{B}\otimes k(\sqrt{-1})(C) \not\sim 1$. We obtain a contradiction in view of $\mathcal{A}\otimes k(\sqrt{-1})(C) \sim 1$. Thus $[\mathcal{A}] \in res(\mathrm{Br}\, k(\sqrt{-1})/k)$.

If $|\mathrm{Br}\ k(\sqrt{-1})/k| = 2$, then the group $\mathrm{Br}\ k(\sqrt{-1})/k$ consists of $[(1,-1)]$, $[(-1,-1)]$. Since these elements are in the kernel of res, then $[\mathcal{A}]=0$. Thus $\mathcal{A} \sim 1$ and hence g is a sum of two squares. Hence $p(k(C))=2$.

If the cardinality of the group $\mathrm{Br}\ k(\sqrt{-1})/k$ is bigger then 2, we can take an element $u \in k^*$ such that $res([(-1,u)]) \neq 0$. Hence u is not a sum of two squares. Since $s(k(C))=2$, then $2 \leq p(k(C)) \leq 3$. Thus $p(k(C))=3$.

REFERENCES

[1] Becker, E. *Hereditarily-Pythagorean fields and orderings of higher level,* Monografias de Mathematica 29, Rio de Janeiro, 1978.

[2] Grothendieck, A. Le groupe de Brauer III: exemples et compléments, in *Dix exposés sur la cohomologie des schémas,* North-Holland, Amsterdam, 1968.

[3] Saltman, D. J. *Lectures on division algebras*, Published by American Mathematicas Society, Providence, RI, 1999.

[4] Tikhonov, S. V. and Yanchevskiĭ, V. I., Pythagoras numbers of function fields of conics defined over heriditerily pythagorean fields (in Russian), *Dokl. Nats. Akad. Nauk Belarusi* 47 (2003), n° 2, 5–8.

MATHEMATICS ON ITS WAY INTO INFORMATION SOCIETY

B. Wegner,
Fakultaet II, Institut füer Mathematik,
TU Berlin, Sekr. MA 8-1 Strasse des 17. Juni 135, D-10623, Berlin, Germany
email: wegner@math.tu-berlin.de

INTRODUCTION

The increasing development of electronic support for the publication of articles and books in mathematics lead to a drastic change in the communication facilities between authors and editors, new ways of distributing mathematics publications to researchers - like electronic journals -, and an enhancement of the information dissemination on mathematical research to the scientific community.

As a consequence of the permanently growing number of research articles the search for appropriate information on new results and the access to the corresponding publications is becoming more and more difficult. In addition to this, even libraries with a comparatively high budget cannot afford to subscribe to all journals they should offer to their users. Thinking of the high subscription rates for some important mathematical journals libraries with low budgets will not be able to provide the minimum of indispensable journals to their mathematicians. Electronic publishing with its diversity of alternative presentations and integrated access systems could offer better solutions to these problems. Initially there was the hope that this would reduce the publication costs, though several publishers gave arguments that these savings have to be used for the compensation of new costs emerging from electronic publishing. Clearly this would not apply if we only think about electronic versions of printed publications. But in the ideal case an electronic publication would provide additional semantics mark-up, interactive components and a multiple linkage with other articles on the same subject, and including these facilities requires enhanced efforts.

The aim of this article is to describe the facilities provided by EMIS (European Mathematical Information Service) as an example for an information gateway to electronic offers in mathematics. It could be seen as a model for integrated access to such offers, not only caring about installation and dissemination of electronic publications, but also implementing advanced tools of mathematics knowledge management.

1. THE GENERAL CONCEPT OF EMIS

EMIS was discussed by the Executive Committee of the EMS in 1994 and founded as a cooperation with FIZ Karlsruhe in 1995. See the article in the references for the first steps of EMIS. In addition to the common society services several enhanced services have been developed during the life time of EMIS. The most prominent ones currently are: the Electronic Library ElibM, links to reference data bases, projects dealing with information and communication in mathematics like EULER, EMANI, MoWGLI, MONET and the digital archive in classical mathematics ERAM. Most important is the free access to ElibM and ERAM.

Another important aspect is the distribution of EMIS to more than 40 mirror sites world wide. The mirroring is organized automatically by a tree like forwarding system. The mirrors are supporting the safety of the data, being able to regenerate one mirror site from another in the case of failure. In Europe almost every country has its own mirror, but also world wide distribution like in Latin America is in good progress. Admittedly there are not so many in the orient up to the copy in Ankara. But a mirror site in Kuwait is just on the way of installation and copies in Iran and Egypt are under discussion. This is very important for the local and regional availability of the services of EMIS.

2. THE ELECTRONIC LIBRARY

The main objective of the Electronic Library in EMIS is to provide free access to a high quality collection of electronic publications which is complete as possible. This refers to standard publications like journals, proceedings volumes and monographs, but also to innovative offers like the collection of geometric models. To monitor the quality of publications included in ElibM the Electronic publishing Committee of the EMS is employed. In particular as a general rule only peer reviewed publications will be included. This was very important at the beginning of EMIS there were some reservations that electronic publishing may damage the quality control and the peer reviewing system. After ten years these preoccupations have more or less vanished. Some electronic journals even succeeded to enter the closed circle of the Science Citation Index. There still are few pure electronic journals without printed version now. But most printed journals developed an electronic version in the recent past.

In ElibM pure electronic journals and electronic versions of printed journals are available at equal ratio. Pure electronic journals are mirrored from the web-site where they are produced. For some electronic versions of printed journals the

production and installation is supported by EMIS, but preferably EMIS only cares about bundling these offers and distributing them to the EMIS mirror sites. The journals in ElibM, which are not fully sponsored by an academic institution, need some modest income from the printed version. They are low-budget journals anyway. To protect that income the so-called moving wall model is applied. The free access will be only available for the back volumes having passed that wall. Generally a period of three years is proposed for this.

This solution worked very well in the first years of EMIS. But very soon many publishers started to offer journals online first: the electronic version is posted as soon as the final editing was completed, the source code is converted to Postscript and PDF, and the article could be read in the web months before the printing of a volume could be completed. This had impact on the electronic offers in EMIS, because online first gives a considerable advantage to the electronic version, and income could not be expected from the printed version in the same way as before, when the electronic version could be read for free so much in advance. Here EMIS had to adjust its policy. So the part before the moving wall will be installed including the online first period, but the new material is open for subscribers only until it will have passed the moving wall. Control for that is done by IP addresses at the EMIS master site. In a first period only metadata will be distributed to the mirror sites, and after having passed the moving wall the full article will be distributed to all mirror sites.

More than 60 journals are freely accessible in ELibM at present. They cover most of the 4 Gigabytes needed to store EMIS at present. The total number of articles offered in ElibM is about 13.000. The journals, which look back to a period where only a printed version was available, have been retrodigitised for that period in the ERAM project. Links from EMIS to these offers are provided, such that these journals a fully electronically available. They are also accessible through links from several offers in EMIS like ERAM directly, or the EULER engine or the reference databases Zentralblatt MATH and JFM (Jahrbuch).

The direct access to the journals in EMIS still is organised quite conventionally. On the top page of EMIS the choice of an appropriate mirror site of EMIS could be made. Then, going to the journals section in the ELibM two choices appear, a short list of journals having an entry of at most to lines for each journals and enabling a quick selection, or a long list of journals with more information on the journal and links to more details like editorial policy, list of editors, information for authors, style files etc. After having chosen the journal the next stations will

be volume, list of contents, article and finally he choice of file. Standard offers are Postscript and PDF, but in many cases also the DVI and the TEX source code will be available. HTML is not considered as an appropriate format for the presentation of mathematical formulas. Search facilities in ElibM are possible using the reference databases in EMIS.

Access to Proceedings in EMIS is organised in a similar way, but this section does not show the same growth as the journals section. One highlight in the ElibM are the complete collected works by Riemann and Hamilton, which have been collected, organised on the electronic level and provided by Professor Wilkins.

3. THE REFERENCE DATABASES ACCESSIBLE THROUGH EMIS

At present five reference databases are accessible through EMIS: Zentralblatt MATH, the Jahrbuch database, MATHDI, Compuscience and MPRESS. Another one dealing with logics is under construction. There offers in EMIS underline the involvement of the European Mathematical Society EMS in these services. EMS directly takes part in the editorial system of Zentralblatt MATH.

Zentralblatt MATH was founded 1931 as a comprehensive literature documentation service in mathematics and its applications. Clearly, at that time it was available in printed form only. Now the name stands for three services, the traditional printed version and the two electronic versions, one as a web service and the other as a stand-alone version on CD-ROM. Providing a lot of good search facilities and links to the full articles the online version became the most important option for the users. It can be reached through EMIS under the URL http://www.emis.de/cgi-bin/ZMATH, or directly at one of the international mirror sites for the database. Just recently the items from the Jahrbuch über die Fortschritte der Mathematik have been added to the Zentralblatt database. As a result all mathematics going back to 1868 is covered by this offer now.

The database itself is not part of the EMIS mirroring system, nor is full usage of the service for free. But there is the free offer of three hits per search for everybody. This enables good navigation in mathematical publications through reference links for example, where searches initiated by these links will lead to one-item hit lists and may activate the access to the full text of the publication of interest, if that will be electronically available and the user will have a reading licence from the provider of the full text. Links to the full text are integrated in Zentralblatt MATH as far as available. This refers to digitally born material as well as to scanned articles from

digitisation projects.

A similar distribution and access model applies to MATHDI, which covers a lot of literature in the field of education in and popularisation of mathematics. Here completeness cannot be reached, because the field is too large and in many case the articles are of national or local interest only. A global core part may be didactics in mathematics, which was the original subject of MATHDI. But having in mind that the community in didactics of mathematics is rather small and that there is a big interest in more general publications, the extension to education and popularisation was a natural consequence.

The Jahrbuch database captures the content of the the Jahrbuch über die Fortschritte der Mathematik electronically, enhanced by modern indexation and links to digitised versions of the papers from the Jahrbuch period (1868 - 1943). As mentioned above it is available for integrated search with the Zentralblatt database, but having been established in a project funded by DFG the Jahrbuch data are freely accessible in this separate offer.

Compuscience is still in development. It covers an extension of the data provided by Zentralblatt in the subject area 68 (Computer Science related to mathematics) adding articles which are not of main interest for mathematicians. A first prototype is visible in EMIS. As a next step a more complete version as far as Computer Science is regarded will be available in the near future.

MPRESS is another free offer, which can be reached in EMIS. It is an index covering electronically available pre-prints in mathematics world-wide. The full text of the preprints can be obtained through a link from MPRESS. The information is gathered automatically by robots visiting websites where preprints are offered more or less regularly. The harvesting of data is organised by some mathematicians at the University of Osnabrück. The quality of the search depends on the quality of the metadata provided at the site of the preprint server. Only some core data could be expected in general. The full index is generated by harvesting after well-defined periods. As a consequence, preprints, which are removed from a server do not appear in MPRESS anymore. The system is far from being complete, but major preprints servers like the arXive are included all.

As a full text database of a highly innovative character also the Electronic Geometric Models have to be mentioned here. This is a collection of small peer-reviewed articles edited by Konrad Polthier (ZIB Berlin) and Michael Joswig (TU Berlin).

For each model an explanatory text is provided. This is accompanied by images and animations trying to give a good idea about the model and its applications and enabling the user to make his own experiments with the subject. Precautions for long-term archiving are taken by the editors. Freely available viewing software has been chosen for theses offers.

4. ACCESS TO MATHEMATICS THROUGH PROJECTS IN EMIS

The projects displayed in EMIS dealing with improved access to mathematics in the web are as follows: ERAM - the Electronic Research Archive in Mathematics, EULER - searching mathematics in distributed heterogeneous sources, LIMES - developing and offering services like Zentralblatt based on a distributed infrastructure, MoWGLI - Mathematics on the Web, Get it by Logics and Interfaces, EMANI - long-term preservation of digital documents in mathematics using a network of libraries, and MONET - a broker system for getting access to mathematical software.

As mentioned above one aim of ERAM was to capture the content of the Jahrbuch über die Fortschritte der Mathematik (1868- 1943) in a database. This was combined with the request to make a lot of the documents described in the Jahrbuch digitally available. The digital archive is installed at the SUB Göttingen, the holdings are linked with the databases in EMIS and as a consequence of the funding from DFG free access could be provided for all mathematicians world-wide. The construction of the content is still in progress though the major part of the capacity of 1.200.000 pages has been spent already. Famous parts of the content are Mathematische Zeitschrift, Mathematische Annalen, Inventiones Matematicae, Kleiné Encyclopaedia et al. This was achieved in cooperation with EMANI where also other digitisation projects are participating.

The EULER project was financed by the EU within the program Telematics for Libraries. It had developed a prototype for searching mathematics in distributed heterogeneous sources. That prototype had been brushed up to a service by an additional take-up project. EULER now is provided as a service by a consortium. Members are mainly libraries which make adjustments of their information available in their OPACS to be able to apply the EULER search. The consortium still is growing by adding more libraries from Europe to the system.

Still three projects have to be mentioned, all of them being supported by the 5th Framework Program. In LIMES the architecture and tools have been developed for producing the input for reference databases like Zentralblatt MATH remotely

within a distributed system of editorial units. This enables European countries to care about their own contribution to the database in a quite autonomous way. On the other side the system allows them to connect national or regional electronic offers of mathematics with the database. The project ended in March 2004. Currently several efforts are made to implement the system step by step for Zentralblatt MATH.

General goal of MoWGLI is to contribute to the implementation of the semantic web in the case of mathematics. More concretely unified XML encodings of formulas are considered, search facilities for mathematics knowledge bases are developed, links to and communication with proof assistants are enabled, facilities for human understandable rendering of automated proofs are installed. This is integrated on a test basis in interactive textbooks to demonstrate that these offers can be used for an improved understanding of mathematics. The project is still in progress.

The same applies to MONET. People from the Mathematics Knowledge Management are involved and some selected software providers are in that group. Main goal is to develop a communication system where in an automated way professionals and researchers needing mathematical software can be guided to software providers having appropriate offers for them. This worked before for special software libraries, but to have a broker system, which integrates as many offers as possible, is a very ambitious goal, which finally may provide an extremely useful access tool to mathematics.

5. CONCLUDING REMARKS

Over the period of ten years EMIS has developed its services and links in a way, which provides one of the most comprehensive gateways to heterogeneous offers in mathematics. This has been shown above by just describing three main sections of the EMIS web page. Additional information could be obtained from other sections like the conference calendar. Furthermore links lists to external full text offers could also be consulted. Some facilities could be found in different ways. But in spite of this, there is plenty of electronic material in mathematics, to which no access through EMIS could be found.

Hence EMIS still is far from being a portal to mathematics, nor should it have the ambition to develop itself into such a portal. The next step for enhancing the system further should be a cooperation with other services of the same kind in order to have a bundle which may give a better approximation to a comprehensive portal to mathematics. A good navigation facility has to be established for that bundle.

At least such a solution seems to be a more realistic approach than what is pursued by other systems, which just establish an ambitious and authoritarian governance structure with having almost nothing to govern.

6. REFERENCES AND LINKS

Bernd Wegner: EMIS (European Mathematical Information Service), III Seminario Sistema Informativo Nazionale per la Matematica, Lecce 1997,
URL: http://siba2.unile.it/sinm/int_wegner.htm

Electronic Geometric Models — http://www-sfb288.math.tu-berlin.de/eg-models/

ElibM — http://www.emis.de/ELibM.html

EMANI — http://www.emani.org/

EMIS — http://www.emis.de/

ERAM — http://www.emis.de/projects/JFM

EULER — http://www.emis.de/projects/EULER/

LIMES — http://www.emis.de/projects/LIMES/

MONET — http://monet.nag.co.uk/cocoon/monet/index.html

MOWGLI — http://www.mowgli.cs.unibo.it/

ALGEBRAIC STRUCTURES ARISING FROM EULER'S AND PYTHAGOREAN EQUATIONS

M. Wójtowicz
Institute of Mathematics
Bydgoszcz Academy, Pl. Weyssenhoffa 11, 85-072 Bydgoszcz, Poland
e-mail: mwojt@ab.edu.pl

1. INTRODUCTION

It is known (see [5], p. 202) that the set $\mathcal{P}_d$ of all integer solutions (x, y) of the Pell equation $x^2 - dy^2 = 1$, where d is a square-free positive integer, is a multiplicative group on two generators. More exactly, every pair $(x, y) \in \mathcal{P}_d$ corresponds uniquely to the element $\xi = x + y\sqrt{d}$, and there is a smallest pair $(x_0, y_0) \in \mathcal{P}_d$ with $x_0, y_0 > 0$; defining ξ^* as $x - y\sqrt{d}$, we have that every ξ is of the form $\xi = \pm\delta^n$ for some $n \in \mathbf{Z}$, where $\delta = x_0 + y_0\sqrt{d}$ and $\delta^{-1} = \delta^*$.

In this paper, we address natural algebraic structures that can be defined on the set E_n of all solutions of the Diophantine equation considered by Euler

$$(E) \qquad a^2 + b^2 = t^n,$$

where n is a fixed positive integer. Thus, $E_n := \{(a, b, t) : \ (a, b, t) \ fulfils \ (E)\}$. For $n = 2$ equation (E) becomes the classical Pythagorean equation, and in this case we put $\mathcal{P}$ instead of E_n; the elements of $\mathcal{P}$ are called Pythagorean triples. This section deals with the semigroup structures on E_n, and the next one is devoted to the ring structures on $\mathcal{P}$ (a new result is presented in Theorem 1).

Let $\mathbf{Z}$ be the set of all integers, and let $\mathbf{Z}(i)$ denote the Gauss ring $\mathbf{Z} + i\mathbf{Z}$. It is well known that a triple (a, b, t) is an element of $\mathcal{P}$ if and only if $a = \mathrm{Re}(z^2)$, $b = \mathrm{Im}(z^2)$ (or $a = \mathrm{Im}(z^2)$, $b = \mathrm{Re}(z^2)$), $|t| = |z|^2$ for some $z \in \mathbf{Z}(i)$. The form of (a, b, t) suggests that a natural multiplication on $\mathcal{P}$ can be defined by means of multiplication on $\mathbf{Z}(i)$. Indeed, if $\alpha_j = (a_j, b_j, t_j) = (\mathrm{Re}(z_j^2), \mathrm{Im}(z_j^2), \varepsilon_j|z_j|^2)$ for some $z_j \in \mathbf{Z}(i)$ and $\varepsilon_j = \pm 1$, $j = 1, 2$, then the triple $(\mathrm{Re}((z_1z_2)^2), \mathrm{Im}((z_1z_2)^2), \varepsilon_1\varepsilon_2|z_1z_2|^2)$ lies in $\mathcal{P}$ and has the form $(a_1a_2 - b_1b_2, a_1b_2 + a_2b_1, t_1t_2)$. It is now easy to check that the latter expression defines a commutative and associative operation $\circ$ on the *whole* set $\mathcal{P}$:

$$(*) \qquad \alpha_1 \circ \alpha_2 = (a_1a_2 - b_1b_2, a_1b_2 + a_2b_1, t_1t_2).$$

Moreover, the triple $e := (1, 0, 1)$ is the neutral element of $\mathcal{P}$ with respect to $\circ$, and hence $(\mathcal{P}, \circ, e)$ *is a monoid* (i.e., a semigroup with unit).

For every $n \geq 1$ formula $(*)$ defines also $\circ$ on E_n, and $(E_n, \circ, e)$ is a commutative monoid called an Euler monoid [4]. For some n's the monoids E_n are free.

Proposition 1. (see [4]). *We have that E_n is free if and only if n is odd. More exactly, if R denotes the set of all prime elements of the Gauss ring $\mathbf{Z}(i)$ then*

(i) *for every odd integer $n \geq 1$, the set $X_n = \{(\mathrm{Re}(z^n), \mathrm{Im}(z^n), |z|^2) : \ z \in R\}$ is a base of E_n;*

(ii) *if $n = 2k \geq 2$, then the element $\alpha = (5^n, 0, 25) \in E_n$ has two different factorizations by irreducible elememnts of E_n: $\alpha = (5^k, 0, 5) \circ (5^k, 0, 5) = (A_+, B_+, 5) \circ (A_-, B_-, 5)$, where $A_\pm = \mathrm{Re}((1 \pm i2)^n)$, $B_\pm = \mathrm{Im}((1 \pm i2)^n)$.*

Thus, for $n = 2$ we have that $E_2 = \mathcal{P}$ is non-free. However, for a narrower (and more natural) subset of $\mathcal{P}$ the situation changes drastically. Let $\mathcal{P}^o$ denote the subset of all *primitive* elements (a, b, t) of $\mathcal{P}$, that is, such that a and t are positive and co-prime with $a, b, t \geq 0$. One can slightly modify operation $\circ$ for $\mathcal{P}^o$ to be a free group. The result presented below is due to Eckert [2].

Proposition 2. *The set $\mathcal{P}^o$ is a free abelian group under the operation $\circ$ defined by*

$$(a, b, c) \circ (A, B, C) = (aA - bB, bA + aB, cC) \ \text{ when } \ aA - bB > 0,$$

$$(a, b, c) \circ (A, B, C) = (bA + aB, -aA + bB, cC) \ \text{ when } \ aA - bB < 0.$$

The identity in $\mathcal{P}^o$ is $(1, 0, 1)$, the inverese of (a, b, c) is (b, a, c), and the base of $\mathcal{P}^o$ is the set $X = \{(a, b, p) : \ p \text{ prime}, \ p \equiv 1 \bmod 4, \ a > b\}$.

2. THE RING STRUCTURES ON $\mathcal{P}$

Let us divide $\mathcal{P}$ into infinite and pairwise disjoint subsets $\mathcal{P}_k := \{(a, b, t) \in \mathcal{P} : t - b = k\}$, for $k \neq 0$, and $\mathcal{P}_0 = \{(0, j, j) : j \in \mathbf{Z}\}$:

$$\mathcal{P} = \bigcup_{k \in \mathbf{Z}} \mathcal{P}_k \ .$$

(Thus, $(3, 4, 5) \in \mathcal{P}_1$ and $(4, 3, 5) \in \mathcal{P}_2$.) In 1994 Dawson [1] defined the ring operations on every set $\mathcal{P}_k$ and then extended them to $\mathcal{P}$ in such a way as to give $\mathcal{P}$ a ring structure, but both the addition an multiplication were given in inconvenient form. In 1997 Grytczuk [3] gave the construction of more natural ring operations

$\oplus_k$ and $\odot_k$ on $\mathcal{P}_k$, and this construction was studied in details in [6], [7]. Grytczuk showed that every $(a,b,t) \in \mathcal{P}_k$, $k \neq 0$, is of the form

$$b = \frac{a^2 - k^2}{2k}, \quad t = \frac{a^2 + k^2}{2k}, \tag{1}$$

where a and k are the same parity. Now we shall describe the Grytczuk's operations (G-operations, for short), and we shall show below how this description can be applied to define more natural pairs of ring operations on $\mathcal{P}$.

Since the elements of each $\mathcal{P}_k$, $k \neq 0$, depend on the first coordinate only, it is enough to define "basic" operations just on this coordinate and, by the use of (1), to carry them to b and t. The operations defining $\oplus_k$ and $\odot_k$ were obtained in [3] from the kth shifts on $\mathbf{Z}$ (see [6]): $S_k(x) = x - k$; then the "basic" operations $+^*$ and $\cdot^*$ on $\mathbf{Z}$ are given by the formulas

$$\begin{aligned} x +^* y &= S_k^{-1}(S_k(x) + S_k(y)) = x + y - k, \\ x \cdot^* y &= S_k^{-1}(S_k(x) \cdot S_k(y)) = (x-k)\cdot(y-k) + k, \end{aligned} \tag{2}$$

and G-operations on $\mathcal{P}_k$ are of the form: if $k \neq 0$ and $\alpha_j = (a_j, (a_j^2 - k^2)/2k, (a_j^2 + k^2)/2k)$, $j = 1, 2$, then

$$\begin{aligned} \alpha_1 \oplus_k \alpha_1 &= (a_1 +^* a_2, \frac{(a_1 +^* a_2)^2 - k^2}{2k}, \frac{(a_1 +^* a_2)^2 + k^2}{2k}), \\ \alpha_1 \odot_k \alpha_1 &= (a_1 \cdot^* a_2, \frac{(a_1 \cdot^* a_2)^2 - k^2}{2k}, \frac{(a_1 \cdot^* a_2)^2 + k^2}{2k}), \end{aligned} \tag{3}$$

and coordinatewise for $k = 0$. One should note that G-operations have nothing to do with formula $(*)$: for example, if $\alpha_1, \alpha_2 \in \mathcal{P}_k$, $k \neq 0$, then $\alpha_1 \circ \alpha_2 \notin \mathcal{P}_k$, in general. Moreover, for every $k \neq 0$ the ring $(\mathcal{P}_k, \oplus_k, \odot_k)$ has no unit.

The ring structure of $\mathcal{P}$, through the structure of $\mathcal{P}_k$'s, was also studied by the present author in [7], where the key role played the form of a in the triple $(a,b,t) \in \mathcal{P}_k$. Let us define the number $x(k)$, $k \in \mathbf{Z} \setminus \{0\}$ as follows (see [7], pp. 17-18). If $p_1^{d_1} \cdot \ldots p_s^{d_s}$ is the prime factorization of $|k|$, then the integer $\sqrt[*]{k}$ is defined as $p_1^{e_1} \cdot \ldots p_s^{e_s}$, where $e_j = [(d_j + 1)/2]$, $j = 1, \ldots, s$, and $[x]$ denotes the integer part of x. Moreover, if $|k| = 2^q \cdot r_0$, where q, r_0 are non-negative integers with $r_0 \geq 1$ and odd, then $d_{ev}(k)$ is defined as 1 for q odd, and 2 for q even or $q = 0$. Then we put $x(k) = \sqrt[*]{k} \cdot d_{ev}(k)$. (Thus, $x(k) = x(2k) = 2k$ for every odd and square-free positine integer k.) The result presented below is another form of ([7], Proposition 1).

Lemma 1. *Let $\alpha = (a,b,t) \in \mathcal{P}_k$, $k \neq 0$. Then there is uniquely determined $l \in \mathbf{Z}$ such that $a = l \cdot x(k) + k$.*

The above uniqueness allows us to describe G-operations more precisely. We have (see (2)):

$$(l_1x(k)+k) +^* (l_2x(k)+k) = (l_1+l_2)x(k)+k, \text{ and}$$
$$(l_1x(k)+k) \cdot^* (l_2x(k)+k) = (l_1l_2x(k))x(k)+k,$$

for all $l_1, l_2 \in \mathbf{Z}$ and fixed $k \neq 0$. We put $A(l,k) := l \cdot x(k) + k$ to define a function acting from $\mathbf{Z} \times \mathbf{Z}$ into $\mathbf{Z}$, and we see that for every $l \in \mathbf{Z}$ and k as above we have $(A(l,k), b, t) \in \mathcal{P}_k$, where b, t are as in (1) with $a = A(l,k)$. Moreover,

$$A(l_1,k) +^* A(l_2,k) = A(l_1+l_2,k), \text{ and}$$
$$A(l_1,k) \cdot^* A(l_2,k) = A(l_1l_2x(k),k); \text{ thus}$$

$G-$operations on $\mathcal{P}_k$ correspond to the ring operations on the ring ideal $x(k)\mathbf{Z}$. (4)

The lemma allows us also to define a natural bijection from $\mathbf{Z} \times \mathbf{Z}$ onto $\mathcal{P}$. For this purpose we define two auxiliary functions: $B(l,k) := (A(l,k)^2 - k^2)/2k$, and $T(l,k) := B(l,k) + k$, where $l, k \in \mathbf{Z}$. From the above remarks and formulas (1) we obtain that every $(a,b,t) \in \mathcal{P}_k$, with $k \neq 0$, has the form $(a,b,t) = (A(l,k), B(l,k), T(l,k))$ for some unique $l \in \mathbf{Z}$. It is now obvious that the function $\lambda : \mathbf{Z} \times \mathbf{Z} \to \mathcal{P}$ defined by

$$\lambda(l,k) = \begin{cases} (A(l,k), B(l,k), T(l,k)) & for \quad k \neq 0 \\ (0,j,j) & for \quad k = 0 \end{cases}$$

is a bijection, and hence we can transfer the ring structures (both coordinatewise and complex) from $\mathbf{Z} \times \mathbf{Z}$ onto $\mathcal{P}$ (see [7], Theorem 7):

Proposition 3. *The set $\mathcal{P}$ is a commutative ring with unit under the following pairs of addition and multiplication*:

(i) $\lambda(l_1,k_1) \oplus \lambda(l_2,k_2) := \lambda(l_1+l_2, k_1+k_2)$, *and* $\lambda(l_1,k_1) \odot \alpha(l_2,k_2) := \lambda(l_1l_2, k_1k_2)$,

with the additive zero $(0,0,0) = \lambda(0,0)$ *and the multiplicative unit* $(3,4,5) = \lambda(1,1)$;

(ii) $\lambda(l_1,k_1) \widetilde{\oplus} \lambda(l_2,k_2) := \lambda(l_1+l_2, k_1+k_2)$, *and* $\lambda(l_1,k_1) \widetilde{\odot} \lambda(l_2,k_2) := \lambda(l_1l_2 - k_1k_2, l_1k_2 + l_2k_1)$,

with the additive zero $(0,0,0)$ *and the multiplicative unit* $(0,1,1) = \lambda(1,0)$.

The above pairs of addition and multiplication do not extend G-operations because they do not leave $\mathcal{P}_k$'s invariant: for $k \neq 0$ and $\alpha, \beta \in \mathcal{P}_k$ we have that $\alpha \oplus \beta \in \mathcal{P}_{2k} \neq \mathcal{P}_k$ (similarly for $\widetilde{\oplus}$).

It is readily seen that every addition $\widehat{\oplus}$ on $\mathcal{P}$, defined by means of the function λ and leaving each $\mathcal{P}_k$ invariant, leads to an idempotent semigroup operation $\#$ on the second coordinate. Indeed, for any $l, k \in \mathbf{Z}$ we then have

$$\lambda(2l, k\#k) = \lambda(l, k)\widehat{\oplus}\lambda(l, k) = \lambda(2l, k),$$

and hence, by injectivity of λ, we obtain that every $k \in \mathbf{Z}$ must be $\#$-idempotent.

The remaining part of this paper is devoted to the ring operations on $\mathbf{Z} \times \mathbf{Z}$ defined coordinatewise with *exactly one* idempotent operation on the second coordinate (a general construction is presented in the lemma below, and its easy proof is omitted). These structures will be transferred onto $\mathcal{P}$ in Theorem 1 below.

Lemma 2. *Let $X = (X, +, \cdot)$ be a ring with unit e_X, and let $Y = (Y, \#)$ be a commutative and idempotent monoid with unit e_Y. Then the product $X \times Y$, enowed with the addition $\oplus$ and the multiplication $\odot$ defined by the formulas*:

$$(x_1, y_1) \oplus (x_2, y_2) := (x_1 + x_2, y_1\#y_2), \text{ and } (x_1, y_1) \odot (x_2, y_2) := (x_1 \cdot x_2, y_1\#y_2)$$

is a ring with unit $e = (e_X, e_Y)$. Moreover, for all $x_1, x_2 \in X$, $y \in Y$ we have

$$(x_1, y) \oplus (x_1, y) = (x_1 + x_2, y) \text{ and } (x_1, y) \odot (x_1, y) = (x_1 \cdot x_2, y).$$

Now we give two sample operations that make $\mathbf{Z}$ an idempotent monoid with unit element $e = 1$. For this purpose, let $s(k)$ denote the sign of $k \in \mathbf{Z}$, and let $[m, n]$ denote the least common multiple of two positive integers m, n. Then the operations are defined as follows:

$$k_1 \clubsuit k_2 := \min\{s(k_1), s(k_2)\} \cdot [|k_1|, |k_2|] \text{ for } k_1k_2 \neq 0, \text{ and } k_1 \clubsuit k_2 := 0 \text{ otherwise};$$

$$k_1 \spadesuit k_2 := \min\{s(k_1), s(k_2)\} \cdot \max\{|k_1|, |k_2|\}.$$

From Lemma 2 and the above two examples it follows that $\mathbf{Z} \times \mathbf{Z}$ possess two additional ring structures, and that these structures can be transferred onto $\mathcal{P}$ by means of the function λ, with $\lambda(l_1, k) \oplus \lambda(l_2, k) = \lambda(l_1 + l_2, k)$ and $\lambda(l_1, k) \odot \lambda(l_2, k) = \lambda(l_1 l_2, k)$, where $\oplus$ and $\odot$ are defined as in Lemma 2. We thus have obtained the following result.

Theorem 1. *Let* $\# \in \{\clubsuit, \spadesuit\}$ *be fixed. Then the set* $\mathcal{P}$ *is a commutative ring with unit* $e = (3, 4, 5) = \lambda(1, 1)$ *under the following addition* $\oplus$ *and multiplication* $\odot$:

$$\lambda(l_1, k_1) \oplus \lambda(l_2, k_2) := \lambda(l_1 + l_2, k_1 \# k_2),$$

$$\lambda(l_1, k_1) \odot \lambda(l_2, k_2) := \lambda(l_1 l_2, k_1 \# k_2),$$

and these operations leave each $\mathcal{P}_k$ *invariant.*

By comparing Theorem 1 and the claim in (4), we see that the problem of extending (in a natural way) G-operations to the whole set $\mathcal{P}$ remains open.

ACKNOWLEDGEMENTS.

I would like to thank the Organizers of ICMA 2004 both for the financial support of my participation in the Conference and their warm hospitality during my stay in Kuwait.

REFERENCES

[1] Dawson, B., A Ring of Pythagorean Triples, *Missouri J. Math. Sci.*, 6 (1994), 72-77.

[2] Eckert, E.J., The Group of Primitive Pythagorean Triangles, *Math. Magazine*, 57 (1984), 22-27.

[3] Grytczuk, A., Note on a Pythagorean Ring, *Missouri J. Math. Sci.*, 9 (1997), 83-89.

[4] Grytczuk, A. and Wójtowicz, M., The Problem of Freeness for Euler Monoids and Möbius groups, *Semigroup Forum*, 61 (2000), 277-282.

[5] LeVeque, W.J., *Fundamentals of Number Theory*, Addison-Wesley 1977, Massachusetts-London-Amsterdam.

[6] Wójtowicz, M., Algebraic Structures of Some Sets of Pythagorean Triples. I, *Missouri J. Math. Sci.*, 12 (2000), 31-35.

[7] Wójtowicz, M., Algebraic Structures of Some Sets of Pythagorean Triples. II, *Missouri J. Math. Sci.*, 13 (2001), 17-23.

IMPLICIT TAYLOR METHODS FOR PARABOLIC EQUATIONS WITH JUMPS IN DATA

M. A. Al-Zanaidi[1], C. Grossmann[2] and A. Noack[2]
[1]Department of Mathematics & Computer Science
Kuwait University, P.O. Box 5969, Safat 13060, Kuwait
e-mail: zanaidi@mcs.sci.kuniv.edu.kw
[2]Institute of Numerical Mathematics, TU Dresden, 01062 Dresden, Germany
e-mail: grossm&noack@math.tu-dresden.de

1. INTRODUCTION

In various applications, e.g. semi-discretized boundary control problems, parabolic problems with piecewise defined boundary conditions generically occur, in particular, with jumps at the boundary. Since discontinuous data lead to a reduced smoothness of the solution, adapted numerical methods have to be chosen. As model problem in the present paper we consider linear parabolic problems with a finite number of jumps in boundary data. For a spatial semi-discretization we apply piecewise linear finite elements with mass lumping. This method of lines approach generates initial value problems (IVP) which are stiff and have to be treated by an efficient method for time discretization. In case of smooth boundary data the trapezoidal rule, corresponding to the Crank-Nicolson method, works efficiently for such a class of equations. In case of jumps in data, however, applying Crank-Nicolson leads to high frequency oscillations over a long time horizon. To avoid these spurious oscillations one could ensure the discrete maximum principle, but this would result in serious restrictions on the time steps similar to the explicit Euler integration. On the other hand the commonly used implicit Euler scheme, though it generates stable discretizations without restrictions on the step size, is only of first order convergent.

The focus of the present paper is to analyze in detail certain implicit Taylor methods. Numerical experiments showed that these techniques are quite well adapted to the particular situation of linear IVP with nonsmooth data. They possess higher accuracy properties as well as provide a damping of unwanted oscillations. Assuming smoothness of the solution standard consistency and convergence analysis has been given e.g. in Scholz [13]. However, the derived consistency error explicitely contains estimates of derivatives of the solution. In case these terms are not bounded no conclusion can be derived from this type of consistency errors. An important aim of our study is to obtain estimates that depend only upon the chosen step sizes and original data of the considered parabolic problem.

Furthermore, in the investigated implicit Taylor methods at each discrete time level a linear system has to be solved. Unlike in the implicit Euler method or Crank-Nicolson the coefficient matrix of the generated linear systems contains higher powers of the original stiffness matrix. Conjugate gradient (CG) methods provide an efficient tool for the numerical treatment of these systems because only repeated evaluations of the type stiffness matrix times a vector are required. However, the increasing ill-conditioning of the system with decreasing spatial steps sizes would lead to a high number of necessary iterations if no appropriate preconditioner is applied. We propose a special preconditioner for the CG algorithm which bases on linear systems that have the same sparsity structure as the original stiffness matrix of the FEM spatial discretization. The convergence of the corresponding preconditioned conjugate gradient algorithm (PCG) is studied. It turns out that this preconditioner is highly efficient; due to the rapid convergence of only a few iteration steps are needed. Moreover, a truncated and a further simplified version of this PCG method are considered, their convergence is discussed and the claimed damping properties concerning higher order frequencies are justified.

The paper is organized as follows. In Section 2 we introduce the parabolic boundary value problem and its semi-discretization by finite elements. Implicit Taylor methods are studied in Section 3 and are applied to the considered model problem to obtain a corresponding discrete problem. Then a convergence analysis is provided, where unlike the general nonlinear consistency analysis, there we derive truncation errors directly depending on the given problem data. That way no estimates of higher derivatives of the solution are required. In Section 4 a CG algorithm with preconditioning to solve the generated linear systems is constructed and its convergence properties are investigated. Further, a truncated and a simplified versions of this PCG method are discussed. Several test examples are given in Section 5 to illustrate the efficiency of the proposed methods.

2. THE PARABOLIC PROBLEM AND SEMI-DISCRETIZATION

Let $\Omega \subset \mathbb{R}^2$ denote a bounded domain with a piecewise Lipschitz boundary Γ. Further, let $T > 0$ be fixed. We consider the parabolic boundary value problem

$$\begin{aligned} \frac{\partial w}{\partial t} - \Delta w &= f \quad \text{in} \quad Q := \Omega \times (0,T], \\ \gamma_D\, w + \frac{\partial w}{\partial n} &= b \quad \text{on} \quad \Gamma_T := \Gamma \times (0,T], \\ w(\cdot,0) &= 0 \quad \text{on} \quad \overline{\Omega}. \end{aligned} \tag{1}$$

Here $\gamma_D \geq 0$ is a given coefficient, $f \in L_\infty(Q)$ and $b \in L_\infty(\Gamma_T)$.

Concerning existence and uniqueness of the solution of (1) various results can be found in literature. We refer e.g. to Casas [5], where existence and uniqueness results for more general classes of parabolic problems are proved. For our model problem it provides the existence of a unique weak solution $w(\cdot,\cdot) \in W(0,T) \cap C(\overline{Q})$ of (1) satisfying the variational equation

$$\int_Q \frac{\partial w}{\partial t}\,\varphi\,dQ + \int_Q \nabla w \circ \nabla\varphi\,dQ + \gamma_D \int_{\Gamma_T} w\,\varphi\,d\Gamma_T = \int_Q f\,\varphi\,dQ + \int_{\Gamma_T} b\,\varphi\,d\Gamma_T \quad (2)$$

$$\forall\,\varphi \in V := L_2(0,T;H^1(\Omega))$$

and $w(\cdot,0) = 0$. Here

$$W(0,T) := \{v \in V\ :\ \frac{\partial v}{\partial t} \in V^*\}, \quad (3)$$

and V^* denotes the dual space corresponding to V. Notice that this result is valid for any right-hand side $f \in V^*$. The type of Robin or Neumann boundary conditions can be also replaced by those of Dirichlet type. However, in this case of non-matching conditions further singularities occur (cf. Grossmann, Noack and Vanselow [10]).

Obviously, in the considered parabolic problem (1) the boundary conditions are naturally included in the variational equation (2).

To solve the equations (1) numerically appropriate discretizations of the states w and the appearing differential operators in (1) as well as of the boundary function b are required. We apply the principle of semi-discretization in space (MOL) with piecewise linear triangular finite elements and mass-lumping. To avoid additional errors due to the discretization of the boundary of Γ we assume Ω to be a polyhedron and consider such discretizations only which take into account the sub-structuring of its boundary Γ. Furthermore, we assume that the discretization at the boundary, i.e. for te function b, relates to macro-elements of the discretization of (1) w.r.t. space and time.

Let Ω be covered by triangles which satisfy the standard assumptions of finite element methods (cf. Ciarlet [7]). The related grid points in space we denote by x_j, $j = 1,\ldots,N$ and $\varphi_j \in C(\bar{\Omega})$ denote the Lagrange basis functions of piecewise linear C^0 finite elements. The related conforming finite element discretization of the Sobolev space $H^1(\Omega)$ is then given by the subspace $V_h := \operatorname{span}\{\varphi_i\}_{i=1}^N$.

We apply the Ritz-Galerkin semi-discretization with mass lumping, i.e. integrals $\int_\Omega \varphi_j\,\varphi_i\,d\Omega$ are replaced by the lumping operator

$$D_L(\varphi_j,\varphi_i) := \mu_i\,\delta_{ij}, \quad i,j = 1,\ldots,N \quad (4)$$

with $\mu_i := \sum_{j=1}^N \int_\Omega \varphi_j \, \varphi_i \, d\Omega$. As shown in Thomée [14] this kind of mass lumping is equivalent to the evaluation of the mass integrals by the trapezoidal rule. Introducing the semi-discrete solution $w_h \in V_h$ in the spatial basis $\{\varphi_j\}_j$ by

$$w_h(x,t) \;=\; \sum_{j=1}^{N} w_j(t)\; \varphi_j(x) \tag{5}$$

with coefficient functions $w_j : [0,T] \to \mathbb{R}$, $j = 1, \ldots, N$ the Ritz-Galerkin technique applied to (2) leads to a finite dimensional IVP for the coordinate function $\mathbf{w} = (w_1, \ldots, w_N) : [0,T] \to \mathbb{R}^N$ as given by

$$\mathbf{D}\,\mathbf{w}'(t) \;=\; \mathbf{A}\mathbf{w}(t) \;+\; \mathbf{f}(t), \quad t \in (0,T], \qquad \mathbf{w}(0) = \mathbf{0}, \tag{6}$$

where the elements of the matrix $\mathbf{A} := (a_{ij})_{i,j=1}^N$ are defined by

$$a_{ij} := -\int_\Omega \nabla\varphi_j \circ \nabla\varphi_i \, d\Omega - \int_\Gamma \gamma_D \, \varphi_j \varphi_i \, d\Gamma\,, \qquad i,j = 1, \ldots, N, \tag{7}$$

$\mathbf{D} := \mathrm{diag}(\mu_i)$ denotes the matrix representation of D_L and $\mathbf{f} := (f_i)_{i=1}^N$ is given by

$$f_i(t) := \int_\Omega f(x,t)\, \varphi_i(x)\, d\Omega + \int_\Gamma b(x,t)\, \varphi_i(x)\, d\Gamma\,, \quad i = 1, \ldots, N. \tag{8}$$

Let us indicate that existence and uniqueness of solutions of system (6) are covered by standard theory and piecewise defined solutions.

By means of the transformation $\mathbf{D}^{1/2}\mathbf{w}$, and simply renaming all modified matrices and functions by their former names, (6) is equivalent to the IVP

$$\mathbf{w}'(t) \;=\; \mathbf{A}\mathbf{w}(t) \;+\; \mathbf{f}(t), \quad t \in (0,T], \qquad \mathbf{w}(0) = \mathbf{0} \tag{9}$$

with a symmetric, negative definite matrix $\mathbf{A}$.

Next we consider a restriction of the space of boundary functions b to the subspace of piecewise constant functions w.r.t. some given time grid

$$0 = t^0 < t^1 < \cdots < t^{M^c - 1} < t^{M^c} = T \tag{10}$$

by $\mathcal{P}_{0,\rho} \subset L^\infty(\Gamma_T)$, i.e. it holds

$$b^\rho \in \mathcal{P}_{0,\rho} \quad \Longleftrightarrow \quad b(x,t) = b^k(x),\ \forall t \in (t^{k-1}, t^k],\ k = 1, \ldots, M^c,\ x \in \Omega \tag{11}$$

with functions $b^k \in L^\infty(\Gamma)$. Here M^c denotes the number of grid points. In case of boundary control problems this can be interpreted as simple time discretization

of controls. In the sequel we use b^ρ in (8) if the function b belongs to $\mathcal{P}_{0,\rho}$, i.e. by $b^\rho(x,t) \equiv b(x,t^k)$ for $t \in (t^{k-1}, t^k]$. Similarly, we write

$$\mathbf{w}'(t) \;=\; \mathbf{A}\mathbf{w}(t) \;+\; \mathbf{f}^\rho(t), \quad t \in (0,T], \qquad \mathbf{w}(0) = \mathbf{0} \tag{12}$$

if f is generated via boundary data g^ρ. Since b^ρ is piecewise constant the right-hand side of (12) is discontinuous, so we cannot expect a higher smoothness in the classical sense as it would be in case of, say, continuous boundary values $b \in C(\Gamma_T)$.

But according to the structure (11) of b the solution of (12) can be received recursively on the subintervals $(t^{k-1}, t^k]$ by

$$\begin{aligned} \mathbf{w}'(t) &= \mathbf{A}\mathbf{w}(t) + \mathbf{f}^\rho(t), \quad t \in (t^{k-1}, t^k], \\ \mathbf{w}(t^{k-1}+0) &= \mathbf{w}(t^{k-1}), \end{aligned} \tag{13}$$

for $k = 1, \ldots, M^C$ with $\mathbf{w}(t^0) = \mathbf{0}$. Taking into account the linearity of the above equations and the smoothness of $\mathbf{f}^\rho$ and b^ρ in the subintervals there exists a classical solution of (13) which can be represented in integral form by

$$\mathbf{w}(t) = e^{\mathbf{A}(t-t^{k-1})}\mathbf{w}(t^{k-1}) + \int_{t^{k-1}}^{t} e^{\mathbf{A}(t-s)}\mathbf{f}^\rho(s)ds, \; t \in (t^{k-1}, t^k], \; k = 1, \ldots, M^C. \tag{14}$$

This representation will be utilized in the next section.

3. IMPLICIT TAYLOR METHODS

In this section we briefly describe implicit Taylor methods (for further details cf. Griewank et al [6], Scholz [13]) applied to N-dimensional linear initial value problems of the type

$$\mathbf{w}'(t) \;=\; \mathbf{A}\mathbf{w}(t) \;+\; \mathbf{f}(t), \quad t \in (0,T], \qquad \mathbf{w}(0) = \mathbf{g} \tag{15}$$

with a symmetric, negative definite matrix $\mathbf{A}$, a given function $\mathbf{f} : (0,T] \to \mathbb{R}^N$ and some vector $\mathbf{g} \in \mathbb{R}^N$. As shown in Section 2 above, the chosen semi-discretization of the parabolic problem (1) leads to a finite dimensional IVP of this type. In view of this semi-discretization with jumps in data we restrict our attention to right-hand sides $\mathbf{f}$ which may possess discontinuities, as finite jumps, at the grid points of some given time grid.

Unlike Scholz [13], where a general study of implicit Taylor methods in terms of derivatives of the solution is given, here we investigate the special case (15) more

in detail for the case of linear IVPs with a finite number of discontinuities. In the convergence analysis, which we provide in this section, truncation errors are derived which only depend on the given problem data such that no estimates of the solution and its derivatives need to be involved. This convergence behaviour and certain damping properties, which shall be discussed later, show that for semi-discretized parabolic problems with discontinuous data, as considered in Section 2, a time discretization with implicit Taylor methods prove to be suitable.

Now we consider the first step of the implicit Taylor method, i.e. from $t_0 = 0$ to $t_1 = \tau$. For simplicity we take some equidistant time grid $t_j = j\tau, j = 0, \ldots, M$, where M denotes the number of grid points and $\tau = T/M$. The idea is to approximate the solution $\mathbf{w}$ of (15) in the interval $[0, \tau]$ by a function $\mathbf{w}_\tau$ of the form

$$\mathbf{w}_\tau(t) = \sum_{j=0}^{q} \frac{\alpha_j}{j!} (\tau - t)^j, \quad t \in [0, \tau] \tag{16}$$

with a fixed $q \in \mathbb{N}$ and vectors $\alpha_j \in \mathbb{R}^N$, $j = 0, 1, \ldots, q$, which are uniquely determined by the conditions

$$\mathbf{w}_\tau(0) \;=\; \mathbf{w}_0 \tag{17}$$

$$\mathbf{w}_\tau(\tau) \;=\; \mathbf{w}_0 \,+\, \int_0^\tau \Big([\mathbf{A}\,\mathbf{w}_\tau](t) + \mathbf{f}(t) \Big)\, dt \tag{18}$$

and the conditions for the derivatives

$$\mathbf{w}_\tau^{(j)}(\tau) \;=\; \mathbf{A}^j \mathbf{w}_\tau(\tau) \,+\, \sum_{l=0}^{j-1} \mathbf{A}^{j-1-l} f^{(l)}(\tau),, \quad j = 1, \ldots, q-1, \tag{19}$$

where due to (15) the inital vector $\mathbf{w}_0 = \mathbf{g}$ is chosen. In the sequel we denote the implicit Taylor method (16) - (19) by ITM-q.

Let us now restrict our attention to the vector $\mathbf{w}_1 := \mathbf{w}_\tau(\tau)$ at the grid point t_1. After eliminating the coefficients α_j (cf. Al-Zanaidi and Grossmann [1], Grossmann and Horváth [9] for details) the above conditions yield that $\mathbf{w}_1$ arises from $\mathbf{w}_0$ according to

$$\begin{aligned}\left(\mathbf{I} + \sum_{j=0}^{q-1} \frac{(-1)^j}{j!} \left(\frac{1}{q+1} - \frac{1}{j+1} \right) \tau^{j+1}\, \mathbf{A}^{j+1} \right) \mathbf{w}_1 = \left(\mathbf{I} + \frac{\tau}{q+1}\, \mathbf{A} \right) \mathbf{w}_0 \\ + \sum_{j=1}^{q-1} \frac{(-1)^j \tau^{j+1}}{j!} \left(\frac{1}{j+1} - \frac{1}{q+1} \right) \sum_{l=0}^{j-1} \mathbf{A}^{j-l} \mathbf{f}^{(l)}(\tau) + \int_0^\tau \mathbf{f}(t)\, dt.\end{aligned} \tag{20}$$

ITM-q consists in applying the above rules, which we presented for the first time step, analogously to each of the following time steps of the given time grid. That way in the grid points an approximation $\mathbf{w}_j \sim \mathbf{w}(t_j)$, $j = 1, \ldots, N$, of the solution of the IVP (15) is generated. Now we consider this method in each of the subintervals where the right-hand side $\mathbf{f}$ is smooth, and for simplicity omit the notation of possibly different step sized τ for different subintervals. Then by introducing the matrices

$$\begin{aligned} \mathbf{B} &:= \mathbf{I} + \sum_{j=0}^{q-1} \frac{(-1)^j}{j!} \left(\frac{1}{q+1} - \frac{1}{j+1} \right) \tau^{j+1} \mathbf{A}^{j+1}, \\ \mathbf{S} &:= \mathbf{I} + \frac{\tau}{q+1} \mathbf{A} \end{aligned} \tag{21}$$

and the linear operator

$$\mathbf{R}\,\mathbf{f} := \sum_{j=1}^{q-1} \frac{(-1)^j \tau^{j+1}}{j!} \left(\frac{1}{j+1} - \frac{1}{q+1} \right) \sum_{l=0}^{j-1} \mathbf{A}^{j-l} \mathbf{f}^{(l)} \tag{22}$$

ITM-q is equivalently described by the recursive equations

$$\begin{aligned} \mathbf{B}\mathbf{w}_k &= \mathbf{S}\mathbf{w}_{k-1} + [\mathbf{R}\,\mathbf{f}](t_k) + \int_{t_{k-1}}^{t_k} \mathbf{f}(t)\, dt, \qquad k = 1, \ldots, M \\ \mathbf{w}_0 &= \mathbf{g}. \end{aligned} \tag{23}$$

Remark 1 In general, to avoid a heavy reduction of sparsity caused by higher powers of $\mathbf{A}$ in formula (20) we focus on the cases $q = 1$, 2, 3. Observe that for $q = 1$ the corresponding Implicit Taylor method forms a Crank-Nicolson-like method and for $q = 2$ it coincides with ETF (extended trapezoidal formula) as studied in Chawla and Al-Zanaidi [3].

To derive stability results for ITM-q we indicate that the stability function of ITM-q for values $q = 2$ and $q = 3$ coincides with that of the Implicit Runge-Kutta methods Radau IA and Lobatto III C, respectively. Hence, according to the stability behaviour of these Runge-Kutta methods (cf. Hairer, Nørsett and Wanner [11]) we conclude that ITM-q are A-stable for values of $q = 1$, 2, 3 and even L-stable for values of $q = 2$, 3. Stability as well as consistency properties of Implicit Taylor methods in case of sufficiently smooth data can be found e.g. in Scholz [13].

Notice further that for $q = 2$, 3 the corresponding Implicit Taylor methods are advantageous over Crank-Nicolson in case of dominantly occurring high eigenfrequencies. This fact rests on the L-stability of these methods and will be discussed more in detail in the next section.

Theorem 1 Let be given a function $\mathbf{f} : (0,T] \to \mathbb{R}^N$ which is sufficiently smooth on each of the subintervals of a given time grid. Then ITM-q applied to (15) is convergent with order of convergence $q+1$ and the local error of one step of the method, w.l.o.g. for the first one, is of the form

$$\mathbf{e}_{loc} = \frac{(-1)^q}{(q+2)!}\left(\frac{1}{q+1}\mathbf{A}^{q+2}\mathbf{w}(\tau) - \mathbf{A}^{q+1}\mathbf{f}(\tau) + \mathbf{f}^{(q+1)}(\tau)\right)\tau^{q+2} + O(\tau^{q+3}). \tag{24}$$

Proof: We sketch only the main steps of the proof and refer to Al-Zanaidi and Grossmann [1] for the further details of underlying expansions. According to the structure of the implicit Taylor method (23) it is obvious that we have to consider just one step in the consistency analysis.

First notice that we may increase the upper index $q-1$ to q in all the appearing summations without changing the method. Hence, considering the first interval, scheme (20) can be written equivalently as

$$\begin{aligned}\Big(\mathbf{I} + \sum_{j=0}^{q} \frac{(-1)^j}{j!}\Big(\frac{1}{q+1} - \frac{1}{j+1}\Big)\tau^{j+1}\,\mathbf{A}^{j+1}\Big)\mathbf{w}_1 &= \Big(\mathbf{I} + \frac{\tau}{q+1}\,\mathbf{A}\Big)\,\mathbf{w}_0 \\ &+ \sum_{j=1}^{q} \frac{(-1)^j\tau^{j+1}}{j!}\Big(\frac{1}{j+1} - \frac{1}{q+1}\Big)\sum_{l=0}^{j-1}\mathbf{A}^{j-l}\mathbf{f}^{(l)}(\tau) + \int_0^\tau \mathbf{f}(t)\,dt.\end{aligned} \tag{25}$$

Now, replacing $\mathbf{w}_\tau$ by the exact solution $\mathbf{w}$ of (13) and using a Taylor expansion and the expression $\mathbf{w}(\tau) = e^{\mathbf{A}\tau}\mathbf{w}_0 + \int_0^\tau e^{\mathbf{A}(\tau-s)}\mathbf{f}(s)\,ds$ (compare (14)) we obtain that the left-hand side of (25) equals

$$\begin{aligned}&\Big(\mathbf{I} + \sum_{j=0}^{q} \frac{(-1)^j}{j!}\Big(\frac{1}{q+1} - \frac{1}{j+1}\Big)\tau^{j+1}\,\mathbf{A}^{j+1}\Big)\mathbf{w}(\tau) \\ &\qquad = \Big(\mathbf{I} + \frac{\tau}{q+1}\,\mathbf{A}\Big)\,\mathbf{w}_0 + r_q + \Big(\mathbf{I} + \frac{\tau}{q+1}\,\mathbf{A}\Big)\int_0^\tau e^{-\mathbf{A}s}\mathbf{f}(s)\,ds\end{aligned} \tag{26}$$

with

$$\mathbf{r}_q = \frac{(-1)^q}{(q+2)!}\frac{\tau^{q+2}}{q+1}\mathbf{A}^{q+2}\mathbf{w}(\tau) + O(\tau^{q+3}) \tag{27}$$

On the other hand, after some calculations, it follows

$$\begin{aligned}&\Big(\mathbf{I} + \frac{\tau}{q+1}\mathbf{A}\Big)\int_0^\tau e^{-\mathbf{A}s}\mathbf{f}(s)\,ds \;\; = \\ &\qquad \int_0^\tau \mathbf{f}(s)\,ds + \sum_{j=1}^{q}\tau^{j+1}\frac{(-1)^j}{j!}\Big(\frac{1}{j+1} - \frac{1}{q+1}\Big)\sum_{l=0}^{j-1}\mathbf{A}^{j-l}\mathbf{f}^{(l)}(\tau) + s_q,\end{aligned} \tag{28}$$

where

$$\mathbf{s}_q = \frac{(-1)^{q+1}}{(q+2)!}\left(\mathbf{A}^{q+1}\mathbf{f}(\tau) - \mathbf{f}^{(q+1)}(\tau)\right)\tau^{q+2}. \tag{29}$$

Thus for the defect appearing when replacing the approximation by the exact solution in the Implicit Taylor formula we obtain

$$\begin{aligned}\mathbf{e}_{loc} &= \mathbf{r}_q + \mathbf{s}_q \\ &= \frac{(-1)^q}{(q+2)!}\left(\frac{1}{q+1}\mathbf{A}^{q+2}\mathbf{w}(\tau) - \mathbf{A}^{q+1}\mathbf{f}(\tau) + \mathbf{f}^{(q+1)}(\tau)\right)\tau^{q+2} + O(\tau^{q+3})\end{aligned} \tag{30}$$

Since $O(\tau^{-1})$ subintervals occur, the above results imply that the ITM-q (23) possesses the overall consistency order $q+1$ provided that $\mathbf{f}$ is sufficiently smooth on each of the subintervals. Together with the known stability of one-step methods this proves the assertion. □

Remark 2 One should be aware of the fact that the term $\tau^{q+2}\mathbf{A}^{q+2}$ and consequently $(\tau/h^2)^{q+2}$ occurs in the error term. In case of smooth data this effect will be compensated by the smoothness of the solution - not so for nonsmooth data. Thus, in case of discontinuities to guarantee convergence a coupling condition for spatial and time steps of the form

$$\tau/h^2 \to 0 \tag{31}$$

has to be required.

Returning to the semi-discretized parabolic problem (1) with discontinuous boundary data $b^\rho \in \mathcal{P}_{0,\rho}$, notice that in Section 2 we derived the recursive formula (13) on subintervals. Hence, assuming a sufficiently smooth right-hand side f in each of the subintervals and choosing a probably finer time grid for ITM-q containing all jumps at the boundary, Theorem 1 can be applied and gives a characterization of the convergence behaviour of ITM-q for semi-dicretizations of linear parabolic problems with jumps in boundary data.

4. EFFICIENT NUMERICAL REALIZATION OF ITM-q

In each step of ITM-q linear systems

$$\begin{aligned}\mathbf{B}\mathbf{w}_k &= \mathbf{S}\mathbf{w}_{k-1} + [\mathbf{R}\,\mathbf{f}](t_k) + \int_{t_{k-1}}^{t_k} \mathbf{f}(t)\,dt, \qquad k = 1, \ldots, M \\ \mathbf{w}_0 &= \mathbf{g}.\end{aligned} \tag{32}$$

have to be solved. In particular, we are interested in the method for values of $q = 2, 3$. The related coefficient matrices $\mathbf{B} := \mathbf{B}_q$ for the considered values $q = 2$ and $q = 3$ have the specific form

$$\mathbf{B}_2 \;=\; \mathbf{I} \,-\, \frac{2}{3}\,\tau\,\mathbf{A} \,+\, \frac{1}{6}\,\tau^2\,\mathbf{A}^2 \tag{33}$$

and

$$\mathbf{B}_3 \;=\; \mathbf{I} \,-\, \frac{3}{4}\,\tau\,\mathbf{A} \,+\, \frac{1}{4}\,\tau^2\,\mathbf{A}^2 \,-\, \frac{1}{24}\,\tau^3\,\mathbf{A}^3, \tag{34}$$

respectively. It is well known that the semi-discretization of the parabolic state equations (1) by linear finite elements generates a stiffness matrix $\mathbf{A}$ that is rather sparse. However, this sparsity is partially destroyed in $\mathbf{A}^2$ and $\mathbf{A}^3$. Hence, appropriate numerical realizations of (23) that avoid this effect should be applied. To maintain the structural properties of $\mathbf{A}$ our method of choice is PCG, the preconditioned conjugate gradient method, with a preconditioner of the type

$$\mathbf{P}_q \;:=\; \prod_{j=1}^{q} \left(\mathbf{I} \,-\, \sigma_j\,\tau\,\mathbf{A}\right), \quad q = 2, 3 \tag{35}$$

with appropriately chosen constants $\sigma_j > 0$, $j = 1, \ldots, q$. For these matrices we obtain spectral bounds which are independent of the spatial and time discretization parameters h and τ. To cancel the highest order term of $\mathbf{B}_q$ and, that way, to obtain good contraction properties of PCG for relatively large time steps $\tau > 0$ we impose upon the parameters σ_j the condition

$$\prod_{j=1}^{q} \sigma_j \;=\; \frac{1}{q!\,(q+1)}. \tag{36}$$

Theorem 2 Independent of the discretization parameters $h > 0$ and $\tau > 0$ with the constant

$$c_2 \;:=\; c_2(\sigma_1, \sigma_2) \;=\; \frac{6 + 2\sqrt{6}}{6 \,+\, 3\sqrt{6}\,(\sigma_1 + \sigma_2)} \tag{37}$$

it holds

$$c_2\,\mathbf{v}^T\mathbf{P}_2\,\mathbf{v} \;\le\; \mathbf{v}^T\mathbf{B}_2\,\mathbf{v} \;\le\; \mathbf{v}^T\mathbf{P}_2\,\mathbf{v}, \quad \forall\,\mathbf{v} \in \mathbb{R}^N. \tag{38}$$

Moreover, the constant c_2 is optimal, i.e. maximal, in case of $\sigma_1 = \sigma_2 = \dfrac{1}{\sqrt{6}}$.

Proof: To obtain the stated spectral bound we study the generalized symmetric eigenvalue problem

$$\mathbf{B_2}\,\mathbf{v} \;=\; \nu\,\mathbf{P_2}\,\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}. \tag{39}$$

Since $\mathbf{A}$ is symmetric and negative definite due the structure of $\mathbf{B_2}$ and $\mathbf{P}_2$ we obtain that $\nu \in \mathbb{R}$ is an eigenvalue of the problem (39) iff an eigenvalue λ of A exists such that

$$\nu \;=\; \frac{1 - \frac{2}{3}\tau\lambda + \frac{1}{6}(\tau\lambda)^2}{(1-\sigma_1\tau\lambda)(1-\sigma_2\tau\lambda)}\,, \tag{40}$$

holds. With the abbreviation $s := -\tau\lambda$, $\sigma := \sigma_1 + \sigma_2$ and (36) as function of the parameter $s > 0$ we define

$$\nu_2(s) \;:=\; \nu \;=\; \frac{1 + \frac{2}{3}\,s + \frac{1}{6}\,s^2}{1 + \sigma\,s + \frac{1}{6}\,s^2} \;=\; 1 \;+\; \Big(\frac{2}{3} - \sigma\Big)\,\frac{s}{1 + \sigma\,s + \frac{1}{6}\,s^2}\,. \tag{41}$$

By the estimate $\frac{1}{2}(\sigma_1 + \sigma_2) \;\geq\; \sqrt{\sigma_1\sigma_2} \;=\; \frac{1}{\sqrt{6}}$ between arithmetic and geometric means we have $\sigma \geq \frac{2}{\sqrt{6}}$. Hence, $\frac{2}{3} - \sigma < 0$ for any feasible value of σ and from (41) it follows the upper bound

$$\nu \;\leq\; 1 \tag{42}$$

for any eigenvalue ν of (39). To obtain a lower bound we use

$$\max_{s\geq 0}\; \frac{s}{1 + \sigma\,s + \frac{1}{6}\,s^2} \;=\; \frac{\sqrt{6}}{2 + \sigma\,\sqrt{6}}. \tag{43}$$

in (41) which proves the left inequality in (38). It remains to derive the optimal parameters σ_1, σ_2. Obviously $c_2(\sigma_1, \sigma_2)$ is monotone in σ, thus using $\sigma \geq \dfrac{2}{\sqrt{6}}$ again results in

$$\max_{\sigma\geq 2/\sqrt{6}}\; c_2(\sigma_1,\sigma_2) \;=\; c_2(1/\sqrt{6}, 1/\sqrt{6}) \;=\; \frac{3+\sqrt{6}}{6} \;\approx 0.90825. \tag{44}$$

□

A direct consequence of the theorem is that the optimal preconditioning of the form (35) is attained for $\sigma_1 = \sigma_2 = 1/\sqrt{6}$ which yields the preconditioner

$$\mathbf{P}_2 \;=\; (\,\mathbf{I} \;-\; \frac{1}{\sqrt{6}}\,\tau\,\mathbf{A}\,)^2. \tag{45}$$

For the case $q = 3$ the study can be similarly done. In the following we concentrate upon $\sigma_1 = \sigma_2 = \sigma_3 = 1/\sqrt[3]{24}$, i.e. upon the preconditioner

$$\mathbf{P}_3 \;=\; (\,\mathbf{I} \;-\; \frac{1}{\sqrt[3]{24}}\,\tau\,\mathbf{A}\,)^3. \tag{46}$$

For this particular preconditioner P the eigenvalues of the generalized symmetric eigenvalue problem

$$\mathbf{B_3}\,\mathbf{v} \;=\; \nu\,\mathbf{P_3}\,\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}. \tag{47}$$

are positive and can be expressed by

$$\nu_3(s) := \frac{1 + \frac{3}{4}\,s + \frac{1}{4}\,s^2 + \frac{1}{24}\,s^3}{(1 + s/\sqrt[3]{24}\,)^3} \tag{48}$$

with $s = -\tau\lambda$ and λ an eigenvalue of A. Further elementary calculus yields

$$c_3\,\mathbf{v}^T\mathbf{P}_3\,\mathbf{v} \;\le\; \mathbf{v}^T\mathbf{B}_3\,\mathbf{v} \;\le\; \mathbf{v}^T\mathbf{P}_3\,\mathbf{v}, \quad \forall\,\mathbf{v} \in \mathbb{R}^N \tag{49}$$

with $c_3 \approx 0.7803$.

The scheme of ITM-q can shortly be written as linear system

$$\mathbf{B}\,\mathbf{v} \;=\; \mathbf{b} \tag{50}$$

with $\mathbf{B} = \mathbf{B}_q$ and $\mathbf{b}$ the right-hand side of ITM-q. Then in both cases, $q = 2$ and $q = 3$, the corresponding estimates (38) and (49), respectively, guarantee that the related PCG-method when applied to (50) starting from any $\mathbf{v}^0 \in \mathbb{R}^N$ generates vectors $\mathbf{v}^l \in \mathbb{R}^N$, $l = 1, 2, \ldots$ which converge to its solution $\mathbf{v}$ according to (compare e.g. Axelsson [2])

$$\|\mathbf{v}^l - \mathbf{v}\|_{\mathbf{P}^{-1}\mathbf{B}} \;\le\; 2\,\left(\frac{1 - \sqrt{c_q}}{1 + \sqrt{c_q}}\right)^l \|\mathbf{v}^0 - \mathbf{v}\|_{\mathbf{P}^{-1}\mathbf{B}}, \quad l = 1, 2, \ldots. \tag{51}$$

Here $\|\cdot\|_{\mathbf{P}^{-1}\mathbf{B}}$ denotes the related discrete energy norm. In the considered cases this provides the estimate

$$\|\mathbf{v}^l - \mathbf{v}\|_{\mathbf{P}^{-1}\mathbf{B}} \;\le\; 2\,\gamma^l\,\|\mathbf{v}^0 - \mathbf{v}\|_{\mathbf{P}^{-1}\mathbf{B}}, \quad l = 1, 2, \ldots \tag{52}$$

with $\gamma \approx 0.024$ and $\gamma \approx 0.062$ for $q = 2$ and $q = 3$, respectively. If, as a rule, $\tau/h^2 \gg 1$ then the convergence is even better since $\lim\limits_{s \to +\infty} \nu_q(s) = 1$, and since this limit is reached rather rapidly (see Fig. 1 and Fig. 2).

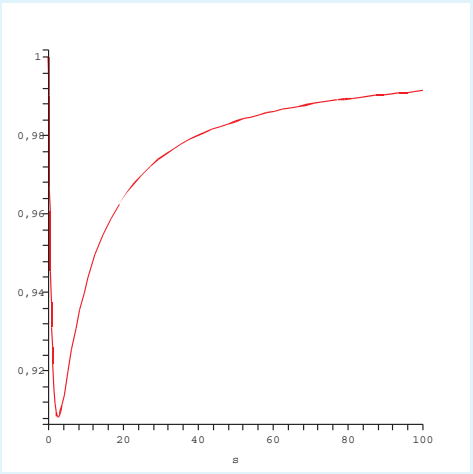


Fig. 1: Graph of ν_2

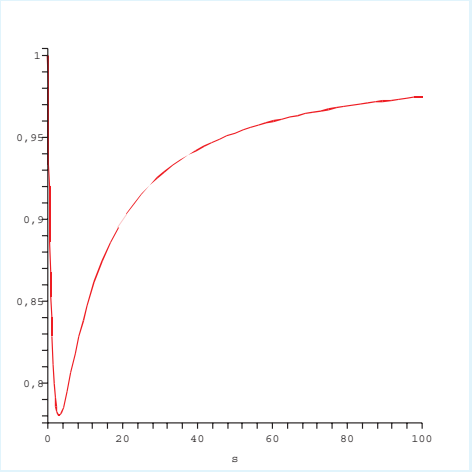


Fig. 2: Graph of ν_3

As a consequence of the fast reduction of the error only a few iteration steps of the PCG-method are required to solve problem (50) approximately, even up to high accuracy. Moreover, in case the linear systems arise from the time discretization of a semi-discrete parabolic problem good starting iterates are available from the previous time level provided that the solution does not change too rapidly. In addition, rapid changes correspond to a dominant influence of larger eigenvalues, but these are damped quite fast. In computational experiments which are reported later beside the discussed ITM-q methods we studied numerically the convergence behaviour of a simplified algorithm considering ITM-2 with only one PCG step at each time level. Given an initial guess $\mathbf{v}^0 \in \mathbb{R}^N$ this truncated version of PCG yields

$$\tilde{\mathbf{v}} \;=\; \mathbf{v}^0 \;+\; \alpha\,\mathbf{p}\,, \tag{53}$$

where the search direction $\mathbf{p} \in \mathbb{R}^N$ and the step size $\alpha > 0$ are defined by

$$\mathbf{p} \;:=\; \mathbf{P}^{-1}\mathbf{d}, \qquad \alpha \;:=\; \frac{\mathbf{p}^T\mathbf{d}}{\mathbf{p}^T\mathbf{B}\mathbf{p}} \qquad \text{with } \mathbf{d} \;:=\; \mathbf{b} - \mathbf{B}\mathbf{v}^0\,. \tag{54}$$

¿From (54) it follows

$$\alpha = \frac{\mathbf{p}^T\mathbf{P}\mathbf{p}}{\mathbf{p}^T\mathbf{B}\mathbf{p}}, \tag{55}$$

and applying Theorem 2 this leads to the estimate $1 \le \alpha \le 1/c_q$ and thus $\alpha \sim 1$. This suggests the following further simplification of the truncated ITM-2 method

$$\tilde{\mathbf{v}} \;=\; \mathbf{v}^0 \;+\; \mathbf{P}^{-1}(\mathbf{b} - \mathbf{B}\mathbf{v}^0) \;=\; \mathbf{P}^{-1}\mathbf{b} \;+\; (\mathbf{I} - \mathbf{P}^{-1}\mathbf{B})\mathbf{v}^0\;. \tag{56}$$

Applying the structure of $\mathbf{B} = \mathbf{B}_2$ and $\mathbf{P} = \mathbf{P}_2$ this method determines the approximate solution $\tilde{\mathbf{v}}$ of (50) as solution of the linear system

$$\mathbf{P}\,\tilde{\mathbf{v}} \;=\; \mathbf{b} \;+\; \tau\,(\frac{2}{3} - \frac{2}{\sqrt{6}})\,\mathbf{A}\,\mathbf{v}^0. \tag{57}$$

It is easily seen that for the spectral radius of the related iteration matrix the estimate

$$\rho(\mathbf{I} - \mathbf{P}^{-1}\mathbf{B}) \;\leq\; \frac{1}{2}\,(1 - \frac{\sqrt{6}}{3}\,) \;\approx\; 0.092 \tag{58}$$

holds which is independent of the discretization parameters $h > 0$ and $\tau > 0$. Hence, a good improvement of an approximate solution of (50) relative to the choice of the starting value is achieved.

For this simplified PCG method (56) the announced damping behaviour is rather easily to be shown. Consider an orthonormal eigensystem $(\mathbf{w}_j)_{j=1}^N$ of $\mathbf{A}$. It is a system of eigenvectors for $\mathbf{B}$ and $\mathbf{P}$ as well, due to the structure of these matrices, as announced in the proof of Theorem 2. Let $\mathbf{v}$ denote the exact solution of $\mathbf{B}\mathbf{v} = \mathbf{b}$. Then from (56) we obtain

$$\tilde{\mathbf{v}} - \mathbf{v} = (\mathbf{I} - \mathbf{P}^{-1}\mathbf{B})\mathbf{v}^0 + \mathbf{P}^{-1}\mathbf{B}\mathbf{v} - \mathbf{v} = (\mathbf{I} - \mathbf{P}^{-1}\mathbf{B})(\mathbf{v}^0 - \mathbf{v}). \tag{59}$$

Introducing the representations

$$\tilde{\mathbf{v}} = \sum_{j=1}^{N} \tilde{\gamma}^j \mathbf{w}_j, \quad \mathbf{v}^0 = \sum_{j=1}^{N} \gamma_0^j \mathbf{w}_j, \quad \mathbf{v} = \sum_{j=1}^{N} \gamma^j \mathbf{w}_j, \tag{60}$$

this leads to

$$\sum_{j=1}^{N} (\tilde{\gamma}^j - \gamma^j)\mathbf{w}_j = \sum_{j=1}^{N} (1 - \nu_2(\lambda_j))(\gamma_0^j - \gamma^j)\mathbf{w}_j, \tag{61}$$

and thus

$$\tilde{\gamma}^j - \gamma^j = (1 - \nu_2(\lambda_j))(\gamma_0^j - \gamma^j) \tag{62}$$

holds, where $\nu_2^j, j = 1, \ldots, N$, as defined in (41) denote the eigenvalues of $\mathbf{P}^{-1}\mathbf{B}$. As indicated above $\lim_{s\to+\infty} \nu_2(s) = 1$, and this convergence is rather fast. Hence, for larger values of s the term $1 - \nu_2(s)$ becomes quite small. This property causes the rapid damping of components corresponding to larger eigenvalues.

If we choose on each time level the value of the solution at the previous time step, $\mathbf{v}^0 := \mathbf{w}^{k-1}$, as initial value of the PCG step then simplified ITM-2 (56) becomes

$$\begin{aligned} \mathbf{P_2}\mathbf{w}_k &= \big(\mathbf{S_2} + (\frac{2}{3} - \frac{2}{\sqrt{6}})\tau A\big)\mathbf{w}_{k-1} + [\mathbf{R_2}\,\mathbf{f}](t_k) + \int_{t_{k-1}}^{t_k} \mathbf{f}(t)\,dt,\; k = 1, \ldots, M \\ \mathbf{w}_0 &= \mathbf{g}. \end{aligned} \tag{63}$$

Because of this single inner iteration step it is to be expected that the order of convergence of this modified method is reduced compared to the original ITM-2 method. Analogously to the proof of Theorem 1 we obtain the following assertion about the consistency error.

Theorem 3 Let be given a function $\mathbf{f} : (0, T] \to \mathbb{R}^N$ which is sufficiently smooth on each of the subintervals of a given time grid. Then simplified ITM-2 as given in (63) applied to (15) is of first order convergent with the local truncation error

$$\mathbf{e}_{loc} = \frac{2 - \sqrt{(}6)}{3}\tau^2\mathbf{A}^2\mathbf{w}(0) + \frac{2 - \sqrt{6}}{6}\tau^2\mathbf{A}\mathbf{f}(\tau) + O(\tau^3). \tag{64}$$

□

Remark 3 Notice that other values of α in the range $1 \leq \alpha \leq 1/c_q$ could be chosen. But a similar consistency analysis as for Theorem 3 shows that only for the value $\alpha = 1$ convergence of first order is achieved, for any $\alpha \neq 1$ error terms of first order appear.

5. NUMERICAL RESULTS

Example 1. We consider the parabolic initial-boundary value problem (1) with

$$\Omega = (-1, 1) \times (-1, 1) \subset \mathbb{R}^2, \qquad f \equiv 0, \qquad \gamma_D = 1, \qquad T = 1. \tag{65}$$

Further, the boundary conditions are defined by the function g that is piecewise constant along the boundary and takes the values $g = 1$ or $g = -1$, alternating w.r.t. subintervals of length 0.5 in spatial as well as in time direction.
For a spacial discretization we choose a uniform triangulation with totally 4225 grid points. Concerning time integration we compared the performance of the Euler implicit method (EI), Crank-Nicolson (CN) and ITM-2 in all cases with equidistant time steps $\tau = 0.05$. Euler implicit as well as ITM-2 lead to acceptable results where ITM-2 guarantees a higher accuracy. However, the Crank-Nicolson method generates oscillating approximations, which due to maximum principle cannot occur as features of solutions of the considered continuous problem.

Example 2. In this example we consider again the parabolic initial-boundary value problem (1), but now with an anchor shaped domain $\Omega \subset \mathbb{R}^2$ and select $T = 1$. Further, we assume that no source term occurs, i.e. $f \equiv 0$.

For the boundary conditions we choose $\gamma_D = 5$ and

$$g(x, y) = \begin{cases} 0, & \text{if } y \geq 0.3, \\ -30, & \text{if } y < 0.3. \end{cases} \tag{66}$$

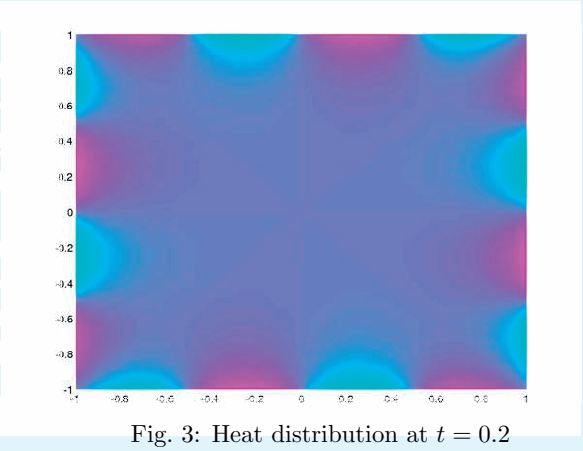


Fig. 3: Heat distribution at $t = 0.2$

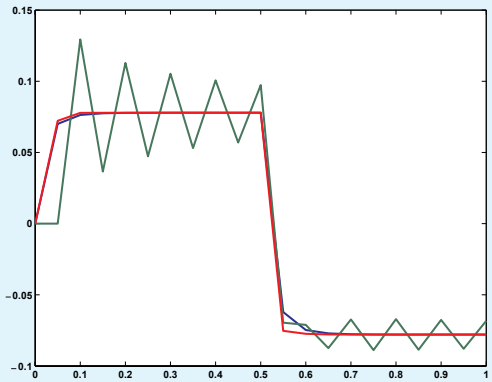


Fig. 4: Behaviour of EI, CN and ITM-2 at the point $x = (-1, 0.0938)$

In our calculations we applied the following triangulation generated by the pde-toolbox from MATLAB.

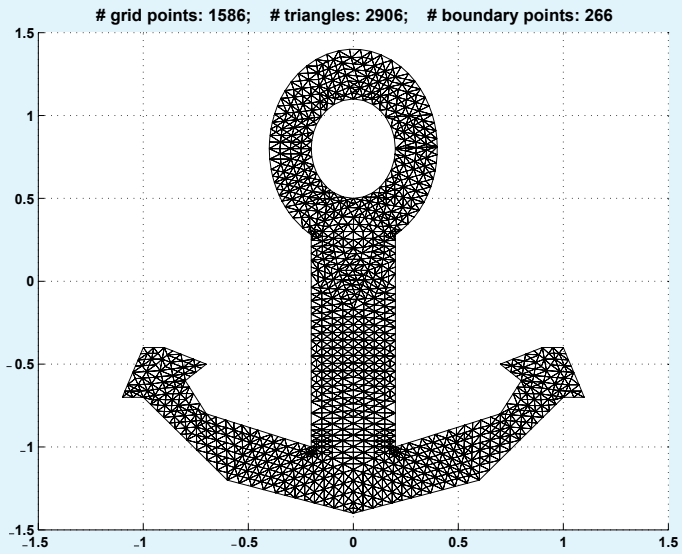


Fig. 5:Anchor with grid

In the figures given below the numerical results for a step size $\tau = 0.2$ obtained by EI, CN and ITM-2, respectively, at the time levels $t = 0.2$ and $t = 0.4$ are sketched as a 3D-plot.

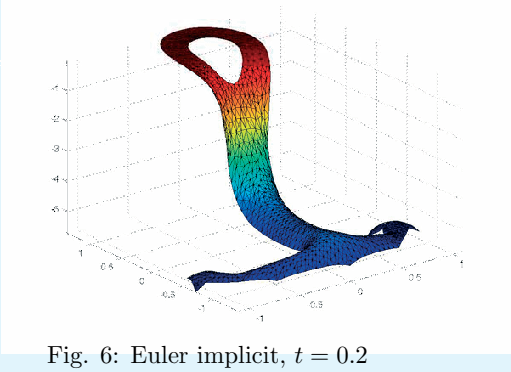


Fig. 6: Euler implicit, $t = 0.2$

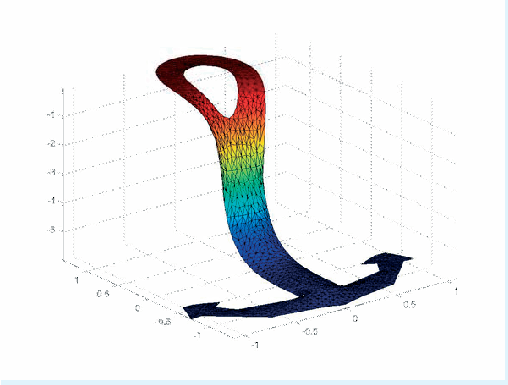


Fig. 7: Euler implicit, $t = 0.4$

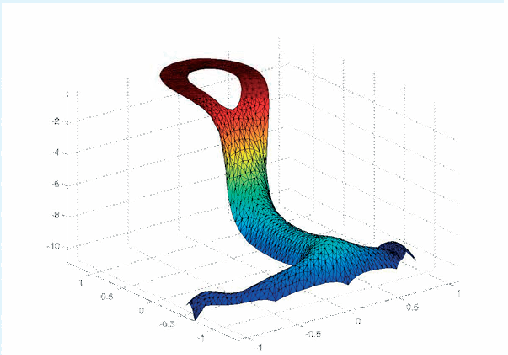


Fig. 8: Crank-Nicolson, $t = 0.2$

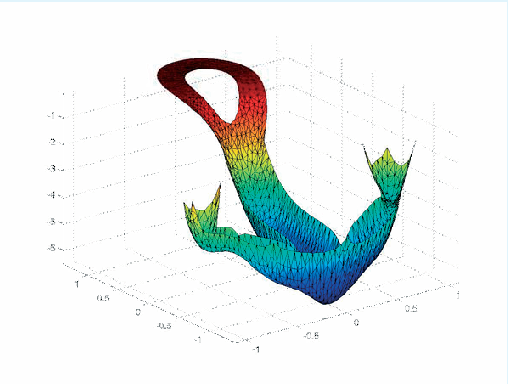


Fig. 9: Crank-Nicolson, $t = 0.4$

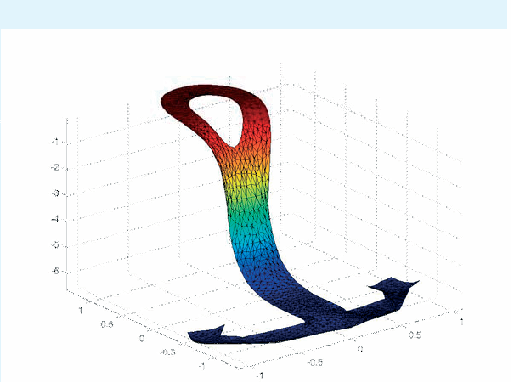


Fig. 10: ITM-2, $t = 0.2$

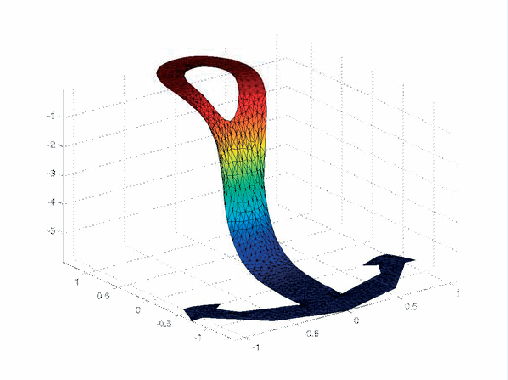


Fig. 11: ITM-2, $t = 0.4$

Again, the Crank-Nicolson method leads to physically unacceptable approximations of the wanted temperature profile as one can observe in Figure 10. Despite the cooling at the lower part of the boundary the numerically obtained temperature is internally lower than at the nearby boundary - which, indeed, cannot occur in reality.

Finally, Fig. 12 shows the behaviour of Euler implicit, Crank-Nicolson, ITM-2 and the simplified ITM-2 in time direction for a step size $\tau = 0.2$.

Here again, the already mentioned spurious oscillatory behavior of the Crank-Nicolson method is clearly visible. For comparison the solution obtained with high accuracy is sketched. The severe damping behavior of EI can be recognized as well as the good approximation via ITM-2. The simplified ITM-2 (dashed line) shows

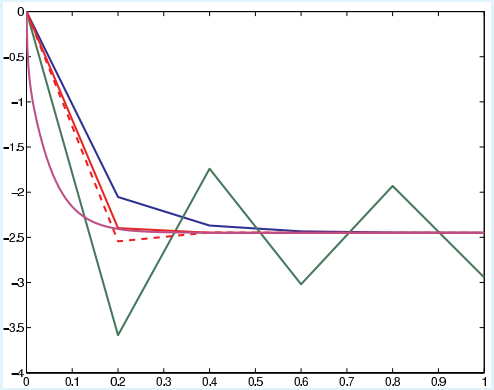


Fig. 12: Temperature at $(x, y) = (-0.2, 0.3)$

some mild overshooting in the first time step.

All computations have been implemented in MATLAB, Version 6.5.

AKNOWLEDGMENTS

We are grateful for the stimulating discussions with A.Griewank and R.Vanselow as well as for the support given by DFG grant GR 1777.2-2 and by Kuwait University.

REFERENCES

[1] Al-Zanaidi, M.A. and Grossmann, C., Implicit Taylor Methods and Parabolic Boundary Control. *Preprint MATH-NM-21-01, TU Dresden*, 2001.

[2] Axelsson, O. *Iterative solution methods.* Cambridge Univ. Press, 1996.

[3] Chawla, M.M. and Al-Zanaidi, M.A., An extended trapezoidal formula for the diffusion equation. *Computers & Math. Appl.* 38 (1999), 51-59.

[4] Chawla, M.M. and Al-Zanaidi, M.A., An extended trapezoidal formula for the diffusion equation in two space dimensions. *Computers & Math. Appl.* 42 (2001), 157-168.

[5] Casas, E. Pontryagin's principle for state-constraint boundary control problems of semilinear parabolic equations. *SIAM J. Control Optim.* 35 (1997), No. 4, 1297-1327.

[6] Corliss, G.F., Griewank, A., Henneberger, P., Kirlinger, G., Potra, F.A. and Stetter, H.J., High-order stiff ODE solvers via automatic differentiation and rational prediction. *Preprint IOKOMO-02-1996, TU Dresden*, 1996.

[7] Ciarlet, P.G. *The finite element method for elliptic problems.* North Holland, Amsterdam, 1978.

[8] Crouzeix, M. and Thomee, V., On the discretization in time of semilinear equations with nonsmooth initial data. *Math. Comput.* 49 (1987), 359-377.

[9] Grossmann, C. and Horváth, Z., Construction of two-sided bounds for initial-boundary value problems. *APNUM* 42 (2002), 177-187.

[10] Grossmann, C., Noack, A. and Vanselow, R., On numerical problems caused by discontinuities in controls. In: Sachs, E.W.; Tichatschke, R. (eds.): *System modeling and optimization XX.* Kluwer, Boston, 2003, 255-270.

[11] Hairer, E., Nørsett, S.P. and Wanner, G., *Solving ordinary differential equations I.* Springer, Berlin, 1993.

[12] Hemker, P.W. and Shishkin, G.I., Approximation of parabolic PDEs with a discontinuous initial condition. *East-West J. Numer. Math.* 1 (1993), 287-302.

[13] Scholz, H.-E. The examination of nonlinear stability and solvability of the algebraic equations for the implicit Taylor series method. *Appl. Numer. Math.* 28 (1998), 439-458.

[14] Thomée, V. *Galerkin finite element methods for parabolic problems.* Springer, Berlin, 1997.

AUTHOR INDEX

M. Lehn, Institute of Practical Mathematics, Universität Karlsruhe (TH), Germany, **318**
M. Lénárd, Kuwait University, Kuwait, **329**
J. Mahmoodi, Qom University, Iran, **27**
R. Mallier University of Western Ontario, Canada **33**
V. D. Mazurov, Institute of Mathematics, Russia, **343**
M. S. S. Mohamed, Universiti Tenaga Nasional, Kajang, **13**
M. Momeni, Tabriz University, Iran, **27**
P. Miškinis, Nordic Institute of Theoretical Physics, Denmark, **357**
S. Naik, Indian Institute of Technology, India, **370**
A. Noack, Institute of Numerical Mathematics, Germany, **458**
A. Ødegard, Simula Research Laboratory, Norway, **92**
M. A. Pathan, Aligarh Muslim University, India, **375**
S. Ponnusamy, Indian Institute of Technology, India, **370**
K. Rost, University of Chemnitz, Germany, **285**
M. Savsar, Kuwait University, Kuwait, **388**
M. Sayed, Kuwait University, Kuwait, **424**
R. Scherer, Institute of Practical Mathematics, Universität Karlsruhe (TH), Germany, **105, 318**
S. Sobhanian, Tabriz University, Iran, **27**
S. V. Tikhonov, Institute of Mathematics of the National Academy of Sciences, Belarus, **438**
B. Wegner, Institut füer Mathematik, Germany, **444**
M. Wójtowicz, Bydgoszcz Academy, Poland, **452**
V. I. Yanchevskiĭ, National Academy of Sciences of Belarus, Belarus, **438**

جامعة الكويت

مؤسسة الكويت للتقدم العلمي

وقائع
المؤتمر الدولي في الرياضيات وتطبيقاتها
(ICMA 2004)

5-7 أبريل 2004

دولة الكويت