گزارش
همایش ماهانهٔ انجمن ریاضی ایران

جلد دوم

به کوشش:

ارسلان شادمان

شابک: ۹۶۴-۹۵۲۶۸-۳-۸
قیمت: ۲۵۰۰ تومان

PREFACE

The aim of this foreword is to give a short presentation of the Monthly Colloquium of the Iranian Mathematical Society (IMS). It will be best understood through a brief history of the IMS activities and the role of conferences therein.

The IMS originated its effective life from a meeting in 1970 and has organized several conferences thereafter. The Annual Mathematics Conference in Iran (AMCI) has undoubtedly the largest evaluating number among them. This is reflected from both scientific and social aspects. Its scientific importance is due to the general talks given by the main invited speakers, the number and the length of its refereed papers, and the fact that the proceedings of the past 34 AMCI have been almost regularly published and are now reviewed in both Mathematical Reviews and MathSci. Its social importance is credited to the number of participants, the enthusiastic panels concerning some local or universal problems in mathematical education or research, the distribution of prizes to winners in its opening ceremonies, the traditional annual session of the General Assembly of the IMS held simultaneously, the announcement of some decisions made by the General Assembly in the closing ceremonies, cultural activities containing visits of a few tourist sites in the country offered to the foreign speakers, folkloric or classical concerts, etc.

These annual conferences remain one of the main concerns of the IMS. However, from the early days of the IMS's life, other needs have motivated the organization of a lot of different seminars and colloquia as well as the publication of periodical and non periodical issues.

During the first decade (up to 1979), a Weekly Seminar has also been held in Tehran at different universities following the speaker's affiliation. From the end of that period, many specialized seminars have been organized by the IMS and held at different universities not necessarily in Tehran. For instance, the first Algebra Seminar was held in Ahwaz (December 31st 1977 - January 1st 1978). Due to a lot of circumstances, the proceedings of this seminar has not been published. There are lots of other annual or biannual well-established seminars in algebra, analysis, geometry and topology, differential equations and dynamical systems, etc. They are mostly publishing their proceedings. A detailed analysis of these conferences and seminars requires a serious study, which is quite interesting, but goes beyond the scope of this preface.

Let me now return back to the philosophy of the Monthly Colloquium as I feel its spirit. The general topics addressed in AMCI presented by main invited speakers to a large audience are of great importance. However, their number are limited by the restrictions naturally imposed by such a program with a big variety in a short period of time (4 days only). The main speeches presented at specialized seminars have also the same shortage of time (2 days only) and, in plus, their participants are almost exclusively scholars involved in the field. By the way, the participants of the AMCI and the specialized seminars are practically the only persons who are receiving the proceedings. But a general talk or a good survey article is of interest to a larger public. Anyone of our graduate students and us needs to obtain information about what is done by others even outside his specialties. So, the IMS has decided to organize a Monthly Colloquium presenting mainly survey articles by research leaders in the country or from abroad. The articles presented for the moment at Tehran are addressed to a reasonable audience, around one hundred graduate students and professors. Ideally, a first draft of the papers will be distributed to the audience the day of the colloquium. Moreover, they will potentially be recorded in videotapes, for the use of universities and research centers outside the capital, etc. Finally, they will be published, in a final form, annually or biannually, when a lot of papers are ready for printing.

During the period 2000-2003, the organizing committee of the Monthly Colloquium that I had the honor to manage has published the first volume containing 8 articles written in Persian, and is publishing now the second volume containing 11 articles in English. These 2 volumes are available from the IMS head office and can be obtained by the members of the society as well as the interested general public.

The third and next volumes will be published by our successors. I am confident that they will continue to fulfill the requirements of the mathematical community more perfectly. The monthly colloquium of the Iranian Mathematical Society is an important step in the propagation of mathematics in our country and it is worth all the efforts done for it.

For the final word, I present my deepest gratitude to the authors who have accepted our invitation and submitted their priceless manuscripts for publication in volume one and this volume. Personally, I have learned a lot from them. The publication of this issue would not have been possible without the endeavors of the secretaries of the IMS whom I thank hereby. Specifically, I would like to thank Mrs. F. Samadian who adapted the format of the Farhang-va-Andishe-ye-Riazi for the first volume and that of the Bulletin of the Iranian Mathematical Society for the present volume.

Arsalan Chademan
Tehran, May 10, 2004

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 1-20 (2004)

Numerical Solution of Integral Equations

E. Babolian

Faculty of Mathematical Sciences and Computer Engineering
Teacher Traning University, Tehran, Iran
babolian@saba.tmu.ac.ir

Abstract: Mathematical models of a wide class of problems are integral equations. Most of integral equations have no analytic solution. In this talk I introduce different kinds of integral and integro-differential equations, their origins and applications. I also explain, in some detail, the numerical methods for solving integral equations.

1. Introduction

In some sense integral equations must be felt to be either more advanced or of less practical interest than differential equations. We turn more

MSC (2000): 45A05, 45B05, 45D05, 45E05, 45J05

Keywords: Integral equations, Fredholm integral equations, Volterra integral equations, Integro-differential equations, Nystrom method, Galerkin method, Expansion method.

readily to a differential formulation of a problem than to an integral formalism. Yet the theory of linear nonsingular integral equations is at least as well developed as that of differential equations and it is in many respects rather simpler. The corresponding operators are *bounded* rather than unbounded, leading to a very straightforward existence theory (the Fredholm theory). There is a much tighter link between the theory and practice of integral equations than is the case for with differential equations. Most of the convergence proofs are constructive in nature and all or nearly all of the constructions have been used as the basis of algorithms for the numerical solution of the underlying equations (although not always with any great success).

For those who require detailed mathematical theory of Integral Equations (I. E.), a number of books are available:

1. S. Smithies, *Integral Equations*, Combridge University Press (1958).
2. R. P. Kanwal, *Linear Integral Equations*, Academic Press (1971).
3. C. D. Green, *Integral Equation Methods,* Nelson (1969).
4. J. A. Cochran, *The Analysis of Linear Integral Equations,* McGraw-Hill (1972).

The books which you may find useful in connection with the numerical solution of integral equations are listed below:

1. S. G. Mikhlin and J. L. Smolitsky, *Approximate Methods for the Solution of Differential and Integral Equations,* Elsevier (1967).
2. L. M. Delves and J. Walsh (eds.), *Numerical Solution of Integral Equations*, Oxford University Press (1974).
3. C. T. H. Baker, *The Numerical Treatment of Integral Equations*, Oxford University Press (1977).
4. C. T. H. Baker and G. F. Miller (eds.), *Treatment of Integral Equations by Numerical Methods,* Academic Press (1982).

5. L. M. Delves and J. L. Mohamed, *Computational Methods for Integral Equations*, Cambridge University Press (1985).

2. Classification

i) *Linear Fredholm Integral Equations of the First and Second Kind, respectively*

$$
\begin{aligned}
y(s) &= \lambda \int_a^b K(s,t)x(t)dt, \\
x(s) &= y(s) + \lambda \int_a^b K(s,t)x(t)dt,
\end{aligned}
$$

where $x(t)$ is the unknown function and $K(s,t)$ is the kernel of the equation.

ii) *Linear Volterra Integral Equations of the First and Second Kind*

$$
\begin{aligned}
y(s) &= \lambda \int_a^s K(s,t)x(t)dt, \\
x(s) &= y(s) + \lambda \int_a^s K(s,t)x(t)dt.
\end{aligned}
$$

iii) *Homogeneous equation, eigenvalue equation,* or a *Fredholm equation of the third kind*

$$x(s) = \lambda \int_a^b K(s,t)x(t)dt$$

those values of λ for which nontrivial solutions exist are called *characteristic values* of the equation.

iv) *Nonlinear Integral Equations*

In (i) to (iii) if you replace $K(s,t)$ with $K(s,t,x(s),x(t))$ you obtain a nonlinear integral equation.

v) Integral equation involving unknown functions $x(s_1, s_2, \ldots, s_m)$ also occur regularly in practice but lie outside the scope of this talk!

vi) *Integro-differential equations*

$$Q(s)x'(s) + R(s)x(s) = y(s) + \lambda \int_a^b K(s,t)x(t)dt, \quad x(a) = x_0.$$

3. Origin of Integral Equations

i) Consider the following initial value problem

$$\frac{dx(s)}{ds} = f(s,x), \quad x(0) = x_0,$$

under suitable continuity conditions on $f(s,x)$, we have:

$$x(s) = x_0 + \int_0^s f(t, x(t))dt,$$

which is an integral equation for $x(s)$.

ii) Consider the second order differential equation

$$\begin{aligned} \frac{d^2x}{ds^2} &= f(s,x) \\ x(0) &= x_0, \frac{d(x(0))}{ds} = x_1, \end{aligned}$$

this equation converts to

$$x(s) = x_0 + x_1 s + \int_0^s (s-t)f(t, x(t))dt$$

which is an integral equation for $x(s)$.

iii) Two-point boundary value problems may be converted to an integral equation. Consider the problem

$$\begin{cases} \frac{d^2x}{ds^2} = f(s,x), \\ x(0) = \alpha, x(l) = \beta, \end{cases}$$

this problem converts to

$$x(s) = y(s) - \int_0^l K(s,t)f(t,x(t))dt,$$

where

$$y(s) = \alpha + \frac{\beta - \alpha}{l}s,$$
$$K(s,t) = \begin{cases} \frac{t(l-s)}{l}, & 0 \leq t \leq s, \\ \frac{s(l-t)}{l}, & s \leq t \leq l. \end{cases}$$

4. Applications

We take a random selection of research papers in a variety of fields.

i) *Currents in a superconducting strip* (Rhoderick and Wilson, 1962 [18])

$$y(s) = \frac{1}{\pi}\int_0^1 \frac{t-s}{(t-s)^2 + a^2}x(t)dt.$$

ii) *Flow round a hydrofil* (Kershaw, 1974 [15])

$$x(s) = y(s) + \lambda \int_C K(s,t)x(t)dt,$$

subject to: $\int_C x(t)dt = 0, C$ a closed contour.

iii) *Population competition* (Dounham and Shah, 1976 [12])

$$\mathbf{x}(s) = \mathbf{y}(s)\int_a^b \mathbf{k}(s,t,\mathbf{x}(s,t))\mathbf{x}(t)dt, \mathbf{x}^T = (x_1, x_2, x_3).$$

iv) *Quantum scattering: closed coupled calculations* (Horn and Frazer, 1975 [14])

$$x(s) = y(s) - \lambda\int_0^\infty K_1(s,z)\int_0^\infty K_2(z,t)x(t)dtdz.$$

5. Existence of Solution to Integral Equations

Smithies [19], defined the concept of *relatively uniformly convergence* of sequences of functions and kernels.

Definition. Suppose $\{f_n(s)\}$ be a sequence of $\mathcal{L}^2$ functions. We say that this sequence converges relatively uniformly (c.r.u.) to $f(s)$ if there exists a non-negative $\mathcal{L}^2$ function $P(s)$ such that

$$\forall\epsilon\exists N(\epsilon)\forall n\forall s \ \ (n \geq N(\epsilon) \Longrightarrow |f_n(s) - f(s)| \leq \epsilon P(s)).$$

It can be shown that if $\{f_n(s)\}$ (c.r.u.) to $f(s)$ then it converges pointwise, but not necessarily uniformly. Absolutely relatively uniformly convergence of sequences and series of functions and kernels can be defined similarly.

It is proved that if for a fixed λ there exists kernel $H_\lambda(s,t) \in \mathcal{L}^2[a,b]\times[a,b]$ such that

$$K(s,t) - H_\lambda(s,t) = \lambda H_k K = \lambda K H_\lambda,$$

then $x = y + \lambda K H_\lambda$ is a solution of the equation

$$x = y + \lambda K x.$$

It is also shown that if $\|\lambda K\| < 1$ then the series

$$K(s,t) + \lambda K^2(s,t) + \lambda^2 K^3(s,t) + \ldots,$$

known as Neumann series, is absolutely relatively uniformly convergent to $H_\lambda(s,t)$. (For a detailed discussion of the above concepts see Smithies [19]).

Values of λ for which $H_\lambda(s,t)$ exist are called *regular values of the kernel* K. It is shown that the set of regular values of K is open.

6. Numerical Solution of Fredholm Integral Equations

6.1 Iterative method and the Neumann series

The Fredholm I. E. of the second kind

$$x(s) = y(s) + \lambda \int_a^b K(s,t)x(t)dt \tag{1}$$

can be written in operator form as

$$x = y + \lambda K x.$$

The Neumann series is

$$x = \sum_{i=0}^{\infty} \lambda^i K^i y, \tag{2}$$

and in truncated form

$$x \simeq x_n = \sum_{i=0}^{n} \lambda^i K^i y.$$

If the series (2) converges, then $\lim_{n\to\infty} \|x - x_n\| = 0$. The approximate solution x_n is most easily produced iteratively via the obvious recurrence relation:

$$\begin{aligned} x_{n+1} &= y + \lambda K x_n \\ x_0 &= y, \end{aligned}$$

provided that the function $Kx_n = \int_a^b K(s,t)x_n(t)dt$ can be computed.

6.2 The Nyström (or quadrature) method

In equation (1) if you use a quadrature rule you find

$$\int_a^b K(s,t)x(t)dt \simeq \sum_{j=1}^{n} w_j K(s,t_j)x(t_j),$$

where w_j and t_j are weights and nodes of the quadrature rule. So (1) converts to (approximately)

$$x(s) = y(s) + \lambda \sum_{j=1}^{n} w_j K(s, t_j) x(t_j), \tag{3}$$

if you put $s = t_i$, $i = 1, \ldots, n$, in (3) you get the following system of linear equations for the unknowns $x(t_1), \ldots, x(t_n)$,

$$x(t_i) = y(t_i) + \lambda \sum_{j=1}^{n} w_j K(t_i, t_j) x(t_j), \quad i = 1, 2, \ldots, n. \tag{4}$$

If you define $\mathbf{x}^t = (x(t_1), \ldots, x(t_n))$, $\mathbf{y}^t = (y(t_1), \ldots, y(t_n))$ and $\mathbf{k}_{ij} = w_j K(t_i, t_j)$ the system (4) becomes

$$(I - \lambda \mathbf{k})\mathbf{x} = \mathbf{y}, \tag{5}$$

this equation has a solution if $I - \lambda\mathbf{k}$ is not singular, and this may happen even if $\|\lambda\mathbf{k}\| > 1$. In general the matrix $I - \lambda\mathbf{k}$ is well-conditioned and the system (5) can be solved easily by direct methods.

6.3 Expansion method for Fredholm Equations

6.3.1 *Methods based on an expansion of the solution of Fredholm I. E. of the second kind*

Suppose that the sequence of functions $\{h_n(s)\}$ is orthogonal and complete in $\mathcal{L}^2(a, b)$ and

$$x(s) = \sum_{j=1}^{\infty} b_j h_j(s),$$

then for sufficiently large N, we may approximate x as closely as we please by x_N:

$$x(s) \simeq x_N(s) = \sum_{j=1}^{N} a_j h_j(s).$$

An *expansion method* is then an algorithm for determining the a_i for either an arbitrary or a specified choice of the sequence $\{h_n(s)\}$. There are many possible algorithms and we consider only the most common.

1. Residual minimisation methods Assuming $L = I - \lambda K$, the equation (1) can be written as

$$Lx = y.$$

Now if $\epsilon_N = x - x_N$, $r_N = y - Lx_N$ then $r_N = L\epsilon_N$ and computing r_N requires no knowledge of x. We now choose the vector $\mathbf{a} = (a_1, \ldots, a_n)^t$ from the minimisation criterion: $\min_{\mathbf{a}} \|r_N\|$.

It can be shown that

$$\frac{\|r_N\|}{1 + \|K\|} \leq \|\epsilon_N\| \leq \frac{\|r_N\|}{1 - \|K\|}, \quad \text{provided that} \quad \|K\| < 1.$$

There are many norms avialable, the choice of norm is influenced by the analytical properties of the kernel K.

Chebyshev norm. For continuous kernels a possible choice is the Chebyshev (L_∞) norm:

$$\|x\| = \max_{a \leq x \leq b} |x(s)|, \quad \|K\| = \max_{a \leq s \leq b} \int_a^b |K(s,t)|dt.$$

We seek to compute

$$\|r_N\| = \min_{\mathbf{a}} \max_{a \leq s \leq b} |y(s) - x_N(s) + \int_a^b K(s,t)x_N(t)dt|,$$

with $x_N = \sum_{i=1}^N a_i h_i(s)$, setting $k_i(s) = \int_a^b K(s,t)h_i(t)dt$,

$$\|r_N\| = \min_{\mathbf{a}} \max_{s} |y(s) - \sum_{i=1}^{n} a_i(h_i(s) - k_i(s))|,$$

usually $k_i(s)$ can be estimated by numerical quadrature (for fixed s) and to search, not over $[a,b]$, but over a discrete point set $\{\xi_i, i = 1, \ldots, q\}$, i.e.

$$\|r_N\| \simeq \min_{\mathbf{a}} \max_{k=1}^{q} |y(\xi_k) - \sum_{i=1}^{n}(h_i(\xi_k) - k_i(\xi_k))|.$$

L_2 norm: Least squares approximation

Alternatively, we may choose to minimise $\|y - Lx\|_2$:

$$\min_{\mathbf{a}} \int_a^b \left| \left[x_N(s) - y(s) - \int_a^b K(s,t) x_N(t) dt \right] \right|^2 ds,$$

whence on inserting x_N we find the system

$$\mathbf{L}_{ls}\mathbf{a} = \mathbf{y}_{ls}$$

where

$$\begin{aligned} (\mathbf{L}_{ls})_{ij} &= \int_a^b (Lh_i)^*(s) h_j(s) ds, \quad i,j = 1,\ldots,N \\ (\mathbf{y}_{ls})_i &= \int_a^b y(s)(Lh_i)^*(s) ds, \quad i = 1,\ldots,N \\ Lh_i(s) &= h_i(s) - \int_a^b K(s,t) h_i(t) dt. \end{aligned}$$

In this norm we have to evaluate N^2 apparently triple integrals of the form

$$\int_a^b [\int_a^b K^*(s,t) h_i^*(t) dt][\int_a^b K(s,t') h_j(t) dt'] ds.$$

2. Galerkin method. Our aim is to make $r(s)$ zero. Now if the set $\{h_i\}$ is complete and orthonormal in $\mathcal{L}^2(a,b)$, the statement $r(s) = 0$ is equivalent to

$$(r(s), h_i(s)) = 0, \quad i = 1, 2, \ldots$$

Now with only N parameters $a_i, i = 1, \ldots, N$, at our disposal the best we can do is to make the residual $r_N(s)$ orthogonal to the first N functions $h_1(s), \ldots, h_N(s)$, i.e.

$$\int_a^b h_i^*(s)[y(s) - (Lx_N)(s)] ds = 0, \quad i = 1, \ldots, N.$$

this leads to the system $\mathbf{L}_G\mathbf{a} = \mathbf{y}_G$ where now

$$(\mathbf{L}_G)_{ij} = \int_a^b h_i^*(s)(Lh_j(s))ds, \quad i,j = 1,\ldots,N$$
$$(\mathbf{y}_G)_i = \int_a^b h_i^*(s)y(s)ds, \quad i = 1,\ldots,N.$$

3. The Fast Galerkin algorithm Delves [10,11] and his Ph.D. students used Chebyshev sequence of polynomials $\{T_i(s)\}_{i=0}$ and set $x(s) \simeq \sum_{i=0}^N a_i T_i(s)$ and for

$$x(s) = y(s) + \int_{-1}^1 K(s,t)x(t)dt$$

obtained

$$(\bar{\mathbf{D}} - \bar{\mathbf{B}})\mathbf{a} = \bar{\mathbf{y}}$$

where

$$\bar{B}_{ij} = \int_{-1}^1 \frac{T_i(s)}{(1-s^2)^{\frac{1}{2}}}\left(\int_{-1}^1 K(s,t)T_j(t)dt\right)ds, \quad i,j = 0,1,\ldots,N$$
$$\bar{y}_i = \int_{-1}^1 \frac{T_i(s)}{(1-s^2)^{\frac{1}{2}}}y(s)ds, \quad i = 0,1,\ldots,N.$$

They used Fast Fourier Transformation (FFT) and evaluated approximate values B_{ij} and y_i for $\bar{B}_{ij}$ and $\bar{y}_i$ in $\mathcal{O}(N^2 \log N)$ operations. Another advantage of this method is that

$$|\bar{y}_i| \le c\hat{i}^{-p}, \hat{i} = \max\{1,i\}, \quad i = 0,1,\ldots$$

where p depends on the smoothness of the function $y(s)$. Similarly they showed that $|\bar{B}_{ij}| \le c''\hat{i}^{-(r+1)}\hat{j}^{-2}$ if $|k_{ij}| \le c'\hat{i}^{-(r+1)}\hat{j}^{-(q+1)}$ where $K(s,t) = \sum_{i=0}^{\prime\infty}\sum_{j=0}^{\prime\infty} k_{ij}T_i(s)T_j(t)$, again r, q depend on the countinuity properties of $K(s,t)$ as functions of s and t, respectively. They introduced *asymptotically lower diagonal* matrix M of type A and degree p as

$$\frac{|M_{ij}|}{|M_{ii}||M_{jj}|} \le ci^{-p}, \quad i > j, c > 0, p \ge 0,$$

and showed that the matrix $I - \lambda \bar{B}$ is asymptotically lower diagonal of type A and degree r [13]. This was the basis of an iterative method for solving $\mathbf{L}_G\mathbf{a} = \mathbf{y}_G$ with $\mathcal{O}(N^2)$ operations. Some authors have used splines or wavelet functions to solve Fredholm I. Es. of the second kind (Razzaghi and Ordokhani [17]).

7. Fredholm Integral Equations of the First Kind

All of the numerical methods discussed for solving second kind equations apply formally also to first kind equation

$$y(s) = \int_a^b K(s,t)x(t)dt, \tag{6}$$

with the single exeption of the Neumann iterations.

Fredholm integral equations of the first kind are among ill-posed problems, i.e.

i) They may have no solution;

ii) They may have not a unique solution;

iii) The solution does not depend on $y(s)$ continuously, i.e. small changes on $y(s)$ may cause dramatic changes on the solution $x(s)$.

For example the equation $\int_0^{2\pi} \cos s \sin t x(t)dt = e^s$ has no solution. If x is a solution of $Kx = y$ and $\phi \neq 0$ is such that $K\phi = 0$ then $x + \alpha\phi$ is also a solution of $Kx = y$ for arbitrary $\alpha \in \mathbb{C}$.

7.1 Eigenfunction expansion

The first method we consider (Baker *et al.*, 1964; Turchin, Kozlov and Malkevich, 1971; Hanson, 1971 [9]) is a direct extension of the numerical method for finite rank kernels.

Theorem. Let K be of finite rank n and have the representation:

$$K = \sum_{v=1}^{n} a_\nu \oplus b_\nu$$

then (6) has no solution unless $y \in \text{span}\{a_1, \ldots, a_n\}$, i.e.

$$\exists a_1, \ldots, a_n : y(s) = \sum_{v=1}^{n} \alpha_\nu a_\nu .$$

In this case (6) has the solution

$$x = \sum_{v=1}^{n} \beta_\nu a_\nu ,$$

if and only if the following system of linear equations is nonsingular

$$\sum_{v=1}^{n} (b_\mu, a_\nu)\beta_\nu = a_\mu , \quad \mu = 1, \ldots, n.$$

Hermitian kernels

If K is Hermitian, i.e. $\overline{K(t,s)} = K^*(s,t) = K(s,t)$, it has real eigenvalues λ_i, with corresponding eigenfunctions v_i, satisfying the equation $\int_a^b K(s,t)v_i(t)dt = \lambda_i v_i(s)$. If the $\{v_i\}$ form a complete set, any function, and in particular the solution x and driving term y, have expansions of the form

$$x = \sum_{i=1}^{\infty} \beta_i v_i , \quad y = \sum_{i=1}^{\infty} \alpha_i v_i$$

giving $Kx = \sum_{i=1}^{\infty} \beta_i K v_i = \sum_{i=1}^{\infty} \beta_i \lambda_i v_i = y = \sum_{i=1}^{\infty} \alpha_i v_i$. So, $\beta_i = \alpha_i / \lambda_i$ and $x = \sum_{i=1}^{\infty} (v_i, y)\lambda_i^{-1} v_i$. [$v_i$, $i = 1, 2, \ldots$ are assumed to be orthonormal]. The eigenvalues λ_i tend to zero as i increases, giving the sum

$$\sum_{i=1}^{\infty} (v_i, y)\lambda_i^{-1} v_i , \tag{7}$$

a dangerous look. Now if $\lim_{i\to\infty}(v_i, y)\lambda_i^{-1} \neq 0$ the sum (7) does not represent an $\mathcal{L}^2$ function; that is the equation has no solution, a numerical test for consistency. However, even if $(v_i, y)\lambda_i^{-1} \underset{i\to\infty}{\longrightarrow} 0$, the series (7) is very unstable. Suppose we replace y by $y + \epsilon v_k$. Then the solution x changes to $x + \epsilon\lambda_k^{-1} v_k$ and the "response ratio", defined as $\|\delta x\|/\|\delta y\|$ is $|\lambda_k^{-1}|$. Since $\lim_{k\to\infty}|\lambda_k| = 0$, this response ratio can be made arbitrary large, showing that the solution x does not depend continuously on y.

Non-Hermitian kernels

If K is not Hermitian K^*K and KK^* are Hermitian. Singular values μ_i and singular functions $\{u_i, v_i\}$ satisfy the relations

$$u_i = \mu_i K v_i, \quad v_i = \mu_i K^* u_i, \quad (u_i, u_j) = (v_i, v_j) = \delta_{ij}.$$

Further, $\lim_{i\to\infty}\mu_i^{-1} = 0$.

Now suppose we make the expansions

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

Then it follows that $\beta_i = \mu_i(u_i, y)$ and

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

The method of regularisation

In many practical applications the driving term $y(s)$ represents some measured function, the integral operator K represents a model of the instrument used to do the measuring and $x(s)$ represents the "true"measured quantity, $y(s)$ being a "smeared"version of $x(s)$ as seen through the instrument being used. Now $y(s)$ will be known to some finite accuracy

ϵ; we should therefore only expect to find $||Kx - y|| \leq \epsilon$. Of all the functions x satisfying this relation, we seek that which is the smoothest, in the sence that for some linear operator L, $||Lx||$ has the minimum value. This yields the constrained minimisation problem

$$\text{minimise} \quad ||Lx||$$

$$\text{subject to}$$

$$||Kx - y|| \leq \epsilon. \tag{8}$$

This problem can be solved directly, in any norm. It is however relatively complicated to do so. It is therefore usual to solve, not (8), but a related problem, which we develop as we note that the minimum value of $||Lx||$ in (8) will decrease as ϵ increases, that is as the constrains weaken. Therefore, at the minimum of (8), the constrains will be binding, that is $||Kx - y|| = \epsilon$. Now, if we solve the unconstrained problem (for fixed α).

$$\text{minimise}_x \quad ||Kx - y||^2 + \alpha||Lx||^2, \tag{9}$$

we will find the minimum some value η for $||Kx - y||$. As $\alpha \longrightarrow 0$, $\eta \longrightarrow 0$ provided that a solution of (6) exists, and for some value of $\alpha, \eta = \epsilon$. This makes plausible the following results, due to Tikhonov [20,21].

For some α (which depends on ϵ) the solution of problem (9) is identical to that of problem (8). Problem (9) is easier to solve and it is refered to as the *regularised problem.* The choice of operator L must be made on qualitative grounds, the usual choices are: $L = I$ or $\frac{d}{as}$ or $\frac{d^2}{ds^2}$. When $L = I$ it can be shown that the solution to (9) is the solution of the following Fredholm I. E. of the second kind,

$$\int_a^b \hat{K}(s,t)x(t)dt + \alpha x(s) = \hat{y}(s) \tag{10}$$

where

$$\hat{K}(s,t) = \int_a^b K^*(\xi,s)K(\xi,t)d\xi, \quad \hat{y}(s) = \int_a^b K^*(t,s)y(t)dt.$$

Unfortunately the value of α for which $\eta = \epsilon$ is very small and the solution of (10) is very sensitive to α. This makes this kind of regularisation very expensive and impractical.

Regularisation method was also used by Phillips [16] and Twomey [22], they used $L = \frac{d^2}{ds^2}$.

The augmented Galerkin method

We now turn to the use of expansion methods for (6), and recall that (6) converts to

$$\bar{B}\mathbf{a} = \bar{\mathbf{y}} \tag{11}$$

where $\bar{B}$ is very ill-conditioned, and even singular when K is of finite rank. Recall that we assumed

$$x(s) = \sum_{i=0}^{\infty} b_i h_i(s)$$

and for $x(s) \in \mathcal{L}^2(a,b)$ we should have

$$\sum_{i=0}^{\infty} |b_i|^2 < \infty.$$

Hence we may assume that there exist constants $c > 0$, $r \geq \frac{1}{2}$ such that

$$|b_i| \leq C_b \hat{i}^{-r} = \delta_i, \quad i = 0, 1, \ldots,$$

again r depends on the smoothness of the solution $x(s)$.

So, instead of solving the ill-conditioned system (11) we solve the well-conditioned problem:

$$\begin{array}{ll} \text{minimise} & \|\bar{B}\mathbf{a} - \bar{\mathbf{y}}\| \\ \text{subject to:} & |a_i| \leq \delta_i, i = 0, 1, \ldots, N, \end{array} \tag{12}$$

where $\delta_i = C_f \hat{i}^{-r}$.

The problem is now to estimate C_f, r. There are a number of *heuristic* and practical ways to set C_f, r (see Babolian & Delves (1979) [8] and Babolian (1980) [7]). The solution of (12) is not sensitive to the values of C_f, r, but r should not be set very large unless x is very smooth.

Problem (12) has been solved in $\|\cdot\|_\infty$ and $\|\cdot\|_2$ (Belward (1982)). In 1997, Abbasbandy & Babolian [1,2] proposed some fully automatic algorithms to set C_f and r. Then some papers were published using *automatic augmented Galerkin algorithms* [3,4]. Some authors have used direct regularisation, by multiplying both sides of $Kx = y$ by a function $Q(s,t)$ (Maleknejad & Rostami) or used preconditioning to reduce the condition number of the resulting system of equations. But none of them have the general success of *automatic augmented Galerkin algorithms.*

8. Integro-differential Equations

Integro-differential equations are also considered using expansion methods with complete error analysis (see Babolian & Delves (1981) [5]).

References

[1] Abbasbandy, S. & Babolian, E. (1996) *Automatic Augmented Galerkin Algorithms for Hammerstein First Kind Integral Equations,* Journal of Science, Teacher Training University, 7, Nos. 1,2.

[2] Abbasbandy, S. & Babolian, E. (1997), *Automatic Augmented Galerkin Algorithms for Fredholm Integral Equations of the First Kind*, ACTA Math. Sci. (English Ed.) 17, 1, 69-84.

[3] Abbasbandy, S. & Babolian, E. (1997) *An Automatic Augmented Galerkin Method for Singular Integral Equations with Hilbert Kernel*, Scientia Iranica, **4**, Nos. 1,2, 60-64.

[4] Abbasbandy, S. & Babolian, E. (1995) *Automatic Augmented Galerkin Algorithms for Linear First Kind Integral Equations: Non-Singular and Weak Singular Kernels*, Bulletin of the Iranian Mathematical Society, **21**, No. 1.

[5] Babolian, E. & Delves, L. M. (1981) *A Fast Galerkin Scheme for Linear Integro-differential Equations*, IMA Journal of Numerical Analysis.

[6] Babolian, E. & Delves, L. M. (1989) *Parallel Solution of Fredholm Integral Equations.* Journal of Parallel Computing, **12**, 95-106.

[7] Babolian, E. (1980), *Galerkin Methods for Integral and Integro-Differential Equations.* Ph. D. Thesis, Liverpool University.

[8] Babolian, E. & Delves, L. M. (1979) *An Augmented Galerkin Method for First Kind Fredholm Equations,* J. Inst. Maths. Applics. **24**, 157-174.

[9] Delves, L. M., & Walsh, J. (1974) *Numerical Solution of Integral Equations*, Oxford University Press.

[10] Delves, L. M. (1977a), *A Fast Method for the Solution of Fredholm Integral Equations*, J. Inst. Maths. Applics. **20**, 173-182.

[11] Delves, L. M. (1977b), *On the Solution of the Sets of Linear Equations Arising From Galerkin Methods*, J. Inst. Maths. Applics. **20**, 163-171.

[12] Downham, D. Y. & Shah, S. M. M. (1976) *The Integral equation approach for models of clines.* Preprint, Liverpool University.

[13] Freeman, T. L. and Delves, L. M. (1974), *On the Convergence Rates of Variational Methods III, Unsymmetric Systems,* J. Inst. Maths. Applics. **14**, 311-323.

[14] Horn, S. & Fraser, P. A. (1975), *Low-energy ortho-positronium scattering by hydrogen atoms.* J. Phys. B, **8**, 2472-5.

[15] Kershaw, D. (1974), *Singular integral and boundary value problems,* In [9].

[16] Phillips, D. L. (1962), *A Technique for the Numerical Solution of Certain Integral Equations of the First Kind.* J. Ass. Comput. Mach. **9**, 84-97.

[17] Razzaghi, M. & Ordokhani, Y. (2001), *Solution of Differential Equations via Rationalized Haar Functions.* Journal of Mathematics and Computers in Simulation, **56**, 235-246.

[18] Rhoderick, E. H. & Wilson, E. M. (1962), *Current distribution in thin super conducting films,* Nature, **194**, 1167-9.

[19] Smithies, *Integral Equations,* (1958), Cambridge University Press.

[20] Tikhonov, A. N. (1963a), *On the Solution of Incorrectly Posed Problems and the Method of Regularisation,* Soviet Math. **4**, 1624-1627.

[21] Tikhonov, A. N. (1963b), *Regularisation of Incorrectly Posed Problems,* Soviet Math. **4**, 236-247.

[22] Twomey, S. (1963), *On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature,* J. Ass. Comput. Mach. **10**, 97-101.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 21-38 (2004)

Factorizations of Finite Groups

M. R. Darafsheh

Department of Mathematics and Computer Science
Faculty of Science, University of Tehran, Tehran, Iran
Daraf@khayam.ut.ac.ir

Abstract: Let G be a finite group and A, B proper subgroups of G. If $G = AB$, then we say that G is a factorizable group and A, B are called factors of this factorization. In this case G is also called the product of two subgroups A and B. The problem of which finite groups are factorizable is still an open probelm. Certain conditions on the structure of G or the factors of the factorization yields some information about G. In this paper we will give a survey of the results obtained so far on the product of finite groups.

1. Introduction

Let G be a group and A, B subgroups of G. If $G = AB = \{ab | a \in A, b \in B\}$, then we say that G is a factorizable group and A, B are called

MSC (2000): 20D40

Keywords: Factorization, Simple group, Product of groups

factors of the factorization. We also say that G is the product of its subgroups A and B. We always have $G = AG$ which is called the trivial factorization of G. Therefore we call a factorization $G = AB$ proper or non-trivial if both factors are proper subgroups of G. If $G \cong A \times B$ is the external direct product of two non-trivial groups A and B, then $G \cong \bar{A}\bar{B}$, where $\bar{A} \cong A$ and $\bar{B} \cong B$ and so G has a proper factorization. Also factorizations of groups as product of 3 or more subgroups may be of interest, but it is not considered here. In the book [AFD], page 13, the authors ask the following question:

Question: *Describe all groups that have a proper factorization.*

The above question is an open problem in group theory and it seems to be a difficult problem in general. Not every group has a factorization, for example if G is an infinite group with all proper subgroups finite, then G is not factorizable. And if G is the cyclic group of prime order or one of the sporadic finite simple groups:

$$M_{22}, J_1, J_3, J_4, M^cL, L_y, O'N, Co_2, Co_3, Fi_{23}, Fi'_{24}, HN, TH, BM \text{ or } M$$

then G does not have a proper factorization. The other 11 sporadic groups have proper factorization, for example:

$M_{11} = L_2(11)M_{10}$, $M_{12} = M_{11}M_{11}$, $M_{23} = F_{23}^{11}M_{22}$, $M_{24} = M_{23}(M_{12}.2)$, $J_2 = U_3(3)(A_5 \times D_{10})$, $HS = M_{22}(U_3(5).2)$, $He = (SP_4(2).2)(7^2.SL_2(7))$, $Ru = L_2(29).^2F_4(2)$, $Suz = G_2(4).U_5(2)$, $Fi_{22} =^2 F_4(2)'.(2.U_6(2))$, $Co_1 = Co_2.(3.Suz.2)$.

In the above F_{23}^{11} denotes a Frobenius group with kernel isomorphic to Z_{23} and complement isomorphic to Z_{11}.If L is an exceptional group of Lie type except $G_2(q), q = 3^n, F_4(q), q = 2^n$ or $G_2(4)$, then L does not have a factorization. The simple unitary groups $U_{2m+1}(q)$ don't have a factorization except when $(m, q) = (1, 3), (1, 5)$ or $(4, 2)$.

To see why the sporadic simple group J_1 is not a factorizable group we recall from [CCNPW] that J_1 has order $2^3.3.5.7.11.19$ with maximal subgroups of the form:

$$L_2(11),\ 2^3 : F_7^3,\ Z_2 \times A_5,\ F_{19}^6,\ F_{11}^{10},\ D_6 \times D_{10},\ F_7^6.$$

If $J_1 = AB$ is a proper factorization of J_1, then A is contained in a maximal subgroup M of J_1, therefore $J_1 = MB$. We will consider the 7 possible cases for M. If $M \cong L_2(11), 2^3 : F_7^3, Z_2 \times A_5, F_{11}^{10}, D_6 \times D_{10}$ or F_7^6, then we have $19 \times 7 || B|$. But then B must be contained in a maximal subgroup of J_1 with order divisible by 19×7, which is not the case. If $M \cong F_{19}^6$, then $7 \times 11 || B|$ and from the list of maximal subgroups of J_1 we see that J_1 does not have a maximal subgroup with order divisible by 7×11. This final contradiction shows that J_1 is not a factorizable group.

Looking at that factorization of finite groups, the problem goes back to Burnside [Bu] who proved any finite group G of order $p^a q^b$ is solvable. In the case if p and q are distirct primes and P and Q are p-Sylow and q-Sylow subgroups of G respectively, then $G = PQ$. N.Ito [It1] proved that if $G = AB$ and if A and B are abelian subgroups of G, then G is metabelian. H. Wielandt [Wi] generalized the result of Ito and proved that if $G = AB$ is a finite group, where A and B are nilpotent subgroups of G, and $(|A|, |B|) = 1$, then G is solvable. O.Kegel in [Ke] generalized the result obtained by Wielandt without the assumption $(|A|, |B|) = 1$. Chernikov in [Ch] without the condition of finiteness on G proved that if $G = AB$, A and B nilpotent subgroups of G satisfying the minimum condition on groups, then G is a solvable group.Ore in [Or] proved that if G is a solvable group and if A and B are two non-conjugate maximal subgroups of G, then $G = AB$ and conversely any maximal factorization of G arises in this way. We recall that a factorization $G = AB$ of a group G is called maximal if both A and B are maximal subgroups of G.

Although there are interesting problems concerning factorizations of infinite groups, but here we are concerned with finite groups. Therefore from now on all groups are assumed to be finite although the general results stated may be true for infinite groups as well. Our aim in this paper is to review results on factorizable finite groups and to state our recent results concerning factorizations $G = AB$, where one of the factors is the alternating group.

2. Preliminary Results.

To start with, first we will mention some elementary results.

Theorem 2.1. *Let G be a finite group with subgroups A and B, then the following statements are equivalent:*

(i) $G = AB$,

(ii) A acts transitively on the set of right cosets of B in G,

(iii) B acts transitively on the set of right cosets of A in G,

(iv) If π_1 and π_2 are the permutation characters of G obtained from (ii) and (iii) respectively, then $(\pi_1, \pi_2) = 1$.

Theorem 2.2. *If G acts transitively on a set Ω and H is a transitive subgroup of G, then for any $\alpha \in \Omega$ we have $G = G_\alpha H$. In general if G acts k-homogeneously on a set Ω and H is a k-homogeneous subgroup of G, then $G = HG_{(\Delta)}$ where Δ is a subset of size k in Ω.*

To see how theorem 2.2 can be used to obtain factorization of some groups, we mention that any finite group G is a factor of some symmetric group. Because, if G is a group of order m, then we can assume G as a transitive subgroup of S_n and therefore by Theorem 2.2 we get $S_n = S_{n-1}G$.

If we consider the simple group $L_2(7)$ of order $2^3.3.7$, then it is easy to see that $L_2(7)$ has subgroups of orders $1, 2, 4, 8, 3, 6, 12, 24, 7, 21, 168$. Therefore $L_2(7)$ has faithful transitive actions on sets of cardinalities: $168, 84, 42, 21, 56, 28, 14, 7, 24, 8$. Now consider the transitive action of

$L_2(7)$ on 24 letters. Since the sporadic group M_{24} has a subgroup isomorphic to $L_2(7)$, hence by Theorem 2.2 we will get $M_{24} = M_{23}.L_2(7)$. Now we consider the action of $L_2(7)$ on 8 points. Since A_8 has a subgroup isomorphic to $L_2(7)$ which is 2-transitive on 8 points, therefore by the first and the second parts of Theorem 2.2 we have the following factorizations of the alternating group $A_8, A_8 = A_7L_2(7) = A_6L_2(7) = S_6L_2(7)$.

Using part (iv) of Theorem 2.1 and information given in [CCNPW] about permutation characters of certain groups on maximal factorizations of sporadic groups, for instance: $J_2 = U_3(3).(A_5 \times D_{10}), HS = M_{22}(U_3(5)2)$.

M.W.Liebeck, C.E.Praeger and J.Saxl in [LPS] determined completely the maximal factorizations of all the finite simple groups and their antomorphism groups. As we mentioned earlier the maximal factorizations of solvable groups were obtained by Ore in [Or]. The main Theorem of [LPS] is the following.

Theorem 2.3. ([LPS]) *Let L be a finite simple group and let G be a group such that $L \trianglelefteq G \trianglelefteq Aut(L)$. Suppose $G = AB$, where A and B are maximal subgroups of G not containing L. Then the triple (G, A, B) is explicitly known.*

The following two results are also obtained about the factorizations of the alternating and symmetric groups.

Theorem 2.4. ([LPS]). *Let $L = A_n (n \geq 5)$ acting naturally on a set Ω of n points, and let $L \trianglelefteq G \leq Aut(L)$. Suppose that $G = AB$ where A and B are arbitrary subgroups of G not containing L. Then one of the following holds:*

(i) $A_{n-k} \trianglelefteq A \leq S_{n-k} \times S_k$ for some k with $1 \leq k \leq 5$, and B is k-homogeneous on Ω.

(ii) $n = 6, 8, 10$. If $n = 6$ the groups A and B have the following property that $A \cap L = L_2(5), B \cap L = S_3 \wr S_2$. If $n = 8$, then $A =$

$2^3.L_3(2)$ and $B \geq Z_5 \times Z_3$. If $n = 10$, then $A = L_2(8)$ or $L_2(8).3$ and $A_5 \times A_5 \trianglelefteq B \leq S_5 \wr S_2$ and B is transitive on Ω.

Of course in the above theorem the roles of A and B may be interchanged.

Corollary 2.5. ([LPS]) *Let G be A_n or $S_n (n > 5)$ and suppose that $G = AB$ with A and B maximal subgroups of G not containing A_n. Then one of the following holds.*

(i) $A = (S_{n-k} \times S_k) \cap G, 1 \leq k \leq 5$, and B is k-homogeneous.

(ii) $n = 6, A = PGL_2(5) \cap B, B = (S_3 \wr S_2) \cap G$.

If $k \geq 2$, then combining the results of W.M.Kantor [Ka] and P.J. Cameron [Ca] it is possible to find the group B mentioned in Theorem 2.4 and Corollary 2.5. If $k = 1$, then B is a transitive subgroup of S_n or A_n and in general it is not difficult to find B in this case.

Factorizations of the symmetric and alternating groups are considered in [WW] as well. We call a factorization $G = AB$ of G exact if $A \cap B = 1$. In [WW] all the exact factorizations of the alternating and symmetric groups are found.

3. General result on factorizations of group.

Much of recent research is concerned with factorization of finite simple groups, or factorizations involving simple groups. The so called Szep conjecture was proved by E.Fisman and Z.Arad in [FA]. The conjecture states: No simple group has a factorization $G = AB$ with $Z(A) \neq 1, Z(B) \neq 1$.

Z.Arad and E.Fisman in [AF] considered factorizations of simple groups $G = AB$ such that $(|A|, |B|) = 1$. They proved:

Theorem 3.1. ([AF]) *Let G be a simple group and $G = AB$ where A and B are subgroups of G with $(|A|, |B|) = 1$. Then one of the followings* *holds:*

(i) $G = A_r$ *with* $r \geq 5$ *a prime, and* $A \cong A_{r-1}, |B| = r$.

(ii) $G = M_{11}$ *and either* A *is solvable or* $A \cong M_{10}$.

(iii) $G = M_{23}$ *and either* B *is a Frobenius group of order* 11.23 *or* B *is cyclic of order* 23 *and* $A \cong M_{22}$.

(iv) $G = PSL_2(q)$ *where either* $q \in \{11, 29, 59\}$ *and* $A \cong A_5$ *or* $q \not\equiv 1(mod \quad 4), q > 3$ *and* A *is solvable.*

(v) $G = PSL_r(q), r$ *an odd prime such that* $(r, q-1) = 1$ *and either* $G \cong PSL_5(2)$ *and* $|B| = 5.31$ *or* A *is maximal parabolic subgroup such that* $PSL_{r-1}(q)$ *is involved in* A. *In particular* B *is either cyclic or Frobenius.*

As a cosequence of the above theorem we obtain the following corollary.

Corollary 3.2. ([AF]) *Let* G *be a group such that* $G = AB$, $(|A|, |B|) = 1$. *Let* D *be a composition factor of* G. *Then either* D *is of type* $(i)-(v)$ *in Theorem 3.1 or* $\pi(D) \subset \pi(A)$ *or* $\pi(D) \subset \pi(B)$. *In particular in the latter case* D *is either a section of* A *or of* B, *respectively. Here* $\pi(G)$ *denotes the set of primes involved in the prime factorization of the order of the finite group* G.

Before stating the next result we introduce some definitions. Let p be a prime. An elementary abelian p-group E_p^n is the direct product of n copies of Z_p and in this case n is called the rank of E_p^n. If G is a finite group, then the p-rank of G is the maximum p-rank of an elementary abelian p-subgroup of G and is denoted by $m_p(G)$.

Now considering the factorizations $G = AB$ and putting restrictions on the structures of A and B, especially on the Sylow 2-subgroups of these groups leads to the following results due to U.Preiser [Pr].

Theorem 3.3. ([*Pr*]) *Let* G *be a finite group that contains subgroups*

A and B with $G = AB$. Assume that a Sylow 2-subgroup of A has rank 2 and a Sylow 2-subgroup of B is elementary abelian. Then one of the followings holds:

(i) $G \cong A$ or B,

(ii) $G \cong A \times B$,

(iii) $G \cong A_6, A \cong A_5 \cong B$,

(iv) $G \cong A_7$, $A \cong PSL_2(7)$, $B \cong A_5$,

(v) $G \cong A_9$, $A \cong A_6$ or A_7, $B \cong PSL_2(8)$,

(vi) $G \cong M_{12}$, $A \cong M_{11}$, $B \cong A_5$ or $PSL_2(11)$,

(vii) $G \cong G_2(3^m)$, $A \cong PSL_3(3^m)$, $B \cong^2 G_2(3^m)$, m odd, $m \neq 1$.

Also in the paper [Pr] by U. Prieser the following result is proved which asserts certain finite groups are not factorizable.

Theorem 3.4. ([Pr]) *(i) The groups $PSP_4(q)$, q odd, don't admit a factoriziation with proper simple subgroups,*

(ii) ${}^3D_4(q^3), q$ odd, don't admit a factorization with two proper simple subgroups,

(iii) $PSL_5(q), q \equiv 3(mod\ 4)$ and $PSU_5(q^2), q \equiv 1(mod\ 4)$, don't admits factorizations with proper simple subgroups,

(iv) $G_2(q), q = p^f, p > 3$, don't admit a factorization with proper simple subgroups.

In [HOS] finite groups G with factorization $G = AB$, where A and B have small Sylow 2-subgroups ard found.

Theorem 3.5. ([HOS]) *Let $G = AB$ be a factorization of G and $G \ncong A \times B$. If A and B are finite simple groups such that a Sylow 2-subgroup of B has order 4 or 8, then $G \cong A_6$ or A_7.*

There are several results by G.Walls in [Wal] concerning Sylow 2-subgroups in a factorization of a finite group and we mention some of them here.

Theorem 3.6. ([Wa1]) *Let $G = AB$ where A and B are simple subgroups of G such that a Sylow 2-subgroup of G is dihedral. Then*

(i) $G \cong A$ *or* B,

(ii) $G \cong A_6$, $A, B \cong A_5$,

(iii) $G \cong A_7, A \cong A_6, B \cong L_2(7)$ *or* $A \cong A_5$, $B \cong L_2(7)$.

(iv) $G \cong A \times B$.

Theorem 3.7. ([Wa1]) *Let $G = AB$ with A and B simple subgroups of G. If a Sylow 2-subgroup of G is abelian or quasi-dihedral, then $G \cong A, B$ or $A \times B$.*

Theorem 3.8. ([Wa1]) *If $G = AB, A$ and B simple, and if the order of a Sylow 2-subgroup of G is at most 32, then the case $(i) - (iv)$ in Theorem 3.6 occur.*

Factorizations of simple groups as the product of two simple subgroups first appeared in the following paper of Ito.

Theorem 3.9. ([It2]) *If $G = L_2(q) = AB$ with A and B simple subgroups of G, then*

(i) $G = A$ *or* B,

(ii) $q = 9$ *and* $G \cong L_2(9) \cong A_6 \cong A_5A_5$.

Extending the above result we have the following results of Gentchev.

Theorem 3.10. ([Ge1]) *Let G be a simple group of Lie type of Lie rank 1 or 2 over $GF(q)$. Let $G = AB$ with A and B simple non-abelian subgroups of G. Then one of the followings occurs.*

(i) $G \cong L_2(9), A, B \cong A_5$,

(ii) $G \cong G_2(4), A \cong J_2, B \cong U_3(4)$,

(iii) $G \cong G_2(q), q = 3^{2n+1} > 3, A \cong L_3(q), B \cong^2 G_2(q)$,

(iv) $G \cong G_2(q), q = 3^n, A \cong L_3(q), B \cong B_3(q)$,

(v) $G \cong U_4(3), A \cong L_3(4), B \cong PSP_4(3)$,

(vi) $G \cong U_4(q), q \not\equiv -1 \ (mod 3)$, $A \cong U_3(q), B \cong PSP_4(q)$.

Theorem 3.11. ([Ge2]) *Let G be a sporadic simple group such that $G = AB$ where A and B are simple non-abelian subgroups of G. Then*

(*i*) $G \cong M_{24}, A \cong L_2(23), B \cong M_{22}$ *or* M_{23} *or* $A \cong L_2(7), B \cong M_{23}$,

(*ii*) $G \cong M_{12}, A \cong A_5, L_2(11)$ *or* M_{11} *and* $B \cong M_{11}$,

(*iii*) $G \cong Suz, A \cong U_5(2), B \cong G_2(4)$,

(*iv*) $G \cong Co_1, A \cong G_2(4), B \cong Co_2$.

4. Factorizations with one factor being an altrnating or a symmetric group.

Now we turn to factorizations where one of the factors is isomorphic to an alternating or a symmetric group. O.Kegel and H.Luneberg [KL] proved the following.

Theorem 4.1. ([KL]) *Let $G = AB$ be a factorizable group with $A, B \cong A_5$. Then $G \cong A, B, A \times B$ or A_6.*

W.R. Scott in [So] found all finite groups which are equal to the product of A_5 and a finite non-abelian simple group.

Theorem 4.2. ([So]) *Let $G = AB$ be a finite group, A and B subgrops of G with A a finite non-abelian simple group and $B \cong A_5$ such that $Out(A) \not\geq A_5$. Then one of the following holds:*

(*i*) $G \cong A$ *or* B,

(*ii*) $G \cong A \times B$,

(*iii*) $G \cong A_n, A \cong A_{n-1}, n = 6, 10, 12, 15, 20, 30, 60$,

(*iv*) $G \cong A_7, A \cong L_2(7)$,

(*v*) $G \cong M_{12}, A \cong M_{11}$.

G.Walls in [Wa2] extended the main results of [So] by taking B to be a non-simple group.

Theorem 4.3. ([Wa2]) *Suppose that G is a simple group and that*

$G = AB$, where A is a non-abelian simple group and $B \cong S_5$. Then one of the following occurs:

(i) $G \cong A_n, A \cong A_{n-1}, n = 10, 12, 15, 20, 24, 30, 40, 60, 120$,

(ii) $G \cong M_{12}, A \cong M_{11}$,

(iii) $G \cong A_7$ or $A_8, A \cong L_2(7)$,

(iv) $G = A, A \geq A_5$.

Theorem 4.4. ([Wa2]) *Suppose that $G = AB$ is not a simple but A is a simple group and $B \cong S_5$. Then one of the following cases occurs:*

(i) $G \cong A \times B$,

(ii) $G \cong (A \times A_5)\langle\tau\rangle, \tau$ acting as an outer automorphism on both factors,

(iii) $G \cong (A_5 \times A_5)\langle\tau\rangle, A \cong A_5$ and $A_5 \times A_5$ is a minimal normal subgroup of G,

(iv) $G \cong S_6, A \cong A_5$,

(v) $G \cong A_n \times Z_2, A = A_{n-1}, n = 10, 12, 15, 20, 24, 30, 40, 60, 120$,

(vi) $G \cong M_{12}\langle\tau\rangle, A \cong M_{11}, \tau$ acts as an outer automorphism of order 2 on M_{12},

(vii) $G \cong A\langle\tau\rangle$ where $A \geq A_5$ and τ is an outer automorphism of the simple group A which acts as an outer automorphism on the copy of A_5 in A.

In Theorem 4.4 we observe that one of the factors of the factorizable group G is a non-simple group. Walls in [Wa2] proved the following general result.

Theorem 4.5. ([Wa2]) *Suppose that $G = AB$ is a finite group with subgroups A and B such that $A \cong S_n, n \geq 5$ and $B \cong A_5$. Then one of the following occurs.*

(i) $G \cong A \times B$,

(ii) $G \cong S_n = A, n \geq 5$,

(iii) $G \cong S_{n+1}, n = 5, 9, 11, 14, 19, 29, 59$,

(iv) $G \cong (A_n \times A_5)\langle\tau\rangle, \tau$ *acting as an automorphism of order 2 on both factors,*

(v) $n = 5, G \cong (A_5 \times A_5)\langle\tau\rangle, \tau$ *an automorphism of order 2,* $A_5 \times A_5$ *a minimal normal subgroup of* G.

To prove Theorem 4.5 one needs a special treatment of factorizable groups with a factor being non-simple. Lemma 3 of [Wa2] is adjusted to deal with the situations as above.

Lemma 4.6. *Let* $G = AB$ *be a factorization of a finite group and* A, B *subgroup of* G *such that* A *is a simple group and* B *has a unique proper normal subgroup* N *which is simple non-abelian. Let* $G \ncong A \times B$ *and let* M *be a minimal normal subgroup of* G. *Then one of the following holds:*

(i) $G = AB = M$ *is simple.*

(ii) $G = MB, M \cong A \times N, N \cong A$,

(iii) $G = MB, M \cong NA$ *is simple,*

(iv) $M = A$ *or* N, $[G : AN] = [B : N], AN \cong A \times N$,

(v) $M \cap X = 1$, $|X| | [X : A \cap B]$ *for* $X \in \{A, B\}$ *and* $|M| | A \cap B| = |AM/M \cap BM/M|$.

Garry Walls in [Wa3] considers factorizable non-simple groups. Some examples of non-simple groups which can be factored as product of two proper subgroups are:

$$2.PSU_4(3^2) = PSP_4(3)PSL_3(4),$$
$$3.PSU_4(3^2) = PSP_4(3)PSU_3(3^2),$$
$$3.Suz = PSU_5(2^2)G_2(4).$$

Motivated by recent results of Walls, We were interested in finding the structure of finite factorizable groups with one factor isomorphic to an alternating group. In [DR] we proved the following results:

Theorem 4.7. ([DR]) *Let G be a finite group such that $G = AB$ where $A \cong A_6$ and $B \cong A_n, n \geq 5$, then either $G \cong A \times B$ or G is a simple altenating group as follows:*

(i) $G \cong A_{n+1}, n = 5, 9, 29, 35, 39, 44, 59, 71, 89, 119, 179, 359$,

(ii) $G \cong A_n, n \geq 6$,

(iii) $G \cong A_{10}, n = 8$.

Theorem 4.8. ([DR]) *Let G be a finite group such that $G = AB$, $A \cong A_6$ and $B \cong S_n$, $n \geq 5$. Then one of the following occurs:*

(i) $G \cong A_6 \times S_n$,

(ii) $G \cong A_{10} \cong A_6 S_8, n = 8$,

(iii) $G \cong (A_6 \times A_6)\langle\tau\rangle, \tau$ *an automorphism of order* 2 *and* $A_6 \times A_6$ *is the minimal normal subgroup of* $G, n = 6$,

(iv) $G \cong A_{n+1}, n = 5, 9, 14, 19, 35, 39, 44, 59, 71, 89, 119, 179, 359$,

(v) $G \cong S_n, n \geq 6$,

(vi) $G \cong A_{10} \times Z_2, n = 8$,

(vii) $G \cong (A_6 \times A_n)\langle\tau\rangle, n \geq 5$, *where* τ *acts as an automorphism of order 2 on both factors.*

In [DRW] we extended the result obtained in [So] by considering factorizable groups with one factor isomorphic to the alternating group A_6 or S_6. These results are as follows:

Theorem 4.9. ([DRW]) *Let G be a simple group and $G = AB$ be a proper factorization of G with A simple and $B \cong A_6$. Then one of the following occurs:*

(i) $G \cong A_n$, $A \cong A_{n-1}$, $n = 10, 15, 20, 30, 36, 40, 45, 60, 72, 90, 120,$ $180, 360$,

(ii) $G \cong A_{10} \cong A_8 A_6$,

(iii) $G \cong A_8 \cong L_2(7) A_6$,

(iv) $G \cong A_9 \cong L_2(8) A_6$,

(*v*) $G \cong A_7 \cong L_2(7)A_6$,

(*vi*) $G \cong O_8^+(2) \cong S_6(2)A_6$.

Theorem 4.10. ([DRW]) *Let G be a simple group with non-trivial factorization $G = AB$, where A and B are subgroups of G with A simple and $B \cong S_6$. Then we have one of the following cases:*

(*i*) $G \cong A_n \cong A_{n-1}S_6$, $n = 10, 12, 15, 20, 30, 36, 40, 45, 60, 72, 80, 90, 120, 144, 180, 240, 360, g720$,

(*ii*) $G \cong A_{10} \cong A_8S_6$,

(*iii*) $G \cong A_8 \cong L_2(7)S_6$,

(*iv*) $G \cong A_9 \cong L_2(8)S_6$,

(*v*) $G \cong S_6(2) \cong U_3(3)S_6$,

(*vi*) $G \cong S_4(4) \cong L_2(16)S_6$,

(*vii*) $G \cong S_8(2) \cong O_8^-(2)S_6$,

(*viii*) $G \cong O_8^+(2) \cong S_6(2)S_6$.

Theorem 4.11. ([DRW]) *If G is not a simple group and has a factorization $G = AB$ with A simple and $B \cong S_6$, then one of the following occurs:*

(*i*) $G \cong A \times B$,

(*ii*) $G \cong A\langle\tau\rangle$ *where A is a simple group containing A_6 and τ is an outer automorphism of both A and A_6.*

(*iii*) $G \cong S_n$ *or* $A_n \times Z_2, A \cong A_{n-1}, n = 10, 15, 20, 30, 36, 40, 45, 60, 72, 90, 120, 180, 360$,

(*iv*) $G \cong S_{10}$ *or* $A_{10} \times Z_2, A \cong A_8$,

(*v*) $G \cong S_p$ *or* $A_8 \times Z_2, A \cong L_2(7)$,

(*vi*) $G \cong S_9$ *or* $A_9 \times Z_2, A \cong L_2(8)$,

(*vii*) $G \cong S_7, A \cong L_2(7)$,

(*viii*) $G \cong O_8^+(2)\langle\tau\rangle$ *or* $O_8^+(2) \times Z_2, A \cong S_6(2), \tau$ *an outer automorphism of* $O_8^+(2)$, *(ix)* $G \cong (A_6 \times A_6)\langle\tau\rangle, A \cong A_6$, *where* $A_6 \times A_6$ *is the minimal normal subgroup of* G,

(x) $G \cong (A \times A_6)\langle\tau\rangle, \tau$ *acting as an outer automorphism on both factors.*

In general, the structure of groups which are the product of an alternating and a symmetric group is given in [Da]. This paper is a generalization of [DR].

Theorem 4.12. ([Da]) *Let* G *be a finite group with two subgroups* A *and* B *such that* $G = AB$, *where* A *is isomorphic to some alternating group* A_r *and* B *is isomorphic to some symmetric group* $S_n, r, n \geq 5$. *Then one of the following occurs:*

(i) $G \cong A_r$ *or* S_n, *the trivial factorization,*

(ii) $G \cong A_r \times S_n$,

(iii) $G \cong A_{10}, S_{10}$ *or* $A_{10} \times Z_2$, *where* $A \cong A_6$ *and* $B \cong S_8$;

(iv) $G \cong A_{15}, S_{15}$ *or* $A_{15} \times Z_2$, *where* $A \cong A_7$ *or* A_8 *and* $B \cong S_{13}$,

(v) $G \cong A_{r+1}$ *(or* S_{n+1}*),* $A \cong A_r, B \cong S_n$, *where* A_{r+1} *and* $A_{r+1} \times Z_2$ *(or* S_{n+1}*) have a transitive subgroup isomorphic to* S_n *(or* A_r*),*

(vi) $G \cong (A_r \times A_r)\langle\tau\rangle, \tau$, *an outer automorphism of order* 2 *of* A_r *interchanging the two* A_r*'s and* $A_r \times A_r$ *is the minimal normal subgroup of* G,

(vii) $G \cong (A_r \times A_n)\langle\tau\rangle, \tau \neq n$, *where* τ *acts as an automorphism of order* 2 *on both factors.*

(viii) $G \cong S_r$ *or* S_{r+1}.

At the end of this paper we would like to mention some interesting problems concerning factorizable finite groups.

Q1: Find the structure of all finite groups $G = L_2(7)B$ where B is a non-abetian simple group.

Q2: Find all finite groups G such that $G = L_2(7)B$ where B is any finite group with the property that $B' = N$ is a simple group which is the unique normal subgroup of B and $[B : N] = 2$.

Q3: Let G be a simple factorizable group such that $G = AB$, where A and B are proper subgroups of G. If A and B are perfect groups is it true that one of A or B must be a simple group? (A is called a perfect group if $A' = A$). The answer to this question in the case of alternating G is yes.

Q4: Obtain general results about factorizations of infinite alternating or symmetric groups.

Q5: Find the structure of all finite groups $G = L_2(q)B$, where B isomorphic to the symmeteric group $S_n, n \geq 5$.

Q6: In general find the structure of all finite groups G such that $G = AB$ where A is a simple group and B is isomorphic to a symmetric group.

References

[AFD] B. Amberg, S.Fransiosi and F.De Giovanni, Products if groups, Oxford University Press (1992).

[CCNPW] J.H. Conway, R.T. Curtis, S.P.Norton, R.A. Párker, R.A. Wilson, *Atlas of finite groups*, Oxford University Press, 1985.

[Bu] W. Burnside,*On groups of order $p^a q^b$*, Proc.London Math. Soc. **1**(1904), 388-392.

[It2] N. Ito, *Uber das product von zwei abelschen gruppen*, Math. Z. **62**(1955), 400-401.

[Wi] H. Wielandt, *Uber produkte nilpotenten gruppen*, Illinois J.Math **2**(1958), 611-618.

[Ke] O.Kegel, *Produkte nilpotenter gruppen*, Arch. Math.(Basel) **12**(1961) 90-93.

[Ch] N.S. Chernikov, *Infinite groups that are products of nilpotent groups*, Dokl.Akad. Nauk SSSR 252,(1980),57-60 (Soviet Math. Dokl.) 21,(1980),
(701-703)

[Or] O.Ore, *Contribution to the theory of groups of finite order*, Duke Math. J.5(1938), 431-460.

[LPS] M.W.Liebeck, C.E. Praeger and J.Saxl, *The maximal factorization of the finite simple groups and their automorphism groups*, AMS 86 No.432(1990).

[Ka] W.M.Kantor, *k-homogenous groups*, Math. Z. 124(1972), 261-265.

[Ca] P.J.Cameron, Permutation groups, Cambridge University Press (1999).

[WW] J.Wiegold and A.G.Williamson, *The factorization of alternating and symmetric groups*, Math. Z. 175,171-179(1980).

[FA] E.Fisman and Z.Arad, *A proof of Szep's conjecture on non-simplicity of certain finite groups*, J. Alg. 108(1987)340-354.

[Pr] Udo Preiser, *Factorizations of finite groups*, Math. Z., 185 (1984)373-402.

[HOS] H.Hanes, K.Olson and W.R. Scott, *Product of simple groups*, J. Alg. 36(1975)167-184.

[Wal] G.Walls, *Products of finite simple groups*, J. Alg. Vol.48 No.1 (1977)68-88.

[It2] N.Ito, *On the factorization of the linear fractional groups $LF_2(p^n)$*, Acta. Sci. Math. (Szeged)15(1953),79-85.

[Ge1] T.R. Gentchev, *Factorizations of the groups of Lie type of Lie rank* 1 *or* 2, Arch. Math. 47(1986)493-499.

[Ge2] T.R. Gentchev, *Factorization of the sporadic simple groups*, Arch. Math. 47(1986)97-102.

[KL] O.Kegel and H.Luneberg, *Uber die kleine reidemeister bedingungen*, Arch. Math.,14 (Basel) 1963,7-10.

[So] W.R. Scott, *Products of A_5 and a finite simple group*, J. Alg., 37(1975)165-171.

[Wa2] G.Walls, *Products of simple groups and symmetric groups*, Arch. Math. Vol.58
(1992) 312-321.

[Wa3] G.Walls, *Non-simple groups which are the product of simple groups*, Arch. MAth. Vol. 53(1989)209-216.

[DR] M.R.Darafsheh and G.R. Rezaeeadeh, *Factorizations of groups involving symmetric and alternating groups*, IJMMS 27:3(2001) 161-167.

[DRW] M.R.Darafsheh, G.R.Rezaeezadeh and G.Walls, *Groups which are the product of S_6 and a simple group*, to appear in Alg. Colloq.

[Da] M.R.Darafsheh, *Finite groups which factor as product of an alternating group and a symmetric group*, submitted.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 39-57 (2004)

Algebra in a Category: Injectivity in Equational Classes

M. Mehdi Ebrahimi

Department of Mathematics
Shahid Beheshti University, Tehran, Iran
m-ebrahimi@cc.sbu.ac.ir

Abstract: Much of classical set based universal algebra can be and, if we may say so, should be, studied in a general topos rather than in the topos Set of sets.

Each topos carries its own logic, which is not necessarily the same as the classical logic of the topos Set. Thus, the classical notions of mathematics may behave differently in different topoi.

The notion from classical universal algebra we have been chosen for this talk is injectivity, which is one of the most central and useful notion in many mathematical disciplines.

The main purpose of this talk is to describe the relationship between the class $mod\Sigma$ of models of a set Σ of equations in the category Set and the corresponding class $mod(\Sigma, \mathbb{E})$ of models of Σ in a suitable category $\mathbb{E}$ with respect to injectivity and some related notions.

MSC (2000): 08B30, 08C05, 18C05

Keywords: Universal algebra, Grothendieck topos, essential boundedness, injectivity.

The basic nature of our results is that, for any given Σ, whatever holds in **Set**, concerning this notion, also holds in $\mathbb{E}$, provided $\mathbb{E}$ has some special properties, in particular when $\mathbb{E}$ is a Grothendieck topos.

The talk is general and the results have already been published. For the details of the proof, see [10].

1. Introduction

1.1 Universal Algebra: "Much of the beauty of mathematics is derived from the fact that it affords abstraction. Not only does it allow one to see the forest rather than the individual trees, but it offers the possibility for the study of the structure of the entire forest, in preparation for the next stage of abstraction-comparing forests."

Mathematicians at the beginning of the twentieth century were confronted with a large number of algebraic systems such as groups, rings, quaternions, Lie algebras, number rings, algebraic number fields, vector spaces, Boolean algebras, lattices, etc. A. N. Whitehead [1898] tried to place these diverse algebraic systems within a common frame work. Universal algebra as understood today goes back to the 1930's and it emerged as a natural development of the abstract approach to algebra initiated by Emmy Noether.

Universal Algebra studies features common to familiar algebraic systems mentioned above. Although, "one can become a very good mathematician without being a professional logician even though logical thought is central to mathematics". This goes for algebraists with regard to Universal Algebra. But, such a study places the algebraic notions in their proper setting. It often reveals connections between seemingly different concepts and helps to systematize one's thoughts. However, it does not usually solve the whole problem for us, tidies up a mass of rather trivial details and thus allows us to concentrate our powers on the hard core of the problem.

Universal algebra has grown very rapidly in the last thirty or forty years. Not only the literature expanded rapidly, but also the problems have become more sophisticated and the results deeper. Young mathematicians entering this field today are indeed fortunate, for there are hard and interesting problems to be attacked and sophisticated tools to be used.

In the last two decades universal algebra has become useful and important in combinatorics and theoretical computer sciences. In particular, structural aspects such as syntax and semantics, data abstraction, etc., are mainly investigated by methods of universal algebras [20, 21, 23, 24, 25, 34].

One of the fundamental ideas of universal algebra is the representation of logical notions in nonlogical terms. The famous Birkhoff Variety Theorem

states that a class of algebras is equationally definable iff it is closed under subalgebras, homomorphic images and products; such a class is called a variety. This characterization result was the starting point of universal algebra.

The general theory of algebras borrows techniques and ideas from lattice theory, logic, and category theory and derives inspirations from older, more specialized branches of algebras such as the theories of groups, rings, and modules.

1.2 Category Categorical methods of speaking and thinking has become widespread in mathematics because they achieve a unification of parts of different mathematical fields; frequently they bring simplifications and provide the impetus for new developments.

Categories, initially a convenient way of formulating exact sequences, and axiomatic homology theory, obtained independent life in the works of Ehresman, Kan, Maclane, Eilenberg, Barr, Freyd, Gray, Lawvere, Linton, Tierney, and others.

A category may be thought of in the first instance as a universe for a particular kind of mathematical discourses. Such a universe is determined by specifying a certain kind of "objects" and a certain kind of "arrows" that links different objects. The most general universe of current mathematical discourse is the category Set of sets with functions between them. Many basic properties of sets and set theoretic operations can be described by reference to the arrows in Set, and these descriptions can be interpreted in any category by means of its arrows. So the question that arises is "when does a category look and behave like Set ?" A vague answer is "when it is (at least) a topos".

1.3 Topos In 1963, Lawvere tried to find a purely categorical foundation for all mathematics, beginning with an appropriate axiomatization of the category of sets. When Lawvere heard of the properties of Grothendieck topoi, he soon observed that such a topos admits basic operations of set theory such as the formation of sets Y^X of all functions from X to Y and of power sets $\mathcal{P}(X)$. Subsequently, Lawvere and Tierney, working together at Dalhousie University, defined in an elementary way, free of all set-theoretic assumptions, the notion of an "elementary topos".

A **topos** is formally a category which has finite limits, exponentiations (abstracting the function set B^A) and subobject classifier (abstracting the truth set $2 = \{0, 1\}$).

Recall that, for a category $\mathcal{C}$ with finite limits, we say that

$\mathcal{C}$ has *exponentiations* (exponentials) if for every objects A and B, there is an object B^A together with an arrow $ev : B^A \times A \to B$ (called *evaluation*) such that for every arrow $g : C \times A \to B$ there is a unique arrow $\hat{g} : C \to B^A$ with $ev \circ (\hat{g} \times id_A) = g$.

We also say that C has *subobject classifier* if there exists an object Ω with an arrow $t : 1 \to \Omega$ (called *the truth arrow*) such that for every monomorphism $f : B \to A$ there is a unique arrow $\chi_f : A \to \Omega$ (called *the classifying arrow*) making the square

$$\begin{tikzcd}
B \arrow[r, "f"] \arrow[d, "!"'] & A \arrow[d, "\chi_f"] \\
1 \arrow[r, "t"] & \Omega
\end{tikzcd}$$

a pullback square.

In fact, a topos is informally a category which looks and behaves very much like the category of sets. One of the first examples of a topos is $Set^{C^{op}}$.

A **Grothendieck topos** $\mathbb{E}$ is a reflective subcatgory of $Set^{C^{op}} = \hat{C}$ ($i : \mathbb{E} \rightleftarrows \hat{C} : R$, $R \dashv i$), for some small category C, whose reflection functor R preserves finite limits. For example, for a topological space X, the category $PreshX$, of presheaves on X, and the category ShX, of sheaves on X, are Grothendieck topoi.

Again, one can become a good mathematician without being a professional category theorist. But

"How can you do "new maths" problems with an "old math" mind?"- Charlie Brown.

"Virtually all algebraic notions in category theory are parodies of their parents in the most "classical" of categories ··· the category of left A-modules." H. Bass.

Here we briefly study injectivity, the most central notion in classical universal algebra, modelled in a topos, rather than in the category *Set* of sets. It is intended to provide a deeper understanding of the real features of this algebraic notion and to show that how a classical set-based Universal Algebra can be, as we may say so, it should be, studied in a topos (or category) theoretic setting. For more information about this approach see [3, 5, 6, 7, 10, 11, 12, 13, 14, 18, 30].

2. Algebra in a category

Let $I\!\!K$ be a finitely complete category (that is, a category which has finite products and equalizers, in particular, it has a terminal object $\mathbf{1}$).

2.1 Definition Given a family $\tau = (n_\lambda)_{\lambda\in\Lambda}$ of finite cardinal number n_λ, indexed by a set Λ, an *algebra* in $I\!\!K$ is an entity $(A, (\lambda_A)_{\lambda\in\Lambda})$, where A is an object of $I\!\!K$ and, for each $\lambda \in \Lambda$, the λ th operation $\lambda_A : A^{n_\lambda} \to A$ is a morphism in $I\!\!K$. The family $\tau = (n_\lambda)_{\lambda\in\Lambda}$ is called the type of this algebra. The algebra $(A, (\lambda_A)_{\lambda\in\Lambda})$ is simply denoted by A.

In the sequel, all algebras are of the same type τ.

2.2 Definition A *homomorphism* from an algebra A to an algebra B in $I\!\!K$ is a morphism $h : A \to B$ in $I\!\!K$ such that, for each $\lambda \in \Lambda$, $\lambda_B \circ h^{n_\lambda} = h \circ \lambda_A$. That is, the following diagram commutes, for each $\lambda \in \Lambda$:

$$\begin{tikzcd}
A^{n_\lambda} \arrow[r, "h^{n_\lambda}"] \arrow[d, "\lambda_A"'] & B^{n_\lambda} \arrow[d, "\lambda_B"] \\
A \arrow[r, "h"] & B
\end{tikzcd}$$

Clearly the identity morphism on the algebra A in $I\!K$ is a homomorphism $1_A : A \to A$ and for composable homomorphisms $A \xrightarrow{g} B \xrightarrow{f} C$, $f \circ g$ is a homomorphism.

As a result, the collection of all algebras (of the type τ) in $I\!K$ and homomorphisms between them forms a category denoted by $Alg(\tau)I\!K$ (or by $Alg(\tau)$ if $I\!K = Set$).

2.3 Remark For $A \in Alg(\tau)I\!K$ and any natural number n, the set $Hom_K(A^n, A)$ of all morphisms in $I\!K$ from A^n to A can easily be made into an algebra of the type τ in Set, by defining the λth operation as

$$\lambda(\phi_1, \cdots, \phi_{n_\lambda}) = \lambda_A \circ \prod_{i=1}^{n_\lambda} \phi_i$$

for any $\phi : A^n \to A$ $(i = 1, \cdots, n_\lambda)$ in $I\!K$, where $\prod_{i=1}^{n_\lambda} \phi_i : A^n \to A^{n_\lambda}$ is the morphism determined by ϕ_i's. That is, $\lambda_{Hom_K(A^n,A)}$ takes $(\phi_1, \cdots, \phi_{n_\lambda})$ to

$$\lambda \circ \prod_{i=1}^{n_\lambda} \phi_i : A^n \to A^{n_\lambda} \to A$$

Let FX be the absolutely free algebra in Set of the type τ on a set $X = \{x_1, \cdots, x_n\}$ of n elements. Extend the map

$$\begin{gathered} X \to Hom_K(A^n, A) \\ x_i \mapsto p_i : A^n \to A \end{gathered}$$

(p_i the i th projection), freely to

$$\phi : F \to Hom_K(A^n, A)$$

and denote $\phi(p)$ by p_A for any $p \in F$.

2.4 Definition A *law* (*identity* or *equation*) over Λ in the set X of variables is any pair $\sigma = (p, q) \in F \times F$, denoted by $p = q$.

We say A *satisfies the equation* $p = q$, written $A \models (p = q)$, if $p_A = q_A$.

The full subcategory of $Alg(\tau)\mathbb{K}$ given by the class of all algebras in $\mathbb{K}$ satisfying every equation of a set Σ of equations is denoted by $mod(\Sigma, \mathbb{K})$, or by $mod\Sigma$, if $\mathbb{K} = Set$, and is called an *equational category of algebras.*

2.5 Examples

1) A group in the category $\mathbb{K} = ShX$, the category of all sheaves on a topological space X, is an entity $(G; *, ()^{-1}, e)$, where G is a sheaf on X and

$$* : G \times G \to G \;\; , \;\; ()^{-1} : G \to G \;\; , \;\; e : 1 \to G$$

are morhpisms in ShX (that is, natural transformations) such that the following diagrams are commutative:

i) (associativity of *)

$$\begin{tikzcd}
G \times G \times G \arrow[r, "* \times 1_G"] \arrow[d, "1 \times *"'] & G \times G \arrow[d, "*"] \\
G \times G \arrow[r, "*"] & G
\end{tikzcd}$$

ii) (the identity condition)

$$\begin{tikzcd}
1 \times G \arrow[d, "e \times 1"'] & G \arrow[l, "{(!,1)}"'] \arrow[r, "{(1,!)}"] \arrow[d, "1"'] & G \times 1 \arrow[d, "1 \times e"] \\
G \times G \arrow[r, "*"] & G & G \times G \arrow[l, "*"']
\end{tikzcd}$$

iii) (the inverse condition)

$$\begin{tikzcd}
G \arrow[r, "!"] \arrow[d, "{(1,()^{-1})}"'] & 1 \arrow[d, "e"'] & G \arrow[l, "!"'] \arrow[d, "{(()^{-1},1)}"] \\
G \times G \arrow[r, "*"] & G & G \times G \arrow[l, "*"']
\end{tikzcd}$$

Notice that the commutativity of the above diagrams in ShX means that the diagrams are commutative at each $U \in O(X)$ in Set. In fact, a sheaf G is a group in ShX iff GU is a group in Set (see also 2.8).

A homomorphism between groups in ShX, is a sheaf morphism (natural transformation) which is a group homomorphism at each $U \in O(X)$.

2) A ring in the category $MSet$, of all M-sets (that is, sets with an action of the monoid on it), is an entity $(R; +, ., -, 0)$, where R is an M-set and

$$+ : R \times R \to R, \quad . : R \times R \to R, \quad - : R \to R, \quad 0 : \{\bullet\} \to R$$

are action preserving maps satisfying the ring axioms.

In fact, a ring in $MSet$ is an M-set which is also a ring in Set whose ring operations are action preserving maps.

A homomorphism between rings in $MSet$ is an action preserving ring homomorphism.

3) A monoid in the category $R - Mod$, of all left R-modules, is an entity $(M; ., e)$, where M is an R-module and

$$. : M \times M \to M \quad , \quad e : \{\bullet\} \to M$$

are R-module homomorphisms satisfying the following identities:

i) $(m.n).k = m.(n.k)$

ii) $e.m = m = m.e$

So, an R-module which is also a monoid whose monoid operations are R-module homomorphisms is a monoid in $R - Mod$.

A homomorphism between monoids in $R - Mod$ is an R-module homomorphism which is also a monoid homomorphism.

2.6 Remark Let $k : \mathbb{K} \to \mathbb{L}$ be a functor, preserving finite limits. Then k induces another functor

$$\bar{k} : Alg(\tau)\mathbb{K} \to Alg(\tau)\mathbb{L}$$

defined on objects by

$$\overline{k}A = (A, (k\lambda_A)_{\lambda \in \Lambda})$$

and on homomorphisms $f : A \to B$, by

$$\overline{k}(f) = k(f)$$

If $\sigma = (p, q)$ is an equation, then $A \models (p = q)$ implies that $p_A = q_A$, and hence $kp_A = kq_A$ which implies that $p_{\overline{k}A} = q_{\overline{k}A}$; and thus $\overline{k}A \models (p = q)$. We thus get a functor

$$\overline{k} : mod(\Sigma, I\!\!K) \to mod(\Sigma, I\!\!L)$$

for any given set Σ of equations.

Recall that a set Φ of objects of a category $I\!\!K$ is said to be a set of *generators* if for every pair of morphisms $f, g : A \to B$ with $f \neq g$ there exists $G \in \Phi$ and a morphism $h : G \to A$ such that $fh \neq gh$.

2.7 Lemma *Let $I\!\!K$ have a set Φ of generators. Then, for any $A \in Alg(\tau)I\!\!K$ and a set Σ of equations, $A \in mod(\Sigma, I\!\!K)$ iff $\overline{h}_G(A) \in mod\Sigma$ for each $G \in \Phi$, where $h_G = Hom_K(G, -)$.*

Proof Applying remark 2.6 to the functor h_G, we get that for $A \in mod(\Sigma, I\!\!K)$, $\overline{h}_G(A) \in mod\Sigma$. Conversely, let $A \in Alg(\tau)I\!\!K$ and $\overline{h}_G(A) \in mod\Sigma$ for each $G \in \Phi$. Let $\sigma = (p, q)$ be an equation in Σ. By hypothesis, for all $G \in \Phi$, $\overline{h}_G(A) \models \sigma$, that is, $p_{\overline{h}_G(A)} = q_{\overline{h}_G(A)}$. So, $h_G(p_A) = h_G(q_A)$ for all $G \in \Phi$. Since Φ is a set of generators, the preceding equalities yield that $p_A = q_A$. Thus $A \models \sigma$, and hence $A \in mod(\Sigma, I\!\!K)$.

2.8 Corollary *The category $Alg(\tau)\hat{C}$ is isomorphic to the category of all $Alg(\tau)$-valued presheaves on C. And, for a Grothendieck topos $\mathbb{E}$, and a set Σ of equations, $A \in mod(\Sigma, \mathbb{E})$ iff $AU \in mod\Sigma$ for all $U \in C$.*

3. Injectivity in equational classes

A natural question to ask would be, what is the relationship between the behaviour of a certain classical algebraic notion in $mod\Sigma$ and in $mod(\Sigma, \mathbb{K})$. In the following, we briefly consider the notion of injectivity, and show that the properties of $mod\Sigma$, regarding injectivity, survive the passage to $mod(\Sigma, \mathbb{E})$, for a set Σ of equations and an arbitrary Grothendieck topos $\mathbb{E}$ (fixed from now on); for the details and some particular cases see for example [10], [13], [30]. For the case of equational classes of algebras in **Set** see for example [2], [32].

3.1 Definition An object E in a category $\mathbb{K}$ is called *injective* if, for any monomorphism $B \xrightarrow{i} C$ and any morphism $B \xrightarrow{f} E$ there exists a morphism $C \xrightarrow{\overline{f}} E$ with $\overline{f} \circ i = f$; that is the following diagram commutes:

$$\begin{tikzcd}
B \arrow[r, "i"] \arrow[d, "f"'] & C \arrow[dl, "\overline{f}"] \\
E &
\end{tikzcd}$$

Replacing B, in the above definition, by E and f by the identity morphism on E, we get the definition of an *absolute retract* object E in a category.

3.2 Definition A monomorphism $h : A \to B$ in a category $\mathbb{K}$ is called **essential** if, for any morphism $g : B \to C$, wherever $g \circ h$ is a monomorphism, then so is g.

3.3 Lemma *In $mod(\Sigma, \mathbb{E})$, for a Grothendieck topos $\mathbb{E}$, we have*

i) *any composite of essential monomorphisms is an essential monomorphism, and*

ii) *any direct limit of essential monomorphisms is an essential monomorphism.*

3.4 Lemma *In $mod(\Sigma, \mathbb{E})$, for any monomorphism $h : A \to B$ there exists a homomorphism $g : B \to C$ with $g \circ h$ an essential monomorphism.*

Proof Take Θ_0 to be the maximal congruence on B such that $B/\Theta_0 \in mod(\Sigma, \mathbb{E})$ and $A \to B \to B/\Theta_0$ is a monomorphism. This composition is then essential.

3.5 Corollary *In $mod(\Sigma, \mathbb{E})$, an algebra A is an absolute retract iff it has no proper essential extension.*

3.6 Definition A category $\mathbb{K}$ is called *essentially bounded* if, for each $A \in \mathbb{K}$ there exists, up to isomorphism, only a set of essential extensions in $\mathbb{K}$.

3.7 Proposition *$mod(\Sigma, \mathbb{E})$ is essentially bounded iff $mod\Sigma$ is essentially bounded.*

3.8 Definition In any category $\mathbb{K}$, *pushouts transfer monomorphisms* if, for any pushout diagram

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[d, "u"'] & B \arrow[d, "v"] \\ C \arrow[r, "g"] & D \end{tikzcd}$$

whenever f is a monomorphisms, then g is also a monomorphism.

If $I\!\!K$ has pushouts, the above is equivalent to say that, any diagram

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[d, "u"'] & B \\ C & \end{tikzcd}$$

with f a monomorphism can be completed to a commutative diagram

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[d, "u"'] & B \arrow[d, "v"] \\ C \arrow[r, "g"] & D \end{tikzcd}$$

with g a monomorphism.

3.9 Proposition *Pushout transfer monomorphisms in* $mod(\Sigma, I\!\!E)$ *iff they do in* $mod\Sigma$.

3.10 Lemma *The category* $mod(\Sigma, I\!\!E)$ *has enough injectives iff it is essentially bounded and pushouts transfer monomorphisms.*

3.11 Proposition *The category* $mod(\Sigma, I\!\!E))$ *has enough injectives iff* $mod\Sigma$ *has enough injectives.*

4. Behaviour of injectivity in $mod(\Sigma, I\!\!E)$

Banaschewski in [2] calls the notion of injectivity in a category $I\!\!K$ *properly behaved* iff the following three propositions hold:

(I) For any $A \in \mathbb{K}$, the following conditions are equivalent:

(I1) A is injective.

(I2) A is an absolute retract.

(I3) A has no proper essential extensions.

(E) Every $A \in \mathbb{K}$ has an injective hull, unique up to isomorphisms.

(H) For any monomorphism $f : A \to B$, the following conditions are equivalent:

(H1) $f : A \to B$ is an injective hull of A.

(H2) $f : A \to B$ is a maximal essential extension.

(H3) $f : A \to B$ is a minimal injective extension.

For $\mathbb{K} = mod(\Sigma, \mathbb{E})$, we now have the following.

4.1 Proposition *For $mod(\Sigma, \mathbb{E})$, the following are equivalent:*

(i) *Injectivity is properly behaved.*

(ii) *$mod(\Sigma, \mathbb{E})$ has enough injectives.*

(iii) *$mod(\Sigma, \mathbb{E})$ is essentially bounded and pushout transfer monomorphisms.*

In particular, one has, by proposition 3.11:

4.2 Corollary *Injectivity is properly behaved in $mod(\Sigma, \mathbb{E})$ iff it is properly behaved in $mod(\Sigma)$.*

4.3 Examples

1) Recall that, the category Set has enough injectives. So, if we take Σ to be the empty set, then $mod\Sigma$ as the full subcategory of all algebras

of the type $\tau = \emptyset$, is Set and hence has enough injectives. Using 3.11, this implies that for any Grothendieck topos $\mathbb{E}$, $mod(\Sigma, \mathbb{E}) = \mathbb{E}$ has enough injectives. In particular, the category $MSet$, of M-sets have enough injectives. Thus injectivity is properly behaved in such categories.

2) The category Boo, of Boolean algebras has enough injectives (the power set of each set is injective in Boo). So, the category of Boolean algebras in any Grothendieck topos has enough injectives.

3) The category Ab, of abelian groups has enough injectives (recall that here injectivity means divisibility). So, the category of abelian groups in any Grothendieck topos, in particular in ShX, has enough injectives.

4.4 Note For certain Σ, one has characterization of the injective $A \in mod\Sigma$ by properties of A in terms of its elements and subsets. For example: divisibility for abelian groups, completeness for Boolean algebras (the Sikorski Theorem), and completeness and Booleanness for distributive lattices. A good question to ask is to what extent, that is for what $\mathbb{E}$, such characterizations remain valid in $mod(\Sigma, \mathbb{E})$. Banaschewski [3] shows that divisibility = injectivity for abelian groups in the category ShL of sheaves on a frame L iff L is Boolean. Also Ebrahimi [13] shows that for Boolean algebras in the topos of M-sets, injectivity implies "internal" completeness, but the converse is not true. Also, Mahmoudi [30] defines internal injectivity for the category $MBoo$ of Boolean algebras in the topos $MSet$ and finds some equivalent conditions in which the internal version of the Sikorski Theorem holds.

References

1. Adamek, J., H.Herrlich and G.E. Strecker, *Abstract and Concrete Categories*, John Wiley and Sons, Inc., 1990.

2. Banaschewski, B., *Injectivity and essential extensions in equational classes of algebras*, Queen's Paper in Pure and Applied Mathematics, No. 25 (1970), 131-147.

3. Banaschewski, B., *When are divisible abelian groups injective*, Quaestiones Mathematicae, No. 4 (1981), 285-307.

4. Banaschewski, B., and K. R. Bhutani, *Boolean algebras in a localic topos*, Math. Proc. Cambridge Philo. Soc. **100** (1986), 43-55.

5. Bhutani, K. R., *Injectivity and injective hulls of abelian groups in a localic topos*, Bull. Austral. Math. Soc. Vol. 37 (1988), 43-59.

6. Bhutani, K. R., *Abelian groups in a topos of sheaves: Torsion and essential Extensions*, Intenat. J. Math. & Math. Sci. Vol. 12, No. 1 (1989), 89-98.

7. Bhutani, K. R., *Stability of Abelian groups in a topos of sheaves*, J. Pure and Applied Algebra, No. 68 (1990), 47-54.

8. Burris, S. and H. P. Sankapanavar, *A Course in Universal Algebra*, Graduate Texts in Math. No. 78, Springer-Verlag, 1981.

9. Burris. S. and M. Valeriote, *Expanding varieties by monoids of endomorphisms*, Alg. Universalis, Vol. 17, No. 2 (1983), 150-169.

10. Ebrahimi, M. Mehdi, *Algebra in a Grothendieck Topos: Injectivity in quasi-equational classes*, J. Pure and Applied Algebra, No. 26 (1982), 269-280.

11. Ebrahimi, M. Mehdi, *Equational compactness of sheaves of algebras on a Noetherian locale*, Algebra Universalis, (16) (1983), 318-330.

12. Ebrahimi, M. Mehdi, *Bimorphisms and tensor products in a topos*, Bull. Iranian. Math. Soc. Vol. 15, No. 1 (1988), 1-31.

13. Ebrahimi, M. Mehdi, *Internal completeness and injectivity of Boolean algebras in the topos of M-sets*, Bull. Austral. math. Soc. Vol. 41, No. 2 (1990), 323-332.

14. Ebrahimi, M. Mehdi, and M. Mahmoudi,*The internal ideal lattice in the topos of M-sets*, J. of Sciences, Islamic Republic of Iran, Vol. 5, No. 3 (1994), 123-128.

15. Ebrahimi, M. Mehdi, and M. Mahmoudi, *When is the category of separated M-sets a quasitopos or a topos?*, Bull. Iranian Math. Soc., Vol. 21, No. 1 (1995), 25-33.

16. Ebrahimi, M. Mehdi, and M. Mahmoudi, *Purity and equational compactness of G-sheaves*, Tech. Rep., Shahid Beheshti University (1997).

17. Ebrahimi, M. Mehdi, and M. Mahmoudi, *Purity and equational compactness of projection algebras*, Appl. Categ. Structures 9 (2001), No. 4, 381-394.

18. Ebrahimi, M. Mehdi, and M. Mahmoudi, *Tensor product and flatness of M-algebras*, to appear in Southeast Asian Bull. of Math.

19. Ebrahimi, M. Mehdi, and M. Mahmoudi, *The category of M-sets*, Italian journal of pure and applied Mathematics, No. 9 (2001), 123-132.

20. Ehrig, H., F. Parisi-Presicce, P. Boehm, C. Rieckhoff, C. Dimitrovici, and M. Grosse-Rhode, *Algebraic data type and process specifications based on projection spaces*, Lecture Notes in Computer Science **332** (1988), 23-43.

21. Ehrig, H., F. Parisi-Presicce, P. Bohem, C. Rieckhoff, C. Dimitrovici, and M. Grosse-Rhode, *Combining data type and recursive process specifications using projection algebras*, Theoretical Computer Science **71** (1990), 347-380.

22. Goldblatt, R., *Topoi: The categorial analysis of logic*, North Holland, 1986.

23. Grosse-Rhode, *Parametrized data type and process specifications using projection algebras*, Lecture Notes in Computer Science **393** (1988), 185-197.

24. Herrlich, H., and H. Ehrig, *The construct* **PRO** *of projection spaces: its internal structure*, Lecture Notes in Computer Science, **393** (1988), 286-293.

25. Hyland, J. M. E., and A. M. Pitts, *The theory of constructions: categorical semantics and topos-theoretic models*, Contemporary Mathematics, Vol. 92, 1989.

26. Johnstone, P., *Topos Theory*, Academic Press, 1977.

27. Joyal, A., and M. Tierney, *An extension of the Galois theory of Grothendieck*, Memories of the American Math. Soc. Vol. 51, N0. 309, 1984.

28. Maclane, S., *Categories for the working mathematician*, Graduate Texts in Mathematics, No. 5, Springer-Verlag, 1971.

29. Maclane, S. and I. Moerdijk, *Sheaves in Geometry and Logic*, Springer-Verlag, 1992.

30. Mahmoudi, M., *Internal injectivity of Boolean algebras in **MSet***, Alg. Univ. 41 (1999), 155-175.

31. Mahmoudi, M., *M-boolean envelope of M-distributive lattices* Italian journal of pure and applied Mathematics, No. 9 (2001), 133-138.

32. Taylor, W., *Residually small varieties*, Alg. Univ. 2 (1972), 33-53.

33. Tennison, B. R., *Sheaf Theory*, Cambridge University Press, 1975.

34. Wechiler, W., *Universal Algebra for Computer Scientists*, EATCS monographs on theoretical computer science, Springer-Verlag, 1992.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 59-104 (2004)

Approximation in Complex and Real Lipschitz Algebras

Taher Ghasemi Honary

Faculty of Mathematical Sciences and Computer Engineering

Teacher Training University, Tehran, Iran

tghonary@hotmail.com

Abstract: For a compact metric space (X,d) and $0<\alpha\le 1$, the algebra of all complex valued functions f on X for which $p_\alpha(f)=\sup\left\{\frac{|f(x)-f(y)|}{d^\alpha(x,y)}:x,y\in X, x\neq y\right\}<\infty$ is denoted by $\mathrm{Lip}(X,\alpha)$ and the subalgebra of those $f\in Lip(X,\alpha)$ for which $\frac{|f(x)-f(y)|}{d^\alpha(x,y)}\longrightarrow 0$ as $d(x,y)\longrightarrow 0$ is denoted by $\ell ip(X,\alpha)$. These algebras are known as the Lipschitz algebras and were first studied by D.R. Sherbert in 1964. The Lipschitz algebras are Banach algebras under the norm $\|f\|=\|f\|_X+p_\alpha(f)$, where $\|f\|_X$ is the uniform norm on X.

The Hedberg's Theorem concerning the density of certain subalgebras of $\ell ip(X,\alpha)$ in $\ell ip(X,\alpha)$ will be discussed and, as a consequence, the density of $Lip(X,1)$ in $\ell ip(X,\alpha)$ for $0<\alpha<1$ will be shown.

When X is a compact subset of $\mathbb{C}^n$ there are interesting subalgebras of Lipschitz algebras which are generated by polynomials, rational functions with poles off X, functions which are analytic in some neighbourhood of X, or functions which are analytic in the interior of X.

MSC (2000): 46J10, 46J15, 46J20, 41A10, 41A20.

Keywords and Phrases: Approximation, Banach function algebras, complex Lipschitz algebras, real Lipschitz algebras, Fréchet Lipschitz algebras,

The approximation problem concerning the equality among the above algebras and the density of these algebras in certain uniform algebras will be investigated. In particular, when X is a compact plane set with planar measure zero, it is shown that $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$, where $\ell ip_R(X,\alpha)$ is the norm closure of rational functions with poles off X. Another important result is the density of $D^1(X)$ in $\ell ip_R(X,\alpha)$, where $D^1(X)$ is the Banach function algebra of functions with continuous derivative on the perfect compact plane set X.

We introduce Lipschitz algebras of differentiable complex functions on perfect compact plane sets and then interesting subalgebras of these algebras are defined and the approximation problem among these kinds of Lipschitz algebras will be discussed.

To introduce the real Lipschitz algebras on a compact metric space (X,d) we assume that the map $\tau : X \longrightarrow X$ is a topological involution on X such that $d(\tau(x),\tau(y)) \leq Cd(x,y)$ for all $x,y \in X$, where C is a positive constant. Let $\sigma : C(X) \longrightarrow C(X)$ be the algebra involution on $C(X)$ induced by τ; i.e. $\sigma h = \bar{h} \circ \tau$. The real Lipschitz algebras are introduced by

$$Lip(X,\tau,\alpha) = \{h \in Lip(X,\alpha) : \sigma(h) = h\} \quad (0 < \alpha \leq 1)$$
$$\ell ip(X,\tau,\alpha) = \{h \in \ell ip(X,\alpha) : \sigma(h) = h\} \quad (0 < \alpha < 1)$$

The extension of the Hedberg's Theorem for the real Lipschitz algebra $\ell ip(X,\tau,\alpha)$ will be discussed and the density of $Lip(X,\tau,1)$ in $\ell ip(X,\tau,\alpha)$ will be shown.

Another important class of Lipschitz algebras is the Fréchet Lipschitz algebras $FLip(X,\alpha)$ and $F\ell ip(X,\alpha)$, where X is a hemicompact metric space. The density of $FLip(X,1)$ in $F\ell ip(X,\alpha)$ and the Hedberg's Theorem for the Fréchet Lipschitz algebra $F\ell ip(X,\alpha)$ will be discussed.

We will try to present a brief historical background on the progress of the subject.

1. Historical Background

Although the notion of Lipschitz functions is very old and these interesting functions have been studied for many decades, interest in the Banach space and Banach algebra theory of Lipschitz functions was not developed until 1955. As it is mentioned in the paper of D.R. Sherbert [30], the only work known to us till 1955, which treats the space of Lipschitz functions as a Banach algebra, was done by S.B. Myers [22] . His paper contains a summary of results only, the proofs never published because of his untimely death in 1955. The proofs of many of the unproved statements of Myers have been supplied by Sherbert in [30] , and in some cases, his results have been extended to more general settings. Sherbert also discussed different aspects of Banach algebras of Lipschitz functions in [29].

Since 1964 some other mathematicians have been working on real and complex Lipschitz algebras and obtained interesting results in the following fields: Structure of ideals, maximal ideal spaces, point derivation, automatic continuity, eventual continuity, amenability and weak amenability, differentiability of Lipschitz functions, the Stone-Weierstrass theorem, extreme points, peak points, Shilov boundary and Choquet boundary, isometries between Banach spaces of Lipschitz functions, closed ideals, Lipschitz algebras of differentiable functions, and the approximation problem.

A good collection of the works in Lipschitz algebras can be found in the recent book of Nik Weaver [32] , though it does not contain many results in this field. For more advanced account on some aspects of the complex Lipschitz algebras, including the Frechet Lipschitz algebras, one can refer to the recent and interesting monograph of H.G. Dales [7; Sec. 4.4].

In this paper we are going to study an interesting aspect of Lipschitz algebras, called the approximation problem. One of the famous results in approximation is the Hedberg's theorem, which is, in fact, the Stone-Weierstrass theorem for Lipschitz algebras and will be discussed here for real and complex Lipschitz algebras as well as the Frechet Lipschitz algebras.

2. Complex Lipschitz Algebras

Let X denote a compact Hausdorff space and $C(X)$ denote the commutative complex unital Banach algebra of all continuous complex-valued functions on X with the pointwise operations and with respect to uniform norm on X.

Definition 2.1. A *complex Banach function algebra* on X, is a complex subalgebra of $C(X)$ which separates the points of X, contains the constant function 1 and it is complete under an algebra norm. If the norm of a complex Banach function algebra is the uniform norm on X, then it is called a *complex uniform (function) algebra* on X.

If A is a complex Banach function algebra on X the space of maximal ideals of A is denoted by M_A, which is, in fact, the space of all non-zero complex (continuous) homomorphisms on A. Clearly, for each $x \in X$ the map $e_x : A \longrightarrow \mathbb{C}$, defined by $e_x(f) = f(x)$ is a homomorphism (character) on A which is called the *evaluation homomorphism (character)* on A at x. It is clear that A is semisimple, since

$$\begin{aligned}\operatorname{rad}(A) &= \{f \in A : f \in \ker(\phi), \text{ for all } \phi \text{ in } M_A\}\\ &\subseteq \{f \in A : f \in \ker(e_x), \text{ for all } x \text{ in } X\}\\ &= \{f \in A : f(x) = 0, \text{ for all } x \text{ in } X\}\\ &= \{0\}.\end{aligned}$$

The map $J : X \longrightarrow M_A$, defined by $J(x) = e_x$, is continuous and injective.

Definition 2.2 The Banach function algebra A is called *natural* if the map J is surjective, that is, every homomorphism (character) on A is an evaluation homomorphism (character) at some $x \in X$.

For example, $C(X)$ is a natural complex uniform algebra on X [4; Sect.17] or [26; Example 11.13]. If $(A, \|\cdot\|)$ is a complex Banach function algebra on X, the uniform norm of each element $f \in A$, does not exceed from its norm, since

$$\begin{aligned} \|f\|_X &= \sup_{x\in X} |f(x)| \le \sup_{\phi\in M_A} |\phi(f)| \\ &= \sup_{\phi\in M_A} |\hat{f}(\phi)| = \|\hat{f}\|_{M_A} = \rho_A(f) = \inf_{n\ge 1} \|f^n\|^{\frac{1}{n}} \le \|f\|. \end{aligned}$$

Definition 2.3 Let (X, d) be a compact metric space and $\alpha > 0$. The algebra of all complex-valued functions f on X for which

$$p_\alpha(f) = \sup_{\substack{x,y\in X \\ x\ne y}} \frac{|f(x) - f(y)|}{d^\alpha(x,y)} < \infty \ ,$$

is denoted by $Lip(X, \alpha)$ and the subalgebra of those functions f for which $\lim\limits_{d(x,y)\to 0} \frac{|f(x)-f(y)|}{d^\alpha(x,y)} = 0$, is denoted by $\ell ip(X, \alpha)$. These are called Lipschitz algebras of order α and were first studied by Sherbert [30].

Clearly $Lip(X, \alpha)$ is a Banach algebra of continuous complex-valued functions on X under the norm $\|f\|_\alpha = \|f\|_X + p_\alpha(f)$, where $\|f\|_X = \sup\limits_{x\in X} |f(x)|$. If $\ell ip(X, \alpha)$ is also equipped with the above norm then it is easily proved that $\ell ip(X, \alpha)$ is a closed subalgebra of $Lip(X, \alpha)$ and hence it is also a Banach algebra.

Lipschitz algebras contain the constants for every positive α. But when X is a connected compact subset of $\mathbb{C}^n$, $d(x, y) = \|x - y\|$ and $1 < \alpha$ $(\alpha \ge 1)$, then $Lip(X, \alpha)$ $(\ell ip(X, \alpha))$ may contain the constant functions only. In general, when $1 < \alpha$, $Lip(X, \alpha)$ may contain the nonconstant functions. for example, if d is the discrete metric on X then for a fixed $a \in X$, the function , $f(x) = d(x, a)$ is an element of $Lip(X, \alpha)$ which is a nonconstant function.

In most parts of this paper X is usually a compact subset of $\mathbb{C}^n$, so we shall concern ourselves with the algebras $Lip(X, \alpha)$ (respectively, $\ell ip(X, \alpha)$) only when $0 < \alpha \le 1$ (respectively, $0 < \alpha < 1$). In this case the algebras contain the family of functions $f(x) = d(x, a)$, as a runs over X, which separates the points of X. Hence $Lip(X,\alpha)$ for $\alpha \leq 1$ and $\ell ip(X,\alpha)$ for $\alpha < 1$, are Banach function algebras on the compact metric space X. Since these Lipschitz algebras are self-adjoint and separate the points of X, they are uniformly dense in $C(X)$, by the Stone-Weierstrass Theorem. Hence they are natural Banach function algebras on X, that is, their maximal ideal spaces coincide with X. For the proof one can either follow the same process as the proof of the naturality of $C(X)$, by using the fact that these algebras are inverse-closed, or apply the main theorem in [11].

Note that the function $f(x) = d^\alpha(x,a)$ is an element of $Lip(X,\alpha)$ but $\lim_{x\to a} \frac{f(x)}{d^\alpha(x,a)} = 1$. Hence f does not belong to $\ell ip(X,\alpha)$ when X is infinite and so $\ell ip(X,\alpha)$ is a proper subalgebra of $Lip(X,\alpha)$ in this case. Moreover, for every $\alpha < 1$ we have the inclusion $Lip(X,1) \subseteq \ell ip(X,\alpha)$. In fact, $Lip(X,1)$ is dense in $\ell ip(X,\alpha)$. This result has been proved in [3], using the measure theory and duality, but it is also followed from a theorem due to Hedberg in [10], which is, in fact, a Stone-Weierstrass theorem for Lipschitz algebras of real-valued functions. However, this theorem does hold even for Lipschitz algebras of complex-valued functions with an extra condition as follows, which we need later for some approximation theorems.

Theorem 2.4 (Hedberg's Theorem) Let (X,d) be a compact metric space, and $0 < \alpha < 1$. Let A be a self-adjoint subalgebra of $\ell ip(X,\alpha)$ which separates the points of X and contains the constant functions. Then A is dense in $\ell ip(X,\alpha)$ if for every $a \in X$ there are numbers M_a and δ_a such that for every $\delta \leq \delta_a$ there is an $f \in A$ with $f(a) = 1$, $f(x) = 0$ on $S_a(\delta) = \{x \in X : d(x,a) = \delta\}$, and $\sup_{y,z\in B_a(\delta)} \frac{|f(y)-f(z)|}{d^\alpha(y,z)} < \frac{M_a}{\delta^\alpha}$, where $B_a(\delta) = \{x \in X : d(x,a) \leq \delta\}$.

For the proof one can set $A_1 = \{f \in A : f \text{ is real-valued}\}$ and see that A_1 is a subalgebra of $\mathrm{Re}\ell ip(X,\alpha)$, the real Lipschitz algebra, which has all properties of Hedberg's theorem in the real case. Hence A_1 is dense in $\mathrm{Re}\ell ip(X,\alpha)$. Now by considering the fact that A is self-adjoint and the real and imaginary parts of Lipschitz functions are also Lipschitz functions, it is easy to show that qA is dense in $\ell ip(X,\alpha)$.

Corollary 2.5 Since $A = Lip(X,1)$ is a self-adjoint subalgebra of $\ell ip(X,\alpha)$ and for every $a \in X$ and arbitrary positive δ the function $g(x) = 1- \frac{d(x,a)}{\delta}$ is an element of $Lip(X,1)$ which satisfies the conditions of Hedberg's theorem, it follows that $Lip(X,1)$ is dense in $\ell ip(X,\alpha)$ for every $\alpha < 1$.

Now we discuss the approximation problem in $Lip = Lip(X,\alpha)$ for $\alpha \leq 1$ and in $\ell ip = \ell ip(X,\alpha)$ for $\alpha < 1$, when X **is a compact subset of** $\mathbb{C}^n$. Since the polynomials and the rational functions with poles off X belong to $Lip(X,1)$, we can define the following subalgebras of $Lip(X,\alpha)$ and $\ell ip(X,\alpha)$. From now on we always consider $Lip(X,\alpha)$ for $\alpha \leq 1$ and $\ell ip(X,\alpha)$ for $\alpha < 1$.

Before presenting the next definition we prove the following useful result.

Theorem 2.6 Let A be a natural Banach function algebra on a compact subset X of $\mathbb{C}^n$, and suppose A contains the polynomials. Then every function which is analytic in a neighbourhood of X is an element of A.

proof. Since $M_A \cong X$, i.e. the maximal ideal space of A is homeomorphic with X, the joint spectrum of the coordinate function is precisely the set X. If f is a function analytic in a neighbourhood of X, then by the Functional Calculus Theorem [8; III, 4.5], there exists $g \in A$ such that $\hat{g} = f(\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n)$ on $M_A \cong X$, where $\hat{z}_k$ is the Gelfand transform of the coordinate function z_k $(1 \leq k \leq n)$. Therefore $g = f$ on X and hence $f \in A$.

By the above theorem the Lipschitz algebras $Lip(X,\alpha)$ and $\ell ip(X,\alpha)$

contain any function which is analytic in some neighbourhood of X.

Definition 2.7 The subalgebra of Lip (ℓip) which is generated by the polynomials in $z_1, z_2, \ldots, z_n$, by the rational functions with poles off X, or by the analytic functions in some neighbourhood of X, is denoted by Lip_P (ℓip_P), by Lip_R (ℓip_R), or by Lip_H (ℓip_H), respectively.

Definition 2.8 The subalgebra of Lip (ℓip) which is generated by those elements of Lip (ℓip) which are analytic in the interior of X, is denoted by Lip_A (ℓip_A).

These subalgebras are all Banach function algebras on X and they satisfy the following inclusions:

$$Lip_P \subseteq Lip_R \subseteq Lip_H \subseteq Lip_A \subseteq Lip,$$

$$\ell ip_P \subseteq \ell ip_R \subseteq \ell ip_H \subseteq \ell ip_A \subseteq \ell ip.$$

Note that continuous functions on X, which are analytic in the interior of X, may not be in $Lip(X, \alpha)$. Since for every $f \in Lip(X, \alpha)$, $\|f\|_X \leq \|f\|_\alpha$, it is easy to see that $Lip_A(X, \alpha) = A(X) \cap Lip(X, \alpha)$ and $\ell ip_A(X, \alpha) = A(X) \cap \ell ip(X, \alpha)$, where $A(X)$ is the algebra of continuous functions on X which are analytic in the interior of X. Hence $Lip_A(X, \alpha) = Lip(X, \alpha)$ *or* $\ell ip_A(X, \alpha) = \ell ip(X, \alpha)$, if and only if, X has empty interior.

Now we investigate the maximal ideal spaces of these subalgebras, and the equality among some of them. For this purpose, we consider the standard uniform algebras $P(X), R(X)$ and $H(X)$ which are the uniform closures of polynomials, rational functions with poles off X and functions which are analytic in a neighbourhood of X, respectively. For the subalgebra B of $C(X)$, we denote its maximal ideal space by M_B and its uniform closure by $\overline{B}$. Clearly $\overline{Lip_P} = \overline{\ell ip_P} = P(X)$, $\overline{Lip_R} = \overline{\ell ip_R} = R(X)$, and $\overline{Lip_H} = \overline{\ell ip_H} = H(X)$.

Also for every $f \in Lip$ we have $\|f^n\|_\alpha \leq \|f\|_X^n + np_\alpha(f)\|f\|_X^{n-1}$, and so $\|\hat{f}\| \leq \|f\|_X$, where $\hat{f}$ is the Gelfand transform of f. Thus as a consequence of the Theorem in [11], we obtain the following results:

Theorem 2.9

I. $M_{Lip_P} = M_{\ell ip_P} \cong \hat{X}$, where $\hat{X}$ is the polynomial convex hull of X.

II. $M_{Lip_R} = M_{\ell ip_R} \cong$ *R-hull* (X), where *R-hull* (X) is the rational convex hull of X.

III. $M_{Lip_H} = M_{\ell ip_H} = M_{H(X)}$.

IV. $\hat{X} =$ *R-hull* (X) if and only if $Lip_P = Lip_R$.

V. If *R-hull* $(X) = X$ then $Lip_R = Lip_H$.

Note that whenever $A(X) = H(X)$ we have $\overline{Lip_A}(X, \alpha) = \overline{\ell ip_A}(X, \alpha) = A(X)$ and so $M_{Lip_A} = M_{\ell ip_A} = M_{H(X)}$. But in general these equalities may not be satisfied. However, if X is an arbitrary compact plane set then $\ell ip_A(X, \alpha)$ and $Lip_A(X, \alpha)$ are natural Banach function algebras on X [20,16, or1], but when X is an arbitrary compact subset of $\mathbb{C}^n$ $(1 < n)$ the maximal ideal spaces of Lip_A and ℓip_A are not known yet. For further results on some other subalgebras of $\ell ip(X, \alpha)$ one can refer to [31].

To establish some results about the approximation problem we need to define the algebra of continuously differentiable functions.

Definition 2.10 Let X be a perfect compact plane set, a complex-valued function f on X is called *differentiable* at $z_0 \in X$, if

$$f'(z_0) = \lim_{\substack{z \in X \\ z \to z_0}} \frac{f(z) - f(z_0)}{z - z_0}$$

exists, and f is called differentiable on X if it is differentiable at each point of X. The complex algebra of complex-valued functions with derivatives of all orders on X is denoted by $D^{\infty}(X)$. Let $D^n(X)$ be the algebra of functions with continuous nth derivatives on X. For $f \in D^n(X)$, we define the norm by

$$p_n(f) = \| f \| = \sum_{k=0}^{n} \frac{\| f^{(k)} \|_X}{k!},$$

This norm is actually an algebra norm on $D^n(X)$. It is interesting to see that $D^{\infty}(X) = \bigcap_{n=1}^{\infty} D^n(X)$.

Definition 2.11 Let $M = \{M_k\}_{k=0}^{\infty}$ be a sequence of positive numbers such that

$$M_0 = 1 \quad and \quad \frac{M_k}{M_r . M_{k-r}} \geq \binom{k}{r} \qquad r = 0, 1, \ldots, k.$$

The algebra of infinitely differentiable functions f on X such that

$$\sum_{k=0}^{\infty} \frac{\| f^{(k)} \|_X}{M_k} < \infty$$

is denoted by *D(X,M)*. For convenience, we regard $D^n(X)$ as being an algebra of the type $D(X, M)$ by setting $M_r = r!$ $(r = 0, 1, \ldots, n)$ and $\frac{1}{M_r} = 0$ $(r = n + 1, n + 2, \ldots)$.

From now on whenever we refer to $M = \{M_k\}$ we mean this sequence satisfies the above conditions.

Now we introduce the type of compact sets which we shall consider next.

Definition 2.12 Let X be a compact plane set which is connected by rectifiable arcs, and suppose $\delta(z, w)$ is the geodesic metric on X, the infimum of the lengths of the arcs joining z and w.

(i) X is called regular if for each $z_0 \in X$ there exists a constant C such that for all $z \in X$, $\delta(z, z_0) \leq C|z - z_0|$.

(ii) X is called uniformly regular if there exists a constant C such that for all $z, w \in X$, $\delta(z, w) \leq C|z - w|$.

If X is a finite union of regular sets then for each $z_0 \in X$ there exists a constant C such that for every $z \in X$ and any $f \in D^1(X)$,

$$| f(z) - f(z_0)| \leq C \ |z - z_0| \ (\|f\|_X + \|f'\|_X).$$

This inequality implies that $D^1(X)$ is complete under the norm $\|f\|_1 = \|f\|_X + \|f'\|_X$ [6,14]. It is also interesting to note that the above condition is, in fact, a necessary and sufficient condition for the completeness of $D^1(X)$. To see this, let $D^1(X)$ be complete and define another norm on $D^1(X)$ by

$$\|f\| = \|f\|_X + \|f'\| + \sup_{\substack{z \in X \\ z \neq z_0}} \frac{| f(z) - f(z_0)|}{|z - z_0|} \qquad (f \in D^1(X)) ,$$

where z_0 is a fixed point in X. Then $D^1(X)$ is also à Banach function algebra on X under this new norm. Thus there exists a constant C such that for all $f \in D^1(X)$ and for every $z \in X$

$$| f(z) - f(z_0)| \leq C \ |z - z_0| \ (\|f\|_X + \|f'\|_X).$$

Note that the completeness of $D^1(X)$ is, in fact, equivalent to the completeness of *D(X,M)*. Moreover, $D^1(X) \subseteq Lip(X, 1)$, and the norm $\|.\|_1$ of $D^1(X)$ and the Lipschitz norm of $Lip(X, 1)$ are equivalent on $D^1(X)$. Note that $D^1(X)$ is a proper closed subalgebra of $Lip(X, 1)$ for every uniformly regular set X.

From now on we assume that X is a perfect compact plane set such that $D^1(X)$ is complete, unless otherwise specified.

As mentioned before for each $\alpha < 1$, $Lip(X, 1)$ is dense in $\ell ip(X, \alpha)$. A question which arises here is that under what conditions $D^1(X)$ is also dense in $\ell ip(X,\alpha)$. If $D^1(X)$ is dense in $\ell ip(X,\alpha)$ then $\ell ip_A(X,\alpha)= \ell ip(X,\alpha)$ and so $intX = \phi$. Therefore, a necessary condition for the density of $D^1(X)$ in $\ell ip(X,\alpha)$ is that $intX = \phi$. But the inverse is not true. However, we shall prove the interesting result that $D^1(X)$ is actually dense in $\ell ip_R(X,\alpha)$.

For this purpose and the next general result we make some preliminaries. Let X be a compact subset of the complex plane and define $V = \{(z,w) \in X \times X : z \neq w\}$ and $W = X \cup V$. We denote the Banach space of bounded continuous functions on W by $C^b(W)$, the closed subspace of functions which vanish at infinity by $C_0(W)$, and the space of all regular complex Borel measures on W by $M(W)$. It follows from the Riesz representation theorem that the dual space of $C_0(W)$ is isometrically isomorphic to $M(W)$. As noted in [3, §3] we extend any function $g \in C(X)$ to a function $\tilde{g}$ on W by defining

$$\begin{cases} \tilde{g}(w) = g(w) & (w \in X), \\ \tilde{g}(\zeta,\eta) = \frac{g(\zeta)-g(\eta)}{|\zeta-\eta|^\alpha} & ((\zeta,\eta) \in V). \end{cases}$$

The map $g \mapsto \tilde{g}$, $Lip(X,\alpha) \to C^b(W)$ is a linear isometry and the image of $\ell ip(X,\alpha)$ is contained in $C_0(W)$.

Theorem 2.13 [13] Let X be a compact subset of the complex plane and suppose $f : X \to \mathbb{C}$ can be extended to a function having continuous partial derivatives in some neighbourhood of X. If either of the following conditions is satisfied, then $f \in \ell ip_R(X,\alpha)$ for every $\alpha < 1$.

(i) $m(X) = 0$, where m is planar measure.

(ii) $\dfrac{\partial f}{\partial \bar{z}} = 0$ on X.

Proof. Suppose f satisfies the hypotheses of the theorem but $f \notin \ell ip_R(X,\alpha)$. By the Hahn-Banach theorem there exists a continuous linear functional ϕ on $\ell ip(X,\alpha)$ for which $\phi(f) \neq 0$ and $\phi = 0$ on $\ell ip_R(X,\alpha)$. Again, by the Hahn-Banach theorem, ϕ has a norm preserving extension to a continuous linear functional ψ on $C_0(W)$, and so there exists a $\mu \in M(W)$ with $\phi(g) = \psi(\tilde{g}) = \int_W \tilde{g}d\mu$, for every $g \in \ell ip(X,\alpha)$.

By hypotheses there exists a function F defined and with continuous partial derivatives on $\mathbb{R}^2$, having compact support, such that $F|_X = f$. By Green's formula

$$F(w) = -\frac{1}{\pi}\iint_{\mathbb{R}^2} \frac{\partial F}{\partial \bar{z}} \cdot \frac{1}{z-w}dxdy \qquad (w \in \mathbb{C}) \ .$$

So we have

$$\begin{aligned}\phi(f) &= \int_W \tilde{f}d\mu = \int_X f(w)d\mu(w) + \int_V \frac{f(\zeta)-f(\eta)}{|\zeta-\eta|^\alpha}d\mu(\zeta,\eta)\\ &= \int_X \left(\frac{-1}{\pi}\iint_{\mathbb{R}^2}\frac{\partial F}{\partial \bar{z}}\cdot\frac{1}{z-w}dxdy\right)d\mu(w)\\ &\quad + \int_V \left(\frac{-1}{\pi}\iint_{\mathbb{R}^2}\frac{\partial F}{\partial \bar{z}}\cdot\frac{\frac{1}{z-\zeta}-\frac{1}{z-\eta}}{|\zeta-\eta|^\alpha}dxdy\right)d\mu(\zeta,\eta).\end{aligned}$$

Since F has a compact support in $\mathbb{R}^2$, there exists a closed disk Δ in $\mathbb{C}$ such that $\frac{\partial F}{\partial \bar{z}} = 0$ on $\mathbb{C}\backslash\Delta$. To show the boundedness of the integral

$$\iint_\Delta \left|\frac{\partial F}{\partial \bar{z}}\right| \frac{|\zeta-\eta|^{1-\alpha}}{|z-\zeta|.|z-\eta|}dxdy$$

over V, we split Δ into three parts, D_ζ,D_η and $\Delta\backslash(D_\zeta \cup D_\eta)$, where D_ζ and D_η are disks with centres ζ and η, respectively, having radius $\frac{|\zeta-\eta|}{2}$. Clearly the integrals on D_ζ and D_η are bounded. If $z \in \Delta\backslash(D_\zeta\cup D_\eta)$ then $|\zeta-\eta| \leq 2|z-\zeta|$ and $|\zeta-\eta| \leq 2|z-\eta|$ and so $|\zeta-\eta|^{1-\alpha} \leq 2^{1-\alpha}|z-\zeta|^{\frac{1-\alpha}{2}}.|z-\eta|^{\frac{1-\alpha}{2}}$. Hence

$$\iint_{\Delta\backslash(D_\zeta\cup D_\eta)} \left|\frac{\partial F}{\partial \bar{z}}\right| \frac{|\zeta-\eta|^{1-\alpha}}{|z-\zeta||z-\eta|}dxdy \leq$$

$$2^{1-\alpha}\iint_{\Delta\backslash(D_\zeta\cup D_\eta)} \left|\frac{\partial F}{\partial \bar{z}}\right| \frac{dxdy}{(|z-\zeta|.|z-\eta|)^{\frac{1+\alpha}{2}}}.$$

Now, using Hölder's inequality the last integral is bounded over V since $\alpha < 1$. Therefore by Fubini's theorem and hypotheses we have

$$\begin{aligned}\phi(f) &= -\frac{1}{\pi}\iint_{\Delta}\frac{\partial F}{\partial \bar{z}}\left(\int_X \frac{d\mu(w)}{z-w} + \int_V \frac{\frac{1}{z-\zeta}-\frac{1}{z-\eta}}{|\zeta-\eta|^{\alpha}}d\mu(\zeta,\eta)\right)dxdy \\ &= -\frac{1}{\pi}\iint_{\Delta\setminus X}\frac{\partial F}{\partial \bar{z}}\phi\left(\frac{1}{z-w}\right)dxdy = 0,\end{aligned}$$

which contradicts our previous assumption. Therefore $f \in \ell ip_R(X,\alpha)$ and this completes the proof of the theorem.

As an interesting result on the approximation problem we conclude an extension of the Hartogs-Rosenthal Theorem, which states that $R(X) = C(X)$, when X has planar measure zero.

Theorem 2.14 [13] If X is a compact subset of the complex plane with planar measure zero, then $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$.

Proof. By Theorem 2.13, $\bar{z} \in \ell ip_R(X,\alpha)$, where $\bar{z}$ is the complex conjugate of the coordinate function z. For every $a \in X$ and $\delta > 0$ the function $f(z) = 1 - \frac{|z-a|^2}{\delta^2}$ is an element of $\ell ip_R(X,\alpha)$. Hence $A = \ell ip_R(X,\alpha)$ satisfies the conditions of Hedberg's theorem, and so A is dense in $\ell ip(X,\alpha)$, that is, $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$.

Remark 2.15 The above theorem is not true for the algebra $Lip(X,\alpha)$, when $\alpha < 1$ and X is a compact subset of $\mathbb{C}^n$. Note that $Lip_R(X,1)$ is also a proper closed subalgebra of $Lip(X,1)$ for every uniformly regular set X.

For similar results, which are related to the second part of Theorem 2.13 and Theorem 2.14, one can refer to [23, 24, 25].

Now we extend a result due to Dales and Davie which states that $D^1(X) \subseteq R(X)$ if X is uniformly regular [6, Lemma 1.5].

Theorem 2.16 [13] For every perfect compact plane set X, $\overline{D^1}(X) = \ell ip_R(X,\alpha)$.

Proof. It is sufficient to prove that $D^1(X) \subseteq \ell ip_R(X,\alpha)$. Let $f \in D^1(X)$. By the Whitney's extension theorem [21], there exists a function F, defined and with continuous partial derivatives on $\mathbb{R}^2$, having a compact support, such that $F|_X = f$, $\frac{\partial F}{\partial \bar{z}}|_X = 0$ and $\frac{\partial F}{\partial z}|_X = f'$. Hence f satisfies the hypotheses of Theorem 2.13, and so $f \in \ell ip_R(X,\alpha)$.

Remark 2.17 [13] If $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$ then $R(X) = C(X)$. But there exists a Swiss cheese X_0,with empty interior and positive planar measure, such that $R(X_0)$ is different from $C(X_0)$, [8; II,1]. Hence $\ell ip_R(X_0,\alpha)$ is different from $\ell ip(X_0,\alpha)$. Since every Swiss cheese is a perfect compact plane set we conclude that $D^1(X_0)$ is not dense in $\ell ip(X_0,\alpha)$.

Finally we extend the result of Theorem 2.16 to obtain an extension of a similar result in uniform algebras for Lipschitz algebras.

Theorem 2.18 [13] Let X be a compact subset of $\mathbb{C}^n$, and let X_j $(1 \leq j \leq n)$ denote the projection of X onto the jth coordinate plane.
(i) If $\ell ip_P(X_j,\alpha) = \ell ip(X_j,\alpha)$ for all j $(1 \leq j \leq n)$, then $\ell ip_P(X,\alpha) = \ell ip(X,\alpha)$.
(ii) If $\ell ip_R(X_j,\alpha) = \ell ip(X_j,\alpha)$ for all j $(1 \leq j \leq n)$, then $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$.

Proof. (i) Clearly the complex conjugate of every polynomial in $z_1, z_2, \ldots, z_n$ is an element of $\ell ip_P(X,\alpha)$. Hence the function $f(z) = 1 - \frac{\|z-a\|^2}{\delta^2}$ is an element of $\ell ip_P(X,\alpha)$ for every $a \in X$ and $\delta > 0$. Therefore by the Hedberg's theorem, $\ell ip_P(X,\alpha)$ is dense in $\ell ip(X,\alpha)$, in other words, $\ell ip_P(X,\alpha) = \ell ip(X,\alpha)$.

(ii) Note that every rational function r is of the form $r(z) = \frac{p(z)}{q(z)}$, where $p(z)$ and $q(z)$ are polynomials in $z_1, z_2, \ldots, z_n$ and $q(z) \neq 0$ on *R-hull* (X), which is the maximal ideal space of $\ell ip_R(X,\alpha)$. Hence $\bar{q}(z)$, the complex conjugate of $q(z)$, as an element of $\ell ip_R(X,\alpha)$, is invertible in $\ell ip_R(X,\alpha)$ and so $\bar{r}(z)$ is an element of $\ell ip_R(X,\alpha)$. The result now follows as in (i).

As a consequence we get the following.

Corollary 2.19 If for each j $(1 \leq j \leq n)$, X_j has planar measure zero, then $\ell ip_R(X,\alpha) = \ell ip(X,\alpha)$. If moreover $\hat{X}_j = X_j$ for all j $(1 \leq j \leq n)$ then $\ell ip_P(X,\alpha) = \ell ip(X,\alpha)$.

Now we are going to discuss the approximation problem for the Lipschitz algebras of differentiable functions on perfect compact plane sets. For further details one can refer to [19].

Definition 2.20 The algebra of functions f on X whose derivatives up to order n exist and for each k $(0 \leq k \leq n)$, $f^{(k)} \in Lip(X,\alpha)$ is denoted by $Lip^n(X,\alpha)$. The algebra $\ell ip^n(X,\alpha)$ is defined in a similar way.

We now equip these algebras with the norm

$$p_n(f) = \| f \| = \sum_{k=0}^{n} \frac{\| f^{(k)} \|_\alpha}{k!} = \sum_{k=0}^{n} \frac{\| f^{(k)} \|_X + p_\alpha(f^{(k)})}{k!}.$$

The above algebras have similar properties to $D^n(X)$. Clearly for each n ≥ 1, $Lip^n(X,1) \subseteq \ell ip^n(X,\alpha) \subseteq Lip^n(X,\alpha) \subseteq D^n(X) \subseteq D^1(X)$. It is also known that $D^1(X) \subseteq R(X)$ [6].

Definition 2.21 The algebra of functions f with derivatives of all orders for which $f^{(k)} \in Lip(X,\alpha)$ $(f^{(k)} \in \ell ip(X,\alpha))$ for all k is denoted by $Lip^\infty(X,\alpha)$ $(\ell ip^\infty(X,\alpha))$.

It is interesting to see that $D^\infty(X) = \bigcap_{n=1}^{\infty} D^n(X)$, $Lip^\infty(X,\alpha) = \bigcap_{n=1}^{\infty} Lip^n(X,\alpha)$ and $\ell ip^\infty(X,\alpha) = \bigcap_{n=1}^{\infty} \ell ip^n(X,\alpha)$. Note that these algebras are not Banach algebras under any norm [7, 12, or 22].

We now introduce certain subalgebras of $Lip^\infty(X,\alpha)$ and $\ell ip^\infty(X,\alpha)$.

Definition 2.22 Let

$$Lip(X,M,\alpha)=\{f\in Lip^{\infty}(X,\alpha):\sum_{k=0}^{\infty}\frac{\|f^{(k)}\|_{\alpha}}{M_k}<\infty\},$$

$$\ell ip(X,M,\alpha)=\{f\in \ell ip^{\infty}(X,\alpha):\sum_{k=0}^{\infty}\frac{\|f^{(k)}\|_{\alpha}}{M_k}<\infty\},$$

and for f in $Lip(X,M,\alpha)$ or in $\ell ip(X,M,\alpha)$ let $\|f\|=\sum_{k=0}^{\infty}\frac{\|f^{(k)}\|_{\alpha}}{M_k}$

Remark 2.23 The above algebras have similar properties to $D(X,M)$. For convenience, we regard $Lip^n(X,\alpha)$ and $\ell ip^n(X,\alpha)$ as being algebras of the type $Lip(X,M,\alpha)$ and $\ell ip(X,M,\alpha)$, respectively, by setting $M_k=k!$ $(k=0,1,\dots,n)$ and $1/M_k=0$ $(k=n+1,\dots)$. Clearly for $\alpha<1$, $Lip(X,M,1)\subseteq \ell ip(X,M,\alpha)\subseteq Lip(X,M,\alpha)\subseteq D(X,M)$, $\ell ip(X,M,\alpha)\subseteq \ell ip^{n+1}(X,\alpha)\subseteq \ell ip^{n}(X,\alpha)$, $Lip(X,M,\alpha)\subseteq Lip^{n+1}(X,\alpha)$ $\subseteq Lip^{n}(X,\alpha)$. Since $Lip(X,\alpha)$ and $\ell ip(X,\alpha)$ are normed function algebras and $\|.\|_{\alpha}$ is an algebra norm on them, for every $f,g\in Lip(X,M,\alpha)$ $(f,g\in \ell ip(X,M,\alpha)$ we have

$$\|f.g\|=\sum_{k=0}^{\infty}\frac{\|(f.g)^{(k)}\|_{\alpha}}{M_k}\leq\sum_{k=0}^{\infty}\frac{1}{M_k}\sum_{r=0}^{k}\binom{k}{r}\|f^{(r)}.g^{(k-r)}\|_{\alpha}$$
$$\leq\sum_{k=0}^{\infty}\sum_{r=0}^{k}\frac{\|f^{(r)}\|_{\alpha}}{M_r}.\frac{\|g^{(k-r)}\|_{\alpha}}{M_{k-r}}=(\sum_{k=0}^{\infty}\frac{\|f^{(k)}\|_{\alpha}}{M_k})(\sum_{k=0}^{\infty}\frac{\|g^{(k)}\|_{\alpha}}{M_k})=\|f\|.\|g\|.$$

Hence $f.g\in Lip(X,M,\alpha)$ ($f.g\in \ell ip(X,M,\alpha)$ and so any of the above algebras are normed function algebras on X, with respect to the above norm. The completeness of $D^1(X)$ or, equivalently, *D(X, M),* implies that $Lip(X,M,\alpha)$ and $\ell ip(X,M,\alpha)$ are Banach function algebras on X.

Remark 2.24 When X is a uniformly regular set we have $D^{n+1}(X)\subseteq$ $Lip^n(X,1)\subseteq \ell ip^n(X,\alpha)$ and for $\alpha<1$ $\ell ip(X,M,\alpha)=Lip(X,M,\alpha)$ if $\frac{1}{M_k}$ is different from zero for infinitely many k.

Remark 2.25 When $P_k=\sqrt[k]{\frac{M_k}{k!}}\to\infty$ as k → ∞, $D(X,M)$ and likewise $Lip(X,M,1)$ contain all rational functions with poles off X.

From now on we assume that X **is a perfect compact plane set such that** $D^1(X)$ **is complete**, unless otherwise specified.

Now we introduce subalgebras of $Lip(X, M, \alpha)$ and $\ell ip(X, M, \alpha)$.

Definition 2.26 The closed subalgebra of $Lip(X, M, \alpha)$ $(\ell ip(X, M, \alpha))$ which is generated by the polynomials, by the rational functions with poles off X that belong to $Lip(X, M, \alpha)$ $(\ell ip(X, M, \alpha))$, or by those functions of $Lip(X, M, \alpha)$ $(\ell ip(X, M, \alpha))$ which are analytic in some neighbourhood of X, is denoted by $Lip_P(X, M, \alpha)$ $(\ell ip_P(X, M, \alpha))$, $Lip_R(X, M, \alpha)$ $(\ell ip_R(X, M, \alpha))$ or $Lip_H(X, M, \alpha)$ $(\ell ip_H(X, M, \alpha))$, respectively.

Remark 2.27 Clearly $Lip_P(X, M, \alpha)$ is uniformly dense in $P(X)$, the uniform closure of polynomials. Moreover, when $P_k = \sqrt[k]{M_k/k!} \to \infty$ as $k \to \infty$, we have $R_0(X) \subseteq Lip(X, M, 1)$, so $Lip(X, M, 1)$ and hence $\ell ip(X, M, \alpha)$ and $Lip(X, M, \alpha)$ are uniformly dense in $R(X)$. In particular, $Lip_R(X, M, \alpha)$ is uniformly dense in $R(X)$. Note that, when $P_k \to \infty$ as $k \to \infty$, $Lip_R(X, M, \alpha)$ is generated by the all rational functions with poles off X, and hence it is a natural Banach function algebra on X. So by the Theorem in [11] we have $M_{Lip_P(X,M,\alpha)} \cong M_{P(X)} \cong \hat{X}$, where M_A is the maximal ideal space of the algebra A. Thus when $P_k \to \infty$ as $k \to \infty$, $Lip_P(X, M, \alpha) = Lip_R(X, M, \alpha)$ if and only if $\hat{X} = X$. Also when $P_k \to \infty$ as $k \to \infty$, by the Functional Calculus Theorem [8; 3.4.5], $Lip_R(X, M, \alpha)$ contains all analytic functions in a neighbourhood of X, and so $Lip_R(X, M, \alpha) = Lip_H(X, M, \alpha)$.

Theorem 2.28 [14] For each $n \geq 0$, $\ell ip^n(X, \alpha)$ and $Lip^n(X, \alpha)$ are natural Banach function algebras on X.

Proof. Since the algebras $\ell ip(X, \alpha)$ and $Lip(X, \alpha)$ are uniformly dense in $C(X)$ and for $n \geq 1$, $\ell ip^n(X, \alpha)$ and $Lip^n(X, \alpha)$ are uniformly dense in $R(X)$, by the naturality of $C(X)$ and $R(X)$, we can show that for each $f \in Lip^n(X,\alpha)$ we have $\|\hat{f}\| \leq \|f\|_X$, where $\hat{f}$ is the Gelfand transform of f. Thus by the Theorem in [11] the result follows.

It is clear that for $\alpha < 1$ $\ell ip_P(X,M,\alpha) = Lip_P(X,M,\alpha)$, $\ell ip_R(X,M,\alpha) = Lip_R(X,M,\alpha)$ *and* $\ell ip_H(X,M,\alpha) = Lip_H(X,M,\alpha)$.

Theorem 2.29 If X is uniformly regular and $\ell ip_P(X,\alpha) = \ell ip_A(X,\alpha)$ then $\ell ip_P^n(X,\alpha) = \ell ip^n(X,\alpha)$ for all $n \geq 1$ and $\alpha < 1$.
Proof. As we know if $f \in D^1(X)$ then $p_\alpha(f) \leq Cd^{1-\alpha}\|f'\|_X$, where $d = diam(X)$. Now let $n \geq 1$ and $f \in \ell ip^n(X,\alpha)$. Since $f^{(n)} \in \ell ip_A(X,\alpha) = \ell ip_P(X,\alpha)$, for every $\epsilon > 0$ there exists a polynomial P_0 such that

$$\|f^{(n)} - P_0\|_{\ell ip(X,\alpha)} = \|f^{(n)} - P_0\|_X + p_\alpha(f^{(n)} - P_0) < \epsilon.$$

Let z_0 be a fixed point in X and P_1 be the antiderivative of P_0 with the initial condition $P_1(z_0) = f^{(n-1)}(z_0)$.
Since $f^{(n-1)} - P_1 \in \ell ip^1(X,\alpha) \subseteq D^1(X)$ we have

$$p_\alpha(f^{(n-1)} - P_1) \leq Cd^{1-\alpha}\|f^{(n)} - P_0\|_X < Cd^{1-\alpha}\epsilon.$$

Continuing in this manner, we obtain polynomials $P_2, P_3, \dots, P_n$ such that $P_k' = P_{k-1}$, $P_k(z_0) = f^{(n-k)}(z_0)$, $\|f^{(n-k)} - P_k\|_X < C^k d^k \epsilon$, and $p_\alpha(f^{(n-k)} - P_k) \leq C^k d^{k-\alpha}\epsilon$, for $k = 1, 2, \dots, n$. Clearly $P_n^{(k)} = P_{n-k}$ on X and

$$\|f - P_n\|_{\ell ip^n(X,\alpha)} \leq \sum_{k=0}^{n-1} \frac{C^{n-k}d^{n-k}\epsilon + C^{n-k}d^{n-k-\alpha}\epsilon}{k!} + \frac{\epsilon}{n!} = \lambda\epsilon,$$

for some constant λ. Hence $f \in \ell ip_P^n(X,\alpha)$.

Now we investigate rational approximation on circles and annuli. We note that when X is uniformly regular and $\alpha < 1$, then $\ell ip(X,M,\alpha) = Lip(X,M,\alpha)$ if $1/M_k \neq 0$ for infinitely many k.

Theorem 2.30 [14] If $T = \{ z \in C : |z - z_0| = R\}$ then $\ell ip_R(T,M,\alpha) = \ell ip(T,M,\alpha)$.

Proof. We assume that $z_0 = 0$ and $R = 1$. Let $f \in \ell ip(T, M, \alpha)$ and $\sum_{-\infty}^{\infty} a_j z^j$ be the Fourier series generated by f, where $a_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(e^{i\theta})e^{-ij\theta}d\theta$. The Cesaro means of this series are

$$\sigma_n(e^{i\theta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(e^{it})K_n(\theta - t)dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(e^{i(\theta-t)})K_n(t)dt,$$

where $K_n(t)$ is the Fejer kernel. It is known that σ_n is a rational function on T with the only pole $z = 0$, and $\|\sigma_n - f\|_T \to 0$ as $n \to \infty$. Since for each $k \geq 0$, $f^{(k)}$ is continuous on T we have

$$\sigma_n^{(k)}(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ikt} f^{(k)}(ze^{-it})K_n(t)dt \qquad (z \in T),$$

and so $\|\sigma_n^{(k)}\|_T \leq \|f^{(k)}\|_T$.

On the other hand,

$$\frac{|\sigma_n^{(k)}(z) - \sigma_n^{(k)}(w)|}{|z-w|^\alpha} \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|f^{(k)}(ze^{-it}) - f^{(k)}(we^{-it})|}{|ze^{-it} - we^{-it}|^\alpha} K_n(t)dt \leq p_\alpha(f^{(k)}).$$

Hence $p_\alpha(\sigma_n^{(k)}) \leq p_\alpha(f^{(k)})$ and so $\sigma_n \in \ell ip_R(T, M, \alpha)$.

Now we prove that $\|\sigma_n - f\|_{\ell ip(T,M,\alpha)} \to 0$ as $n \to \infty$. Since

$$\|\sigma_n - f\|_{\ell ip(T,M,\alpha)} \leq 2\|f\|_{\ell ip(T,M,\alpha)},$$

by the dominated convergence theorem, it is enough to show that for each $k \geq 0$, $\|\sigma_n^{(k)} - f^{(k)}\|_T + p_\alpha(\sigma_n^{(k)} - f^{(k)}) \to 0$ as $n \to \infty$.

By the uniform continuity of each $f^{(k)}$ on T we have $\|\sigma_n^{(k)} - f^{(k)}\|_T \to 0$ as $n \to \infty$.

Since $f^{(k)} \in \ell ip(T, \alpha)$, for $\epsilon > 0$ there exists $\delta > 0$ such that for all $z, w \in T$, if $0 < |z - w| < \delta$ then $|f^{(k)}(z) - f^{(k)}(w)|/|z-w|^\alpha < \epsilon/2$. Let $k \geq 0$ and $z, w \in T$, $(z \neq w)$. If $|z - w| < \delta$, then

$$\frac{|\sigma_n^{(k)}(z) - f^{(k)}(z) - \sigma_n^{(k)}(w) + f^{(k)}(w)|}{|z-w|^\alpha} < \epsilon.$$

If $|z-w| \geq \delta$ and n is large enough, then

$$\frac{|\sigma_n^{(k)}(z) - f^{(k)}(z) - \sigma_n^{(k)}(w) + f^{(k)}(w)|}{|z-w|^\alpha} \leq \frac{2\|\sigma_n^{(k)} - f^{(k)}\|_T}{\delta^\alpha} < \epsilon.$$

Hence $p_\alpha(\sigma_n^{(k)} - f^{(k)}) \to 0$ as $n \to \infty$. This completes the proof of the theorem.

Note that the following results are not satisfied when $\ell ip(X, M, \alpha)$ reduces to $\ell ip(X, \alpha)$.

Theorem 2.31 [14] If $X = \{z \in C : r \leq |z - z_0| \leq R\}$, where $0 < r < R$, then $\ell ip_R(X, M, \alpha) = \ell ip(X, M, \alpha)$. But in these cases we have $\ell ip_R(X, \alpha) = \ell ip_A(X, \alpha)$.

Proof. Without loss of generality we can assume that $z_0 = 0$. Let $f \in \ell ip(X, M, \alpha)$. Since f is analytic in $r < |z| < R$ it has a Laurent series of the form $f(z) = \sum_{-\infty}^{\infty} a_j z^j$ on $r < |z| < R$, where $a_j = (2\pi\rho^j)^{-1} \int_{-\pi}^{\pi} e^{-ijt} f(\rho e^{it})dt$, for $r < \rho < R$. The Cesaro means of the Laurent series of f is

$$\sigma_n(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(ze^{-it}) K_n(t)dt \qquad (r < |z| < R),$$

where $K_n(t)$ is the Fejer kernel. Clearly for each $k \geq 0$ we have

$$\sigma_n^{(k)}(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} f^{(k)}(ze^{-it}) K_n(t)dt \qquad (r < |z| < R),$$

and so $|\sigma_n^{(k)}(z)| \leq \|f^{(k)}\|_X$ for $r < |z| < R$. Since σ_n is a rational function with the only pole $z = 0$, $\sigma_n^{(k)}$ is analytic in $r \leq |z| \leq R$. Therefore the above inequality holds for all z in $r \leq |z| \leq R$. Hence $\|\sigma_n^{(k)}\|_X \leq \|f^{(k)}\|_X$ and $p_\alpha(\sigma_n^{(k)}) \leq p_\alpha(f^{(k)})$ for all $k \geq 0$ and for every positive integer n, and so $\sigma_n \in \ell ip_R(X, M, \alpha)$. Now we can proceed exactly the same as in the proof of theorem 2.30 to show that $\|\sigma_n - f\|_{\ell ip(X,M,\alpha)} \to 0$ as $n \to \infty$. Therefore $f \in \ell ip_R(X, M, \alpha)$ and this completes the proof of the theorem .

If $r \to 0$, the above theorem implies the following result.

Corollary 2.32 [14] If $X = \{z : |z| \leq R\}$ then $\ell ip_P(X, M, \alpha) = \ell ip(X, M, \alpha)$.

Theorem 2.33 [14] Let X be a regular set for which there exists $z_0 \in X$ such that for $0 \leq \beta < 1$, $\beta(z - z_0) + z_0 \in intX$ for all $z \in X$. Or, equivalently, the segment $[z_0, z)$ is contained in the interior of X for all $z \in X$. If $P_k \to \infty$ as $k \to \infty$, then $\ell ip_P(X, M, \alpha) = \ell ip(X, M, \alpha)$.
Proof. Clearly X is star-shaped and so it is polynomially convex. Thus

$$\ell ip_P(X, M, \alpha) = \ell ip_R(X, M, \alpha) = \ell ip_H(X, M, \alpha).$$

Without loss of generality we can assume that $z_0 = 0$. By the hypothesis for each positive integer n and every $z \in X$, $r_n z \in intX$, where $r_n = n/(n+1)$. Let $f \in \ell ip(X, M, \alpha)$ and define the sequence $\{f_n\}$ on X by $f_n(z) = f(r_n z)$. Each f_n is analytic in a neighbourhood of X and so $f_n \in \ell ip_H(X, M, \alpha)$. Moreover for each $k \geq 0$, $f_n^{(k)}(z) = r_n^k f^{(k)}(r_n z)$ and so $\|f_n^{(k)}\|_X \leq \|f^{(k)}\|_X$, $p_\alpha(f_n^{(k)}) \leq p_\alpha(f^{(k)})$ for all $k \geq 0$ and every n. By the uniform continuity of each $f^{(k)}$ on X , $lim_{n \to \infty}\|f_n^{(k)} - f^{(k)}\|_X = 0$. Since $f^{(k)} \in \ell ip(X, \alpha)$ for each $k \geq 0$, $p_\alpha(f_n^{(k)} - f^{(k)}) \to 0$ as $n \to \infty$. Consequently by the dominated convergence theorem $\|f_n - f\|_{\ell ip(X,M,\alpha)} \to 0$ as $n \to \infty$, and so $f \in \ell ip_H(X, M, \alpha)$.

Corollary 2.34 If X is a compact convex set with non-empty interior and $P_k \to \infty$ as $k \to \infty$, then $\ell ip_P(X, M, \alpha) = \ell ip(X, M, \alpha)$.

3. Real Lipschitz Algebras

We first present a brief general description of real uniform algebras and real Banach function algebras.

3.1 Real Uniform (Function) Algebras

Definition 3.1.1 Let X be a topological space. A map $\tau : X \longrightarrow X$ is called a *topological involution* on X if τ is continuous and $\tau(\tau(x)) = x$ for all $x \in X$.

Note that, τ is a homeomorphism of X onto X.

Examples 3.1.2 (i) The identity map on a topological space X is a topological involution on X.

(ii) Let X be the closed interval $[a, b]$. The map $\tau : X \longrightarrow X$, defined by $\tau(x) = b - a - x$ is a topological involution on X.

(iii) Let X be a compact plane set such that X is symmetric about the real axis. The map $\tau : X \longrightarrow X$, defined by $\tau(z) = \bar{z}$, is a topological involution on X.

Let X be a compact Hausdorff space and τ be a topological involution on X. Let

$$C(X,\tau) := \{f \in C(X) : f \circ \tau = \bar{f}\ \}.$$

Then $C(X,\tau)$ is a uniformly closed real subalgebra of $C(X)$ which contains the constant function 1.

Remark 3.1.3 If τ is the identity map on a compact Hausdorff space X then $C(X,\tau) = C_{\mathbb{R}}(X)$, the real uniform algebra of continuous realy valued functions on X. On the other hand, for a given compact Hausdorff space X, we suppose $Y = X \times \{0,1\}$ and the map $\tau : Y \longrightarrow Y$ defines by $\tau(x, j) = (x, (j + 1)$ modulo 2). Identifying X with $X \times \{0\}$, X can be treated as a subspace of Y . Then every f in $C(X)$ can be extended uniquely to Y by requiring that the extension belongs to $C(Y,\tau)$. In fact, the map $\psi : C(X) \longrightarrow C(Y,\tau)$, defined by $\psi(f)(x,0) = f(x)$, $\psi(f)(x,1) = \overline{f(x)}$, is an isometrical isomorphism of $C(X)$, as a real Banach algebra, onto $C(Y,\tau)$. Thus $C(X,\tau)$ is a more general object than $C_{\mathbb{R}}(X)$ and $C(X)$.

Theorem 3.1.4 [18; Theorem 1.3.5]. Let X be a compact Hausdorff space and τ be a topological involution on X. Define the map $\sigma : C(X) \longrightarrow C(X)$ by $\sigma(f) = \bar{f} \circ \tau$. Then:

(i) σ is an algebra involution on $C(X)$ and

$$C(X, \tau) = \{f \in C(X) : \sigma(f) = f\}.$$

(ii) $C(X) = C(X, \tau) \oplus iC(X, \tau)$, that is, every h in $C(X)$ can be expressed uniquely as $f + ig$ with f, g in $C(X, \tau)$.

(iii) σ is an isometry.

Definition 3.1.5 Let X be a compact Hausdorff space and τ be a topological involution on X. The algebra involution σ on $C(X)$ defined by $\sigma(f) = \bar{f} \circ \tau$, is called *the algebra involution on $C(X)$ induced by τ.*

Theorem 3.1.6 [18; Theorem 1.3.5(v)] Let X be a compact Hausdorff space. If σ is an algebra involution on $C(X)$ then:

(i) There exists a topological involution on X such that $\sigma(f) = \bar{f} \circ \tau$, for all f in $C(X)$.

(ii) σ is an isometry.

Theorem 3.1.7 [18; Lemma 1.3.7] Let X be a compact Hausdorff space and τ be a topological involution on X. Let $x, y \in X$ and $x \neq y$.

(i) If $y = \tau(x)$, then there exists a function f in $C(X, \tau)$ such that $f(x) = i$ and $f(y) = -i$.

(ii) If $y \neq \tau(x)$, then there exists a function f in $C(X, \tau)$ such that $f(x) = 1$ and $f(y) = 0$.

In particular, $C(X,\tau)$ separates the points of X.

Definition 3.1.8 Let X be a compact Hausdorff space and τ be a topological involution on X. A *real uniform (function) algebra* on (X,τ), is a uniformly closed subalgebra of $C(X,\tau)$ which contains 1 and separates the points of X.

It is clear that $C(X,\tau)$ is a uniform (function) algebra on (X,τ).

Theorem 3.1.9 [18; Theorem 1.3.20]. Let X be a compact Hausdorff space, τ be a topological involution on X and σ be the algebra involution induced by τ on $C(X)$. Let A be a real subspace of $C(X,\tau)$ and define

$$B := \{f + ig : f,g \in A\}.$$

Then

(i) $\sigma(B) = B$ and $A = \{h \in B : \sigma(h) = h\} = B \cap C(X,\tau)$.

(ii) $B = A \oplus iA$, that is, every $h \in B$ can be written uniquely as $f + ig$ with f,g in A.

(iii) For $f,g \in A$,

$$\max\{\|f\|_X, \|g\|_X\} \leq \|f + ig\|_X \leq \|f\|_X + \|g\|_X.$$

(iv) B is uniformly closed (self-adjoint, separates the points of X, contains 1) if and only if A has the same property, respectively. Moreover, B is a complex algebra if and only if A is a real algebra.

(v) If A is a real uniform algebra on (X,τ) then the map $\alpha : M_A \longrightarrow M_B$, defined by

$$\alpha(\phi)(f + ig) = \phi(f) + i\phi(g),$$

is a homeomorphism of M_A *onto* M_B

By the above theorem, if τ is a topological involution on a compact Hausdorff space X then

$$M_{C(X,\tau)} = \{e_x : x \in X\},$$

where e_x is the evaluation character on X at $x \in X$ [18; Corollary 1.3.21].

In some problems, we need to know when a given complex uniform algebra on (X, τ) can be viewed as a complexification of a real uniform algebra. The following theorem gives a criterion.

Theorem 3.1.10 [18; Theorem 1.3.22] Let X be a compact Hausdorff space, τ be a topological involution on X and σ be the algebra involution induced by τ on $C(X)$. If B is a complex uniform algebra on X with $\sigma(B) = B$ and suppose $A = \{h \in B : \sigma(h) = h\}$ then A is a real uniform algebra on (X, τ) and B can be regarded as the complexification of A.

We now state the Stone-Weierstrass theorem for real subalgebras of $C(X, \tau)$.

Theorem 3.1.11 [18; Proposition 1.2, Corollary 2.1.14] Let X be a compact Hausdorff space and τ be a topological involution on X. If A is a self-adjoint real subalgebra of $C(X, \tau)$ containing 1 and separating the points of X then $\bar{A} = C(X, \tau)$, where $\bar{A}$ is the uniform closure of A.

3.2 Real Banach Function Algebras

Real uniform (function) algebras were first defined and studied by S.H. Kulkarni and B.V. Limaye in 1981. In general, a complex uniform (function) algebra on a compact Hausdorff space X may not be a real uniform (function) algebra on X, with the topological involution τ as the identity function. But, it is interesting to note that every complex uniform (function) algebra can be regarded as a real uniform (function) algebra on a compact Hausdorff space with a suitable topological involution τ. Hence, the class of real uniform algebras is larger than that of complex uniform algebras.

Now we extend the notion of real uniform (function) algebras, by introducing a larger class, which is called the *real Banach function algebras.*

We will show that every complex Banach function algebra can be viewed as a real Banach function algebra with some topological involution τ. Hence the class of real Banach function algebras is larger than the class of complex Banach function algebras. We are going to extend some general properties of the complex Banach function algebras for the real Banach function algebras. Then the real Lipschitz algebras of complex functions are introduced and some properties of this interesting class of real Banach function algebras are discussed.

Let X be a compact Hausdorff space and τ be a topological involution on X. In this section, we first define a real Banach function algebra on (X, τ). We prove that for each complex Banach function algebra $(B, \|\cdot\|)$ on X, there exist a compact Hausdorff space Y, a topological involution τ on Y and a real Banach function algebra $(A, \||\cdot\||)$ on (Y, τ) such that $(B, \|\cdot\|)$, as a real Banach algebra, is isometrically isomorphic to $(A, \||\cdot\||)$. Then we extend the Theorems 3.1.9 and 3.1.10 to real Banach function algebras. Finally we give some results concerning the carrier space (maximal ideal space) of real Banach function algebras.

Definition 3.2.1 Let X be a compact Hausdorff space and τ be a topological involution on X. A real Banach function algebra on (X, τ) is a real subalgebra A of $C(X, \tau)$ which contains the constant function 1 and separates the points of X and there exists an algebra norm $\|\cdot\|$ on A such that $(A, \|\cdot\|)$ is a real Banach algebra.

Note that if the norm of a real Banach function algebra A on (X, τ)

is the same as the uniform norm on X, then A is called a real uniform (function) algebra on (X,τ).

Let X be a compact Hausdorff space and τ be a topological involution on X. We recall that the map $\sigma : C(X) \longrightarrow C(X)$ which is defined by $\sigma(f) = \bar{f} \circ \tau$, is an isometric algebra involution on $C(X)$ and is called the algebra involution on $C(X)$ induced by τ.

Theorem 3.2.2 [2; Theorem 1.1] Let X be a compact Hausdorff space, τ be a topological involution on X and σ be the algebra involution on $C(X)$ induced by τ. If $(B, \|\cdot\|)$ is a complex Banach function algebra on X such that $\sigma(B) = B$ and $A := \{h \in B : \sigma(h) = h\}$, then

(i) A is a real subalgebra of B and $A = B \cap C(X,\tau)$.

(ii) Every h in B can be expressed uniquely as $f + ig$ with $f, g \in A$.

(iii) There exists a constant $C \geq 1$ such that $\|\sigma(h)\| \leq C\|h\|$ for every $h \in B$ and $\max\{\|f\|, \|g\|\} \leq C\|f + ig\|$ for all $f, g \in A$.

(iv) $(A, \|\cdot\|)$ is a real Banach function algebra on (X,τ).

(v) For ϕ in M_A, define $\alpha(\phi)(f + ig) := \phi(f) + i\phi(g)$ for $f, g \in A$. Then $\alpha(\phi) \in M_B$ and the map α is a homeomorphism from M_A onto M_B. In particular, A is natural if and only if B is natural.

Remark 3.2.3 In the above theorem, if B is a complex uniform algebra on X then A is a real uniform algebra on (X,τ) and we can choose $C = 1$. Therefore, the above theorem is a generalization of a similar result for the real uniform algebras.

Theorem 3.2.4 [2; Theorem 1.2] Let X be a compact Hausdorff space, τ be a topological involution on X and σ be the algebra involution on $C(X)$ induced by τ. Let $(A, \|\cdot\|)$ be a real Banach function algebra on (X,τ) and define $B := \{f + ig : f, g \in A\}$. Then:

(i) $\sigma(B) = B$ and $A = \{h \in B : \sigma(h) = h\} = B \cap C(X,\tau)$.

(ii) B is complex subalgebra of $C(X)$ and $B = A \oplus iA$.

(iii) There is an algebra norm $|||\cdot|||$ on B such that $\|f\| = |||f|||$ for all $f \in A$ and

$$\max\{\|f\|, \|g\|\} \leq |||f + ig||| \leq 2\max\{\|f\|, \|g\|\} \quad (f, g \in A).$$

(iv) $(B, |||\cdot|||)$ is a complex Banach function algebra on X.

(v) For $\phi \in M_A$ define $(\alpha(\phi))(f + ig) = \phi(f) + i\phi(g)$ for $f, g \in A$. Then $\alpha(\phi) \in M_B$ and α is a homeomorphism from M_A onto M_B. In particular, A is natural if and only if B is natural.

Remark 3.2.5 In the above theorem, if A is a real uniform algebra on (X, τ) then B is a complex uniform algebra on X. Therefore, this result is a generalization of a similar result for the real uniform algebras.

3.3 Real Lipschitz Algebras of Complex Functions

In this section we first define C-quasicontraction and in particular d-isometric topological involutions on a compact metric space (X, d). Then we show that if σ is the algebra involution on $C(X)$ induced by a C-quasicontraction topological involution τ on a compact metric space (X, d) and $\alpha \in (0, 1]$ then $\sigma(Lip(X, \alpha)) = Lip(X, \alpha)$ and $\sigma(\ell ip(X, \alpha)) = lip(X, \alpha)$. By using this result we define the real Lipschitz algebras of complex functions $Lip(X, \tau, \alpha)$ and $lip(X, \tau, \alpha)$, which are real Banach function algebras on (X, τ). Next, we show that for a compact metric space (X, d), there exists a compact metric space (Y, ρ) and a ρ-isometric topological involution τ on Y such that $Lip(X, \alpha)$ (respectively, $lip(X, \alpha)$), as a real Banach algebra, is isometrically isomorphic to $Lip(Y, \tau, \alpha)$ (respectively, $lip(Y, \tau, \alpha)$) for $\alpha \in (0, 1]$ (respectively, $\alpha \in (0, 1)$). Finally, we study the approximation problem of the real Lipschitz algebras $Lip(X, \tau, \alpha)$ and $lip(X, \tau, \alpha)$ and certain subalgebras of these algebras.

Definition 3.3.1 *Let (X,d) be a compact metric space and τ be a topological involution on X.*

(i) τ is called a *C- quasicontraction on X* if there exists a constant $C>0$ such that $d(\tau(x),\tau(y))\leq Cd(x,y)$, for all x,y in X.

(ii) τ is called a *d-isometric topological involution on X*if $d(\tau(x),\tau(y))=d(x,y)$ for all x,y in X.

Remark 3.3.2 If τ is a C-quasicontraction topological involution on a compact metric space (X,d) then it is easy to see that $C\geq 1$. Moreover if $C=1$ then τ is d-isometric.

Lemma 3.3.3 [2; Lemma 2.4] Let (X,d) be a compact metric space and τ be a C-quasicontraction topological involution on X and σ be the algebra involution induced by τ on $C(X)$. Then

(i) For every $\alpha\in(0,1]$ and $f\in Lip(X,\alpha)$, $p_\alpha(\sigma(f))\leq C^\alpha p_\alpha(f)$.

(ii) For every $\alpha\in(0,1]$, $\sigma(Lip(X,\alpha))=Lip(X,\alpha)$.

(iii) For every $\alpha\in(0,1]$, $\sigma(lip(X,\alpha))=lip(X,\alpha)$.

(iv) For every $\alpha\in(0,1]$ and every $f\in Lip(X,\alpha)$, $\|\sigma(f)\|_\alpha\leq C^\alpha\|f\|_\alpha$.

(v) If τ is a d-isometric topological involution then for every $\alpha\in(0,1]$ and $f\in Lip(X,\alpha)$, $p_\alpha(\sigma(f))=p_\alpha(f)$ and $\|\sigma(f)\|_\alpha=\|f\|_\alpha$. Hence σ is an isometric algebra involution on $Lip(X,\alpha)$.

Proof. (i) Let $\alpha\in(0,1]$ and $f\in Lip(X,\alpha)$. Then

$$\begin{aligned}
p_\alpha(\sigma(f)) &= \sup\{\frac{|\sigma(f)(x)-\sigma(f)(y)|}{d^\alpha(x,y)}:x,y\in X,x\neq y\}\\
&= \sup\{\frac{|f(\tau(x))-f(\tau(y))|}{d^\alpha(x,y)}:x,y\in X,x\neq y\}\\
&\leq \sup\{\frac{|f(\tau(x))-f(\tau(y))|}{C^{-\alpha}d^\alpha(\tau(x),\tau(y))}:x,y\in X,x\neq y\}\\
&= C^\alpha p_\alpha(f).
\end{aligned}$$

(ii) By (i), $\sigma(Lip(X,\alpha))\subseteq Lip(X,\alpha)$. Since $\sigma(\sigma(f))=f$ for every $f\in Lip(X,\alpha)$, $\sigma(Lip(X,\alpha))=Lip(X,\alpha)$.

(iii) Let $\alpha \in (0,1)$ and $f \in lip(X,\alpha)$. Let $\varepsilon > 0$ be given. Since $f \in lip(X,\alpha)$, there is a $\delta_1 > 0$ such that $\frac{|f(x)-f(y)|}{d^\alpha(x,y)} < C^{-\alpha}\varepsilon$, whenever $0 < d(x,y) < \delta_1$. Set $\delta = C^{-1}\delta_1$ and suppose $x, y \in X$ such that $0 < d(x,y) < \delta$. Since τ is a C-quasicontraction topological involution on (X,d), therefore

$$0 < d(\tau(x),\tau(y)) \leq Cd(x,y) < C\delta = \delta_1.$$

Hence $\frac{|f(\tau(x))-f(\tau(y))|}{d^\alpha(\tau(x),\tau(y))} < C^{-\alpha}\varepsilon$. This implies that $\frac{|\sigma(f)(x)-\sigma(f)(y)|}{d^\alpha(x,y)} < \varepsilon$ and so $\sigma(f) \in lip(X,\alpha)$. Therefore, $\sigma(lip(X,\alpha)) \subseteq lip(X,\alpha)$ and since σ is an algebra involution on $C(X)$, $\sigma(lip(X,\alpha)) = lip(X,\alpha)$.

(iv) Let $\alpha \in (0,1]$ and $f \in Lip(X,\alpha)$. By (i) and $C \geq 1$, we have

$$\begin{aligned}\|\sigma(f)\|_\alpha &= \|\sigma(f)\|_X + p_\alpha(\sigma(f))\\ &= \|f\|_X + p_\alpha(\sigma(f))\\ &\leq C^\alpha\|f\|_X + C^\alpha p_\alpha(f) = C^\alpha\|f\|_\alpha.\end{aligned}$$

(v) If τ is a d-isometric topological involution on X then $C = 1$ and so $p_\alpha(\sigma(f)) = p_\alpha(f)$. Hence $\|\sigma(f)\|_\alpha = \|f\|_\alpha$. □

Remark 3.3.4 Let (X,d) be a compact metric space and $\alpha \in (0,1]$. If $B = Lip(X,\alpha)$ or $lip(X,\alpha)$, then the map $f \mapsto \bar{f}$ is an isometric algebra involution on B.

Theorem 3.3.5 [2; Theorem 2.7] Let (X,d) be a compact metric space, τ be a C-quasicontraction topological involution on X and σ be the algebra involution on $C(X)$ induced by τ. We define

$$Lip(X,\tau,\alpha) := \{h \in Lip(X,\alpha) : \sigma(h) = h\} \quad (\alpha \in (0,1]),$$

and

$$lip(X,\tau,\alpha) := \{h \in lip(X,\alpha) : \sigma(h) = h\} \quad (\alpha \in (0,1)).$$

If $A = Lip(X, \tau, \alpha)$ and $B = Lip(X, \alpha)$ ($A = lip(X, \tau, \alpha)$ and $B = lip(X, \alpha)$, respectively), then

(i) $B = A \oplus iA$.

(ii) For every $f, g \in A$, $\max\{\|f\|_\alpha, \|g\|_\alpha\} \leq C^\alpha \|f + ig\|_\alpha$.

(iii) $(A, \|\cdot\|_\alpha)$ is a real Banach function algebra on (X, τ).

(iv) A is self-adjoint and $\bar{A} = C(X, \tau)$, where $\bar{A}$ is the uniform closure of A.

(v) $M_A = \{e_x : x \in X\}$.

Proof. Since $(B, \|\cdot\|_\alpha)$ is a complex Banach function algebra on X and by Lemma 3.3.3, we have $\sigma(B) = B$, therefore (i) and (iii) hold by Theorem 3.2.2.

Since σ is an isometric involution on $C(X)$ and by Lemma 3.3.3, $\|\sigma(h)\|_\alpha \leq C^\alpha \|h\|_\alpha$ for each $h \in B$, we can easily show that (ii) holds.

Since A is self-adjoint, $\bar{A} = C(X, \tau)$. Thus (iv) holds. Since B is a natural Banach function algebra on X, A is also natural by Theorem 3.2.2(v) and hence (v) holds.□

Theorem 3.3.6 [2; Theorem 2.8] Let (X, d) be a compact metric space and $\alpha \in (0, 1]$. Then there exist a compact metric space (Y, ρ), a ρ-isometric topological involution τ on Y such that the complex Lipschitz algebra $Lip(X, \alpha)$ (respectively, $lip(X, \alpha)$), regarded as a real Banach algebra, is isometrically isomorphic to the real Lipschitz algebra $Lip(Y, \tau, \alpha)$ (respectively, $lip(Y, \tau, \alpha)$).

Proof. Let $Y := X \times \{0, 1\}$. We define the map $\rho : Y \times Y \longrightarrow \mathbb{R}$ by

$$\rho((x, j), (y, k)) = \max\{d(x, y), |j - k|\} \quad (x, y \in X, j, k \in \{0, 1\}).$$

Then ρ is a metric on Y and the topology induced by ρ on Y coincides to the product topology on Y. Now, we define the map $\tau : Y \longrightarrow Y$ by

$$\tau(x, 0) = (x, 1), \quad \tau(x, 1) = (x, 0) \quad (x \in X).$$

It is easy to see that τ is a ρ-isomeric topological involution on Y. Now, we define the map $\psi : Lip(X,\alpha) \longrightarrow Lip(Y,\tau,\alpha)$ by

$$\psi(f)(x,0) = f(x), \quad \psi(f)(x,1) = \bar{f}(x) \quad (x \in X).$$

We can easily show that ψ is an isometrically homomorphism from $(Lip(X,\alpha), \|\cdot\|_\alpha)$, as a real Banach algebra, into $(Lip(Y,\tau,\alpha), \|\cdot\|_\alpha)$. Now, let $g \in Lip(Y,\tau,\alpha)$ and define the map $f : X \longrightarrow \mathbb{C}$ by

$$f(x) = g(x,0).$$

It is easy to see that $f \in Lip(X,\alpha)$ and $\psi(f) = g$. Therefore, ψ is an isometrically isomorphism from $(Lip(X,\alpha), \|\cdot\|_\alpha)$, as a real Banach algebra, onto $Lip(Y,\tau,\alpha)$. Moreover, if $f \in lip(X,\alpha)$ then we conclude that $\psi(f) \in lip(Y,\tau,\alpha)$ and if $g \in lip(Y,\tau,\alpha)$, $\psi^{-1}(g) \in lip(X,\alpha)$. Therefore $\psi|_{lip(X,\alpha)}$ is an isometrically isomorphism from $(lip(X,\alpha), \|\cdot\|_\alpha)$, as a real Banach algebra, onto $lip(Y,\tau,\alpha), \|\cdot\|_\alpha)$. □

Remark 3.3.7 If (X,d) is a compact metric space and τ is the identity map on X then τ is a d-isometric topological invol ution on X and

$$Lip(X,\tau,\alpha) = Lip_{\mathbb{R}}(X,\alpha) = \{f \in Lip(X,\alpha) : f \text{ is real-valued}\},$$
$$lip(X,\tau,\alpha) = lip_{\mathbb{R}}(X,\alpha) = \{f \in lip(X,\alpha) : f \text{ is real-valued}\}.$$

Therefore the class of real Lipschitz algebras of complex functions $Lip(X,\tau,\alpha)$ ($lip(X,\tau,\alpha)$, respectively) is larger than the class of the real Lipschitz algebras $Lip_{\mathbb{R}}(X,\alpha)$ ($lip_{\mathbb{R}}(X,\alpha)$, respectively). Also, by Theorem 3.3.6 we conclude that the class of real Lipschitz algebras of complex functions $Lip(X,\tau,\alpha)$ ($lip(X,\tau,\alpha)$, respectively) is larger than the class of complex Lipschitz algebras $Lip(X,\alpha)$ ($lip(X,\alpha)$, respectively), as real Banach algebras.

Theorem 3.3.8 [2; Theorem 2.9] Let (X, d) be a compact metric space and τ be a C-quasicontraction topological involution on X. Then

(i) If $0 < \alpha < \beta \leq 1$, then $Lip(X, \tau, \beta)$ is a subalgebra of $lip(X, \tau, \alpha)$.

(ii) If $0 < \alpha < 1$ and X is an infinite set then $lip(X, \tau, \alpha)$ is a proper subalgebra of $Lip(X, \tau, \alpha)$.

Proof. (i) Since $Lip(X, \tau, \beta) = Lip(X, \beta) \cap C(X, \tau)$, $lip(X, \tau, \alpha) = lip(X, \alpha) \cap C(X, \tau)$, $Lip(X, \beta)$ is a complex subalgebra of $lip(X, \alpha)$ and $C(X, \tau)$ is a real subalgebra of $C(X)$, we conclude that $Lip(X, \tau, \beta)$ is a real subalgebra of $lip(X, \tau, \alpha)$.

(ii) Since $lip(X, \alpha)$ is a proper subalgebra of $Lip(X, \alpha)$, there exists $h \in Lip(X, \alpha) \backslash lip(X, \alpha)$. Since $h \in Lip(X, \alpha)$ and by Theorem 3.3.5 we have

$$Lip(X, \alpha) = Lip(X, \tau, \alpha) \oplus iLip(X, \tau, \alpha),$$

thus h can be expressed uniquely as $h = f + ig$ with f, g in $Lip(X, \tau, \alpha)$. Since $h \notin lip(X, \alpha)$ and by Theorem 3.3.5 we have

$$lip(X, \alpha) = lip(X, \tau, \alpha) \oplus ilip(X, \tau, \alpha),$$

thus $f \notin lip(X, \tau, \alpha)$ or $g \notin lip(X, \tau, \alpha)$ and the proof (ii) is now complete. □

Let (X, d) be a compact metric space and take $\alpha \in (0, 1)$. A type of Stone-Weierstrass theorem in real Lipschitz algebra $lip_{\mathbb{R}}(X, \alpha)$ was first given by L. I. Hedberg in 1969 [10; Theorem 1]. Let τ be a C-quasicontraction topological involution on X. We now extend the Hedberg's theorem in real Lipschitz algebra of complex functions $lip(X, \tau, \alpha)$, without using the complexification technique.

Theorem 3.3.9 (Hedberg's theorem in real Lipschitz algebras of complex functions) [2; Theorem 2.10] Let (X, d) be a compact metric space, τ be a C-quasicontraction topological involution on X, and take $\alpha \in$ (0,1). Let A be a self-adjoint real subalgebra of $lip(X, \tau, \alpha)$ which separates the points of X and contains the real-valued constant functions on X. Then A is dense in $lip(X,\tau,\alpha)$ if for every $a \in X$, there are positive numbers M_a and δ_a such that for every $\delta \leq \delta_a$, there is a $f \in A$, with $f(a) = 1$, $f(x) = 0$ on $S_\delta(a) = \{x \in X : d(x,a) = \delta\}$, and

$$\sup\left\{\frac{|f(y)-f(z)|}{d^\alpha(y,z)} : y,z \in B_\delta(a), y \neq z\right\} < \frac{M_a}{\delta^\alpha},$$

where $B_\delta(a) = \{x \in X : d(x,a) \leq \delta\}$.

Notice that if (X,d) is a compact metric space and τ is the identity map on X then $Lip(X,\tau,\alpha) = Lip_{\mathbb{R}}(X,\alpha)$ and $lip(X,\tau,\alpha) = lip_{\mathbb{R}}(X,\alpha)$. Hence, the theorem 3.3.9 is a generalization of the Hedberg's theorem in real Lipschitz algebra $lip_{\mathbb{R}}(X,\alpha)$. Note that T. G. Honary and H. Mahyar stated the Hedberg's theorem in complex Lipschitz algebras $lip(X,\alpha)$ in [13].

As a consequence of Theorem 3.3.9, the Hedberg's theorem in complex Lipschitz algebra $lip(X,\alpha)$ is obtained, which is stated as follows:

Corollary 3.3.10 [2; Corollary 2.11] Let (X,d) be a compact metric space and take $\alpha \in (0,1)$. Let B be a self-adjoint complex subalgebra of $lip(X,\alpha)$ which separates the points of X and contains the complex-valued constant functions on X. Then B is dense in $lip(X,\alpha)$ if for $a \in X$, there are positive numbers M_a and δ_a such that for every $\delta \leq \delta_a$, there is a $f \in B$ such that $f(a) = 1$, $f(x) = 0$ on $S_\delta(a) = \{x \in X : d(x,a) = \delta\}$, and

$$\sup\left\{\frac{|f(x)-f(y)|}{d^\alpha(x,y)} : y,z \in B_a(\delta), y \neq z\right\} < \frac{M_a}{\delta^\alpha},$$

where $B_\delta(a) = \{x \in X : d(x,a) \leq \delta\}$.

As an application of Theorem 3.3.9, we prove that the real Lipschitz algebra $Lip(X,\tau,1)$ is dense in $lip(X,\tau,\alpha)$ for $\alpha \in$ (0,1) without using the complexification technique.

Corollary 3.3.11 [2; Corollary 2.12] Let (X, d) be a compact metric space and τ be a C-quasicontraction topological involution on X. If $\alpha \in (0,1)$ then $Lip(X, \tau, 1)$ is dense in $(lip(X, \tau, \alpha), \|.\|_\alpha)$.

4. Fréchet Lipschitz Algebras

First we introduce some elementary definitions and known results concerning Fréchet algebras. For further details refer to [12, 27, 15 or 7].

Definition 4.1 A *Fréchet algebra* is an algebra which is a complete metrizable topological vector space and has a neighbourhood basis $(V_n)_{n \in \mathbb{N}}$ of zero consisting of convex sets V_n such that $V_n.V_n \subseteq V_n$ for all $n \in \mathbb{N}$. We always assume that the Fréchet algebra contains the unit 1.

The topology of a Fréchet algebra A can be generated by a sequence $(p_n)_{n \in \mathbb{N}}$ of separating submultiplicative seminorms $(p_n(f.g) \leq p_n(f).p_n(g)$ for each $n \in \mathbb{N}$ and every $f, g \in A$ such that $p_n(f) \leq p_{n+1}(f)$ for all $n \in \mathbb{N}$ and $f \in A$. If A has a unit, p_n can be chosen such that $p_n(1) = 1$ [9]. We may denote the Fréchet algebra A with the above generating sequence of seminorms by $(A, (p_n))$. Clearly a sequence (f_k) in A converges to $f \in A$ if and only if $p_n(f_k - f) \underset{k \to \infty}{\longrightarrow} 0$ for each $n \in \mathbb{N}$.

Definition 4.2 The *spectrum* of a Fréchet algebra A is the set of all non-zero continuous complex-valued homomorphisms on A and it is denoted by M_A. We endow M_A with the Gelfand topology.

It can be shown that a complex homomorphism φ on the commutative Fréchet algebra $(A, (p_n))$ is continuous if and only if there exists $n \in \mathbb{N}$ such that $|\varphi(f)| \leq p_n(f)$ for all $f \in A$ [9; Remark 3.2.2(ii)].

Definition 4.3 The *radical* of a Fréchet algebra A, denoted by $rad(A)$, is the intersection of all maximal left (right) ideals in A. The Fréchet algebra A is called *semisimple* if $rad(A) = \{0\}$.

If A is a commutative Fréchet algebra then $rad(A)$ is the intersection of all closed maximal ideals; i.e. $rad(A) = \bigcap_{\varphi \in M_A} \ker \varphi$ [9; Proposition 8.1.2].

The following interesting result for Fréchet algebras, which is due to Carpenter [5], is similar to the uniqueness theorem of Johnson [17] for semisimple Banach algebras.

Theorem 4.5 If A is a commutative semisimple Fréchet algebra then A has a unique topology as a Fréchet algebra.

Definition 4.6 A Hausdorff space X is called *hemicompact* if there exists a sequence $(X_n)_{n \in \mathbb{N}}$ of compact subsets of X such that $X_n \subseteq X_{n+1}$ for each $n \in \mathbb{N}$ and every compact subset K of X is contained in some X_n. Such sequence $(X_n)_{n \in \mathbb{N}}$ is called an admissible exhaustion of X.

If $(A, (p_n))$ is a Fréchet algebra and A_n is the completion of $A/\ker p_n$ with respect to the norm $p'_n(f + \ker p_n) = p_n(f)$, $f \in A$, then A_n is a Banach algebra.

Definition 4.7 Let X be a hemicompact space, and let A be a subalgebra of $C(X)$ which contains the constants and separates the points of X. Then A is called a *Fréchet function algebra* on X if it is a Fréchet algebra with respect to some topology such that the evaluation homomorphisms are continuous on A, i.e. $\varphi_x \in M_A$ for all $x \in X$.

Clearly with the above definition every uniform Fréchet algebra is a Fréchet function algebra which is equipped with the compact-open topology.

It is easy to see that every Fréchet function algebra is semisimple. Moreover, every Banach function algebra on a compact Hausdorff space X is a Fréchet function algebra on X. For further details see [27].

Let X be a perfect compact plane set and for $n \in \mathbb{N}$, $B_n(X)$ be any of $D^n(X)$, $Lip^n(X, \alpha)$ or $\ell ip^n(X, \alpha)$. Suppose A is any of the algebras $D^\infty(X)$, $Lip^\infty(X, \alpha)$ or $\ell ip^\infty(X, \alpha)$ and $(p_n)_n$ is the corresponding sequence of algebraic norms defined on $D^n(X)$, $Lip^n(X,\alpha)$ *or* $\ell ip^n(X,\alpha)$, respectively, which was defined in the Definitions 2.10 and 2.20. We endow A with the metric topology defined by the sequence $(p_n)_n$ and denote it by $(A,(p_n))$. Then we get the following result, which can be found in [12] or [27]. Moreover, by the Carpenter's Theorem there exists a unique topology for each of these algebras as a Fréchet algebra.

Theorem 4.8 If X is a perfect compact plane set such that $D^n(X)$ is complete, then $(A,(p_n))$ is a Fréchet function algebra on X which is not a Banach algebra.

Theorem 4.9 If X is a uniformly regular space then

$$D^\infty(X) = Lip^\infty(X,1) = \ell ip^\infty(X,\alpha) = Lip^\infty(X,\alpha).$$

Proof. If $f \in D^{n+1}(X)$ then $f^{(k)} \in D^1(X)$ for $0 \le k \le n$. Since X is uniformly regular, $D^1(X)$ is complete and so there exists a constant M such that for all $z, w \in X$ and for each k $(0 \le k \le n)$,

$$|f^{(k)}(z) - f^{(k)}(w)| \le M|z-w| \parallel f^{(k+1)} \parallel_X .$$

Hence for all $k(0 \le k \le n)$,

$$p_1(f^{(k)}) = \sup_{\substack{z,w\in X \\ z\ne w}} \left| \frac{f^{(k)}(z) - f^{(k)}(w)|}{z-w} \right| \le M \parallel f^{(k+1)} \parallel_X < \infty,$$

and so for every $k(0 \le k \le n)$, $f^{(k)} \in Lip(X,1)$. Thus $f \in Lip^n(X,1)$ and it follows that $D^{n+1}(X) \subseteq Lip^n(X,1)$. It is easy to check that

$$Lip^n(X,1) \subseteq lip^n(X,\alpha) \subseteq Lip^n(X,\alpha) \subseteq D^n(X),$$

and so

$$D^\infty(X) \subseteq Lip^\infty(X,1) \subseteq \ell ip^\infty(X,\alpha) \subseteq Lip^\infty(X,\alpha) \subseteq D^\infty(X).$$

Hence the result follows.

Corollary 4.10 If X is uniformly regular then the metric topologies of $D^{\infty}(X)$, $Lip^{\infty}(X,1)$ and $Lip^{\infty}(X,\alpha)$ are equivalent.

Proof. It is immediate by the Carpenter's Theorem [5].

Now we introduce the Fréchet function algebras $FLip(X,\alpha)$ and $F\ell ip(X,\alpha)$ and some important subalgebras of them, where X is a hemicompact metric space. We also extend some known results and theorems about the Banach function algebras $Lip(X,\alpha)$ and $\ell ip(X,\alpha)$ to these Fréchet function algebras.

Let (X,d) be an arbitrary metric space (not necessarily compact) and $0<\alpha\leq 1$. Then, as before, we take $\mathrm{Lip}(X,\alpha)$ as *the space of all bounded Lipschitz functions of order α on X,* which is a Banach algebra under the same norm as defined in the Definition 2.3.

Now we define a new topology τ on $Lip(X,\alpha)$ under which it is an LMC-algebra so that when X is a hemicompact metric space, $(Lip(X,\alpha),\tau)$ is metrizable. For a compact subset K *of* X we define

$$p_{K}(f)=\|f\|_{K}+p_{\alpha}(f|_{K})\quad (f\in Lip(X,\alpha)).$$

Clearly p_{K} is a submultiplicative seminorm on $Lip(X,\alpha)$ so that the family $(\ p_{K})_{K}$ is a separating family of seminorms. Let τ be the topology on $Lip(X,\alpha)$ which is defined by the family $(p_{K})_{K}$ of seminorms. Then we have the following interesting result, which can be found in [27;2.4.1]

Theorem 4.11 Let (X,d) be a hemicompact metric space. Then $(Lip(X,\alpha),\tau)$ is a Fréchet algebra if and only if X is compact (in this case $(Lip(X,\alpha),\tau)$ is indeed a Banach function algebra).

Definition 4.12 For the hemicompact metric space (X,d) the completion of $Lip(X,\alpha)$ with respect to τ is denoted by $\mathrm{FLip}(X,\alpha)$ and it is called the Fréchet Lipschitz algebra of order α on X.

We can characterize the elements of $FLip(X,\alpha)$ as the continuous functions f (not necessarily bounded) such that $f|_K \in Lip(K,\alpha)$ for each compact subset K of X. To prove this result one can refer to [27; Theorem 2.4.3], which is stated partly in the following.

Theorem 4.13 Let (X,d) be an arbitrary metric space and let A be the algebra of all continuous functions f on X for which

$$p_K(f) = \|f\|_K + p_\alpha(f|_K)$$

is finite for each compact subset K of X. We endow A with the topology defined by the family $(p_K)_K$ of seminorms. Then

(i) $Lip(X,\alpha)$ is dense in A with respect to this topology.

(ii) $(A,(p_K))$ is a Fréchet algebra if and only if X is hemicompact.

(iii) $(A,(p_K))$ is a Banach algebra if and only if X is compact.

Remark 4.14 Note that Parts (i) and (iii) of the above theorem also imply Theorem 4.11. From now on we assume that X is a hemicompact metric space and (K_n) is an admissible exhaustion of X. The theorem shows that

$$FLip(X,\alpha) = \{f \in C(X) : f|_{K_n} \in Lip(K_n,\alpha), n \in \mathbb{N}\}.$$

Here it is not required to assume that $f \in C(X)$. Because X is a k-space [9] and so if f satisfies the second condition, which implies the continuity of f on each K_n, then it is continuous on X.

Another interesting property is that the algebra $FLip(X,\alpha)$ is dense in $C(X)$ with respect to the compact-open topology. To see this we argue as follows:

It is clear that for each n, $FLip(X,\alpha)|_{K_n} = Lip(K_n,\alpha)$ and $Lip(K_n,\alpha)$ is dense in $C(K_n)$. So if $f \in C(X)$ and $U=\{g \in C(X):\|g-f\|_{K_n}< \varepsilon\}$ is a neighbourhood of f in $C(X)$ for $n \in \mathbb{N}$, then there exists $g \in Lip(K_n, \alpha)$ with $\|f|_{K_n} - g\|_{K_n} < \varepsilon$. Now we can extend g to a $\tilde{g} \in Lip(X, \alpha) \subseteq FLip(X, \alpha)$ and conclude that $\tilde{g} \in U$.

Now as in the compact case we introduce some interesting subalgebras of $FLip(X, \alpha)$. As before X is assumed to be a hemicompact metric space with the admissible exhaustion (K_n) and $0 < \alpha \leq 1$.

Definition 4.15 For $0 < \alpha < 1$ we denote the set of all $f \in C(X)$ for which $f|_{K_n} \in \ell ip(K_n, \alpha)$ by $F\ell ip(X, \alpha)$.

Clearly $F\ell ip(X, \alpha)$ is a subalgebra of $FLip(X, \alpha)$ and it is easy to see that $F\ell ip(X, \alpha)$ is indeed a closed subalgebra of FLip(X, α). Obviously $F\ell ip(X, \alpha)$ is also a Frechet function algebra (Ff-algebra) on X.

The inclusion $Lip(K_n, 1) \subseteq \ell ip(K_n, \alpha)$, for each n, implies easily that $FLip(X, 1) \subseteq F\ell ip(X, \alpha)$. Likewise the compact case we can show that $FLip(X, 1)$ is dense in $F\ell ip(X, \alpha)$. More generally, we have the following result, which is found in [27].

Theorem 4.16 The algebra $Lip(X, 1)$ and hence $FLip(X, 1)$ is dense in $F\ell ip(X, \alpha)$, for each $\alpha < 1$.

Proof. Clearly $Lip(X, 1) \subseteq FLip(X, 1) \subseteq F\ell ip(X, \alpha)$. Let $f \in F\ell ip(X, \alpha)$ and $U = \{g \in F\ell ip(X, \alpha) : p_{K_n}(g - f) < \varepsilon\}$ for $n \in \mathbb{N}$ be a neighbourhood of f in $F\ell ip(X, \alpha)$. Then $f|_{K_n} \in \ell ip(K_n, \alpha) = \overline{Lip(K_n, 1)}$. Hence there exists $g \in Lip(K_n, 1)$ with $\|f - f|_{K_n}\|_{K_n} + p_\alpha(g - f|_{K_n}) < \varepsilon$. Let $\tilde{g} \in Lip(X, 1)$ be an extension of g to X so that $\tilde{g}|_{K_n} = g$ and therefore $\tilde{g} \in U \cap Lip(X, 1)$.

Now we present a Fréchet algebra version of the Hedberg's theorem for the Ff-algebra $F\ell ip(X, \alpha)$. It is interesting to see that we can also prove the density of $FLip(X, 1)$ in $F\ell ip(X, \alpha)$ by using this theorem.

We recall that for $a \in X$ and $\delta > 0$

$$S_a(\delta) = \{x \in X : d(x,a) = \delta\}$$
$$B_a(\delta) = \{x \in X : d(x,a) \leq \delta\}$$

Theorem 4.17 [27] Let (X,d) be a hemicompact metric space, $0 < \alpha < 1$ and A be a self-adjoint subalgebra of $F\ell ip(X,\alpha)$, which separates the points of X and contains the constants. Then A is dense in $F\ell ip(X,\alpha)$ if for every $a \in X$ there exist numbers M_a and δ_a such that for every $\delta \leq \delta_a$ there is an $f \in A$ with $f(a) = 1$, $f(x) = 0$ on $S_a(\delta)$ and

$$\sup_{\substack{y,z \in B_a(\delta) \\ y \neq z}} \frac{|f(y) - f(z)|}{d^\alpha(y.z)} < \frac{M_a}{\delta^\alpha}.$$

Proof. It is enough to show that for each compact subset K of X, the subalgebra $A|_K = \{f|_K : f \in A\}$ of $\ell ip(K,\alpha)$ is dense in $\ell ip(K,\alpha)$. Since $\overline{A|_{K_n}} = \ell ip(K_n,\alpha)$, for an arbitrary element $f \in F\ell ip(X,\alpha)$ and a neighbourhood $U = \{g \in F\ell ip(X,\alpha) : p_{K_n}(g-f) < \varepsilon\}$ of f, we can choose a $g \in A$ with $p_{K_n}(g-f) < \varepsilon$ so that $g \in U \cap A$.

Now let $K \subseteq X$ be compact. We try to verify the hypothesis of the Hedberg's theorem for the subalgebra $A|_K$ of $\ell ip(K,\alpha)$. Clearly $A|_K$ is a self-adjoint subalgebra of $\ell ip(K,\alpha)$, which separates the points of K and contains the constants. For an arbitrary element $a \in K$, choose M_a and δ_a as in the hypothesis of the theorem. So for every $\delta \leq \delta_a$ we can choose $f \in A$ with $f(a) = 1$, $f(x) = 0$ on $S_a(\delta)$ and

$$\sup_{\substack{y,z \in B_a(\delta) \\ y \neq z}} \frac{|f(y) - f(z)|}{d^\alpha(y,z)} < \frac{M_a}{\delta^\alpha}.$$

Then $f|_K \in A|_K$ has the desired properties, i.e., $f|_K(a) = 1$, $f|_K(x) = 0$ on $S_a(\delta) \cap K = \{x \in K : d(x,a) = \delta\}$ and $\sup \frac{|f(y)-f(z)|}{d^\alpha(y,z)} < \frac{M_a}{\delta^\alpha}$, where the supremum is taken over all distinct elements

$y, z \in B_a(\delta) \cap K = \{x \in K : d(x,a) \leq \delta\}$. Therefore, $\overline{A|_K} = \ell ip(K,\alpha)$ by the Hedberg's theorem in the complex case, and hence the claim is now established.

As an application of this theorem one can prove that $\overline{FLip(X,1)} = F\ell ip(X,\alpha)$ by taking the function $f(x) = 1 - \frac{d(x,a)}{\delta}$ for arbitrary $\delta > 0$ and $M_a = 2^{1-\alpha}$ in the above theorem.

References

1. A. Abdollahi, The maximal ideal space of analytic Lipschitz algebras, Rendiconti del Circolo Mathematico di Palermo, Serie II, Tomo XLVII (1998), pp. 347-352

2. D. Alimohammadi and A. Ebadian, Hedberg's theorem in Lipschitz algebras, Indian J. pure. appl. Math., 32(10) (2001), 1479-1493.

3. W.G. Bade, P.G. Curtis, Jr., and H.G. Dales, Amenability and weak amenability for Beurling and Lipschitz algebras, Proc. London Math. Soc. (3) 55 (1987), 359-377.

4. F.F. Bonsall and J. Duncan, Complete normed algebras, Springer Verlag, 1973.

5. R.L. Carpenter, Uniqueness of topology for commutative semisimple F-algebras, Proc. Amer. Math. Soc. 29(1971), 113-117.

6. H.G. Dales and A.M. Davie, Quasianalytic Banach function algebras, J. Funct. Anal. (1) 13 (1973), 28-50.

7. H.G. Dales, Banach algebras and automatic continuity, Oxford University Press, 2000.

8. T.W. Gamelin, "Uniform algebras,"Chelsea Publishing Company, 1984.

9. H. Goldmann, Uniform Fréchet algebras, North Holland, Amsterdam, 1990.

10. L.I. Hedberg, The Stone-Weierstrass theorem in Lipschitz algebras, Ark. Mat. 8 (1969), 63-72.

11. T.G. Honary, Relations between Banach function algebras and their uniform closures, Proc. Amer. Math. Soc. 109 (1990), 337-342.

12. T.G. Honary, Frechet Lipschitz algebras of infinitely differentiable functions, Proceedings of the 9^{th} Seminar on Analysis and its Applications, Tehran, I.R. Iran, 1998.

13. T.G. Honary and H. Mahyar, Approximation in Lipschitz algebras, Quaestions Mathematicae, 23(1) (2000), 13-19.

14. T.G. Honary and H. Mahyar, Approximation in Lipschitz algebras of infinitely differentiable functions, Bull. Korean Math. Soc. 36(4) (1999), 629-636.

15. T. Husain, Multiplicative functionals on topological algebras, Research notes in Math. 85, Pitmann Publishing, Boston, 1983.

16. K. Jarosz, $\mathrm{Lip}_{Hol}(X, \alpha)$, Proceedings of AMS, Volume 125, Number 10 (1997), 3129-3130.

17. B.E. Johnson, The uniqueness of the (complete) norm topology, Bull. Amer. Math. Soc. 73 (1967), 537-539

18. S.H. Kulkarni and B. V. Limaye, "Real Function Algebras", Marcel Dekker, Inc., 1992.

19. H. Mahyar, Approximation in Lipschitz algebras and their maximal ideal space, Ph.D. Thesis, Institute of Mathematics, University for Teacher Education, Tehran, 1994.

20. H. Mahyar, The maximal ideal space of $\ell ip_A(X,\alpha)$, Proc. Amer. Math. Soc.,122 (1994), 175-181.

21. B. Malgrange, Ideals of differentiable functions, Oxford University Press, London / New York, 1966.

22. S.B. Myers, Differentiation in Banach algebras, in Summer Institute on Set Theoretic Topology, University of Wisconsin, Madison, 1955.

23. A.G. O'Farrell, Annihilators of rational modules, J. Funct. Anal.19 (1975), 373-389.

24. A.G. O'Farrell, Hausdorff content and rational approximation in fractional Lipschitz norms, Trans. Amer. Math. Soc. 228 (1977), 187-206.

25. A.G. O'Farrell, Rational approximation in Lipschitz norms-I, Proc. Roy. Irish Acad. 77 A (1977), 113-115.

26. W. Rudin, "Functional Analysis", McGraw-Hill, New York, Second Edition, 1991.

27. F. Sady, Fréchet function algebras and uniqueness of topology for non-commutative Fréchet algebras. Ph.D. Thesis, University for Teacher Education, 1998.

28. F. Sady, Projective limit of a sequence of Banach function algebras as a Fréchet function algebra, Bull. Korean Math. Soc. 38 (2001).

29. D.R. Sherbert, Banach algebras of Lipschitz functions, Pacific J. Math. 13 (1963), 1387-1399

30. D.R. Sherbert, The Structure of ideals and point derivations in Banach algebras of Lipschitz functions, Trans. Amer. Math. Soc. 111 (1964), 240-272

31. N. Weaver, Subalgebras of little Lipschitz algebras, Pacific J. Math. 173 (1996), 283-293.

32. N. Weaver, "Lipschitz Algebras", World Scientific, New Jersey, 1999.

Taher Ghasemi Honary
Faculty of Mathematical Sciences and Computer Engineering
Teacher Training University,
599 Taleghani Avenue, Tehran 15618, I.R. IRAN

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 105-145 (2004)

An Empirical Bayes View of Inference in Inverse Gaussian Distribution

Mohammad R. Meshkani

Department of Statistics, Shahid Beheshti University
19834, Tehran, Iran
rmeshkani@hotmail.com

Abstract: The Inverse Gaussian distribution has been around for more than a century. It has been treated by classical methods for estimation, hypothesis testing, developing models for analysis of variance and regression by many authors. Its possible usefulness, when treated by Bayesian and empirical Bayes methods, first came to the attention of the present author about a decade ago. It was inspirerd by a short commumication dealing with it via Bayesian approach Achcar and Rosales (1993). Since then we have been involved in model building for this distribution exploiting the empirical Bayes approach. First, the estimation problems were studied. Later, we dealt with two-way analysis of variance, and finally with regression problems. The analysis of covariance is in progress. In this paper, we review the main points of this continued research. some applications are presented as illustrations.

MSC (2000): 62JXX, 62KXX, 62C12

Keywords: Inverse Gaussian distribution, estimation, regression, Bayes, empirical Bayes, prior, posterior.

1. Historical Review Of The Inverse Gaussian Distribution

Robert Brown (1773-1858), one of the greatest botanists of England was the one who began the story of the Inverse Gaussian distribution. Interested in pollens, he found a swimming, dancing motion of pollen particles when these particles were immersed in water. He repeated this experiment with a vide variety of plant pollens, finding a similar motion in every case. Later on, he tried particles from dead plants, fossils, and even mineral specimens of all sorts, in fact virtually everything he could imagine from the soot of London to a fragment from the Sphinx. Brown apparently believed he had discovered a new type of particle, common to all matter, organic and inorganic. Several researchers before him had observed and noted the motion of microscopic organic particles in fluids, but his work led to the realization that it was a physical phenomenon, not a biological one. Whether he himself initially recognized this fact is open to debate. During the rest of the century many researchers conducted experiments and theorized about the nature of the so-called **Brownian motion.** The law which governs the position of a single particle performing one-dimensional Brownian motion was derived by Bachelier in 1900. It turned out as a normal distribution. Wiener in 1923 amended Bachelier's work by providing a measure on the path space. By Einestion's (1905) work who also derived the normal distribution as the model for Brownian motion, the theory of Brownian motion was firmly launched. Schrodinger (1915) considered the Brownian motion with a positive drift and derived the distribution of the first passage time. This distribution was also obtained by Smouluchowski (1915).

Tweedie (1941) and Wald (1944) encountered the first passage distribution. Tweedie noticed that there is an inverse relationship between the cumulant - generating function of the time to cover unit distance and the cumulant - generating function of the distance covered in unit time. Tweedie (1945) also observed this type of relationship between the binomial and the negative binomial, and between the poisson and the exponential distribution. He proposed to call them inverse statistical variates. In 1956, he coined the name Inverse Gaussian for the first passage time distribution of the Brownian motion. In 1957, he published a detailed study of this distribution, establishing many of its important statistical properties. A special case of the distribution was given by Wald in 1947, which is an approximation of the sample size distribution in a sequential probability ratio test. Folks and Chhikara (1978) and Chhikara and Folks (1989) are the main sources for the statistical properties of the Inverse Gaussian model. The Inverse Gaussian model has been applied in reliability, marketing, and analysis of experiments. The main references concerning the classical approach are Fries and Bhatacharrya (1983), Chhikara and Gmttman (1982), Banerjee and Bhattacharyya (1979). On the Bayesian front the noteable works are Banerjee and Bhattacharrya (1979), Achcar et al. (1991), and Achcar and Rosales (1992, 1993).

For regression analysis under the Inverse Gaussian model, we refer to Whitmore (1983, 1986) Seto and Iwase (1985), and Chhikara and Folks (1989). Iwase (1989), Woldie and Folks (1994, 1995) and lately Seshadri (1999) are the relevant material.

The present author began his efforts with an enpirical Bayes approach. In 1994, one of his graduate students was engaged in the parameter estimation, Homayun - Aria, (1996). At the same time he himself developed methodology for analysis of variance, Meshkani (1996). Following this line of research, he later developed a methodology for the regression analysis, Meshkani (1999). Currenty he is working on the analysis of covariance. In this paper, we intend to summarize the findings from these works and show some of its applications in different settings. Some directions for further research will also be mentioned.

2. Empirical Bayes estimation of the parameters in an Inverse Gaussian distribution

2.1 Introduction

A random variable Y with the probability distribution function

$$f(y|\theta,\lambda)=\left[\frac{\lambda}{2\pi y^3}\right]^{\frac{1}{2}}\exp\left\{\frac{-\lambda(y-\theta)^2}{2y\theta^2}\right\}\qquad y,\lambda,\theta>0\qquad(2-1)$$

is called an Inverse Gaussain variate, where θ is the mean and λ is the scale parameter. This function will be denoted by IG(θ,λ). The IG(θ,λ) for various valuse of θ and λ provides a rich family of positively skewed densities which are suitable for modelling many types of positive random variables, Figures 2-1 and 2-2. It has the following cumulant - generating function

$$k_x(t)=\frac{\lambda}{\theta}\left[1-\left(1-\frac{2\theta^2t}{\lambda}\right)^{\frac{1}{2}}\right],\qquad t<\frac{\lambda}{2\theta^2}\qquad(2-2)$$

which provides the moments $E(x)=\theta$, $\quad \mathrm{var}(x)=\frac{\theta^3}{\lambda}$, and generally

$$K_r=1\cdot3\cdot5\cdots(2r-3)\theta^{2r-1}\lambda^{r-1},\quad r\geq2$$

The measures of skewness and kurtosis are $\sqrt{\beta_1}=3\sqrt{\frac{\theta}{\lambda}}$ and $\beta_2=15(\frac{\theta}{\lambda})+3$, which both are positive. For a normal variate, $V\sim N(v,\sigma^2)$, one has the cumulant - generating function

$$k_v(t)=vt-\frac{1}{2}\sigma^2t^2,\quad t\in\mathbb{R}\qquad(2-3)$$

since (2-2) and (2-3) are inverse functions of each orther hence the Inverse Gaussian name for (2-1). Moreover, there are striking anologies between the sampling distributions for IG(θ, λ) and those for $N(v, \sigma^2)$, see Chhikara and Folks (1989). The Inverse Gaussian distribution belongs to the exponential family.

2.2 Estimation

We suppose a random sample of $Y = (Y_1, \ldots, Y_n)$ is available from the IG(θ, λ) distribution. The aim is to obtain an estimator for (θ, λ) based on this sample. Most of estimation procedures exploit the likelihood function which is

$$L(\theta, \lambda | y) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \left(\Pi_{i=1}^{n} y_i^{-\frac{3}{2}}\right) \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i \theta^2} \right\}, \qquad \theta, \lambda > 0 \qquad (2-4)$$

The maximum likelihood estimators are based on maximization of (2-4) with respect to θ and λ, simultaneously. This approach yeilds:

$$\hat{\theta}_{ML} = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad \hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n}} \left(\frac{1}{Y_i} - \frac{1}{\bar{Y}}\right) \qquad (2-5)$$

If we define the sample harmonic mean as

$$\bar{Y}_H = \frac{1}{\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{Y_i}\right]}$$

Them $\hat{\lambda}_{ML}$ in (2-5) is non-negative, since $\bar{Y}_H \leq \bar{Y}$. Among many variants of the Bayes estimators relative to various priors, we report a case where it is assumed that θ and λ are independent and exponentially distributed, viz.

$$\theta \sim \exp(a) \quad \text{and} \quad \lambda \sim \exp(b)$$

Then the joint posterior of (θ, λ) will be

$$P(\theta, \lambda | y) \propto \lambda^{\frac{n}{2}} \exp\left\{ -\frac{\lambda v}{2} - \frac{n\lambda v \bar{Y}}{2} \left(\theta - \frac{1}{\lambda}\right)^2 - a\theta - b\lambda \right\} \qquad (2-6)$$

with $v = \sum_{i=1}^{n} \left(\frac{1}{y_i} - \frac{1}{\bar{y}} \right)$. Upon integration, we can obtain the marginal posteriors of $P(\theta|\boldsymbol{y})$ and $P(\lambda|\boldsymbol{y})$. From (2-6), the Bayes estimators of θ and λ relative to the squared error loss are the posterior means of $P(\theta|\boldsymbol{y})$ and $P(\lambda|\boldsymbol{y})$. That is,

$$\hat{\theta}_B = \hat{\theta}(a,b) = C\Gamma\left(\frac{n}{2}+1\right)\int_0^\infty \frac{\theta e^{-a\theta}}{\frac{v}{2} + \frac{n\bar{Y}}{2}(\theta - \frac{1}{\bar{Y}})^2 + b} d\theta \qquad (2-7)$$

where $C = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \Pi_{i=1}^{n} Y_i^{-\frac{3}{2}}$. Likewise,

$$\hat{\lambda}_B = \hat{\lambda}_B(a,b) =$$

$$\sqrt{2\pi} C e^{-\frac{a}{\bar{Y}}} \int_0^{\frac{a}{n}} \left\{ \frac{\lambda^{\frac{n}{2}+1}}{\sqrt{n\lambda\bar{Y}}} \exp\{-\lambda(\frac{v}{2}+b) - \frac{a^2}{2n\lambda\bar{Y}}\} \Phi\left(\frac{n\lambda - a}{\sqrt{n\lambda\bar{Y}}}\right) \right\} d\lambda$$

$$+ \int_{\frac{a}{n}}^{\infty} \left\{ \frac{\lambda^{\frac{n}{2}+1}}{\sqrt{n\lambda\bar{Y}}} \exp\{-\lambda(\frac{v}{2}+b) - \frac{a^2}{2n\lambda\bar{Y}}\} \Phi\left(\frac{a - n\lambda}{\sqrt{n\lambda\bar{Y}}}\right) \right\} d\lambda \qquad (2-8)$$

The Bayes estimators (2-7) and (2-8) have to be evaluated by numerical methods. Details of these derivations can be found in Homayun-Aria (1996). By a similar operation, we find the posterior variances var$(\theta|\boldsymbol{y})$ and var$(\lambda|\boldsymbol{y})$. The Empirical Bayes estimators are derived from the Bayes estimators by replacing the prior distribution by its estimated version. In case of a parametric prior, this strategy amounts to estimation of the parameters of the prior from the observed data and then using them in the Bayes estimators. This estamation can be done employing either the method of maximum likelihood (ML) or the method moments (MM). For the method of ML, we first obtain the marginal distribution of the sample values under the above exponential priors for (θ, λ):

$$m(y|a,b) = \int_0^\infty \int_0^\infty L(\theta, \lambda|y) P(\theta) P(\lambda)\} d\theta d\lambda$$

$$\frac{ab}{(2\pi)^{\frac{n}{2}} \ddot{y}} \int_0^\infty \int_0^\infty \exp\left\{ \frac{-\lambda}{2} \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i \theta^2} - a\theta - b\lambda \right\} d\theta d\lambda$$

Where $\dot{y} \equiv \Pi_{i=1}^{n} y_i^{\frac{3}{2}}$,

$$m(y|a,b) = \frac{2ab\bar{y}_H}{(2\pi)^{\frac{n}{2}}\dot{y}(2b\bar{y}_H + n)} \int_0^\infty \frac{\theta^2 e^{-a\theta}}{\theta^2 - r\theta + s} d\theta$$

with $r = \frac{2n\bar{y}_H}{2b\bar{y}_H + n} > 0, S = \frac{r\bar{y}}{2} > 0.$

After evaluation of this integral, as I(a,b) say, we obtain

$$m(y|a,b) = \frac{2ab\bar{y}_H}{(2\pi)^{\frac{n}{2}}\dot{y}(2b\bar{y}_H + n)} I(a,b)$$

To obtain the ML estimates of a and b, we have to maximize $m(y|a,b)$ with retpect to a, and b. Suppose $\hat{a}_{ML}$ and $\hat{b}_{ML}$ are the solutions. Then, we obtain the empirical Bayes estimators as

$$\hat{\theta}_{EBML} = \hat{\theta}_B(\hat{a}_{ML}, \hat{b}_{ML}) \qquad (2-9)$$

$$\hat{\lambda}_{EBML} = \hat{\lambda}_B(\hat{a}_{ML}, \hat{b}_{ML}) \qquad (2-10)$$

whose variances are estimated by $\mathrm{var}(\hat{\theta}_{EBML}|y)$ and $\mathrm{var}(\hat{\lambda}_{EBML}|y)$, respectively. The above results do not have closed from and are obtainable only by numerical integrations. To obtain an explicit solution, we may use the method of moments to estimate a and b. This procedure leads to

$$\hat{a}_{MM} = \bar{y} \quad \text{and} \quad \hat{b}_{MM} = \frac{[\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2]}{2\bar{y}^3}.$$

These estimates provide an alternative version of the empirical Bayes estimators of θ and $\lambda, i, e.$,

$$\hat{\theta}_{EBMM} = \hat{\theta}_B(\hat{a}_{MM}, \hat{b}_{MM}) \qquad (2-11)$$

$$\hat{\lambda}_{EBMM} = \hat{\lambda}_B(\hat{a}_{MM}, \hat{b}_{MM}) \qquad (2-12)$$

with corresponding variances $\mathrm{var}(\hat{\theta}_{EBMM}|y)$ and $\mathrm{var}(\hat{\lambda}_{EBML}|y)$.

3. Linear Models under the Inverse Gaussian Model

In scientific activities one encounters problems which are broadly classified as comparing the means of a certain variable among several subpopulations. Examples are the average yields from different process formulas, average number of defectives from various machines, etc. The Technical name for these problems in statistics is the Analysis of Variance (ANOVA). Occasionalyy, one is interested in finding the functional relation betwen a dependent (response) variable and some independent (explanatory) variables. These types of problems are called regression analysis. Yet a third type of problem is encountered when one has the regression and ANOVA at the same time which is referred to Analysis of Covariance (ANCOVA). These three type of problems are subsumed under the general title of linear models, whenever the functional relations imvolve parameters of the first degree. A number of authors have tried to develope statistical linear models for the Inverse Gaussian model akin to those under the normal model. However, the results are limited in the realm of classical approach. But Bayesian and Empirical Bayes approaches seem promising. we report some results obtained for ANOVA and regression.

4. Empirical Bayes Analysis of Variance

4.1 Introduction

There are many types of experimental setups in science and engineering where the normal theory is inappropriate for the analysis of factorial experiments. One important class is related to the highly skewed nature of the data which cannot be removed by the usual transformations. Alternatively, the Inverse Gaussian family of distributions are flexible enough to provide a suitable model for these types of data. Tweedie (1957 a,b) pioneered work in providing an analogue to analysis of variance for nested classifications concerning observations from an Inverse Gaussian model. Despite the quite striking resemblance between normal analysis of variance and what he called, the Inverse Gaussian analysis of reciprocals, in one-way or nested classifications, the possibilities of developing analogous results for other classifications appeared to be limited, Folks and Chhikara (1978), and Chkikara and Folks (1989). However, Shuster and Muira (1972) succeeded in providing tests for balanced two - way classifications. Their approach has the disadvantage that it requires many observations in each cell. Such a requirement is hard to fulfil in most experiments. Fries and Bhattacharyya (1983) treated the analysis of two-factor experiment with no interaction and obtained explicit solutions to the likelihood equations. They also proved asymptotic consistency and normality of their estimators. A few authors have contributed to the Bayesian analysis of the Inverse Gaussian distribution. Achcar and Rosales (1992,1993) are the only records in print for Bayesian analyis of two-factor experiments under an Inverse Gaussian model. They actually follow the approach presented in Fries and Bhattachatyya (1983), assuming a non-informative prior density. Consequently, as one expects, because of heavy reliance on the likelihood, their results are not much different from those obtained by the maximum likelihood method of estimation.

In this paper, we present an empirical Bayes analysis of two-factor experiments under an Inverse Gaussian model. A real-life example previously analyzed by Shuster and Muira(1972), and later by Achcar and Rosales

(1993), is reworked.

4.2 THE MODEL

Consider an experiment with two factors, factor A with I levels indexed by i, and factor B with J levels, numbered by j, with each treatment combination being repeated n times . Observations from this experiment, denoted by y_{ijk}, are assumed to follow an Inverse Gaussian model, $IG(\theta_{ij}, \lambda)$,

$$Y_{ijk} \sim IG(\theta_{ij}, \lambda), \quad i = 1, \ldots, I \quad j = 1, \ldots, J \quad k = 1, \ldots, n.$$

For each i,j , the random variables Y_{ijk} are iid with mean θ_{ij} and shape parameter λ. The two - parameter Inverse Gaussian density is

$$f(y_{ijk}; \theta_{ij}, \lambda) = \left\{\lambda/(2\pi y_{ijk}^3)\right\}^{\frac{1}{2}} \exp\left\{-\lambda(y_{ijk} - \theta_{ij})^2/2y_{ijk}\theta_{ij}^2\right\},$$

$$y_{ijk} > 0, \quad \theta_{ij} > 0, \;\; \lambda > 0, \quad i = 1, ..., I \quad j = 1, ..., J, \quad k = 1, ..., n. \qquad (4-1)$$

In a two - factor experiment with interaction, each cell mean is assumed to be inversely proportional to the drift, while the drift is considered as the sum of factor main effects (α, β) and their interaction(γ). Thus, it is assumed that

$$\theta_{ij}^{-1} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad i = 1, ..., I, \qquad j = 1, ..., J, \qquad (4-2)$$

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{i=1}^{I} \gamma_{ij} \sum_{j=1}^{J} \gamma_{ij} = 0. \qquad (4-3)$$

Here, μ denotes the reciprocal of each cell mean when there is no drift. To incorporate the constraints (4-3) in the model, we can define the $IJ \times 1$ parameter vector Φ as

$$\Phi = [\mu \mid \alpha_1, \ldots, \alpha_{I-1} \mid \beta_1, \ldots, \beta_{J-1} \mid \gamma_{11}, \ldots, \gamma_{1,J-1} \mid \cdots$$
$$\mid \gamma_{I-1,J}, \ldots, \gamma_{I-1,J-1} \|] = [\mu, \alpha, \beta\gamma_1, \ldots \gamma_i, \ldots \gamma_{I-1}]. \qquad (4\text{-}4)$$

Then, the likelihood for the whole experiment can be written as

$$L(\Phi, \lambda \mid y) \propto \lambda^{nIJ/2} \exp\{-\lambda[R_{+++} - 2n\Phi' d + n\Phi' M \Phi]/2\}. \qquad (4-5)$$

In (4-5), the convention used by Fries and Bhattacharyya (1983) has been utilized, where we have set

$$\begin{aligned} y =& [y_{111}, \cdots, y_{11k}, \cdots, y_{11n}, \cdots, y_{ij1}, \cdots, \\ & y_{ijk}, \cdots, y_{ijn}, \cdots, y_{IJ1}, \cdots, y_{IJk}, \cdots, y_{IJn}], \\ \theta_{ij}^{-1} =& \mu + \alpha_i + \beta_j + \gamma_{ij} = X'_{ij}\Phi, \\ X' =& (X_{11}, X_{12}, \cdots, X_{IJ}) = design \quad matrix, \\ D =& \text{diag}\{y_{11.}, y_{12.}, \cdots, y_{IJ.}\}, \qquad y_{ij.} = \sum_{k=1}^{n} y_{ijk}/n, \\ M =& X'DX, \qquad d = \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}, \\ R_{ijk} =& y_{ijk}^{-1}. \end{aligned} \qquad (4\text{-}6)$$

Summing over and index is shown by a plus sign while averaging is denoted by a dot. Thus, we shall use $R_{i++}, R_{+j+}, R_{ij+}, R_{+++}, R_{i..}, R_{.j.}, R_{ij.}$, and $R_{...}$ as sums and averages , respectively.

We intend to use a conjugate prior for λ and Φ. The following priors have been proposed , see Chhikara and Folks(1989), and Banerjee and Bhattacharyya (1979). The prior for λ is chosen from the gamma family and given λ, a normal prior is assumed or Φ. Thus ,

$$\pi(\lambda) \propto \lambda^{a-1} \exp\{-b\lambda/2\}, \qquad \lambda, a, b > 0, \qquad (4-7)$$

and, given λ, elements of Φ are considered independent with either of the following two priors.

Case 1. Unrestricted parameter space.

In the unrestricted case, the prior distribution for Φ is

$$\Phi\Big|_{\lambda} \sim N(\eta, \lambda^{-1}\Delta) \tag{4-8}$$

with

$$\eta = [\eta_1, \dots, \eta_{IJ}], \qquad \Delta = \mathrm{diag}\{\delta_1^2, \dots, \delta_{IJ}^2\}.$$

Then, the posterior is

$$q_1(\Phi, \lambda|y) \propto |\Psi|^{\frac{-1}{2}} \lambda^{\nu-1} \exp\{-\lambda[Q_1(\eta) + Q_2(\Phi)]/2\} \tag{4-9}$$

where

$$\Psi = (n\Delta M + \Delta^{-1})^{-1},$$

$$\nu = a + (n+1)IJ/2,$$

$$Q_1(\eta) = R_{+++} + b + \eta'\Delta^{-1}\eta - \eta^{*\prime}\Psi^{-1}\eta^*,$$

$$\eta^* = (n\Delta M + I)^{-1}(n\Delta d + \eta),$$

$$Q_2(\Phi) = (\Phi - \eta^*)'\Psi^{-1}(\Phi - \eta^*).$$

It is evident from (4-9) that

$$\pi(\lambda|y) \propto \lambda^{k-1} \exp\{-\lambda Q_1(\eta)/2\}, \ \ \lambda > 0, \tag{4-10}$$

$$\kappa = (nIJ + 2a)/2, \qquad Q_1(\eta) > 0,$$

and

$$q_1(\Phi|\lambda, y) \propto |\lambda^{-1}\Psi|^{-1/2} \exp\{-\lambda Q_2(\Phi)/2\},$$

that is, conjugacy holds. Therefore, we can write

$$q_1(\Phi \mid y) \propto \frac{1}{[Q_1(\eta) + Q_2(\Phi)]^{\nu}} \qquad \Phi \in \mathcal{R}^{IJ},$$

$$\propto \left[1 + \frac{(\Phi - \eta^*)' \sum^{-1} (\Phi - \eta^*)}{2a + nIJ}\right]^{-\frac{(2a+nIJ)+IJ}{2}} \tag{4-11}$$

with

$$\sum = \frac{Q_1(\eta)\Psi}{2a + nIJ}$$

which is a multivariate T-type distribution, with $2a + nIJ > 2$ degrees of reedom.

Case 2. Restricted parameter space.
Strictly speaking, one should have $\phi_1 = \mu > 0$. This restriction on ϕ_1 is observed in the prior assigned to ϕ_1. Thus, a normal distribution truncated at zero is considered for ϕ_1, which has density

$$q_2(\phi_1 \mid \lambda) \propto \lambda^{1/2}[\delta_1 \mathcal{N}(\lambda^{1/2}\eta_1/\delta_1)]^{-1} \exp\{-\lambda(\phi_1 - \eta_1)^2/2\delta_1^2\}, \qquad \phi_1 > 0 \qquad (4-12)$$

and the remaining is as in case 1. In (4-12), $\mathcal{N}(.)$ is the standard normal distribution function. The restriction imposed on ϕ_1 results in a posterior proportional to (4-9).

4.3 Bayes Estimates

In case 1, from (4-10) we have

$$E(\lambda^m|y) = [2/Q_1(\eta)]^m[\Gamma(\kappa + m)/\Gamma(\kappa)] \qquad (4-13)$$

which provides the Bayes estimate of λ relative to the squared error loss:

$$\lambda_{B1} = E(\lambda \mid \ y) = nIJ + 2a)/Q_1(\eta), \qquad (4-14)$$

$$V_\lambda = Var(\lambda|y) = 2(2a + nIJ)/[Q_1(\eta)]^2. \qquad (4-15)$$

Upon using (3.11), we arrive at

$$\Phi_{B1} = E(\Phi|y) = \eta^* = (n\Delta M + I)^{-1}(n\Delta d + \eta) \qquad (4-16)$$

$$V_{B1} = Var(\Phi|y) = \frac{Q_1(\eta)\Psi}{2(a-1)+IJ} \qquad (4-17)$$

For case 2, the posterior moments of λ remain unchanged from those for case 1. However, for Φ , the restriction on ϕ_1 renders results on Φ different from those in (4-16) and (4-17) for case 1. To this end, let Φ, η and Ψ be partitioned as

$$\eta = \begin{bmatrix} \eta_1 \\ \eta^{(2)} \end{bmatrix}, \eta^* = \begin{bmatrix} \eta_1^* \\ \eta^{*(2)} \end{bmatrix}, \Phi = \begin{bmatrix} \phi_1 \\ \Phi^{(2)} \end{bmatrix}, \Psi = \begin{bmatrix} \psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}.$$

Then by virtue of the facts $\Psi_{22.1} = \Psi_{22} - \Psi_{21}\psi_{11}^{-1}\Psi_{12}$ and $|\Psi| = \psi_{11}|\Psi_{22.1}|$, $Q_2(\Phi)$ in (4-9) can be written as

$$Q_2(\Phi) = \psi_{11}^{-1}(\phi_1 - \eta_1^*)^2 + (\Phi^{(2)} - \eta_{2.1}^*)'\Psi_{22.1}^{-1}(\Phi^{(2)} - \eta_{2.1}^*) \qquad (4.18)$$

where

$$\eta_{2.1}^* = E(\Phi^{(2)}|\phi_1, \lambda, y) = \eta^{*(2)} + \psi_{11}^{-1}(\phi_1 - \eta_1)\Psi_{21} \qquad (4-19)$$

and

$$Var(\Phi^{(2)}|\phi_1, \lambda, y) = \lambda^{-1}\Psi_{22.1}. \qquad (4.20)$$

In (4-19) , we take successive expectations with respect to ϕ_1 and λ, noting that

$$q_2(\phi_1|\lambda, y) \propto |\lambda^{-1}\psi_{11}|^{1/2}\exp\{-\lambda\psi_{11}^{-1}(\phi_1 - \eta_1^*)^2/2\}, \qquad \phi_1 > 0,$$

which is a normal density truncated at zero. For this distribution,

$$E(\phi_1|\lambda, y) = \eta_1^* + \lambda^{-1}\psi_{11})^{1/2}\omega$$

with

$$\omega = \varphi[\eta_1^*(\lambda^{-1}\psi_{11})^{1/2}]\mathcal{N}[\eta_1^*/(\lambda^{-1}\psi_{11})^{1/2}]$$

where $\varphi(.)$ is the standard normal density function, and

$$Var(\phi_1|\lambda, y) = \lambda^{-1}\psi_{11}[1 - \omega^2] + \eta_1^*\omega(\lambda^{-1}\psi_{11})^{1/2}.$$

These moments, however, are too complicated to be useful for estimation purposes. To simplify them, we observe that $\omega(x) = \varphi(x)/\mathcal{N}(x)$ is a smooth decreasing function of x. In the literature there are a host of approximations to $\mathcal{N}(x)$ [Patel and Read (1996, Chapter 3)], which can be used to approximate $\omega(x)$ with desired precision. Here, we choose to use the simpler one due to Shah(1985),

$$\mathcal{N}(x) = \begin{cases} 0.5 + x(4.4 - x)/10, & 0 \leq x \leq 2.2, \\ 0.99, & 2.2 \leq x < 2.6, \\ 1.00, & x \geq 2.6. \end{cases}$$

Consequently,

$$\omega(x) = \begin{cases} \frac{2(2.2-x)}{5+x(4.4-x)}, & 0 \leq x \leq 2.2 \\ 0, & x > 2.2. \end{cases}$$

In our problem, $0 \leq x \leq \infty$ and $0 \leq \omega(x) \leq 0.8$. Thus, we can approximate $\omega(x)$ by its average value, which is about 0.3. Of course, if one has a better guess of $x = \eta_1^*/(\lambda^{-1}\psi_{11})^{1/2}$, a closer approximation could be obtained. Using this approximation, we have

$$E(\phi_1|\lambda, y) = \eta_1^* + 0.3(\lambda^{-1}\psi_{11})^{1/2}$$

$$Var(\phi_1|\lambda, y) = 0.91\lambda^{-1}\psi_{11} + 0.3\eta_1^*(\lambda^{-1}\psi_{11})^{1/2}.$$

Now, we take expectations with respect to the posterior distribution of λ, employing (4-13) and obtain

$$\phi_{1B2} = E(\phi_1|y) = \eta_1^* + 0.21[\psi_{11}Q_1(\eta)]^{1/2}W \qquad (4-21)$$

with $W = \Gamma(\kappa - 0.5)/\Gamma(\kappa) \simeq [2.72(\kappa - 0.5)/\kappa]^{\frac{1}{2}}$, by Stirling's formula and

$$V_{1B2} = Var(\phi_1|y) = 0.45\psi_{11}Q_1(\eta)/(\kappa-1) + 0.045\psi_{11}Q_1(\eta)[11/(\kappa-1) - W^2]$$

$$+0.21\eta_1^*[\psi_{11}Q_1(\eta)]^{1/2}W. \tag{4-22}$$

Now, we substitute the posterior moments of ϕ_1 into (4-9) to obtain a simpler form as :

$$\Phi_{2B2} = \eta^{*(2)} + \{0.21[\psi_{11}^{-1}Q_1(\eta)]^{1/2}\}\Psi_{21}W \tag{4-23}$$

$$\begin{aligned} V_{2B2} = {} & [Q_1(\eta)/2(\kappa-1)][\Psi_{22} - 0.09(\kappa-1)W^2\Psi_{21}\psi_{11}^{-1}\Psi_{12}] \\ & +0.21W\eta_1^*[\psi_{11}^{-1}Q_1(\eta)]^{1/2}\Psi_{21}\psi_{11}^{-1}\Psi_{12}]. \end{aligned} \tag{4-24}$$

The expressions (4-21) - (4-24) provide the Bayes estimates relative to the restricted prior given in (4-12). As long as the prior distribution can be assessed, the above Bayes estimators, could be put into application for two-way classifications. Unfortunately, the situations where these priors can reasonably be assessed are rare. In such cases, we can utilize the empirical Bayes procedure to estimate the prior distributions from the data. By this, we borrow strength from Bayesian logic and objectivity from classical method.

4.4 Empirical Bayes estimates

To estimate the prior parameters from the marginal distribution of the observations , one can use any method of estimation. Two more common methods are the method of moments and maximum likelihood. To provide explicit expressions for estimates of a,b,η and Δ from the data, y, we shall first use the method of moments. To this end, we have from Chhikara and Folks (1989),

$$V_{ij} = \sum_{k=1}^{n}(Y_{ijk}^{-1} - Y_{ij.}^{-1}) \sim \lambda^{-1}\chi_{n-1}^2.$$

Thus,

$$E(V_{ij}) = E[E(V_{ij})|\lambda)] = (n-1)b/2(a-1),$$

$$\begin{aligned} Var(V_{ij}) &= E[Var(V_{ij})|\lambda)] + Var[E(V_{ij}|\lambda)] \\ &= (n-1)(n-3+2a)b^2/4(a-1)^2(a-2). \end{aligned}$$

Let

$$V = \sum_{i=1}^{I}\sum_{j=1}^{J} V_{ij}/IJ, \qquad S_V^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}[V_{ij} - V]^2/(IJ-1);$$

and let $C = S_V/V$ be the sample coefficient of variation for V_{ij}. Then,

$$a_0 = [2(n-1)C^2 + n - 3]/[(n-1)C^2 - 2], \qquad b_0 = 2(a_0 - 1)V/(n-1)$$

which are valid positive estimates of a and b for $n > 1$, if one has $C^2 > 2/(n-1)$, otherwise, take a and b equal to zero, *i.e.*, use a noninformative prior.

For estimation of η and Δ, we need 2IJ equations. These are provided by the following considerations. Let $Z_i, i = 1, \cdots, n$, be iid andom variables distributed as $IG(\theta, \lambda)$. The distributional relations between Z_i, $\bar{Z}$ and Z_i^{-1} have been found in Chhikara and Folks (1989). That is, $\bar{Z} \sim IG(\theta, n\lambda)$ and Z_i^{-1} has mean and variance as stated below:

$$Z_i^{-1} \sim [\theta^{-1} + \lambda^{-1} = E(Z_i^{-1}), \quad (\lambda\theta)^{-1} + 2\lambda^{-2} = Var(Z_i^{-1})].$$

Using these results in our model (4-1) - (4-6), for the reciprocals and their means , we have :

$$S_{ij}^2 = \sum_{qk=1}^{n} (R_{ijk} - R_{ij.})^2/n(n-1),$$

$$S_i^2 = \sum_{j=1}^{J} (R_{ij.} - R_{i..})^2/J(J-1),$$

$$S_j^2 = \sum_{i=1}^{I} (R_{ij.} - R_{.j.})^2/I(I-1),$$

$$S^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} (R_{ijk} - R_{ij.})^2/nIJ(nIJ-1).$$

We obtain :

$$\eta_1^0 = R_{...} - b_0/2(a_0 - 1)$$

$$\delta^2_{1,0} = 2(a_0 - 1)S^2/b_0 - \eta^0_1/nIj - b_0[2(a_0 - 1) + nIJ](a_0 - 1)(a_0 - 2)$$

For $\quad i = 1, \cdots, I - 1,$

$$\eta^0_{i+1} = R_{i..} - \eta^0_1 - b_0/2(a_0 - 1) = R_{i..} - R_{...},$$

$$\delta^2_{i+1,0} = 2(a_0-1)S^2_i/b_0-(\eta^0_1+\eta^0_{i+1})/nJ-\delta^0_1-b_0[2(a_0-1)+nJ]/2nJ(a_0-1)(a_0-2).$$

For $\quad j = 1, \cdots, J - 1,$

$$\eta^0_{I+j} = R_{.j.} - \eta^0_1 - b_0/2(a_0 - 1) = R_{.j.} - R_{...},$$

$$\delta^2_{I+j,0} = 2(a_0 - 1)S^2_j/b_0 - (\eta^0_1 + \eta^0_{I+j})/nI - \delta^0_1$$
$$-b_0[2(a_0 - 1) + nI]2nI(a_0 - 1)(a_0 - 2).$$

Finally, for $\quad i = 1, \cdots, I - 1,$ and $j = 1, \cdots, J - 1,$

$$\eta^2_{I+i(J-1)+j} = R_{ij.} - \eta^0_1 - \eta^0_{i+1} - \eta^0_{I+j} - b_0/2(a_0 - 1)$$
$$= R_{ij.} - R_{i..} - R_{.j.} + R_{...},$$

$$\delta^2_{I+i(J-1)+j,0} = 2(a_0 - 1)S^2_{ij}/b_0 - [\eta^0_1 + \eta^0_{i+1} + \eta^0_{I+j} + \eta_{I+i(J-1)+j}]/n$$
$$-(\delta^2_{1,0} + \delta^2_{i+1,0} + \delta^2_{I+j,0}) - b_0[2(a_0 - 1) + n]/2n(a_0 - 1)(a_0 - 2)$$

which provide

$$\eta^0 = [\eta^0_1, \eta^0_2, \ldots, \eta^0_{IJ}],$$
$$\Delta^0 = \mathrm{diag}\{\delta^2_{1,0}, \delta^2_{2,0}, \ldots, \delta^2_{IJ,0}\}.$$

Now, we subtitute these estimates into (4-14) and (4-16) to obtain the empirical Bayes estimate relative to unrestricted prior distribution. This gives us

$$\lambda_{EB1} = (nIJ + 2a_0)/Q_1(\eta^0), \qquad (4-25)$$

$$\Phi_{EB1} = (n\Delta^0 M + I)^{-1}(n\Delta^0 d + \eta^0) = \eta^{*^0}. \qquad (4.26)$$

Posterior variances are estimated by

$$Var(\lambda|y) = 2(2a_0 + nIJ)/[Q_1\eta^0)]^2 \tag{4.27}$$

$$Var(\Phi|y) = [Q_1(\eta^0)/2(\kappa-1)]\Psi^0, \qquad \Psi^0 = [n\Delta^0 M + (\Delta^0)^{-1}]^{-1}. \tag{4.28}$$

In case 2, the only difference in prior is that ϕ_1 has a truncated normal prior. Accordingly, we should alter the posterior and the marginal moments for differences in moments of ϕ_1 . In this case,

$$E(\phi|\lambda) = \eta_1 + 0.3\lambda^{-1/2}\delta_1,$$

$$Var(\phi_1|\lambda) = 0.91\lambda_1^{-1}\delta_1^2 + 0.3\eta_1\lambda^{-1/2}\delta_1.$$

These differ from the respective moments of ϕ_1 in case 1. To account for this difference, the previous moment equations should be modifed accordingly. Omitting the details which can be found in Meshkani (1996), we shall give the final results.

Let $\quad k(a_0) = 0.21w(a_0)b_0^{1/2} \qquad w(a_0) = \Gamma(a_0 - 0.5)/\Gamma(a_0)$

$$m_0 = [b_0^2/4nIJ(a_0-1)^2(a_0-2)][2(a_0-1) + nIJ]$$

Then,

$$\eta_1 = R_{...} - b_0/2(a_0-1) - k(a_0)\delta_1$$

$$S^2 = m_0 + [b_0/2(a_0-1)]\delta_1^2 + [b_0/2nIJ(a_0-1)][\eta_1 + k(a_0-1)\delta_1] - k^2(a_0)\delta_1^2$$

Absorbing η_1 into S^2 leads to the quadratic equation

$$A\delta_1^2 + B\delta + D = 0$$

with

$$A = 1 - 4(a_0-1)k^2(a_0)/b_0,$$

$$B = k(a_0)\{(1+2nIJ)[\omega(a_0-1) - \omega(a_0)]/IJ\omega(a_0) + 2(a_0-1)R_{...}/b_0\} - 1,$$

and

$$D = (R_{...}/nIJ) + b_0(a_0 + nIJ)/2nIJ(a_0 - 1)(a_0 - 2) - 2(a_0 - 1)S^2/b_0.$$

Thus, we obtain an estimate for δ_1^2 as

$$\tilde{\delta}_1^2 = \begin{cases} (-B/2A)^2 & \text{if} \quad B^2 - 4AD \geq 0, \\ D/A & \text{if} \quad B^2 - 4AD < 0 \end{cases}$$

This gives $\tilde{\eta}_1 = R_{...} - b_0/2(a_0 - 1) - k_0\tilde{\delta}_1$, while other elements of $\tilde{\eta}$ being equal to those given for case 1. However, for $i = 1, \cdots, I - 1$,

$$\tilde{\delta}_{i+1} = 2(a_0 - 1)S_i^2/b_0 - [\tilde{\eta}_1 + \tilde{\eta}_{i+1} + \tilde{\delta}_1 k(a_0 - 1)]/nJ$$
$$-b_0[2(a_0 - 1) + nJ]/2nJ(a_0 - 1)(a_0 - 2) - \tilde{\delta}_1^2 +$$
$$2k(a_0)[\tilde{\delta}_1 k(a_0) - \tilde{\eta}_1]\tilde{\delta}_1(a_0 - 1)/b_0 - 2\tilde{\delta}_1[k(a_0 - 1) - k(a_0)];$$

for $j = 1, \ldots, J - 1$,

$$\tilde{\delta}_{I+j} = 2(a_0 - 1)S_j^2/b_0 - [\tilde{\eta}_1 + \tilde{\eta}_{I+j} + \tilde{\delta}_1 k(a_0 - 1)]/nI$$
$$-b_0[2(a_0 - 1) + nI]/2nI(a_0 - 1)(a_0 - 2) - \tilde{\delta}_1^2 +$$
$$2k(a_0)[\tilde{\delta}_1 k(a_0) - \tilde{\eta}_1]\tilde{\delta}_1(a_0 - 1)/b_0 - 2\tilde{\delta}_1[k(a_0 - 1) - k(a_0)];$$

and finally, for $i = 1, \cdots, I - 1$, and $j = 1, \cdots, J - 1$,

$$\tilde{\delta}_{I+i(J-1)+j} = 2(a_0-1)S_{ij}^2/b_0 - [\tilde{\eta}_1 + \tilde{\eta}_{i+1} + \tilde{\eta}_{I+j} + \tilde{\eta}_{I+i(J-1)+j} + \tilde{\delta}_1 k(a_0-1)]/nI$$
$$-b_0[2(a_0 - 1) + n]/2n(a_0 - 1)(a_0 - 2) - [\tilde{\delta}_1^2 + \tilde{\delta}_{i+1}^2 + \tilde{\delta}_{I+j}^2]$$
$$+2k(a_0)[\tilde{\delta}_1 k(a_0) - \tilde{\eta}_1]\tilde{\delta}_1(a_0 - 1)/b_0 - 2\tilde{\delta}_1[k(a_0 - 1) - k(a_0)].$$

Thus, we have $\tilde{\eta} = [\tilde{\eta}_1, \ldots, \tilde{\eta}_J]$, $\tilde{\Delta} = \text{diag}\{\tilde{\delta}_1^2, \ldots, \tilde{\delta}_{IJ}^2\}$, which provide the respective empirical Bayes estimates:

$$\lambda_{EB2} = (nIJ + 2a_0)/Q_1(\tilde{\eta}) \qquad (4-29)$$

$$Var(\lambda|y) = 2(nIJ + 2a_0)/[Q_1(\tilde{\eta})]^2 \qquad (4-30)$$

$$\tilde{\phi}_{1,EB2} = \tilde{\eta}_1^* + 0.21\omega(k)[\tilde{\psi}_{11}Q_1(\tilde{\eta})]^{1/2} \qquad (4-31)$$

$$Var(\phi|y) = 0.045\tilde{\psi}_{11}Q_1(\tilde{\eta})[11/(k-1)-\omega^2(k)]+0.21\tilde{\eta}_1^*\omega(k)[\tilde{\psi}_{11}Q_1(\tilde{\eta})]^{1/2} \qquad (4-32)$$

with

$$\tilde{\eta}^* = [\tilde{\eta}_1^*, \tilde{\eta}^*(2)] = [n\tilde{\Delta}M + I]^{-1}[n\tilde{\Delta}d + \tilde{\eta}]$$

$$\tilde{\Psi} = [n\tilde{\Delta}M + \tilde{\Delta}^{-1}] = \begin{bmatrix} \tilde{\psi}_{11} & \tilde{\Psi}_{12} \\ \tilde{\Psi}_{21} & \tilde{\Psi}_{22} \end{bmatrix}$$

Moreover,

$$\Phi_{EB2}^{(2)} = \tilde{\eta}^{*(2)} + \{0.21\omega(k)[\psi_{11}^{-1}Q_1(\tilde{\eta})]^{1/2}\}\tilde{\Psi}_{21} \qquad (4-33)$$

$$\begin{aligned} Var(\Phi|y) = {} & [Q_1(\eta/2(k-1)][\tilde{\Psi}_{22} - 0.09(k-1)\omega^2(k)\tilde{\Psi}_{21}\tilde{\psi}_{11}^{-1}\tilde{\Psi}_{12}] \\ & +0.21\omega(k)\tilde{\eta}_1^*[\tilde{\psi}_{11}^{-1}Q_1(\tilde{\eta})]^{1/2}\tilde{\Psi}_{21}\tilde{\psi}_{11}^{-1}\tilde{\Psi}_{12}. \qquad (4-34) \end{aligned}$$

Although we have used the method of moments to reach explicit solutions, we could have alternatively used the maximum likelihood procedure to obtain estimates of a,b,η and Δ. This method needs numerical maximization which can be done by usual routines. Here, we only outline the procedure and leave the detail for practical data analysis. From (4-5) - (4-8),

$$\begin{aligned} \mathcal{L} = l(y|a,b,\eta,\Delta) = {} & K\Big\{\int_0^\infty \Big[\lambda^{\nu-1}b^a|\Delta|^{-1/2}/\Gamma(a)\Big]\exp\Big\{-\frac{\lambda}{2}Q_1(\eta)\Big\} \\ & \times \int \exp -\frac{\lambda}{2}Q_2(\phi)d\phi\Big\} \\ = {} & \frac{\Gamma(a+nIJ/2)}{\Gamma(a)}.\frac{|\Psi|^{1/2}}{|\Delta|^{1/2}}\frac{(2b)^a}{[Q_1(\eta)]^{nIJ/2+a}} \end{aligned}$$

Maximizing $\mathcal{L}$ with respect to a,b,η and Δ would provide the maximum likelihood estimates, denoted by $\hat{a}, \hat{b}, \hat{\eta}, \text{and}\hat{\Delta}$. Applying them in (4-16) and (4-17) would result in empirical Bayes estimates based on the maximum likelihood procedure. Again, if we observe the estriction on ϕ_1, we shall have the corresponding results. Let the result be xpressed as

$$\phi_{EBL} = (n\hat{\Delta}M + I)^{-1}(n\hat{\Delta} + \hat{\eta}), \qquad (4-35)$$

$$Var(\Phi|y) = [Q_1(\hat{\eta})/2(\kappa - 1)]\hat{\Psi}, \qquad (4-36)$$

with

$$\hat{\Psi} = (n\hat{\Delta}M + \hat{\Delta}^{-1})^{-1}.$$

The above derivation remains valid for the 2-factor **ANOVA** model without interaction, as well as for one-way **ANOVA**. In these cases, one only needs to reduce the order of the vector of parameters and the design matrix, according to the model used and follow the above procedure.

4.5 AN EXAMPLE

To illustrate our proposed estimators and compare them with other estimators, we analyze an experiment originally reported by Ostle(1963) and analyzed by Shuster and Muira (1972), and later by Achcar and Rosales (1993). Data in Table 5.1 ave resulted from a randomized 2 × 5 layout with 10 replicates in each cell. The responses consist of the impact resistance of 5 kinds of insulators to shocks when they are cut lengthwise or widthwise. There are 10 replicates for each combination.

Maximum likelihood estimates:

It can be shown that the maximum likelihood estimates are

$$\hat{\Phi}(M\mathcal{L}) = M^{-1}d, \qquad d = [1, 0, \cdots, 0]$$

$$\hat{\lambda}(M\mathcal{L}) = IJ/[nR_{...} - d' M^{-1}d]$$

whose large - sample variances are

$$Var[\hat{\Phi}(M\mathcal{L})] = (n\hat{\lambda})^{-1}M^{-1}$$

$$Var[\hat{\lambda}(M\mathcal{L})] = 2(nIJ)^{-1}[\hat{\lambda}(M\mathcal{L})]^2$$

$$Cov[\hat{\Phi}(M\mathcal{L}), \hat{\lambda}(M\mathcal{L})] = 0$$

In this example, the diagonal elements of **D** are given in the last column of Table 4.1, and $\boldsymbol{X}'$ is

$$\boldsymbol{X}' = \begin{bmatrix} 1_4' & 1 & 1_4' & 1 \\ 1_4' & 1 & -1_4' & -1 \\ I_4 & -1_4 & I_4 & -1_4 \\ I_4 & -1_4 & -I_4 & 1_4 \end{bmatrix}$$

which yield the estimates and their standard errors (S.E.) given in Table 4.2. The asymptotic 95 percent confidence intervals (CI) are also provided in Table 4.2. It is clear that except for $\boldsymbol{\lambda}$ no parameter can be taken different from zero at 5 percent level. However, only the constant μ is different from zero at 0 percent level. But, due to wide confidence intervals one should feel uncertain about these inferences. Better inferences are possible by the empirical Bayes procedure presented below.

Empirical Bayes estimates of a and b are $a_0 = 6.95$ and $b_0 = 0.26$, espectively.

The estimates of the model parameters for each of the two cases (unrestricted nd restricted) along with the standard errors, are shown in Table 4.3. In Table 4.4, the 95 percent credible intervals (CI) based on the marginal posteriors are given. Again, we observe that only λ and μ are infered to be different from zero at 5 percent level. Although we have reached the same conclusion as the one relative to MLE, but here we have much smaller standard errors which make the inference more precise. In fact, comparing Table 4.2 and 4.3, we note a striking consequence of exploiting the empirical Bayes procedure. The credible intervals in both cases (unrestricted and restricted) are very much shorter than the corresponding confidence intervals given in Table 4.2. The only exception is the intervals for λ which have become somewhat longer for the empirical Bayes procedure.

Table 4.1 Observations from the experiment

K →	1	2	3	4	5	6	7	8	9	10	Mean
(i,j) ↓											$(y_{ij.})$
1,1	1.15	0.84	0.88	0.91	0.86	0.88	0.92	0.87	0.93	0.95	0.919
1,2	1.16	0.85	1.00	1.08	0.80	1.01	1.14	0.87	0.97	1.09	0.999
1,3	0.79	0.68	0.64	0.72	0.63	0.59	0.81	0.65	0.64	0.75	0.690
1,4	0.96	0.82	0.98	0.93	0.81	0.79	0.79	0.86	0.84	0.92	0.870
1,5	0.49	0.61	0.59	0.51	0.53	0.72	0.67	0.47	0.44	0.48	0.551
2,1	0.89	0.69	0.46	0.85	0.73	0.67	0.78	0.77	0.80	0.79	0.743
2,2	0.86	1.17	1.18	1.32	1.03	0.84	0.89	0.84	1.03	1.06	1.022
2,3	0.52	0.52	0.80	0.64	0.63	0.58	0.65	0.60	0.71	0.59	0.623
2,4	0.86	1.06	0.81	0.97	0.90	0.93	0.87	0.88	0.89	0.82	0.899
2,5	0.52	0.53	0.47	0.47	0.57	0.54	0.56	0.55	0.45	0.60	0.526

Table 4.2 ML Estimates of λ and Φ and their symptotic 95 percent confidence intervals

Parameter	MLE	S.E.	95 *percent*	*CI*	*CI length*
λ	0.98	0.09	0.8	1.16	0.36
μ	1.34	0.82	−0.27	2.95	3.22
α_1	0.04	0.82	−1.57	1.65	3.22
β_1	-0.13	1.58	−3.23	2.96	6.19
β_2	-0.35	1.47	−2.27	2.53	4.80
β_3	0.18	1.72	−3.19	3.55	6.74
β_4	-0.21	1.54	−3.23	2.81	6.04
γ_{11}	-0.09	1.58	−3.19	3.01	6.20
γ_{12}	-0.06	1.47	−2.94	2.82	5.76
γ_{13}	-0.03	1.71	−3.38	3.32	6.70
γ_{14}	0.06	1.54	−2.96	3.08	6.04

Table 4.3 Empirical Bayes estimates

Prior	Unrestricted		Restricted	
parameter	Estimated	S.E.	Estimate	S.E.
λ	0.7973	0.1057	1.2198	0.1616
μ	1.1502	0.2324	0.9858	0.1462
α_1	-0.0004	0.1302	-0.0006	0.1239
β_1	0.0216	0.2440	0.0177	0.2406
β_2	-0.0555	0.3223	0.0001	0.3131
β_3	0.0212	0.3733	0.0923	0.2107
β_4	-0.0467	0.3193	-0.0001	0.1025
γ_{11}	0.0005	0.1781	-0.0014	0.3202
γ_{12}	0.0012	0.2324	-0.0003	0.1422
γ_{13}	-0.0001	0.2392	0.0005	0.1552
γ_{14}	0.0016	0.2362	-0.0022	0.1471

Table 4.4 The 95 percent credible intervals for empirical Bayes stimates

Prior	Unrestricted			Restricted		
parameter	95 percent	CI	CI length	95 percent	CI	CI length
λ	0.5901	1.0044	0.4139	0.9031	1.5365	0.6334
μ	0.6947	1.6057	0.9110	0.6992	1.2724	0.5732
α_1	-0.2556	0.2548	0.5104	-0.2434	0.2422	0.4856
β_1	-0.4566	0.4998	0.9564	-0.4539	0.4893	0.9432
β_2	-0.6872	0.5763	0.5950	-0.6136	0.6138	1.2274
β_3	-0.7105	0.7529	1.4664	-0.3207	0.5053	0.8260
β_4	-0.6725	0.5721	1.2516	-0.2010	0.2008	0.4018
γ_{11}	-0.3486	0.3496	0.6982	-0.6290	0.6262	1.2552
γ_{12}	-0.4543	0.4567	0.9110	-0.2790	0.2784	0.5574
γ_{13}	-0.4689	0.4687	0.9376	-0.3037	0.3047	0.6084
γ_{14}	-0.4614	0.4646	0.926	-0.2905	0.2861	0.5766

5. Empirical Bayes regression

5.1 Introduction

In this section we look at the regression analysis under the Inverse Gaussian model. This problem has also been treated mainly from the classical point of view. Whitmore (1983) considers the censored data and estimates the regression coefficient, using the EM algorithm. Seto and Iwase (1985) derive the minimum variance unbiased estimator. However, Hsieh and Korwar (1990) prove that it is inadmissible. Chhikara and Folks (1989) and Seshadri (1999) each devote a chapter to the regression problem.Woldie and Folks (1999, 1994) are the most relevant works, though they adopt the classical approach, too. We intend to extend their results by exploiting the Bayesian and empirical Bayes approaches.

5.2 The model

Suppose we have observed the pairs of observations $(y_i, x_i), i = 1, \ldots, n$ where y_i are realizations of random variable $Y_i$which has the Inverse Gaussian distribution with mean $(\boldsymbol{x}'_i\boldsymbol{\beta})^q$ and the scale parameter λ_i where $(\boldsymbol{x}'_i$ and $\boldsymbol{\beta}$ are p-vectors with $(\boldsymbol{x}'_i\boldsymbol{\beta} > 0$, q is an integer, and $\lambda_i > \theta$. Various assumptions on λ_i and q provide different models such as simple inear regression through origin $\{p = q = 1, \lambda_i = \lambda, \boldsymbol{x}_i = (x_i), \boldsymbol{\beta} = (\beta)\}$, linear regression with intercept $\{p = 2, q = 1, \lambda_i = \lambda, \boldsymbol{x}_i = (1, x_{i2}), \boldsymbol{\beta} = (\beta_1, \beta_2)\}$, non-linear regression $\{q = -1, \lambda_i = \lambda\}$; and general non-linear regression $\{q = -1, \lambda_i = \lambda\}$. The problem is to estimate $\boldsymbol{\beta}$ and λ_i from the observations $(y_i, x_i), i = 1, \ldots, n$. This problem is somewhat similar to the analysis of variance which we discussed in the previous section. The difference is in that here $\boldsymbol{x}_i$ are real vectors while there in ANOVA they were indicators showing the presence (1) or absence (0) of a factor. Thus, the approach adopted there can be helpful in dealing with this problem.

5.3 Bayesian analysis

According to Bayesian way of thinking, we assume

$$Y_i|\boldsymbol{\beta}, \lambda \sim IG\{(\boldsymbol{x}'_i\boldsymbol{\beta})^q, \lambda\}, \qquad i = 1, \ldots, n$$

where the parameters are treated as random variables, having some sort of probability distribution, which are called the prior distribution. To perform a Bayesian analysis, we need to specify the prior distributions $\pi_1(\boldsymbol{\beta}|\lambda)$ and $\pi_2(\lambda)$. While various authors have employed different prior distributions for $\boldsymbol{\beta}$ nd λ_i, we choose to use the conjugate priors for these parameters. Those are explicitly stated as

$$\pi_1(\lambda) = \left(\frac{b}{2}\right)^{\frac{a}{2}} \lambda^{\frac{a}{2}-1} \exp\frac{\{-\frac{b\lambda}{2}\}}{\Gamma(\frac{a}{2})} \quad \lambda, a, b > 0, \qquad (5-1)$$

and

$$\pi_2(\beta|\lambda) = \text{const.}\lambda^{\frac{p}{2}}|\Delta|^{-\frac{1}{2}} \exp\{\frac{-(\beta-\eta')\Delta^{-1}(\beta-\eta)}{2}\} \qquad (5-2)$$

where

$$\eta = [\eta_1, \dots, \eta_p], \eta \in R^p$$

$$\Delta = \text{diag}\{\delta_1, \dots, \delta_p\}, \delta_j \in R^+.$$

The likelihood function is reduced to

$$L(\beta, \lambda|) = \text{constant} + \lambda^{\frac{n}{2}} \exp\{\frac{-\lambda Q_1(\beta)}{2}\} \qquad (5-3)$$

with

$$Q_1(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 / y_i(x_i'\beta)^2$$
$$= [YX\beta - 1_n]' Y^{-1} [YX\beta - 1_n]$$

with

$$X = [x_1, \dots, x_n], Y = \textbf{diag}\{y_1, \cdots, y_n\}, 1_n = [1, \dots, 1].$$

Thus, the posterior distribution of (λ, β) is obtained as

$$q(\lambda, \beta|y) \propto \frac{b^{\frac{a}{2}}\lambda^{v-1}}{\Gamma(\frac{a}{2})|\Delta|^{\frac{1}{2}}} \exp\{\frac{-\lambda}{2}[Q_1(\beta) + Q_2(\eta) + b]\} \qquad (5-4)$$

$$v = \frac{(a+n+p)}{2}, \qquad \lambda > 0, \qquad \beta \in R^p \,.$$

It makes easier to use the identity

$$Q_1(\beta) + Q_2(\eta) + b \equiv g(\eta, \Delta) + h(\eta, \Delta, b)$$

with

$$g(\eta, \Delta) = (\beta - m)' \overset{-1}{\sum} (\beta - m)$$

$$m = [I + \Delta X'YX]^{-1}[\eta + n\Delta\bar{x}]$$

$$\sum = [\Delta^{-1} + X'YX]^{-1}$$

$$h(\eta, \Delta) = \eta'\Delta^{-1}\eta + 1_n'Y^{-1}1_n + b$$

$$-[\eta + n\Delta\bar{x}']\Delta^{-1}[I + \Delta X'YX]^{-1}[\eta + n\Delta\bar{x}]$$

From (5-4), the marginal posterior distribution of λ is obtained by integrating on β. It follows that

$$q(\lambda|y) \propto \lambda^{\frac{a+n}{2}-1} \exp\left\{\frac{-\lambda}{2}[h(\eta + \Delta + b)]\right\}$$

That is,

$$\lambda|y \sim \Gamma\left[\frac{a+n}{2}, \frac{h(\eta, \Delta, b)}{2}\right]. \qquad (5-5)$$

Hence, the Bayes estimator of λ with respect to the squared error loss function is

$$\begin{cases} \hat{\lambda}_B = E(\lambda|y) = \frac{a+n}{h(\eta, \Delta, b)} \\ \\ var(\lambda|y) = \frac{2(a+n)}{h(\eta, \Delta, b)^2} q \end{cases} \qquad (5\text{-}6)$$

For β, we integrate on λ in (5-4) and obtain

$$q(\beta|y) \propto \frac{1}{[(a+n) + (\beta - m)'\Psi^{-1}(\beta - m)]^{\frac{(a+n+p)}{2}}} \qquad (5-7)$$

$$\Psi, = \frac{h(\eta, \Delta, b)}{a+n}[\Delta^{-1} + X'YX]^{-1}$$

That is,

$$\beta|y \sim T_p[(a+n)m, V]$$

Thus, for $a + n > 2$ we have

$$\begin{cases} \hat{\beta}_B = E(\beta|y) = m = [I + \Delta X'YX^{-1}[\eta + n\Delta\bar{x}] \\ \\ var(\beta|y) = V = \frac{a+n}{a+n-2}\Psi \end{cases} \qquad (5\text{-}8)$$

By the distributional properties of the multivariate T-distribution, each subvector of β also has a T-distribution with corresponding mean and variance. Even each linear transfomation of β has a T-distribution, too. These facts prove useful in model selection, below. Therefore, the simple regression is estimated as

$$\hat{\mu}^{-1}_{i,B} = x'_i\hat{\beta}_B, \qquad i = 1, \ldots, n \qquad (5-9)$$

which marginally has a T-distribution. This fact helps us to find a Highest Posterior Density (HPD) region akin to a confidence interval in the classical approach.

5.4 Model selection

Selecting a particular model is tantamount to excluding some independent variables (x_j) from the full regression equation. This task can be accomplished either by testing the significance of various coefficients or by utilizing the Bayes factor. The former is the usual method in the classical approach. The latter method compares the two competing models $M_{j-1}(y)$ and $M_j(y)$ while $M_{j-1}(y)$ is a reduced from of $M_j(y)$, lacking some variables. Then

$$P.O. = posterior\ odds = \frac{\alpha_j}{1-\alpha_j}\frac{q(\beta|y, M_j)}{q(\beta|y, M_{j-1})},\ j = 1, 2, \ldots, p$$

where $0 < \alpha_j < 1$ is the prior probability that the $M_j(y)$ is the correct model. If $P.O. > 1$, the model $M_j(y)$ is more reasonable than $M_{j-1}(y)$, otherwise vice-versa.

5.5 Empirical Bayes analysis

In order to use the result (5-9) in practice, one has to be able to specify the actual values of the parameters in (5-1) and (5-2). This is not an easy task because there may not be a consensus among the users about the specified parameters. One way out of this difficulty is to let the data speak for themselvves and determine those parameters. The operational aspect of this procedure is to estimate the parameters of the prior distributions from the observed data and then follow the Bayesian rules. The marginal distribution of the random sample which is observed is

$$f(y|a,b,\eta,\Delta) = \int \cdots \int \left\{ \int_0^\infty f(y|\beta,\lambda)\pi_2(\beta|\lambda)\pi_1(\lambda)d\lambda \right\} d\beta$$

which by virtue of (5-1)-(5-3) and the properties of the gamma and multivariate T distributions reduces to

$$f(y|a,b,\eta,\Delta) = constant \frac{b^a[h(\eta,\Delta,b)]^{\frac{p}{2}}\Gamma(\frac{a+n}{2})}{[a+n]^{\frac{a+n+p}{2}}\Gamma(a)|I+\Delta X'YX|^{\frac{1}{2}}}$$

$$y \in [\mathbb{R}^+]^n \qquad (5-10)$$

That is, (5-10) is in fact the likelihood function of (a,b,η,Δ) which contains $2(p+1)$ parameters. Now, similar to the previous section, we can maximize (5-10) in terms of (a,b,η,Δ) to obtain the MLE of (a,b,η,Δ) which are denoted by $(\hat{a},\hat{b},\hat{\eta},\hat{\Delta})$. Next we subsitute these estimates in (5-6)-(5-8) to obtain the $\hat{\lambda}_{EBML}$ and $\hat{\beta}_{EBML}$ which in turn yield

$$\hat{\mu}^{-1}_{i,EBML} = x_i'\hat{\beta}_{EBML} \qquad (5-11)$$

which being a linear function of $\hat{\beta}_{EBML}$ has the variance as

$$var(\hat{\mu}^{-1}_{i,EBML}) = x_i'var(\hat{\beta}_{EBML})x$$

Since the asymptotic variance of $\hat{\beta}_{EBML}$ is obtianed by the inverse of the Fisher information matrix $I^{-1}(\hat{\beta}_{EBML})$, it follows that

$$var(\hat{\mu}_{i,EBML}) = x_i' I^{-1}(\hat{\beta}_{EBML}) x.$$

This martrix is used in finding the HPD region, or confidence bands for the regression equation (5-11).

The above procedure needs numerical computations to find $(\hat{a}, \hat{b}, \hat{\eta}, \hat{\Delta})$ from (5-10). Thus he solution has no closed from. In order to provide an explicit solution we may use the method of moment to find the estimatesof (a, b, η, Δ). If $Y \sim IG(\mu, \lambda)$ with the density $f(y|\mu, \lambda)$ then $R_i = Y_i^{-1}$ has the density

$$k(r|\mu, \lambda) = \mu r f(r|\mu^{-1}, \lambda\mu^{-2}), \quad , \mu, \lambda > 0$$

with moments of all order, Chhikara and Folks (1989, P. 43). Using this fact we find the first two marginal moments of R_i, *i.e.*

$$E(R_i) = \gamma + x_i'\eta \qquad (5-12)$$

and

$$var(R_i) = \gamma^2\xi + \gamma x_i'\eta + x_i'\Delta x_i', \quad i = 1, \ldots, n \qquad (5-13)$$

with $\gamma = \dfrac{b}{2(a-2)}$, $\xi = \frac{2(a-1)}{(a-4)}$. The equation (5-12) implies that

$$R_i = \gamma + x'\eta + \epsilon_i, \quad i = 1, \ldots, n$$

which is similar to the multiple linear regression and can be expressed as

$$R = Z\eta + \epsilon \qquad (5-14)$$

with $R = [R_1, \ldots, R_n]$, $Z = [1_n, X]$, $X = (x_{ij})$ $\eta = [\gamma, \eta]$, and $\epsilon = [\epsilon_1, \ldots, \epsilon_n]$. Then the least square solution of (5-13) is

$$\tilde{\eta}_1 = (Z'Z)Z'R \qquad (5-15)$$

A similar treatment of (5-13) leads to the system of equations

$$S_i^* = S_{(i)}^2 - \tilde{\gamma} x_i' \tilde{\eta} = \tilde{\gamma}^2 \xi + \sum_{j=1}^{p} x_{ij}^2 \cdot \delta_j, \quad i = 1, \ldots, n \qquad (5-16)$$

where, for $\bar{R}_{(i)} = \sum_{j \neq i}^{n} \frac{R_j}{(n-1)}$, we have defined

$$S_{(i)}^2 = \sum_{j \neq i}^{n} \frac{[R_j - \bar{R}_{(i)}]^2}{(n-2)}, \quad i = 1, \ldots, n.$$

Now, we define

$$S^* = [S_1^*, \cdots, S_n^*],$$

$$T = [\tilde{\gamma}^2 1_n, X^*], \;\; X^* = (x_{ij}^2)$$

$$\epsilon = [\epsilon_1, \ldots, \epsilon_n], \;\; \delta_c = \Delta 1_p = [\delta_1, \ldots, \delta_p]$$

and $\boldsymbol{\xi}_1 = [\xi, \delta_c]$. Thus, (5-16) is in fact as

$$S^* = T x i_1 + \epsilon \qquad (5-17)$$

From (5-17), through the least squares method, one obtains

$$\tilde{\boldsymbol{\xi}}_1 = (T'T)^{-1} T' S^* \qquad (5-18)$$

If it happens that in (5-15) and (5-18) the inverse matrices do not exist, their Moor-Penrose inverse would be used. We have, therefore, the estimates of γ, ξ, η and Δ. The pair γ and ξ gives us $a = \frac{2(2\xi-1)}{(\xi-2)}$ and $b = \frac{2\gamma(\xi+1)}{(\xi-2)}$. These estimates are denoted by $(\hat{a}, \hat{b}, \hat{\eta}, \hat{\Delta})$ which will be used n the Bayes estimates already obtained as (5-9). Thus, as an alternative to (5-11) we have

$$\mu_{i,EBMM}^{-1} = x_i' \hat{\beta}_{EBMM} \qquad (5-19)$$

with

$$var(\hat{\mu}_{i,EBMM}) = x_i' \mathrm{var}(\hat{\beta}_{EBMM}) x. \qquad (5-20)$$

The estimales obtained above either by using the ML or MM method are in fact Bayes estimates relative to the estimated proir distributions. Thus any inference concerning λ and β should be based on thier posterior distributions. For example in the case of the MM method we have

$$(\lambda|y) \sim \Gamma\left[\frac{(\tilde{a}+n)}{2}, \frac{h(\tilde{\eta}, \tilde{\Delta}, \tilde{b})}{2}\right], \qquad (5-21)$$

$$(\beta|y) \sim T_P[(\tilde{a}+n), \tilde{m}, \tilde{V}] \qquad (5-22)$$

Hence, the variance in (5-20) is found from (5-22). Likewise the distribution, and the HPD region for $\mu_{i,EBMM}^{-1}$ is found from (5-20). Table 5.1 shows the data set relsuted from an experiment on the turnip plant. It is believed that the explanatory variables. X_1=sunlight, X_2=soil humidity, and X_3=air temperature affect y =the content of vitamin B_2 in the leaves of the turnip plant. We would like to establish a relation between μ_i and these 3 variables in the from of $\mu_i = [x_i'\beta]^{-1}$. To this end, we define $Z = [1_{27}, X]$, where X is the last three coumns of Table 5.1. Using (5-15), we abtain the estimates

$$[\tilde{\gamma}, \tilde{\eta}_1, \tilde{\eta}_2, \tilde{\eta}_3] = 10^{-6}[12581, 2, 134, -28].$$

To obtain $\hat{\beta}$, we first compute the statistics

$$\bar{R}_{(i)} = \frac{(0.339 - R_i)}{26}, S_{(i)}^2 = 0.04[0.005 - R_i^2 - 26\bar{R}_{(i)}^2], \quad i = 1, \ldots, 27.$$

From these, we find S^* and T.Finally (5-18) gives us $\bar{\xi}$ which in turn provides

$$(\beta|y) \sim T_3[28, \tilde{m}, \tilde{V}]$$

with

$$\hat{\beta}_{EBMM} = \widetilde{m} = 10^{-7}[-87, 4762, 318]$$

$$\tilde{V} = \text{var}(\beta|y) = 10^{-6}\begin{bmatrix} 43505 & -205205 & 36960 \\ & 21945 & 2695 \\ & & 23870 \end{bmatrix}$$

and finally,from (5-19), we arrive at

$$\tilde{\mu}_{i,EBMM}^{-1} = 10^{-7}[-87x_{i1} + 4762x_{i2} + 318x_{i3}]$$

The use and interpretation of this regression equation is like the usual regression model. Hence we do not elaborate on it.

6. Concluding remarks

We have journyed a relatively long path starting from the genesis of the Inverse Gaussian distribution and heading towards inference about its parameters. Along the way we used the Bayesian and empirical Bayes vehicle to reach estimates, to do analysis of variance and to construct a regression equation. It is not the end of journey. we could have visited the subject of analysis of covariance, or could have studied the reliability problems under this model. the study of asymptotic properties of various estimators is yet another common ground to work on. The posibilities are vast and the problems endless. One could conjecture that there are as many problem under the Inverse Gaussian model as there are under the normal (Gaussian) model. Most of them are not yet explored and await for eager and perseverant workers.

Table 5.1: the data regarding the vitamin B_2 content of the turnip leaves and sunlight, soil humidity and air emperature

plant	Vitamin B_2	Sunlight	Soil Hmidity	Air temperature
number	Y	X_1	X_2	X_3
1	110.4	176	7.0	78
2	102.8	155	7.0	89
3	101.0	273	7.0	89
4	108.4	273	7.0	72
5	100.7	256	7.0	84
6	100.3	280	7.0	87
7	102.2	280	7.0	74
8	93.7	184	7.0	87
9	98.9	216	7.0	88
10	96.6	198	2.0	76
11	99.4	59	2.0	65
12	96.2	80	2.0	67
13	99.0	80	2.0	62
14	88.4	105	2.0	70
15	75.3	180	2.0	73
16	92.0	180	2.0	65
17	82.4	177	2.0	76
18	77.1	230	2.0	82
19	74.0	203	47.4	76
20	65.7	191	47.4	83
21	56.8	191	47.4	82
22	62.1	191	47.4	69
23	61.0	76	47.4	74
24	53.2	213	47.4	76
25	59.4	213	47.4	69
26	58.7	151	47.4	75
27	58.0	205	47.4	76

Source: Draper and Smith (1981, P.406)

References

[1] Achcar, R. J.; Bolfarin, H. and Rodrigues, J. (1991). Inverse Gaussian distribution: a Bayesian approach. *Brazilian Journal of Statistics*, 5, 8194.

[2] Achcar, R.J. and Rosales, O.L.A. (1992). A Bayesian approach for accelerated life test assuming an Inverse Gaussian distribution, *Estadistica*, 2,2532.

[3] —(1993). Use of Bayesian methods in the analysis of two-factor experiments under an Inverse Gaussian distribution. ASA-IASI 3rd school of Regression Models, Brazil.

[4] Babu, Gutti Jogesh and Chaubey, Yogendra, P. (1996). Asymptotics and Bootstrap for inverse Gaussian Regression. Annals of the Institute of Statistical Mathematics. 48, pp. 75-88.

[5] Bachelier, L. (1900). Theorie de la speculation. *Ann. Sci.Ec. Norm. Super.*, Paris, 17(3):21-86.

[6] Banerjee, A.K. and Bhattacharyya, G.K. (1979). Bayesian results for the Inverse Gaussian distribution with an application. *Technometrics*, 21, 247251.

[7] — (1976). A purchase incidence model with Inverse Gaussian interpurchase times. *Jour of Amer. Statist Assoc.* 71,823-829.

[8] Bhattacharyya, G.K. and Fries, A. (1983). Analysis of two-factor experiments under an Inverse Gaussian model. *Jour. of Amer. Statist. Assoc.* 78,820-826.

[9] Bhattacharyya, G.K., and A. Fries (1982). Inverse Gaussian regression and accelerated life tests, in *Survival Analisis,* edited by John

Crowly and Richard A. Johnson, IMS Lecture Notes, Monograph Series, pp.101-118.

[10] Brown, R. (1828). A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Phil. Mag., Series 2,* 4:161-173

[11] Chhikara, R.S. and Folks. L.(1989). *Inverse Gaussian distribution, Theory, methodology, and applications. Marcel Dekker, New York.*

[12] Chhikara, R.S. and Irwin Guttman(1982). Prediction limits for the inverse Gaussian distribution. *Technometrics*, 24:319-324.

[13] Draper, N. and Smith, H.(1981). *Applied Regression Analysis*, 2nd ed. New York, Wiley.

[14] Eaton, W.W., and G.A. Whitmore (1977). Length of stay as a stochastic process: a general approach and application to hospitalization for schizofrenia, *J. Math. Sociol.*, 5, pp. 273-292.

[15] Einestein, A.(1905). Investigations on the theory of Brownian movement, edited with notes by R. Furth, translated by A. D. Cowper, 1956ed. New York: Dover.

[16] Folks, J.L. and Chhicara R.S.(1978). The Inverse Gaussian distribution and its statistical application-a review. *J.R. Statist. Soc.* B,40, 263-275

[17] Homayun-Aria, Shaheen (1996). Empirical Bayes estimation of the parameters of the Inverse Gaussian distribution. M.S. Thesis Shahid Beheshti University, Iran.

[18] Hsieh, H.K. and R.M. Korwar (1990). Inadmissibility of the UMVU estimators of the Inverse Gaussian Variance. *COMMU. STATIST. THEORY METH.* 19(7),pp. 2509-2516.

[19] Iwase, K. (1989). Linear regression through the origin with constant coefficient of variation for the Inverse Gaussian distribution. *COMMUN. STATIST. THEOR. METH.*, 18 (10)PP 3587-3593.

[20] Lancaster, Tony (1972). A stochastic model for the duration of a strike *J. Roy. Statist.* Soc. Ser. A, 135,pp. 257-271.

[21] Meshkani, M.R(1996).*One-way and two-way analysis of variance for Inverse Gaussian distribution by empirical Bayes procedure.* Report of a Research Project. Shahid Beheshti University, (in Persian).

[22] Meshkani, M. R.(1999). Empirical Bayes analysis of regression under the Inverse Gaussian model. Report of a Research Project. Shahid Beheshti University (in persian)

[23] O'Hagan, A. (1994). Kendall's Advanced Theory of Statistics Vol. 2B, Bayesian London. Inference. Edward Arnold.

[24] Patel, J. K. and Read, C. B. (1996). *Handbook of the Normal Distribution, 2^{nd}ed.* Marcel Dekker, New York.

[25] Press, S.J. (1972).*Applied Multivariate Analysis.* New York, Holt, Rinehart, and Wilson, Inc.

[26] Schrodinger, E.(1915). Zur theorie der fall-und Steigversuche an teilchen mit Brownscher bewegung. *phys. ze.,* 16: 289-295.

[27] Seshadri, V. (1999). Inverse Gaussian distribution: Statistical theory and applications. New York. Springer-Verlag.

[28] Seto, N. and Iwase, K. (1985). UMVU estimators of the model and limits of an interval for the Inverse-Gaussian distribution *COMMUN. STATIST-THEORY. METH.*, 14(5), pp. 1151-1161.

[29] Shah, A. K. (1985). A Simpler Approximation for Areas Under the Standard Normal Curve, *American Statistician* 39, 80, 327.

[30] Smoluchowski, M.V.(1915). Notiz uber die berechnung der Browschen molekular-bewegung bei der ehrenhaft-milikanschen versuchsanord-nung- *Phys. Ze.*, 16:318-321.

[31] Shuster, J.J. and Muira, C.(1972). Two-way analysis of resiprocals. *Biometrika.* 59,478-481.

[32] Tweedie, M.C.K.(1941). A mathematical investigation of some electrophoretic measurements on Colloids, M.Sc. Thesis, University of Reading, England.

[33] Tweedie, M.C.K.(1945). Inverse statistical variates. *Nature,* 155:453.

[34] Tweedie, M.C.K. (1957) Statistical properties of inverse Gaussian distributions I. Ann. Math. Statist., 28pp. 362-377.

[35] Tweedie, M.C.K. (1957 a). Statistical properties of inverse Gaussian Distributions I. *Ann. Math. Statist.*, 28, 362-377.

[36] Wald. A. (1944). on cumulative sum of random variables. *Ann. Math. statist.*, 15: 283-296.

[37] Whitmore, G.A. (1979). An Inverse Gaussian model for labour turn over. J. Roy. Statist. Soc. Ser. A, 142, pp. 468-478.

[38] Whitmore, G.A. (1983). A regression method for censored Inverse-Gaussian data. *The Candian Journal of Statistics.* Vol. 11, No. 4, pp. 305-315.

[39] Whitmore, G.A. (1986). Inverse Gaussian ratio estimation. *Appl. Statist.,* 35, pp. 8-15.

[40] Wise, M.E. (1966). Tracer dillution curves in cardiology and random walk and lognormal distributions. *Acta Physiologica pharamalogica Neerlandica,* 14, pp. 175-204.

[41] Woldie, M. and J.L. Folks (1992). Power function for Inverse Gaussian regression models. ASA *Business and Stat. Section Proc. 38th ed.* pp. 187-191.

[42] Woldie, M. and J. L. Folks (1994). Inverse Gaussian regression methodlogy- A review. ASA *Business and Stat. Section Proc.* 40 th ed. pp. 445-450.

[43] Woldie, M. and J.L. Folks (1995). Calibration for Inverse Gaussian regression. *COMMU. STATIST. THEORY METH.*, 24(10), pp. 2609-2620.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 147-154 (2004)

A Note on the Concept of Limit

M.R. Fadaee and M. Radjabalipour

*Department of Mathematics,
University of Kerman, Kerman, Iran
radjab@arg3.uk.ac.ir*

Abstract: The paper investigates the problems that students usually have when beginning the study of the notion of limit in its understanding and usage.

I. Convergence in terms of sequences

A very common problem that most beginners will have with the definition of limit is to adjust the static nature of the definition with the dynamic nature of the words "tending", "converging", that they have in their minds.

Another problem is the difficulty of establishing the inequality $|x_n - x| < \varepsilon$ for large values of n in order to prove $\{x_n\}$ converges to x as $n \to \infty$.

MSC (2000): 26A03, 97C90

Keywords: Limit, Continuity, $\epsilon - \delta$ definition, Sequential approach, Heuristic definition of limit

A common method to solve the latter inequality for n is to prove first that $|x_n - x| \leq a_n^{-1}$ for some unbounded increasing sequence $\{a_n\}$ and then replace the inequality $|x_n - x| < \varepsilon$ by the sufficient condition $a_n > \varepsilon^{-1}$. The new inequality has a solution of the form $n > N_\varepsilon$ for some natural number N_ε. In most practical examples the sequence $\{a_n\}$ and the number N_ε can be obtained through algebraic or other simple methods. For example, for $\lim_{n\to\infty} \frac{n + \sin n}{2n - \cos n + 5} = \frac{1}{2}$, we have

$$|\frac{n + \sin n}{2n - \cos n + 5} - \frac{1}{2}| \leq (\frac{n}{2} + 1)^{-1} < \varepsilon,$$

and hence

$$N_\varepsilon = \max\{1, [[2(\varepsilon^{-1} - 1)]]\},$$

where $[[t]]$ denotes the greatest integer in t.

In general, the sequence $\{a_n\}$ may not have a simple expression but, as the following proposition shows, it always exists.

I.1. Proposition. Let $\{x_n\}$ be a sequence in a metric space X converging to $x \in X$; i.e., for every $\varepsilon > 0$ there exists $N_\varepsilon \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n > N_\varepsilon$, where d denotes the metric on X. Then there exists an unbounded increasing sequence $\{a_n\}$ of positive numbers such that $d(x_n, x) \leq a_n^{-1}$ $(n = 1, 2, \ldots)$.

Proof. Let N_1 be a positive integer such that $d(x_n, x) < 1$ for all $n > N_1$ and define

$$a_1 = a_2 = \cdots = a_{N_1} = (\max\{d(x_1, x), d(x_2, x), \ldots, d(x_{N_1}, x), 1\})^{-1}.$$

Assume by induction that $N_1, N_2, \ldots, N_k$ and $a_1, a_2, \ldots, a_{N_k}$ are defined for some $k \geq 1$. Let $N_{k+1} > N_k$ be a positive integer such that $d(x_n, x) < 1/(k+1)$ for all $n > N_{k+1}$. Define

$$a_{N_k+1} = a_{N_k+2} = \cdots = a_{N_{k+1}} = k.$$

In view of the Archimedes principle or the axiom of completeness, the increasing sequence $\{a_n\}$ thus obtained is unbounded. Moreover, $d(x_n, x) \leq a_n^{-1}$ $(n = 1, 2, \dots)$.■

It is clear that if a sequence $\{x_n\}$ in a metric space (X, d) satisfies $d(x_n, x) \leq a_n^{-1}$ $(n = 1, 2, \dots)$ for some $x \in X$ and some unbounded increasing sequence $\{a_n\}$, then $\lim\limits_{n \to \infty} x_n = x$. Thus, one can see that the axiom of completeness and the definition of $\lim\limits_{n \to \infty} x_n$ are equivalent to the following Axiom I.2 and Definition I.3.

I.2. Axiom. If $y_1 < y_2 < \dots$ is bounded in $\mathbb{R}$, then there exists $y \in \mathbb{R}$ such that $(y - y_n)^{-1}$ is unbounded.

I.3. Definition. A sequence $\{x_n\}$ in a metric space X is said to be convergent to some $x \in X$ if $d(x_n, x) \leq 1/a_n$ $(n = 1, 2, \dots)$ for some unbounded increasing sequence $\{a_n\}$ of positive numbers.

Axiom I.2 guarantees the unboundedness of the increasing sequences $\{n\}$; if $\{n\}$ is bounded then there exists $y \in \mathbb{R}$ such that $\{(y - n)^{-1}\}$ is unbounded. Hence $1 = (y - n) - (y - n - 1) \leq y - n$ for all $n \in \mathbb{N}$; a contradiction. This shows that, in the light of Axiom I.2, Definition I.3 is not vacuous. The inequality $0 \leq d(x_n, x) \leq a_n^{-1}$ reflects the compression of a sequence $\{x_n\}$ to x by the boundary of a shrinking sphere of radius a_n^{-1} centered at x. The terms "compress" and "shrink" have dynamical natures which one expects in the notion of limit.

If the standard $\varepsilon - N$ definition of $\lim\limits_{n \to \infty} x_n = x$ is replaced by the new definition, then one can successfully prove and discuss all necessary results in a calculus or advanced calculus course related to limits of sequences. (The definition of $\lim\limits_{n \to +\infty} x_n = \pm\infty$ can be easily given as $x_n > a_n$ or $x_n < -a_n$ for some unbounded increasing sequence $\{a_n\}$.)

Moreover, limit, continuity, and differentiability of functions can be given in terms of sequences. In fact $\lim\limits_{x \to c} f(x) = L$ can be defined as the requirement $\lim\limits_{n \to \infty} f(x_n) = L$ for all sequences $\{x_n\}$ in the domain of f

such that $x_n \neq c$ and $\lim_{n\to\infty} x_n = c$. (For continuity of f at c we require $\lim_{n\to\infty} f(x_n) = f(c)$ for any sequence $\{x_n\}$ in the domain of f converging to c.)

Thus, to prove the composition fog of continuous functions f and g is continuous, we simply observe $\lim_{n\to\infty} f(g(x_n)) = f(\lim_{n\to\infty} g(x_n)) = f(g(c))$ whenever $\{x_n\}$ is in the domain of g converging to c. The intermediate value theorem, the compactness of a closed interval, and the convergence of a Cauchy sequence, etc. can be proved in similar sequential fashions. For example, if a continuous function $f : [a,b] \to \mathbb{R}$ satisfies $f(a) \leq k \leq f(b)$, one can construct a nest of intervals $[a,b] \supset [a_1,b_1] \supset [a_2,b_2] \supset \ldots$ such that $b_n - a_n = (b-a)/2^n$ and $f(a_n) \leq k \leq f(b_n)$. Hence, if $c = \lim a_n = \lim b_n$, then $0 \leq (k - f(c))^2 = \lim_{n\to\infty} (k - f(a_n))(k - f(b_n)) \leq 0$ and thus $k = f(c)$.

II. Shortcomings of the sequential definitions

As we mentioned in the previous section, the sequential definition of convergence provides a dynamic spirit for the concept of limit. Even in the case of functions, students expect a dynamical approach. In fact, most textbooks and instructors of calculus respond to this need by beginning the subject of limit with tables showing the values of a function as simple as x^2 at various points $x_1, x_2, \ldots$ which cluster around a given point $x = c$ and demonstrate that as x_n approaches c, x_n^2 approaches c^2. This means that while the mathematicians of the nineteenth century regarded continuity of a curve as the free movement of a point on a plane, the twentieth century mathematicians regard the continuity of a function f as the convergence of the sequence $\{f(x_n)\}$ to $f(c)$ for any sequence $\{x_n\}$ converging to c in the domain of f. It is well known that this definition of continuity is equivalent with the standard $\varepsilon - \delta$ definition: For all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, c) > 0$ such that $|f(x) - f(c)| < \varepsilon$

whenever $|x - c| < \delta$. (Absolute values can be replaced by distances in metric spaces.) If δ is independent of the point c, the continuity of f is said to be uniform. It is not easy to give a definition of uniform continuity purely in terms of sequences; even if it is done, it is not handy to be used in the proofs based on uniform continuity. For example, one may define uniform continuity as having $\lim\limits_{n\to\infty} |f(x_n) - f(y_n)| = 0$ whenever $|x_n - y_n| \to 0$. But such a definition is not convenient for proving the integrability of continuous functions. Similarly, uniform convergence of sequences of functions are easily defined and handled by the $\varepsilon - N$ definition; we found it quite artificial to use sequential definitions to prove, for instance, that uniform limit of a sequence of continuous functions is a continuous function.

In general, we believe the $\varepsilon - N$ or $\varepsilon - \delta$ definitions of limit (for sequences or functions) are indispensable parts of mathematics. They are not only suitable for problems involving uniformity, but are the gates to important subjects such as topology. The $\varepsilon - \delta$ definition introduces the concept of neighborhoods and intimates the beginner with the manipulation of abstract topological concepts.

With all these, we still believe the (dynamic) sequential definitions of limit and continuity are worth to be included in calculus textbooks. This is what a student expects when for the first time he/she encounters the notion of convergence. Proposition I.1 and the paragraph preceding Axiom I.2 showed that the $\varepsilon - N$ definition is equivalent to Definition I.3. Thus, a beginner mathematician may begin the subject of limit by Definition I.2 and then observe the standard $\varepsilon - N$ definition as a theorem. The difference between something to be regarded as a "definition" or as a "theorem" is that in the first case one has to accept a metafore of a concept which may differ from one's expectation of that concept, while in the second case one defends a claim that one has successfully proved.

However, once the equivalence of the two methods of approach are established, it is easy to switch from one method to another. The following section shows that how combination of the two methods may help to shorten the proofs of the theorems.

III. Uniformity and integrals

The Riemann integral in calculus is a concept which benefits lot from uniformity. As mentioned before, a function f is uniformly continuous if and only if $d(f(x_n), f(y_n)) \to 0$ whenever $d(x_n, y_n) \to 0$ and $\{x_n\}$ and $\{y_n\}$ are sequences in the domain of f. A sequence $\{f_n\}$ is said to be uniformly convergent to a function f on a set A, if for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $n > N$ implies that $|f_n(x) - f(x)| < \varepsilon$ for all $x \in A$. This is equivalent with saying that $|f_n(x) - f(x)| \leq a_n^{-1}$ for all $x \in A$, where $\{a_n\}$ is an unbounded increasing sequence (independent of x).

The Riemann integral of f on $[a, b]$ exists if, by definition, there exists an increasing sequence $\{\pi_n\}$ of partitions of $[a, b]$ such that $\lim\limits_n \ l(\pi_n) = \lim\limits_n[\overline{S}(f, \pi_n) - \underline{S}(f, \pi_n)] = 0$, where $l(\pi) := \max\limits_k(x_k - x_{k-1})$, $\overline{S}(f, \pi) := \sum\limits_k M_k(\pi)(x_k - x_{k-1})$, $\underline{S}(f, \pi) := \sum\limits_k m_k(\pi)(x_k - x_{k-1})$, $M_k(\pi) := \sup\{f(x) : x_{k-1} \leq x \leq x_k\}$, and $m_k(\pi) := \inf\{f(x) : x_{k-1} \leq x \leq x_k\}$ for a given partition $\{a = x_0 < x_1 < \cdots < x_n = b\}$. Then $\displaystyle\int_a^b f$ is defined to be the limit of the increasing sequence $\{\underline{S}(f, \pi_n)\}$. As an immediate consequence, $|f|$ is integrable and $\displaystyle|\int_a^b f| \leq \int_a^b |f|$, if $\displaystyle\int_a^b f$ exists.

Note. If π_n' is a refinement of π_n and if $\{\pi_n\}$ is as in the definition of $\int_a^b f$, it is easy to see that $\{\pi_n'\}$ also defines the integrability of f and yields the same value for $\int_a^b f$. Thus $\int_a^b f$ is independent of the choice of $\{\pi_n\}$. For a partition π as above we further set $\mu(\pi) = \max\limits_k[M_k(\pi) - m_k(\pi)]$.

The proof of the following results would be lengthy if one tries to stick to one type of definitions. The proofs given here seem to be very short.

III.1. Theorem. *Let $f:[a,b]\to \mathbb{R}$ be bounded and let $\{\pi_n\}$ be an increasing sequence of partitions of $[a,b]$ such that $\lim\limits_{n\to\infty} l(\pi_n)=0$. Then $\int_a^b f$ exists if either the sequence $v_n:=\sum\limits_k [M_k(\pi_n)-m_k(\pi_n)]$ is bounded or $\lim\limits_n \mu(\pi_n)=0$. In particular, if f is monotone or continuous, then it is Riemann integrable.*

Proof. If f is monotone or, more generally, if $v_n\le D$ for all $n\in \mathbb{N}$, then $\overline{S}(f,\pi_n)-\underline{S}(f,\pi_n)\le l(\pi_n)D$ for all $n\in\mathbb{N}$ and hence $\int_a^b f$ exists. If f is continuous or, more generally, if $\lim\limits_n \mu(\pi_n)=0$, then $\overline{S}(f,\pi_n)-\underline{S}(f,\pi_n)\le (b-a)\mu(\pi_n)$ for all $n\in\mathbb{N}$ and hence $\int_a^b f$ exists.■

III.2. Theorem. *Let f be the uniform limit of a sequence of Riemann integrable functions $f_n:[a,b]\to\mathbb{R}$. Then $\int_a^b f$ exists.*

Proof. Let $\{\pi_n\}$ be an increasing sequence of partitions of $[a,b]$ with $\lim\limits_n l(\pi_n)=0$. Assume $\int_a^b f$ does not exist. Then $\overline{S}(f,\pi_n)-\underline{S}(f,\pi_n)\ge \varepsilon_0$ for all $n\in\mathbb{N}$ and some $\varepsilon_0>0$. Let $\{a_n\}$ be an unbounded increasing sequence establishing the uniform convergence of $\{f_n\}$. Since

$$M_k(f,\pi_n)-m_k(f,\pi_n)\le M_k(f_i,\pi_n)-m_k(f_i,\pi_n)+2a_i^{-1},$$

it follows that

$$0<\varepsilon_0\le \overline{S}(f_i,\pi_n)-\underline{S}(f_i,\pi_n)+2a_i^{-1}(b-a) \quad , \quad (i=1,2,\cdots).$$

Letting $n\to\infty$, it follows that $0<\varepsilon_0\le 2a_i^{-1}(b-a)$ and hence $a_i\le 2(b-a)/\varepsilon_0$, $(i=1,2,\ldots)$; a contradiction.■

We conclude this topic with another famous theorem.

III.3. Theorem. *If $\int_a^b f$ exists and if φ is continuous on some closed interval containing $f([a,b])$, then $\int_a^b \varphi\circ f$ exists.*

Proof. The $\varepsilon-\delta$ definition of continuity implies that φ is the uniform limit of a sequence $\{\varphi_n\}$ of piecewise linear continuous functions which can be expressed in the form

$$\varphi_n(x) = a + bx + \sum_{k=0}^{m} c_k|x - x_k|,$$

where the constant $m, a, b, c_0, c_1, \cdots, c_m, x_0, x_1, \cdots, x_m$ depend on n. The rest of the proof follows from Theorem III.2 and the fact that $\varphi_n \circ f$ is the sum of integrable functions.■

Bibliography

For a history of the concepts discussed in this paper we may refer to [1]. The details of this work is given in [2]. For a similar approach we refer to [3].

1. Edwards, Jr., C. H., The Historical Development of the Calculus, Springer-Verlag, New York, Heidelberg, Berlin 1979.

2. Fadaee, M. R., On the Concept of Limit; New Reformulations, Ph. D. Thesis, Univ. of Kerman, Kerman, Iran (under preparation)

3. Gillman, L., Rigor in calculus, Notices AMS, 44, No. 8(1997), 932-934.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 155-177 (2004)

Orthogonal Functions in the Calculus of Variations and Optimal Control

Mohsen Razzaghi

Department of Applied Mathematics,
Amirkabir University of Technology, Tehran, Iran
razzaghi@aut.ac.ir

Abstract: The solution of problems in the calculus of variations is obtained by using hybrid functions. The properties of the hybrid functions which consist of block-pulse functions plus Legendre polynomials and block-pulse functions plus Chebyshev polynomials are presented. Two examples are considered, in the first example the brachistochrone problem is formulated as a nonlinear optimal control problem, and in the second example an application to a heat conduction problem is given. The operational matrix of integration in each case is introduced and is utilized to reduce the calculus of variations problems to the solution of algebraic equations. The method is general, easy to implement and yields very accurate results.

1. Introduction

There has been a considerable renewal of interest in the classical problems of the calculus of variations both from the point of view of mathematics and of applications in physics, engineering, and applied mathematics .

Finding the brachistochrone, or path of quickest decent, is a historically interesting problem that is discussed in virtually all textbooks dealing with the calculus of variations. In 1696, the brachistochrone problem was posed as a challenge to mathematicians by John Bernoulli. The solution of the brachistochrone problem is often cited as the origin of the calculus of variations as suggested in [1]. The classical brachistochrone problem deals with a mass moving along a smooth path in a uniform gravitational field. A mechanical analogy is the motion of a bead sliding down a frictionless wire. The solution to this problem has been obtained by various methods such as the gradient method [2], successive sweep algorithm [3-4] , the classical Chebyshev method [5] and multistage Monte Carlo method [6].

Orthogonal functions (OF's) have received considerable attention in dealing with various problems of dynamic systems. The main characteristic of this technique is that it reduces these problems to those of solving a system of algebraic equations; thus greatly simplifying the problem . The approach is based on converting the underlying differential equations into an integral equation through integration, approximating various signals involved in the equation by truncated orthogonal series and using the operational matrix of integration P, to eliminate the integral operations. The form of P depends on the particular choice of the orthogonal functions. Special attention has been given to applications of Walsh functions [7], block-pulse functions [8], Laguerre series [9], Legendre polynomials [10] and Chebyshev polynomials [11].

There are three classes of sets of OF's which are widely used. The first includes sets of piecewise constant basis functions (PCBF'S) (e.g., Walsh, block-pulse, etc.). The second consists of sets of orthogonal polynomials (OP's) (e.g., Laguerre, Legendre, Chebyshev, etc.). The third is the widely used sets of sine-cosine functions (SCF's) in Fourier series. While OP's and SCF's together form a class of continuous basis functions, PCBF's have inherent discontinuities or jumps. The inherent features(continuity or discontinuity) of a set of OF's largely determine their merit for application in a given situation. References [12] and [13] have demonstrated the advantages of PCBF spectral methods over Fourier spectral techniques. If a continuous function is approximated by PCBF's, the resulting approximation is piecewise constant. On the other hand if a discontinuous function is approximated by continuous basis functions the discontinuities are not properly modeled. Signals frequently have mixed features of continuity and jumps. These signals are continuous over certain segments of time, with discontinuities or jump occuring at the transitions of the segments. In such situations, neither the CBF's nor PCBF's taken alone would form an efficient basis in the representation of such signals.

The direct method of Ritz and Galerkin in solving variational problems has been of considerable concern and is well covered in many textbooks [14], [15]. Chen and Hsiao [7] introduced the Walsh series method to variational problems. Due to the nature of the Walsh functions, the solutions obtained were piecewise constant. Hwang and Shih [9], Chang and Wang [10] and Horng and Chou [11], used Laguerre polynomials, Legendre polynomials and Chebyshev polynomials respectively to derive continuous solutions for the first example in [7]. Furthermore, Razzaghi and Razzaghi [16], [17] applied Fourier series and Taylor series respectively to derive continuous solution for the second example in [7] which is an application to the heat conduction problem. It is shown in Razzaghi and Razzaghi [17] that, to obtain the Taylor series coefficient, an ill-conditioned matrix commonly known as the Hilbert matrix is used. Hence the Taylor series is not suitable for the solution of the second example in [7].

In the present paper we introduce a new direct computational method to solve problems of the calculus of variations. The method consists of reducing the variational problems into a set of algebraic equations by first expanding the candidate functions as hybrid functions with unknown coefficients. The hybrid functions, which consists of block-pulse functions plus

a) Legendre polynomials and

b) Chebyshev polynomials

are first introduced. The operational matrix of integration in each case is given and is used to evaluate the coefficients of hybrid functions in such a way that the necessary conditions for extremization are imposed. Two examples are considered. In example 1, the brachistochrone problem is formulated as an optimal control problem and in the second example we will demonstrate the application of operational matrix of integration for hybrid functions by considering the second example in [7]. It is shown that the hybrid functions of block-pulse and Legendre polynomials approach produces an exact solution for the heat conduction problem.

2. Properties of Hybrid Functions of Block-Pulse and Legendre Polynomials

Hybrid functions $b(n, m, t), n = 1, 2, \cdots, N, m = 0, 1, \cdots, M-1$, have three arguments; n is the order of block-pulse functions, m is the order of Legendre polynomials, and t is the normalized time. They are defined on the interval $[0,t_f)$ as

$$b(n,m,t)=\begin{cases}P_m(\frac{2N}{t_f}t-2n+1), & t\in[(\frac{n-1}{N})t_f,\frac{n}{N}t_f)\\ 0, & \text{otherwise.}\end{cases} \tag{1}$$

Here $P_m(t)$ are the well-known Legendre polynomials of order m which are orthogonal with respect to the weight function $w(t)=1$ and satisfy the following recursive formula.

$$P_o(t)=1,\quad P_1(t)=t$$

$$P_{m+1}(t)=\left(\frac{2m+1}{m+1}\right)tP_m(t)-\left(\frac{m}{m+1}\right)P_{m-1}(t)\ ,\quad m=1,2,3,\cdots$$

Since $b(n,m,t)$ consists of block-pulse functions and Legendre polynomials, which are both complete and orthogonal, the set of hybrid functions of block-Pulse and Legendre polynomials is a complete orthogonal set.

2.1. Function Approximation

A function $f(t)$, defined over the interval 0 to t_f may be expanded as

$$f(t)=\sum_{n=1}^{\infty}\sum_{m=0}^{\infty}c(n,m)b(n,m,t), \tag{2}$$

where

$$c(n,m)=(f(t),b(n,m,t))$$

in which $(.,.)$ denotes the inner product. If the infinite series in Eq. (2) is truncated, then Eq. (2) can be written as

$$f(t)\simeq\sum_{n=1}^{N}\sum_{m=0}^{M-1}c(n,m)b(n,m,t)=C^TB(t), \tag{3}$$

where

$$C=[c(1,0),\cdots,c(1,M-1),c(2,0),\cdots,c(2,M-1),\\ \cdots,c(N,0),\cdots,c(N,M-1)]^T, \tag{4}$$

and

$$B(t) = [b(1,0,t),\cdots,b(1,M-1,t)|b(2,0,t),\cdots,b(2,M-1,t)| \\ \cdots|b(N,0,t),\cdots,b(N,M-1,t)]^T. \qquad (5)$$

2.2. The Operational Matrix of the Hybrid of Block-pulse and Legendre Polynomials.

The integration of the vector $B(t)$ defined in Eq. (5) can approximated by

$$\int_0^t B(t')dt' \simeq PB(t) \qquad (6)$$

where P is the $NM \times NM$ operational matrix for integration and is given by

$$P = \begin{pmatrix} E & H & H & \cdots & H \\ 0 & E & H & \cdots & H \\ 0 & 0 & E & \cdots & H \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & E \end{pmatrix}. \qquad (7)$$

In Eq. (7)

$$H = \frac{t_f}{N} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

and E is operational matrix of integration for Legendre polynomials on the interval $[(\frac{n-1}{N})t_f, \frac{n}{N}t_f]$ given in [18] by

$$E = \frac{t_f}{2N}\begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{-1}{3} & 0 & \frac{1}{3} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{-1}{5} & 0 & \frac{1}{5} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{-1}{2M-3} & 0 & \frac{1}{2M-3} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{-1}{2M-1} & 0 \end{pmatrix}.$$

2.3. The Approximation of $B(t)B^T(t)C$.

The following property of the product of two Legendre polynomial vectors will also be used.

Let

$$P(t) = [P_0(t), P_1(t), \dots, P_{M-1}(t)]^T,$$
$$A = [a_0, a_1, \cdots, a_{M-1}]^T.$$

Then we have

$$P(t)P^T(t)A = \tilde{A}P^T(t), \tag{8}$$

where $\tilde{A}$ is an $M \times M$ matrix given in [18].

Let

$$B_n(t) = [b(n,0,t), b(n,1,t), \dots, b(n,M-1,t)]^T, \quad n = 1,2,\dots,N,$$

$$\bar{C}_n = [c(n,0), c(n,1), \dots, c(n,M-1)]^T, \quad n = 1,2,\dots,N.$$

Then using Eqs. (4) and (5) we get

$$B(t) = [B_1(t), B_2(t), \cdots, B_N(t)]^T, \tag{9}$$

$$C = [\bar{C}_1, \bar{C}_2, \cdots, \bar{C}_N]^T. \tag{10}$$

By using Eqs. (9) and (10) we obtain

$$B(t)B^T(t)C = \begin{pmatrix} B_1(t)B_1^T(t)\bar{C}_1 & 0 & \cdots & 0 \\ 0 & B_2(t)B_2^T(t)\bar{C}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_N(t)B_N^T(t)\bar{C}_N \end{pmatrix}. \tag{11}$$

Similarly to Eq. (8) we have

$$B_n(t)B_n^T(t)\bar{C}_n = \tilde{\bar{C}}_n B_n(t), \qquad n = 1, 2, \cdots, N. \tag{12}$$

Using Eqs. (11) and (12), we get

$$B(t)B^T(t)C = \tilde{C}B(t), \tag{13}$$

where $\tilde{C}$ is an $NM \times NM$ diagonal matrix given by

$$\tilde{C} = \begin{pmatrix} \tilde{\bar{C}}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\bar{C}}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\bar{C}}_N \end{pmatrix}.$$

2.4. Integration of $B(t)B^T(t)$.

The integration of the cross product of two hybrid Legendre vectors can be obtained as

$$D = \int_0^{t_f} B(t)B^T(t)dt. \tag{14}$$

where D is a diagonal matrix, given by

$$D = \begin{pmatrix} L & 0 & \cdots & 0 \\ 0 & L & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L \end{pmatrix}, \tag{15}$$

with L the $M \times M$ diagonal matrix given by

$$L = \frac{t_f}{N}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{2M-1} \end{pmatrix}.$$

3. Properties of Hybrid Functions of Block-Pulse and Chebyshev Polynomials.

Hybrid functions $\hat{b}(n,m,t), n = 1,2,\cdots,N, m = 0,1,\cdots,M-1$, have three arguments; n is the order of block-pulse functions, m is the order of Chebyshev polynomials, and t is the normalized time. They are defined on the interval $[0,t_f)$ as

$$\hat{b}(n,m,t) = \begin{cases} T_m(\frac{2N}{t_f}t - 2n + 1), & t \in [(\frac{n-1}{N})t_f, \frac{n}{N}t_f) \\ 0, & \text{otherwise.} \end{cases} \qquad (16)$$

Here $T_m(t)$ are the well-known Chebyshev polynomials of order m which are orthogonal with respect to the weight function $w(t) = \dfrac{1}{\sqrt{1-t^2}}$ and satisfy the following recursive formula.

$T_0(t) = 1, \quad T_1(t) = t$

$T_{m+1}(t) = 2tT_m(t) - T_{m-1}(t) \ , \quad m = 1,2,3,\ldots$

Since $\hat{b}(n,m,t)$ consists of block-pulse functions and Chebyshev polynomials, which are both complete and orthogonal, the set of the hybrid functions of block-pulse and Chebyshev polynomials is a complete orthogonal set.

3.1. Function Approximation.

A function $f(t)$, defined over the interval 0 to t_f may be expanded as

$$f(t) = \sum_{n=1}^{\infty} \sum_{m=0}^{\infty} c(n,m)\hat{b}(n,m,t), \tag{17}$$

where

$$c(n,m) = (f(t), \hat{b}(n,m,t))$$

in which $(.,.)$ denotes the inner product. If the infinite series in Eq. (17) is truncated then Eq. (17) can be written as

$$f(t) \simeq \sum_{n=1}^{N} \sum_{m=0}^{M-1} c(n,m)\hat{b}(n,m,t) = C^T \hat{B}(t), \tag{18}$$

where

$$C = [c(1,0), \cdots, c(1,M-1)|c(2,0), \cdots, c(2,M-1) \\ | \cdots |c(N,0), \cdots, c(N,M-1)]^T,$$

$$\hat{B}(t) = [b(1,0,t), \ldots, b(1,M-1,t)|b(2,0,t), \ldots, b(2,M-1,t)| \\ \ldots |b(N,0,t), \ldots, b(N,M-1,t)]^T. \tag{19}$$

3.2. The Operational Matrix of the Hybrid of Block-pulse and Chebyshev Polynomials

The integration of the vector $B(t)$ defined in Eq. (19) can approximated by

$$\int_0^t B(t')dt' \simeq \hat{P}\hat{B}(t)$$

where $\hat{P}$ is the $NM \times NM$ operational matrix for integration and is given by

$$\hat{P} = \begin{pmatrix} \hat{E} & \hat{H} & \hat{H} & \cdots & \hat{H} \\ 0 & \hat{E} & \hat{H} & \cdots & \hat{H} \\ 0 & 0 & \hat{E} & \cdots & \hat{H} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \hat{E} \end{pmatrix}.$$

In the above matrix

$$\hat{H} = \frac{t_f}{N} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \frac{-1}{3} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & 0 \\ \frac{(-1)^{M-1}}{2M(M-2)} & 0 & 0 & \cdots & 0 \end{pmatrix},$$

and E is operational matrix of integration for Chebyshev polynomials on the interval $[(\frac{n-1}{N})t_f, \frac{n}{N}t_f]$ given in [11] by

$$\hat{E} = \frac{t_f}{N} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{-1}{8} & 0 & \frac{1}{8} & 0 & \cdots & 0 & 0 & 0 \\ \frac{-1}{6} & \frac{-1}{4} & 0 & \frac{1}{12} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \frac{(-1)^{M-1}}{2(M-1)(M-3)} & 0 & 0 & 0 & \cdots & \frac{-1}{4(M-3)} & 0 & \frac{1}{4(M-1)} \\ \frac{(-1)^{M}}{2M(M-2)} & 0 & 0 & 0 & \cdots & 0 & \frac{-1}{4(M-2)} & 0 \end{pmatrix}.$$

4. Hybrid Functions Direct Method

For now, we will use hybrid of block-pulse and Legendre polynomials, similar results can be obtained by using hybrid of block-pulse and Chebyshev polynomials. Consider the problem of finding the extremum of the functional

$$J(x) = \int_0^1 F[t, x(t), \dot{x}(t)]dt. \tag{20}$$

The necessary condition for $x(t)$ to extremize $J(x)$ is that it should satisfy the Euler-Lagrange equation

$$\frac{\partial F}{\partial x} - \frac{d}{dt}(\frac{\partial F}{\partial \dot{x}}) = 0 \tag{21}$$

with appropriate boundary conditions. However, the above differential equation can be integrated easily only for simple cases. Thus numerical and direct methods such as the well-known Ritz and Galerkin methods have been developed to solve variational hybrid functions. problems. Here we consider a Ritz direct method for solving Eq. (21) using the

Suppose, the rate variable $\dot{x}(t)$ can be expressed as

$$\dot{x}(t) = C^T B(t). \tag{22}$$

Using Eq. (6), $x(t)$ can be represented as

$$\begin{aligned} x(t) &= \textstyle\int_0^t \dot{x}(t')dt' + x(0) \\ &= C^T P B(t) + [x(0), 0, \cdots, 0, x(0), 0, \cdots, 0, \cdots, x(0), 0, \cdots, 0]^T B(t). \end{aligned} \tag{23}$$

We can also express t in terms of $B(t)$ as

$$\begin{aligned} t = [&\frac{1}{2N}, \frac{1}{2N}, 0, \cdots, 0, \frac{3}{2N}, \frac{1}{2N}, \cdots, 0, \cdots, 0, \cdots, \\ &\frac{2N-1}{2N}, \frac{1}{2N}, 0, \cdots, 0]B(t) = d^T B(t) \end{aligned} \tag{24}$$

Substituting Eqs. (22-24) in Eq. (20), the functional $J(x)$ becomes a function of $c(n, m)$, $n = 1, 2, \cdots, N$, $m = 0, 1, 2, \cdots, M-1$. Hence to find the extremum of $J(x)$ we solve

$$\frac{\partial J}{\partial c(n, m)} = 0, \qquad n = 1, 2, \cdots, N, \qquad m = 0, 1, \cdots, M-1. \tag{25}$$

The above procedure is now used to solve the following examples.

5. Illustrative Examples

In this section two problems of the calculus of variations are considered. Example 1 is the classical brachistochrone problem, where as example 2 is an application to the heat conduction problem taken from [7].

5.1. Example 1: The Brachistochrone Problem.

5.1.1. The Brachistochrone Problem as an Optimal Control Problem

As an optimal control problem, the brachistochrone problem may be formulated as [5].

Minimize the performance index J,

$$J = \int_0^1 \left[\frac{1 + U^2(t)}{1 - X(t)}\right]^{\frac{1}{2}} dt, \tag{26}$$

subject to

$$\dot{X}(t) = U(t), \tag{27}$$

with

$$X(0) = 0, \quad X(1) = -0.5. \tag{28}$$

Eqs. (26), (27) and (28), describe the motion of a bead sliding down a frictionless wire in a constant gravitational field. The minimal time transfer expression is obtained from the law of conservation of energy. Here X and t are dimensionless and they represent respectively the vertical and horizontal coordinates of the sliding bead.

As is well known the exact solution to the brachistochrone problem is the cycloid defined by the parametric equations

$$x = 1 - \frac{\beta}{2}(1 + \cos 2\alpha), \qquad t = \frac{t_0}{2} + \frac{\beta}{2}(2\alpha + \sin 2\alpha), \tag{29}$$

where

$$\tan\alpha = \frac{dX}{dt} = U.$$

With the given boundary conditions, the integration constants are found to be

$$\beta = 1.6184891, \quad t_0 = 2.7300631.$$

5.1.2. *The Numerical Method*

Suppose, the rate variable $\dot{X}(t)$ can be expressed approximately as

$$\dot{X}(t) = C^T B(t). \tag{30}$$

Using Eqs. (6) and (28), $X(t)$ can be represented as

$$\begin{aligned} X(t) &= \textstyle\int_0^t \dot{X}(t')dt' + X(0) \\ &= C^T P B(t), \end{aligned} \tag{31}$$

and by using Eqs. (27) and (30) we have

$$U^2(t) = C^T B(t) B^T(t) C. \tag{32}$$

Equation (32) can be simplified by using the property of the product of two hybrid Legendre function vectors given in Eq. (13).

5.1.3. *The Performance Index Approximation*

Using Eqs. (26), (31) and (32) the performance index J can be approximated as follows:

$$J = \int_0^1 \left(\frac{1 + C^T \tilde{C} B(t)}{1 - C^T P B(t)} \right)^{\frac{1}{2}} dt. \tag{33}$$

Divide the interval $[0, 1]$ into N equal subintervals, we have

$$J = \sum_{n=1}^{N} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left(\frac{1 + C^T \tilde{C} B(t)}{1 - C^T P B(t)} \right)^{\frac{1}{2}} dt. \tag{34}$$

In order to use the Gaussian integration formula we transform the t-interval $(\frac{n-1}{N}, \frac{n}{N})$ into the τ-interval $(-1, 1)$ by means of the transformation

$$t = \frac{1}{2}(\frac{1}{N}\tau + \frac{2n-1}{N}). \tag{35}$$

The optimal control problem in Eqs. (26-28) is then restated as follows: Minimize

$$J = \frac{1}{2}\int_{-1}^{1}\left[\frac{1+u^2(\tau)}{1-x(\tau)}\right]^{\frac{1}{2}} d\tau, \tag{36}$$

subject to

$$\frac{dx}{d\tau} = \frac{1}{2}u(\tau), \tag{37}$$

with

$$x(-1) = 0, \quad x(1) = -0.5. \tag{38}$$

Using Eqs. (34) and (35) we get

$$J = \sum_{n=1}^{N}\frac{1}{2N}\int_{-1}^{1}\left(\frac{1+C^T\tilde{C}B(\frac{1}{2}(\frac{1}{N}\tau+\frac{2n-1}{N}))}{1-C^TPB(\frac{1}{2}(\frac{1}{N}\tau+\frac{2n-1}{N}))}\right)^{\frac{1}{2}} d\tau. \tag{39}$$

Using the Gaussian integration formula, Eq.(39) can be approximated as

$$J \approx \sum_{n=1}^{N}\frac{1}{2N}\sum_{j=0}^{k}\left(\frac{1+C^T\tilde{C}B(\frac{1}{2}(\frac{1}{N}\tau_j+\frac{2n-1}{N}))}{1-C^TPB(\frac{1}{2}(\frac{1}{N}\tau_j+\frac{2n-1}{N}))}\right)^{\frac{1}{2}} w_j, \tag{40}$$

where τ_j, $j = 0, 1, \ldots, k$ are the $k+1$ zeros of Legendre polynomial P_{k+1}, and w_j are the corresponding weights, given in [19]. The idea behind the above approximation is the exactness of the Gaussian integration formula for polynomials of degree not exceeding $2k+1$.

5.1.4. *Evaluating the Vector C*

The optimal control problem has now been reduced to a parameter optimization problem which can be stated as follows.

Find $c(n,m)$, $n = 1, 2, \cdots, N$, $m = 0, 1, \ldots, M-1$ that minimizes Eq.(40) subject to

$$x(-1) = 0, \quad x(1) = -0.5. \tag{41}$$

We now minimize Eq. (40) subject to Eq. (41) using the Lagrange multiplier technique. Suppose

$$J^* = J + \lambda_1 x(-1) + \lambda_2[x(1) + 0.5].$$

The necessary conditions for a minimum are

$$\frac{\partial J^*}{\partial c(n,m)} = 0 \quad n = 1, 2, \cdots, N, \quad m = 0, 1, \ldots, M-1 \tag{42}$$

and

$$\frac{\partial J^*}{\partial \lambda_1} = 0, \qquad \frac{\partial J^*}{\partial \lambda_2} = 0. \tag{43}$$

Eqs. (42) and (43) give $(NM + 2)$ non-linear equations with $(NM + 2)$ unknowns which can be solved for $c(n,m)$, λ_1 and λ_2 using Newton's iterative method. The initial values required to start Newton's iterative method have been chosen by taking $x(\tau)$ as a linear function between $x(-1) = 0$ and $x(1) = -0.5$.

In Table 1 the results for hybrid Legendre approximation with $N = 2$, $k = 5$ and $M = 3, 4, 5$ together with $N = 2$, $k = 8$ and $M = 5$ are listed, we compare the solution obtained using the proposed method with other solutions in the literature together with the exact solution.

Methods	$x(1)$	$u(-1)$	J
Dynamic programming gradient methoc[2]	-0.5	-0.7832273	0.9984988
Dynamic programming successive sweep method[3,4]	-0.5	-0.7834292	0.9984989
Chebyshev solutions[5]			
$M = 4$	-0.5	-0.7844893	0.9984982
$M = 7$	-0.5	-0.7864215	0.99849815
$M = 10$	-0.5	-0.7864406	0.9984981483
Hybrid Legendre, $N = 2, k = 5$			
$M = 3$	-0.5	-0.7852418	0.9984989
$M = 4$	-0.5	-0.7864397	0.9984983
$M = 5$	-0.5	-0.7864402	0.9984981
Hybrid Legendre $N = 2, k = 8$ and $M = 5$	-0.5	-0.7864408	0.99849814829
Exact Solution[4]	-0.5	-0.7864408	0.99849814829

Table 1. The hybrid Legendre and other solutions in the literature.

5.2. Example 2: Application to The Heat Conduction Problem

Consider the extremization of

$$J=\int_0^1 [\frac{1}{2}\dot{x}^2-xg(t)]dt=\int_0^1 F(t,x,\dot{x})dt, \tag{44}$$

where $g(t)$ is a known function satisfying

$$\int_0^1 g(t)dt=-1,$$

with the boundary conditions

$$\dot{x}(0)=0 \quad , \quad \dot{x}(1)=0. \tag{45}$$

Schechter [20] gave a physical interpretation for this problem by noting an application in heat conduction and Chen and Hsiao [7] considered the case where $g(t)$ is given by

$$g(t)=\begin{cases}-1, & 0\le t<\frac{1}{4}, \quad \frac{1}{2}\le t<1,\\ 3, & \frac{1}{4}\le t<\frac{1}{2},\end{cases} \tag{46}$$

and gave an approximate solution using Walsh functions. The exact solution is

$$x(t)=\begin{cases}\frac{1}{2}t^2, & 0\le t<\frac{1}{4}\\ -\frac{3}{2}t^2+t-\frac{1}{8}, & \frac{1}{4}\le t<\frac{1}{2}\\ \frac{1}{2}t^2-t+\frac{3}{8}, & \frac{1}{2}\le t<1.\end{cases}$$

Here of we solve the same problem using hybrid of Legendre and block-pulse functions with $M=3$ and $N=4$. First we assume

$$\dot{x}(t)=C^TB(t). \tag{47}$$

In view of Eq. (46), we write Eq. (44) as

$$J=\frac{1}{2}\int_0^1 \dot{x}^2(t)dt+4\int_0^{\frac{1}{4}} x(t)dt-4\int_0^{\frac{1}{2}} x(t)dt+\int_0^1 x(t)dt,$$

or

$$J = \frac{1}{2}\int_0^1 C^T B(t) B^T(t) C dt + 4C^T P \int_0^{\frac{1}{4}} B(t)dt - 4C^T P \int_0^{\frac{1}{2}} B(t)dt + C^T P \int_0^1 B(t)dt\,.$$

Let

$$w(t) = \int_0^t B(t')dt',$$

then using Eq. (20), we have

$$J = \frac{1}{2}C^T DC + C^T P[4w(\frac{1}{4}) - 4w(\frac{1}{2}) + w(1)]. \qquad (48)$$

where

$$D = \int_0^1 B(t)B^T(t)dt.$$

The boundary conditions in Eq.(45) can be expressed in terms of hybrid of Legendre and block-pulse functions as

$$C^T B(0) = 0 \quad , \quad C^T B(1) = 0\,. \qquad (49)$$

We now minimize Eq. (48) subject to Eq. (49) using the Lagrange multiplier technique. Suppose

$$J^* = J + \lambda_1 C^T B(0) + \lambda_2 C^T B(1), \qquad (50)$$

where λ_1 and λ_2 are the two multipliers. Using Eq. (25) we obtain

$$\frac{\partial J^*}{\partial C} = DC + P[4w(\frac{1}{4}) - 4w(\frac{1}{2}) + w(1)] + \lambda_1 B(0) + \lambda_2 B(1) = 0. \qquad (51)$$

We also have

$$w(1) = \frac{1}{4}[1,0,0,1,0,0,1,0,0,1,0,0]^T,$$
$$w(\frac{1}{2}) = \frac{1}{4}[1,0,0,1,0,0,0,0,0,0,0,0]^T,$$
$$w(\frac{1}{4}) = \frac{1}{4}[1,0,0,0,0,0,0,0,0,0,0,0]^T,$$
$$B(0) = [1,-1,1,0,0,0,0,0,0,0,0,0]^T,$$
$$B(1) = [0,0,0,0,0,0,0,0,0,1,1,1]^T.$$

Equations (49) and (51) define a set of 14 simultaneous linear algebraic equations from which the coefficient vector C and the multipliers λ_1 and λ_2 can be found. The vector $C^T P$ is

$$C^T P = \frac{1}{64}[\frac{2}{3}, 1, \frac{1}{3}, 2, -1, -1, -\frac{10}{3}, -3, \frac{1}{3}, -\frac{22}{3}, -1, \frac{1}{3}]^T \qquad (52)$$

Further, to define $x(t)$ for t in the interval $[0, \frac{1}{4}]$ we map $[0, \frac{1}{4}]$ into $[-1, 1]$ by mapping t into $8t - 1$ and similarly for the other intervals. Using the above equation and $P_0 = 1, P_1 = t$ and $P_2 = \frac{3}{2}t^2 - \frac{1}{2}$, we get

$$x(t) = \begin{cases} \frac{1}{64}[\frac{2}{3} + (8t-1) + \frac{1}{3}[\frac{3}{2}(8t-1)^2 - \frac{1}{2}]] = \frac{1}{2}t^2, & 0 \leq t \leq \frac{1}{4} \\ \frac{1}{64}[2 - (8t-3) - [\frac{3}{2}(8t-3)^2 - \frac{1}{2}]] = -\frac{3}{2}t^2 + t - \frac{1}{8}, & \frac{1}{4} \leq t \leq \frac{1}{2} \\ \frac{1}{64}[-\frac{10}{3} - 3(8t-5) + \frac{1}{3}[\frac{3}{2}(8t-5)^2 - \frac{1}{2}]] = \frac{1}{2}t^2 - t + \frac{3}{8}, & \frac{1}{2} \leq t \leq \frac{3}{4} \\ \frac{1}{64}[-\frac{22}{3} - (8t-7) + \frac{1}{3}[\frac{3}{2}(8t-7)^2 - \frac{1}{2}] = \frac{1}{2}t^2 - t + \frac{3}{8}, & \frac{3}{4} \leq t \leq 1 \end{cases}$$

which is the exact solution. This exact solution can not be obtained either with CBF's or with PCBF's.

6. Conclusion

The aim of present work is to develop an efficient and accurate method for solving problems of the calculus of variations. The problem has been reduced to solving a system of algebraic equations. Illustrative examples are included to demonstrate the validity and applicability of the technique. The advantages of using the hybrid Legendre method are :

(1) The operational matrix P contains many zeros which plays an important role in simplifying the performance index.

(2) The Gaussian integration formula is exact for polynomials of degree not exceeding $2k + 1$.

(3) Only small values of k, N and M are needed to obtain very satisfactory results for the brachistochrone problem.

(4) Hybrid functions approach provides an exact solution for the heat conduction problem.

References

[1] V.M. Tikhomirov, *Stories about maxima and minima*, American Mathematical Society Providence, RI 1990, pp. 265–280.

[2] R. Bellman *Dynamic Programming*, Princeton University Press , NJ 1957.

[3] A. V. Balakrishnan and L.W. Neustadt *Computing methods in optimization problems* Academic Press, New York, 1964.

[4] A.E. Bryson and Y.C. Ho, *Applied optimal control,* Blaisdell Waltham, 1969.

[5] J. Vlassenbroeck and R. Van Dooren, *A new look at the Brachistochrone problem*, Journal of Applied Mathematics and Physics (ZAMP) **31** (1980) 785–790.

[6] D.S. Szarkowicz, *Investigating the brachistochrone with a multistage Monte Carlo method*, International Journal of Systems Science **26** (1995), 233–243.

[7] C.F. Chen and C.H. Hsiao, *A Walsh series direct method for solving variational problems*, J. Franklin Inst. **300** (1975), 265–280.

[8] M. Razzaghi and J. Nazarzadeh, *Walsh Functions*, Wiley Encyclopedia of Electrical and Electronics Engineering **23** (1999), 429–440.

[9] C. Hwang and Y.P. Shih, *Laguerre series direct method for variational problems*, Journal of Optimization Theory and Applications (1983), 143–149.

[10] R.Y. Chang and M.L. Wang, *Shifted Legendre direct method for variational problems series* Journal of Optimization Theory and Applications **39** (1983), 299–307.

[11] I.R. Horng and J.H. Chou, *shifted Chebyshev direct method for solving variational problems*, International Journal of Systems Science **16** (1985) 855–861.

[12] T.H. Moulden, and M.A. Scott *Walsh spectral analysis for ordinary differential equations: Part1-Initial value problems*, IEEE Trans. Circuits Syst. **35** (1988), 742–745.

[13] M. Razzaghi, and J. Nazarzadeh, *Optimum pulse-width modulated patterns in induction motors using Walsh functions*, Electric Power Systems Research **35** (1995), 87–91.

[14] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, 1963.

[15] L.E. Elsgolic, *Calculus of Variations, Pergamon Press Ltd.*, 1962.

[16] M. Razzaghi and M. Razzaghi, *Fourier series direct method for variational problems*, International Journal of Control **48** (1988), 887–895.

[17] M. Razzaghi and M. Razzaghi, *Instabilities in the solutions of heat conduction problem using Taylor series and alternative approaches*, Journal of the Franklin Institute **326** (1989), 215–224.

[18] J.H. Chou, *Application of Legendre series to optimal control of integro-differential equations*, International Journal of Control **323** (1987), 269–277.

[19] A. Constantinides, *Applied numerical methods with personal computers, McGraw-Hill, New York*, 1987.

[20] R.S. Schechter, *The Variation Method in Engineering*, McGraw-Hill, New York (1967).

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 179-200 (2004)

The Algebra of Formal Power Series

H. Sharif

Department of Mathematics,
Shiraz University, Shiraz 71456, Iran
sharif@sun01.susc.ac.ir

Abstract: Let K be a field. $K[[X]]$ will denote the ring of formal power series in several commuting variables, $X = (x_1, x_2, \cdots, x_k)$ with coefficients in K. $K((X))$ will denote the field of fractions of $K[[X]]$. An element $f \in K((X))$ is said to be an algebraic function over K if f is algebraic over the field of rational functions $K(X)$. If further $f \in K[[X]]$, then f is said to be an algebraic series over K. A function which is not algebraic is called transcendental (over K).

For the case of one variable G. Christol et. al. (1980) have characterised the algebraic functions over a finite field in terms of automata. We generalise their argument and obtain the corresponding result for the case of several variables over a perfect field of positive characteristic. H. Furstenberg (1967) has shown that if K is a finite field and $f = \sum_{n\geq 0} a_n x^n, g = \sum b_n x^n \in K[[X]]$ are aglebraic, then the Hadamard product of f and $g, f * g = \sum_{n\geq 0} a_n b_n x^n$ is also algebraic. We apply the above characterisation of the algebraic series in several variables to generalise Furstenberg's result and prove that if K is an arbitrary field of positive characteristic and if $f = \sum_l a_l X^l, g = \sum_l b_l X^l \in K[[X]]$ are algebraic series, then $f * g = \sum_l a_l b_l X^l$ is also an algebraic series. As an easy consequence of the above result we give a proof to

MSC (2000): 11J61, 12H25

Keywords: Differentially algebraic formal power series algebraically independent, transcendence basis, Hadamard product.

Deligne's Theorem: If $f \in K[[X]]$ and $f = \sum_{\sigma} a_\sigma X^\sigma$ is an algebraic series in X over K, then $D(f) = \sum_{n \geq 0} a_{n.1} t^n$ is an algebraic series in t over K. Then we consider the Hadamard product of rational formal power series and we show that if L is a field of characteristic zero and if $f = \sum_{l} a_l X^l, g = \sum_{l} b_l X^l$ are rational series in $L[[X]]$, then $f * g = \sum_{l} a_l b_l X^l$ (which is not in general a rational series) is always algebraic only if $k \geq 2$.

We introduce two methods for trying to decide whether or not a given formal power series is an algebraic series over a field. Our first method (in characteristic zero) is based on the reduction process modulo the prime p and our second method (in positive characteristic and so in characteristic zero using reduction) is based on the splitting process for functions.

In 1986, M. Mendes France and A. J. van der Poorten have shown that if $f = \sum_{n=0}^{\infty} a_n x^n \in F[[X]]$ is algebraic, where F is a finite field of characteristic $p > 0, a_0 = 1$ and $f \neq 1$ and if λ is a p-adic integer, then f^λ is algebraic if and only if λ is rational. We generalise their result and prove the following theorem:
Suppose that K is a field of characteristic $p > 0$. Suppose that $f = \sum_{n=0}^{\infty} a_n x^n \in K[[x]]$ is algebraic over K, where $a_0 = 1$ and $a_1 \neq 0$. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be p-adic integers. Then the following conditions are equivalent:

(i) $1, \lambda_1, \lambda_2, \cdots, \lambda_n$ are linearly independent over Q.

(ii) $(1+x)^{\lambda_1}, (1+x)^{\lambda_2}, \cdots, (1+x)^{\lambda_n}$ are algebraically independent over Q.

(iii) $f^{\lambda_1}, f^{\lambda_2}, \cdots, f^{\lambda_n}$ are algebraically independent over $K(x)$.

A formal power series f is called D-Algebraic (respectively, D-finite) if it satisfies an algebraic (respectively, a linear) differential equation with polynomial coefficients. These notions are usually defined only over fields of characteristic zero and are not so significant over fields of characteristic $p > 0$ as $f^{(p)} \equiv 0$. For a formal power series over a perfect field of positive characteristic we make the definition of E-algebraicity (respectively, E-finiteness) which is an analogue of the notion of D-algebraicity (respectively, D-finiteness).

It is slightly surprising that E-finite series are in fact algebraic series. However, the E-algebraic series (which are not all algebraic series) under ordinary addition and multiplication of series, form a field which is algebraically closed in $K((x))$ and has some other natural prrperties. We also study the Hadamard product of two E-algebraic formal power series.

1. Preliminaries and Notations

Let K be a field and $X = (x_1, x_2, \ldots, x_k)$. We will denote the ring of formal power series by $K[[X]]$, the field of fractions of $K[[X]]$ by $K((X))$. An element $f \in K((X))$ is said to be an algebraic function over K if f is algebraic over the field of rational functions $K(X)$. If, further, $f \in K[[X]]$, then f is said to be an algebraic series over K. A function (or series) which is not algebraic is called transcendental over K.

Example 1.1. (i) The series

$$f = \sum_{n=0}^{\infty} \binom{2n}{n} x^n = \sum_{n=0}^{\infty} \binom{-\frac{1}{2}}{n} (-4)^n x^n = (1-4x)^{-\frac{1}{2}}$$

is algebraic over any field.

(ii) The series $f(x) = \sum_{n=0}^{\infty} x^{p^n}$ is algebraic over any field of characteristic p and transcendental over any field of characteristic q, where $q \neq p$.

(iii) The series $\sum_{n=1}^{\infty} x^{n!}$ is transcendental over any field (see Zariski and Samuel [**25**, p. 220]).

(iv) Let p be a prime number and let $S_p(n)$ be the sum of the digits of n is its p-adic expansion. That is, if

$$n = \sum_{i=0}^{\infty} n_i p^i, \qquad 0 \leq n_i \leq p-1,$$

(actually a finite sum for $n \in N$) is the p-adic expansion of n, then $S_p(n) = \sum_{i=0}^{\infty} n_i$. One can show that if

$$f(x) = \sum_{n=0}^{\infty} S_p(n) x^n \in F_p[[x]],$$

then $f(x) = \frac{x^p - 1}{x-1} f(x)^p + \frac{x}{(x-1)^2}$. That is, f is algebraic over F_p (see [**19**]).

2. Some characterisations of algebraic series

The following theorem is a well-known result in the field of complex numbers.

Theorem (Biberbakh). *Let $f = \sum_{n=0}^{\infty} a_n z^n \in C[[z]]$, where a_n just accepts the finite elements $d_1, d_2, \ldots, d_r$. Then f is rational or non-algebraic.*

G. Christol et.al.[4] have the following characterisation of algebraic series in terms of p-automata.

Theorem 2.1. *Suppose that F is a finite field of characteristic p. Suppose that $f = \sum_{n=0}^{\infty} a_n x^n \in F[[x]]$. Then the following are equivalent:*

i) f is algebraic over F.

ii) The sequence (a_n) is generated by a p-automata.

When F is replaced by the ring of p-adic integers, Denef and Lipshitz in [5] have extended Theorem 2.1 to the case of several variables.

We gave a characterisation of algebraic series in several variables over perfect fields of positive characteristic in [15].

Recall that a field K is perfect if every irreducible polynomial over K is separable. Equivalently, every algebraic extension field of K is separable over K. For example, every finite field or every field of characteristic zero is perfect. If K is a field of characteristic p, then K is perfect if and only if the Frobenuis map $\phi : K \longrightarrow K$ is an automorphism.

From now on, K will denote a perfect field of characteristic $p > 0$, unless explicity stated otherwise.

Firstly, we shall recall the splitting process for functions over a perfect field and the associated semilinear operators on the field of fractions of the ring of formal power series which we introduced in [15] and secondly, we shall employ such operators for characterising algebraic functions (for the details of the proofs see [15] or [19]).

Theorem 2.2. *Suppose that K is a field. If $f \in K((X))$ is an algebraic function over L, where L is an extension of K, then f is algebraic over K.*

The above theorem asserts that when we deal with algebraic functions over arbitrary fields of positive characteristic it is enough to consider such functions over perfect fields of positive characteristic, since every field L has a perfect extension, for example, the algebraic closure of L.

Let ι be a non-negative vector, that is, $\iota = (n_1, n_2, \cdots, n_k)$, where $i_j \in N, j = 1, 2, \cdots, k$. Then X^ι will denote the monomial $x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$. We denote by Λ the set of all non-negative vectors, and by Λ_p the set Z_p^k, where $Z_p = \{0, 1, 2, \cdots, p-1\}$.

Lemma 2.3 *If $f(X) \in K[[X]]$ (respectively $K((X))$), then f can be written uniquely as*

$$f = \sum_{\iota \in \Lambda_p} X^\iota f_\iota^p \qquad (2.1)$$

for some $f_\iota \in K[[X]]$ (respectively $K((X))$).

For $\iota \in \Lambda_p$ define $E_\iota : K((X)) \longrightarrow K((X))$ by

$$E_\iota(f) = f_\iota. \qquad (2.2)$$

Now for $f \in K((X))$, by Lemma 2.3 we have

$$f = \sum_{\iota \in \Lambda_p} X^\iota (E_\iota(f))^p. \qquad (2.3)$$

The operator E_ι is semilinear; that is, if $f, g \in K((X))$ and $\lambda \in K$, then

$$E_\iota(\lambda f + g) = \lambda^{\frac{1}{p}} E_\iota(f) + E_\iota(g).$$

Moreover, $E_\iota(g^p f) = gE_\iota(f)$.

Definition 2.4 Suppose that $f, g \in K[[X]]$, say

$$f = \sum_{t \in \Lambda} a_t X^t, \quad g = \sum_{t \in \Lambda} b_t X^t.$$

The Hadamard product of f and g, which will be denoted by $f * g$, is the series which is defined by

$$f * g = \sum_{t \in \Lambda} a_t b_t X^t.$$

Lemma 2.5 If $f, g \in K[[X]]$, then for $t \in \Lambda_p$,

$$E_t(f * g) = E_t(f) * E_t(g).$$

Let Ω be the semigroup generated by the identity operator and the E_ι for $\iota \in \Lambda_p$, with ordinary composition as multiplication. To each $f \in K((X))$ we associate its orbit

$$\Omega(f) = \{E(f) : E \in \Omega\}. \qquad (2.4)$$

Then we have the following:

Theorem 2.6. *Let $f \in K((X))$. Then $< \Omega(f) >$, the K-linear space spanned by $\Omega(f)$, is the smallest K-subspace of $K((X))$ containing f and which is invariant under each E_ι, $\iota \in \Lambda_p$.*

Theorem 2.7. *Let $f \in K((X))$. Then the following are equivalent:*

i) f is algebraic over K.

ii) There exist elements $a_o, a_1, \dots, a_N$ in $K[X]$ such that

$$\sum_{i=0}^{\infty} a_i f^{p^i} = 0,$$

where $a_o \neq 0$.

iii) There exists a finite dimensional K-subspace V of $K((X))$ such that $f \in V$ and $E_\iota(V) \subseteq V$, $\iota \in \Lambda_p$.

iv) $dim_K < \Omega(f) >$ is finite.

Example 2.8. Let $f = \sum_{n \geq 0} x^n y^{n^2}$. We shall show that f is transcendental over $F_2(x, y)$ and hence f will be transcendental with respect to any field of characteristic zero too.

Using the idea of Lemma 2.3 we split f and find the series

$$\begin{aligned} f_1 &= E_{(o,o)}(f) = \sum_{n\geq 0} x^n y^{2n^2}, \\ f_2 &= E_{(o,o)(o,o)}(f) = \sum_{n\geq 0} x^n y^{2^2n^2}, \\ & \quad \ldots, \\ f_k &= E_{(o,o)(o,o)\ldots(o,o)}(f) = \sum_{n\geq 0} x^n y^{2^k n^2} \\ & \quad \ldots \end{aligned}$$

Each series f_i corresponds to the set $\{(n, m) : m = 2^i n^2\}$. These sets are all clearly distinct and hence the set $\{f_1, f_2, f_3, \ldots\}$ is infinite. Therefore, f is transcendental over $F_2(x, y)$ by Theorem 2.7.

Corollary 2.9. *Suppose that $f, g \in K[[X]]$. If f, g are algebraic series over K, then $f * g$ is again an algebraic series over K.*

Theorem 2.10. *Suppose that K is any field. If $h \in K((X))$ is an algebraic function over L, where L is an extension field of K, then h is an algebraic function over K.*

We are now in a position to have the main theorem.

Theorem 2.11. *If K is a field of characteristic $p > 0$ and if f, g are algebraic series over K, then $f * g$ is again an algebraic series over K.*

Deligne's theorem [4] can be proved directly from the main theorem.

Theorem 2.12. *Suppose that K is a field of characteristic $p > 0$. If $f \in K[[X]]$ and $f = \sum_{\sigma\in\Lambda} a_\sigma X^\sigma$ is an algebraic series in X over K, then $I(f) = \sum_{n\geq 0} a_{n.1} t^n$ is an algebraic series in t over K.*

In Theorem 2.7 we considered $< \Omega(f) >$, the K-linear space spanned by $\Omega(f)$. Now we consider $< \Omega(f) >$ as a $K(X)$-linear space spanned by $\Omega(f)$. We shall show that f is algebraic if and only if this space is a finite dimensional subspace of $K((X))$ too.

Theorem 2.13. *Suppose that $f \in K((X))$ and W is the $K(X)$-subspace of $K((X))$ spanned by $\Omega(f)$. Then f is an algebraic function if and only if $dim_{K(X)} W < \infty$.*

Recall that for a field extension $L \subseteq F$, a transcendence base of F over L is a subset S of F which is algebraically independent over L and is maximal (with respect to set-theoretic inclusion) in the set of all algebraically independent subsets of F.

From now on, for simplicity of notations, we just concentrate on series in one variable.

Theorem 2.14. *Let $f \in K((x))$. Then f is algebraic over K if and only if $K(x, \Omega(f))$ is a finitely generated field extension of $K(x)$.*

3. Some methods for transcendency

R.P. Stanley observed in [21] (via analytic techniques) that if $f(x) = \sum_{n=0}^{\infty} \begin{pmatrix} 2n \\ n \end{pmatrix}^t x^n$, $t > 1$, then f is transcendental over $C(x)$ for even t, $t > 1$. He also stated that it is unknown for odd $t > 1$ whether or not it is transcendental.

In this section apart from the method which was introduced in section 2, we introduce another method for deciding whether or not a given formal power series with integer coefficients is transcendental.

Our method (in characteristic zero) is based on the reduction process and looking at the degree of the corresponding explicit equation obtained when the reduction modulo the prime p of the formal power series is algebraic. If the degree of the corresponding polynomial equation is an unbounded function of the prime p, then there is no single polynomial for all primes p and therefore, the formal power series is transcendental.

We introduced our method by dealing with the above particular problem, which was raised by Stanley [21] and we could answer it in [**23**].

First we showed that (the reduction of) f is algebraic over any field of positive characteristic p and we then deduced (from the explicit equation obtained) that f is transcendental over any field of characteristic zero for any integer $t > 1$.

We also gave a generalisation of this result in the case of multinomial coefficients.

Note that if $t = 1$, then f is algebraic of degree at most 2 over any field. In that case, $f = (1-4x)^{-1/2}$ and for $t \geq 1, f = 1$ over any field of characteristic 2. Hence we may suppose that, in the case of positive characteristic $p, p > 2$.

Throughout this section f will denote the series

$$f(x) = \sum_{n=0}^{\infty} \left(\begin{array}{c} 2n \\ n \end{array} \right)^{t} x^{n}$$

for $t \in N$ and $t > 1$.

First we shall see that if a series h with integer coefficients is algebraic over a field of characteristic zero, then the reduction of h is algebraic over F_p.

Proposition 3.1 *Suppose that K is any field of characteristic zero and*

$$h(x) = \sum_{i=0}^{\infty} h_i x^i \in Z[[x]]$$

is algebraic over K of degree N. Then for any prime p,

$$\bar{h}(x) = \sum_{i=0}^{\infty} \bar{h}_i x^i \in F_p[[x]]$$

is algebraic over F_p of degree at most N, where $\bar{a}$ is the image of a in F_p.

Note. Proposition 3.1 can be extended to the case of several variables.

Remark 3.2. Note that $f(x) = \sum_{n=0}^{\infty} \begin{pmatrix} 2n \\ n \end{pmatrix}^t x^n = h * h * \cdots * h$ (t times), where $h(x) = \sum_{n=0}^{\infty} \begin{pmatrix} 2n \\ n \end{pmatrix} x^n$, which is algebraic (with respect to any field) and $*$ denotes the Hadamard product operation.

Now, since over any field of positive characteristic the Hadamard product of two algebraic formal power series is again an algebraic formal power series (see [15, Collorary 5.5]) it follows that (the reduction of) f is algebraic over any field of positive characteristic. However, we shall now prove this directly by using Lucas' Theorem to find the corresponding explicit equation for (the reduction of) f, since we need this equation later.

Theorem 3.3. *(Lucas' Theorem) For $m, n \in N, p$ a prime,*

$$\binom{m}{n}^p \equiv \binom{mp}{np} \equiv \binom{m}{n} \pmod{p}$$

and

$$\binom{mp+i}{np+i} \equiv \binom{m}{n}\binom{i}{j} \pmod{p}$$

for $i, j \in N$ with $0 \leq i, j \leq p-1$.

Proof. See, for example, Dickson [6, p. 271]. ///

Proposition 3.4. *If p is any odd prime, then f is algebraic over Z_p.*

Theorem 3.5. *Let $t \in N, t > 1$. If $f(x) = \sum_{n=0}^{x} \binom{2n}{n}^t x^n \in Z[[X]]$, then f is transcendental over any field of characteristic zero.*

Theorem 3.6. *If $g(x) = \sum_{m=0}^{\infty} \binom{km}{m, m, \cdots, m}^t x^m$ where $t, k \in N$ with $t \geq 1, k \geq 3$, then g is transcedental over any field of characteristic zero.*

4. Rational power series

The following result, which is well known, is due to E. Borel.

Theorem 4.1. Let K be a field of characteristic zero. let $f(x) = \sum_{n\geq 0} a_n x^n, g(x) = \sum_{n\geq 0} b_n x^n \in K[[x]]$. If f, g are rational, then $f * g(x) = \sum_{n\geq 0} a_n x^n$ is again rational.

When the field K has positive characteristic the result is still true.

Lemma 4.2. *Suppose that K is a field of characteristic $p > 0$ which is also algebraically closed. If f and g belong to $K(x)$, then $f * g \in K(x)$.*

Proof. see [16. Lemma 3.2]./ / /

Lemma 4.3. *Suppose that K is a field and L is an extension field of K. Suppose that $f \in K[[x]]$. If $f \in L(x)$, then $f \in K(x)$.*

Proof. See [16, Lemma 3.3]./ / /

Theorem 4.4. *Suppose that K is a field of characteristic $p > 0$. If f and $g \in K(x)$, then $f * g \in K(x)$.*

Proof. Suppose that L is an extension field of K which is algebraically closed. Then f and g belong to $L(x)$. Therefore by Lemma 4.2, $f * g \in L(x)$ and so by Lemma 4.3, $f * g \in K(x)$.

Theorems 4.1 and 4.4 do not hold in the case of several variables. For example if

$$f = \sum_{n,m\geq 0} \begin{pmatrix} n+m \\ n \end{pmatrix} x^n y^m = \frac{1}{1-x-y},$$

which is rational, then (except in characteristic two)

$$f * f = \sum_{n,m\geq 0} \begin{pmatrix} n+m \\ n \end{pmatrix}^2 x^n y^m = \{(1-x-y)^2 - 4xy\}^{-1/2},$$

which is not rational.

A question which arises here is

Suppose tht K is a field and $f, g \in K[[x_1, x_2, \cdots, x_k]]$. Suppose that f, g are rational series. Is the Hadamard product $f * g$ an algebraic series?

The answer is positive when the field K has positive characteristic [see [24, Corollary 5.5]]. If $k > 2$ and K has characteristic zero, then the following example shows that the answer is negative.

Example 4.5. If

$$f = \sum_{n_1,n_2,n_3\geq 0} \begin{pmatrix} n_1+n_2+n_3 \\ n_1, n_2, n_3 \end{pmatrix} x_1^{n_1} x_2^{n_2} x_3^{n_3} = \frac{1}{1-x_1-x_2-x_3}$$

and

$$g = \sum_{n \geq 0} (x_1, x_2, x_3)^n = \frac{1}{1 - x_1 x_2 x_3},$$

which are rational series, then

$$f * g = \sum_{n \geq 0} \binom{3n}{n, n, n} t^n, \quad \text{where } t = x_1 x_2 x_3,$$

which is transcendental over a field of characteristic zero (see, for example, [23]). Note that, by [23], the same series is algebraic over a field of characteristic $p > 0$, but of degree equal to an unbounded function of p.

The only case which is left and still seems to be unknown is the case $k = 2$ and when K has characteristic zero. In [24] we considered this case and proved the following theorem.

Theorem 4.6. *The Hadamard product of two rational series of two variables over a field of characteristic zero is an algebraic series.*

Remark 4.7. We cannot weaken the conditions of Theorem 4.6 to allow one of the series to be algebraic. For example if

$$R(x, y) = \sum_{n,m \geq 0} \binom{n+m}{n}^2 x^n y^m = \{(1 - x - y)^2 - 4xy\}^{-1/2},$$

which is algebraic and

$$S(x, y) = \sum_{n \geq 0} (xy)^n = \frac{1}{1 - xy},$$

which is rational, then

$$R * S = \sum_{n \geq 0} \binom{2n}{n}^2 t^n, \quad \text{where } t = xy,$$

which is transcendental. (See [23].)

Example 6.3. Suppose that $f = \sum_{n\geq 0} x^{n^2}$, and $g = \sum_{n\geq 0} \frac{1}{n!}x^n$, which are D-algebraic. Then $f * g = \sum_{n\geq 0} \frac{1}{(n^2)!}x^{n^2}$, which is not D-algebraic. (See Lipshitz-Rubel [12, Proposition 7.3, p. 1209].)

A sub-algebra of D_L which is closed under the Hadamard product operation is the algebra of differentiably finite power series. A differentiably finite (D-finite, for short) power series is a series which satisfies a linear differential equation. The class of D-finite power series has been subject to extensive investigation, particularly whithin the theory of differential equations. A systematic exposition of their properties from a combinatorial point of view have been given by Stanley [21].

Theorem 6.4. *Suppose that f, g are D-finite power series. Then $f * g$ is again a D-finite power series.*

Proof. See Stanley [21].

Recently, Stanley's notion of D-finiteness has been of interest to several authors such as Gessel, Zeilberger, Lipshitz etc. Theorem 6.4 and many other results concerning D-finiteness have been generalised by Lipshitz [11] to the case of several variables.

Note that in Exampe 6.3, g is D-finite. Hence the Hadamard product of a D-algebraic and a D-finite power series is not D-algebraic. However, we have been able to prove that the Hadamard product of a rational formal power series and a D-algebraic formal power series is D-algebraic. First we need the following Lemma.

Lemma 6.5. *Suppose that $E \subseteq F$ is an algebraic extension of fields. Suppose that $f \in E[[x]]$. If $tr.deg._F F(x, f, f', ..., f^{(n)}, ...) < \infty$, then*

$$tr.deg._E E(x, f, f', ..., f^{(n)}, ...) < \infty.$$

Theorem 6.6. *Suppose that $f, g \in L[[x]]$. If f is rational and g is D-algebraic, then $f * g$ is D-algebraic.*

7. E-algebraic functions

Differentially algebraic functions are usually defined only over fields of characteristic zero and are not so significant over fields of characteristic $p > 0$, as $f^{(p)} \equiv 0$. In this section we shall define an analogue of the concept of a D-algebraic function over a perfect field of characteristic $p > 0$.

From now on K will denote a perfect field of characteristic $p > 0$, unless explicitly stated otherwise. Moreover, all we discuss about can be generalised to the case of several variables.

Definition. 7.1 *Suppose that $f \in K((x))$. We say that f is an E-algebraic function (over $K(x)$) if $tr.deg._{K(x)}K(x, \Omega(f)) < \infty$.*

Notation. We shall denote by Γ_K, the set of all E-algebraic functions.

Example 7.2. Let $K = F_2$ and α be a 2-adic integer. Let $f_\alpha = (1+x)^\alpha \in F_2[[x]]$. If α is rational, then f_α is algebraic ([**20**]) and hence $tr.deg._{F_2(x)}F_2(x, \Omega(f_\alpha)) = 0$. However, if α is not rational, then f_α is not algebraic over F_2. However, we show that $f_\alpha \in \Gamma_K$ and hence the set Γ_K strictly contains the set of all algebraic functions.

Let $\alpha = \sum_{i=0}^{\infty} \alpha_i 2^i$ be the 2-adic expansion of α. Then

$$f_\alpha = (1+x)^\alpha = (1+x)^{\alpha_o} \left[(1+x)^{\frac{\alpha-\alpha_o}{2}}\right]^2.$$

Hence $E_o(f_\alpha) = (1+x)^{\frac{\alpha-\alpha_o}{2}}$ and $E_1(f_\alpha) = \alpha_o E_o(f_\alpha)$. Therefore,

$$E_o(f_\alpha)^2 = (1+x)^{\alpha-\alpha_o} = \frac{f_\alpha}{(1+x)^{\alpha_0}} \in F_2(x, f_\alpha).$$

Similarly, $E_1(f_\alpha)^2 \in F_2(x, f_\alpha)$. One can show that

$$[E_{i_1 i_2 \dots i_r}(f_\alpha)]^{2^r} = \frac{c f_\alpha}{(1+x)^{\alpha_o + 2\alpha_1 + \cdots + 2^{r-1}\alpha_{r-1}}} \in F_2(x, f_\alpha),$$

where $c \in F_2$ and $E_{ij} = E_i o E_j$. Therefore,

$$tr.deg._{F_2(x)} F_2(x, \Omega(f_\alpha)) = 1,$$

since f_α is not algebraic. Thus $f_\alpha \in \Gamma_K$.

Note that if $f \in K((x))$ is algebraic, then

$$tr.deg._{K(x)} K(x, \Omega(f)) = 0.$$

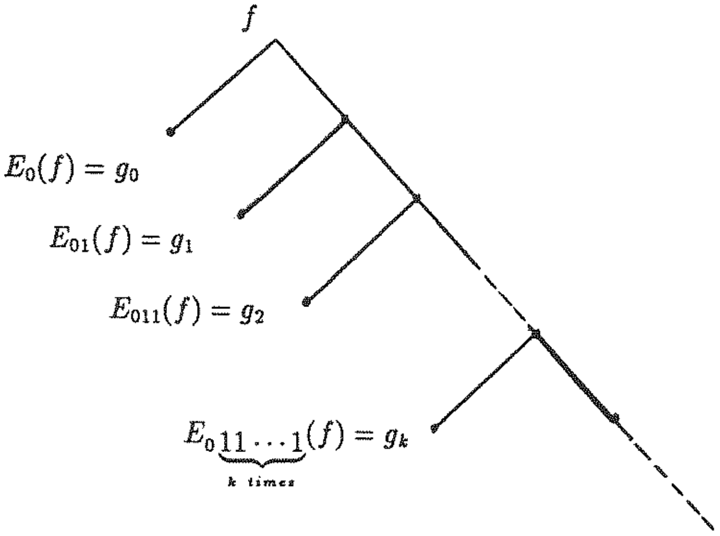
So as in the case of characteristic zero, every algebraic function is E-algebraic.

Now we have that Γ_K is a field.

Theorem 7.3. *Γ_K with ordinary addition and multiplication of series is a field.*

Let $\overline{K(x)}$ be the algebraic closure of $K(x)$ in $K((x))$. In Example 7.2 we showed that $\overline{K(x)} \subset \Gamma_K$. Now we show that $\Gamma_K \subset K((x))$. That is, we construct a power series $f = \sum_{n=0}^{\infty} a_n x^n \in K[[x]]$ such that $f \notin \Gamma_K$.

Example 7.4. Let $K = F_2$ and θ be a 2-adic integer which is transcendental over Q. Then $1, \theta, \theta^2, \cdots$ are linearly independent over Q. Let $f_\theta = (1+x)^\theta \in F_2[[x]]$. Then one can show that $f_\theta, f_{\theta^2}, \cdots$ are algebraically independent over $F_2(x)$ (see, for example, [17]). Let $g_k = f_{\theta^k} = (1+x)^{\theta^k}$, for $k = 0, 1, 2, ...$ and define f as follows.



If we specify a_n, then, since $F_2(g_o, g_1, ...)$ has infinite transcendence degree over $F_2(x)$ and since $F_2(g_o, g_1, ...) \subseteq F_2(x, \Omega(f))$, we conclude that $f \notin \Gamma_K$.

In order to specify a_n, we write the binary expansion of n, $n = \sum_{j=1}^{i} n_j 2^{j-1}$, say

$$n = n_1 n_2 ... n_k n_{k+1} n_{k+2} ... n_i = 11...10 n_{k+2} ... n_i$$

(where $n_i = 1$) and if we use the definition of the E_i operator, then we can get

$$a_n = E_{n_i n_{i-1} ... n_{k+2}}(g_k)(0) = \binom{\theta^k}{m} \,(mod\, 2)$$

where $E_{ij} = E_i o E_j$ and $m = \sum_{j=k+2}^{i} n_j 2^{j-k-2}$ as required. Therefore,

$$\overline{K(x)} \subset \Gamma_K \subset K((x)).$$

As we see from the above example $tr.deg._{K(x)} K((x)) = \infty$.

Barcanescu and Brezuleanu in [2] proved the following proposition.

Proposition 7.5. *Suppose that K is a field of characteristic $p > 0$. Suppose that $K \subseteq M$ is a separable field extension. Let $u \in M\backslash K$.Then the extension $K(u) \subseteq M$ is a separable extension if and only if $u \notin M^pK$.*

We use the above result and have that $K(x) \subseteq \Gamma_K$ is a separable extension.

Proposition 7.6. *i) Γ_K is closed under the E and D (derivative) operators.*

ii) $K(x) \subseteq \Gamma_K$ is a separable extension.

iii) $tr.deg._{K(x)}\Gamma_K = \infty$.

iv) Γ_K is algebraically closed in $K((x))$.

Proof. See[20].

Theorem 7.7. *Γ_K is not closed under the Hadamard product operation.*

Proof. See [20].

References

[1] J.P. Allouche, Somme des chiffres et transcendance, *Bull. Soc. Math. France*, **110** (1982) 279-285.

[2] S. Barcanescu and A. Brezuleanu, Formal power series algebraically independent over polynomials, *Rev. Roum. Math. Pures et Appl.*, **25** (1980) no. 2, 147-155.

[3] G. Christol, T. Kamae, M. Mendes-France and G. Reuzy, Suites algebriques, automates et subsitutions, *Bull. Soc. Math. France*, **108** (1980) 401-419.

[4] P. Deligne, Integration sur un cycle evanescent, *Invent. Math.* **76** (1983) 129-143.

[5] J. Denef and L. Lipshitz, Algebraic power series and diagonals, *J. Number Theory*, **26** (1987) 46-67.

[6] L. E. Dickson, "History of the theory of numbers", Vol. 1, (Carnegie Institution of Washington, Washington, 1919).

[7] H. Furstenberg, Algebraic functions over finite fields, *J. Algebra*, **7** (1967) 271-277.

[8] A. Grothendieck, *Elements de geometrie algebrique*, Inst. Hautes Etudes Sci. Publ. Math., **20** (1964).

[9] T. Harase, Algebraic Dependence of Formal Power Seires, LNM, no. 1434, 133-137.

[10] N. Koblitz, *P-adic analysis; a short course on recent work*, Cambridge U.P.; LMS Lecture Note series **46**, 1980.

[11] L. Lipshitz, The diagonal of a D-finite power series is D-finite, *J. Algebra*, **113** (1988) 373-378.

[12] L. Lipshitz and L. Rubel, A gap theorem for power series solutions of aglebraic differential equations, *Amer. J. Math.,* **108** (1986) 1193-1214.

[13] K. Mahler, *Lectures on Diophantine Approximations*, Ann Arbor, Michigan, 1961.

[14] M. Mendes-France and van der Poorten, Automata and the arithmetic of formal power series, *Acta Arith.* **46** (1986), 211-214.

[15] H. Sharif and C.F. Woodocock, Algebraic functions over a field of positive characteristic and Hadamard products, *J. London Math. Soc.*, **37** (1988) 395-403.

[16] H. Sharif, Childeren products of formal power series, *Math. Japanica*, **38** (1993) 319-324.

[17] —-, Algebraic independence of certain formal power series (I), *J. Sci. I.R.I. Iran*, **2** (1991) 50-55.

[18] —-, Algebraic Independence of Certain Formal Power Series (II), *J. Sci. I. R. Iran* 3 (1992) 148-151.

[19] —-, Algebraic functions, differentially algebraic power series and Hadamard operations, Ph.D. Thesis, Kent, 1989.

[20] —-, E-algebraic functions over fields of positive characteristic an analogue of differentially algebraic functions, *J. Algebra* (1999) 335-366.

[21] R. P. Stanley, Differentiably finite power series, *Europ. J.* Combinatorics **1** (1980), 175-188.

[22] R.J. Walker, *Algebraic curves (Dover, New York, 1950).*

[23] C.F. Woodcock, and H. Sharif, *On the transcendence of certain power series*, J. Algebra **121**, 364-369.

[24] —-, *Hadamard products of rational formal powr series, J. Algebra* **128**, 517-527.

[25] O. Zariski, and P. Samuel, *Commutative algebra* Vol. II (Van Nostrand, New York, 1960).

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 201-225 (2004)

Riemann Extension and Complete Lifts

M.Toomanian

Faculty of Mathematical Sciences,
Tabriz University, Tabriz, Iran
toomanian@tabrizu.ac.ir

Abstract: After some fundamental definitions we define Homogeneous and symmetric spaces, both for Affine and Riemannian cases. Then we introduce s^k-manifolds where, $k = 2$ gives the Cartan symmetric spaces. A generalization of s^k-manifolds are given by Ledger and Obata as s-manifolds (1968). These spaces are defined locally.

Ledger and Graham gave a general and globally defined s-spaces called s-regular spaces. Further works has been done by Ledger, Gray, Sasaki, Vanake, Kowalski, Sekizawa, Razavi and myself, that is Σ-spaces are defined and lifted to the tangent spaces.

This lecture is restricted to the enlargement of homogeneous property to the tangent and cotangent bundles.

MSC (2000): 53022
Keywords: Homogeneous, Symmetric, Riemannian, Affine, s^k-spaces, Σ-spaces

1. Introduction

A topological space M is locally Euclidean if, for every $x \in M$ there exist an integer $n \geq 0$, an open set $U \subseteq M$ of n , an open subset $U' \subseteq I\!R^n$ and a homeomorphism $\varphi : U \longrightarrow U'$.

The integer n is uniquely determined by x and is called local dimension of M at x.

If M is Hausdorff and second countable, then by Brouwer theorem on invariance of domain (if $U \subseteq I\!R^n$ and $V \subseteq I\!R^n$ are open subsets such that U is homeomorphic to V, then $m = n$) then the local dimension is n at every point $x \in M$, and is called the dimension of M.

Definition 1.1 A topological space M is a topological manifold of dimension n if

1. M is locally Euclidean of local dimension n.
2. M is Hausdorff.
3. M is second countable.

The locally Euclidean property allows us to choose local coordinates in any small region of M.

Definition 1.2 A coordinate chart on M is a pair (U, φ) where $U \subseteq M$ is an open subset and $\varphi : U \longrightarrow I\!R^n$ is a homeomorphism onto an open subset of $I\!R^n$. Let $\varphi(p) = (x^1(p), \cdots, x^n(p))$ then the n-tuple $(x^i(p))$ is taken as the coordinate of $p \in M$ and x^i, s as coordinate functions. Relative to such a coordinatization, we can do calculus in the region U of M. The problem is that the point p will generally belong to infinitely many different coordinate charts and calculus in one of these coordinatizations about p might not agree with calculus in another. We need the coordinate systems to be smoothly compatible in the following sense.

Definition 1.3 Two coordinate charts (U, φ) and (V, φ) on M are said to be c^∞-related if either $U \cap V = \emptyset$ or $\varphi o \psi^{-1} : \psi(U \cap V) \longrightarrow \varphi(U \cap V)$ is a diffeomorphism, between open sets of $\mathbb{R}^n$. $\varphi o \psi^{-1}$ is called smooth changes of coordinates on $U \cap V$.

A c^∞ atlas on M is a collection $\mathcal{A} = \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ of coordinate charts such that

(1) $(qU_\alpha, \varphi_\alpha)$ is c^∞-related to (U_β, φ_β) $\quad \forall \alpha, \beta \in I$.

(2) $M = \bigcup_{\alpha \in I} U_\alpha$

Two c^∞ atlases $\mathcal{A}$ and $\mathcal{A}'$ are equivalent if $\mathcal{A} \cup \mathcal{A}'$ is also a c^∞ atlas on M. Equivalence of c^∞ atlases is an equivalent relation. Each c^∞ atlas on M is equivalent to a unique maximal c^∞ atlas on M.

Definition 1.4 A maximal c^∞ atlas $\mathcal{A}$ on M is called smooth structure or differentiable structure or c^∞ structure on M. The pair $(M, \mathcal{A})$ is called a smooth or differentiable or c^∞ manifold of dimension n or simply n-manifold.

By substituting c^k for c^∞ we obtain c^k manifold $1 \leq k < \infty$. The same is for real analytic or c^ω manifold. In all these cases $\mathbb{R}^n$ is called the model space. We can take $\mathbb{C}^n$ as model space (complex manifold) or a Hilbert or Banach spaces, where the manifold will be infinite dimensional.

Definition 1.5 A function $f : M \longrightarrow R$ is said to be smooth if for each $x \in M$, there is a chart $(U, \varphi) \in \mathcal{A}$ such that $x \in U$ and $f o \varphi^{-1} : \varphi(U) \longrightarrow \mathbb{R}^n$ is smooth. The set of all smooth, real valued functions on M will be denoted by $\mathcal{D}(M)$.

Lemma 1.6 The function $f : M \longrightarrow \mathbb{R}^n$ is smooth iff

$fo\varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha) \longrightarrow I\!R^n$ is smooth for every $(U_\alpha, U_\alpha) \in \mathcal{A}$.

Definition 1.7 Let M and N be c^∞ manifold with respective smooth structures $\mathcal{A}$ and $\mathcal{B}$. A mapping $f : M \longrightarrow N$ is said to be smooth if, for each $x \in U_\alpha$, $f(U_\alpha) \subseteq V_\beta$ and

$$\psi_\beta o f o\varphi_\alpha^{-1} : (\varphi_\alpha(U_\alpha)) \longrightarrow \psi_\beta(V_\beta)$$

is smooth.

Lemma 1.8 the map $f : M \longrightarrow N$ is smooth if and only if, for all choices of $(U_\alpha, \psi_\alpha) \in \mathcal{A}$ and $(U_\alpha, \psi_\alpha) \in \mathcal{B}$ such that $f(U_\alpha) \subseteq V_\beta$, then the map

$$\psi_\beta o f o\varphi_\alpha^{-1} : (\varphi_\alpha(U_\alpha)) \longrightarrow \psi_\beta(V_\beta)$$

is smooth. [1;70]

Definition 1.9 Let M be a differentiable manifold and $p \in M$, for a chart or local coordinate neighborhood (U, φ) of p, let
$\alpha(-\varepsilon, \varepsilon) \subset I\!R \longrightarrow U \subset M$ be a differentiable curve such that $\alpha(0) = p$. Let $C(U,p)$ be the set of all differentiable curves on M passing through $p \in U$. If $\alpha, \beta \in C(U,p)$, then α and β are said to be infinitesimally equivalent at p or $\alpha \overset{p}{\simeq} \beta$, if and only if

$$\frac{d}{dt}(f(\alpha(t)))_{t=0} = \frac{d}{dt}(f(\beta(t)))_{t=0}$$

It is easy to check that $\overset{p}{\simeq}$ is an equivalence relation. An equivalence class of $\alpha \in C(U,p)$ is denoted by $[\alpha]_p$, and is called a tangent vector to U (or M) at p, and the set

$$T_pU = \{[\alpha]_p ; \alpha \in C(U,p)\}$$

is called the tangent space at p.

Lemma 1.10 For each $[\alpha]_p \in T_pU$ the operator

$$D[\alpha]_p : \mathcal{D}(U) \longrightarrow I\!R$$

is well defined, by choosing any representative $\alpha \in [\alpha]_p$ and setting

$$D[\alpha]_p(f) = \frac{d}{dt}(f(\alpha(t)))_{t=0} \qquad \forall f \in \mathcal{D}(U),$$

conversely $[\alpha]_p$ is uniquely determined by the operator $D[\alpha]_p$, [1;28].

By the definition of infinitesimal curves there is a natural one to one correspondence between T_pU (or T_pM) and $I\!R^n$ introducing tangent vectors as a linear approximation of nonlinear curves. The key lemma for this, follows:

Lemma 1.11 Let $[\alpha]_p, [\beta]_p \in T_pU$ and $a, b \in I\!R$ then there is a unique infinitesimal curve $[\gamma]_p$ such that the associated derivatives on $\mathcal{D}U$ satisfy

$$D[\gamma]_p = aD[\alpha]_p + bD[\beta]_p$$

Proof: Let $\gamma : (-\varepsilon, \varepsilon) \longrightarrow U$ be a curve on M defined by

$$\gamma(t) = a\alpha(t) = b\beta(t) - (a + b - 1)p$$

Since $\alpha(0) = \beta(0) = p$ then $\gamma(0) = p$. Also

$$\begin{aligned} D[\gamma]_p(f) &= \tfrac{d}{dt}(f(\gamma(t)))_{t=0} \\ &= \tfrac{d}{dt}f(a\alpha(t) + b\beta(t) - (a + b - 1)p)_{t=0} \\ &= a\tfrac{d}{dt}(f(\alpha(t)))_{t=0} + b\tfrac{d}{dt}(f(\beta(t)))_{t=0} \\ &= D[\alpha]_p(f) + D[\beta]_p(f) \\ D[\gamma]_p &= aD[\alpha]_p + bD[\beta]_p \end{aligned}$$

Therefore the operator $D[\alpha]_p$ makes T_pU or T_pM into an n-dimensional vector space over $I\!R$. The zero vector is the infinitesimal curve represented by the constant curve $\alpha : (\varepsilon, \varepsilon) \longrightarrow U$ by $\alpha(t) = p$. If $[\alpha]_p \in T_pU$. then $-[\alpha]_p = [-\alpha]_p$ where $(-\alpha)(t) = \alpha(-t)$, defined for all sufficiently small value of t.

The operator $D[\alpha]_p$ depends only the behavior of f in an arbitrary small neighborhood of p, that is $D[\alpha]_p(f)$ depends only on the " germ " of the t at p.

Definition 1.12 We say that $f, g \in \mathcal{D}(u)$ are germinally equivalent at p and write $f \overset{p}{\sim} g$. if there is an open neighborhood W of p in U such that $f/W = g/W$. Germinally equivalent is an equivalence relation on $\mathcal{D}(u)$, and equivalence class $[f]_p$ of $f \in \mathcal{D}(u)$ is called the germ of f at p. The set $\mathcal{D}(u)/\tilde{p}$ of germs at p is denoted by G_p.

Definition 1.13 For each $\alpha \in C(U, p)$ the operator $D[\alpha]_p : G_p \longrightarrow I\!R$ is defined by $D[\alpha]_p([f]_p) = \frac{d}{dt}(f(\alpha, t))_{t=0}$.

Definition 1.14 G_p together with the operations

1. $a[f]_p = [af]_p \qquad a \in I\!R, \qquad [f]_p \in G_p$ (scalar multiplication)
2. $[f]_p + [g]_p = [\frac{f}{w} + \frac{g}{w}]_p \qquad [f]_p, [g]_p \in G_p$(addition), W is open neighborhood of p in $dom(f) \cap dom(g)$.
3. $[f]_p[g]_p = [(\frac{f}{w})(\frac{g}{w})]_p \qquad [f]_p, [g]_p \in G_p$(multiplication)

Definition 1.15 The evaluation map $e_p : G_p \longrightarrow I\!R$ is defined by $e_p([f]_p) = f(p)$.

A derivative operator (derivative) on G_p is an $I\!R$-linear map $D : G_p \longrightarrow I\!R$ such that

$$\begin{aligned} D([f][g]) &= D[f]e_p([g]) + c_p([p])D[g] \\ &= g(p)D[f] + f(p)D[g]. \end{aligned}$$

A derivative on G_p is also called a tangent vector to U at P. The set of all derivatives on G_p is denoted by $T_p(U)$ or $T_p(M)$ and is called the tangent space to M at p.

Lemma 1.16 The tangent space T_pU or T_pM is a vector space over $I\!R$ under the operations:

1. $(aD)[f] = a(D[f]) \qquad a \in I\!R D \in T_pM, [f] \in G_p$
2. $(D_1 + D_2)[f] = D_1[f] + D_2[f] \qquad D_1, D_2 \in T_pM, [f] \in G_p$

The two definitions of T_PU give canonically isomorphic spaces [1;30]

For example define $(D_i)_pG_p \longrightarrow I\!R$ by

$$(D_i)_p[f]_p = (\frac{\partial t}{\partial x^i})(p)$$

It is proved that the set of tangent vectors $(D_i)_p = (\frac{\partial t}{\partial x^i})(p)$ is a basis of the vector space T_pU or T_pM [1;32].

that is if V is a tangent vector in T_pU, then

$$V = \sum_{i=1}^{n} v^i(D_i)_p = \sum_{i=1}^{n} v^i(\frac{\partial}{\partial x^i})_p$$

or simply by using the summation convention $V = v^i(\frac{\partial}{\partial x^i})_p$. Let $TU = \bigcup_{x \in U} T_p(U)$ be the disjoint union, then there is a one to one correspondence $TU \longleftrightarrow U \times I\!R$ given by

$$v^i(\frac{\partial}{\partial x^i})_p \longleftrightarrow (p; v^1, v^2, \ldots, v^n) \qquad (*)$$

We use $(*)$ to transfer the topology of $U \times I\!R$ to TU.

Definition 1.17 Let $f : M \longrightarrow N$ be a smooth map between differentiable manifolds M and n, where $\dim M = m, \dim N = n$. For any $p \in U \subset M$ and $f(p) \in V \subset N$ we define $(df)_p = (f)_p : T_pU \longrightarrow T_{f(p)}V$ by

$$(df)_p[\alpha]_p = [fo\alpha]_{f(p)} \qquad \forall [\alpha]_p \in T_pU.$$

This is called the differential of f at p.

Lemma 1.18 The differential $(df)_p$ is well defined linear map. Relative to the basis $(\frac{\partial}{\partial x^i})_p$ of T_pU and $(\frac{\partial}{\partial y^j})_{f(p)}$ of $T_{f(p)}V$ the matrix of $(df)_p$ is the Jacobian matrix of f at p. That is $Jf = (\frac{\partial y^j}{\partial x^i})_p \quad 1 \leq i \leq n \quad 1 \leq j \leq m$

.

The differential $(f_\alpha)_p$ computed at all $p \in U$, assemble to a mapping

$$f_* = df : TU \longrightarrow TV$$

given by

$$f_\alpha \left(p, \begin{bmatrix} v' \\ \vdots \\ v^m \end{bmatrix} \right) = \left(f(p), J_f \begin{bmatrix} v' \\ \vdots \\ v^m \end{bmatrix} \right).$$

Here we have identified TU with $U \times I\!R^m$ and TV with $V \times I\!R^n$ [1;35]

2. Tangent and cotangent bundles

Let M be a c^∞, m-manifold with structure $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$. Consider the set $TM = \bigcup_{p \in M} T_pM$ as disjoint union, for each $U_\alpha \quad \alpha \in I$, define

$$T(U_\alpha) = \bigcup_{p \in U_\alpha} T_p(U) \subseteq TM$$

. Then the individual linear map $(d\varphi_\alpha)_p \quad p \in U_\alpha$ unite to define a set map, considering $(*)$

$$d\varphi_\alpha : T(U_\alpha) \longrightarrow T(\varphi_\alpha(U_\alpha)) = \varphi_\alpha(U_\alpha) \times I\!R^n \subseteq I\!R^{2n}$$

More precisely, if V_p denotes a tangent vector to M at $p \in U_\alpha$.

$$(d\varphi_\alpha)(V_p) = (\varphi_\alpha)(p), (d\varphi_\alpha)_p V_p$$

and this defines a bijection of TU_α onto an open subset of $I\!R^{2n}$. If $U_\alpha \cap U_\beta \neq \phi$ consider

$$d\varphi_\alpha . d\varphi_\beta^{-1} : T(\varphi_\beta(U_\alpha \cap U_\beta)) \longrightarrow T(\varphi_\alpha(U_\alpha \cap U_\beta))$$

is a c^∞ diffeomorphism between open sets of $I\!R^{2n}$. We Topologies the set TM such that if

$$W \subseteq d\varphi_\alpha(T(U_\alpha)) = T(\varphi_\alpha(U_\alpha)) \subseteq I\!R^{2n}$$

is an open set, then $d\varphi_\alpha^{-1}(w)$ is to be an open subset of TM.

Lemma 2.1 The above sets from the base of a topology on TM and, in this topology, TM is a topological manifold of dimension $2m$. Furthermore, the system $\{T(U_\alpha), d\varphi_a lpha\}_{\alpha \in I}$ is a c^∞ atlas on TM, determining a maximal atlas $\mathcal{A}$.

$(TM, \mathcal{A})$ is called the tangent bundle of M. The map $\phi : TM \longrightarrow M$ determined by $\pi(p, v) = p \quad p \in M, v \in T_pM$ is smooth. TM is locally a Cartesian product of $U_\alpha \subset M$ and $I\!R^m$ and $\pi^{-1}(p) = T_pM$ is the tangent space of M at p.

Definition 2.2 A vector field on M is a smooth map $X : M \longrightarrow M \quad ; \quad P \longmapsto X_p$ such that $\pi o X = id_M$, the set of all vector fields on M is denoted by $D^1(M)$. If $(U, x^1, x^2, \dots, x^n)$ is a coordinate chart on M, then $X = X^i \frac{\partial}{\partial x^i}$ where $X^i : U \longrightarrow I\!R$ are smooth function, on M for $1 \leq i \leq n$.

In general Let $(U, x^1, x^2, \dots, x^n)$ be a coordinate chart on M, denote by (X^i) the system of Cartesian coordinates on each tangent space $T_p(M)$. Then in any open subset $\pi^{-1}(U)$ of TM we introduce local coordinates $(x^i, X^i) \qquad 1 \leq i \leq n$, which are called induced coordinate on $\pi^{-1}(u)$. We denote

$$\partial_i = \frac{\partial}{\partial x^i} \quad \text{and} \quad \bar{\partial}_i = \frac{\partial}{\partial X^i}$$

3. Covectors and 1-forms

Definition 3.1 Let M be a differentiable manifold $p \in M$. The dual space $T_p^*M = Hom(T_pM, I\!R)$ is called the cotangent space of M at p. Each element $w \in T_p^*M$ is called a cotangent vector to M at p.

A typical cotangent vector is the differential map. Let $U \subseteq M$ be open, $p \in U$, and let $f \in c^\infty(U)$. Since $T(I\!R) = I\!R$, we obtain a linear functional $(df)_p : T_pM \longrightarrow T(R) = R$, so $(df)_p \in T_p^*M$. It is evident that (df_p) depends only on the germ $[f]_p \in G_p$, so we obtain a surjective $I\!R$-linear map $d : G_p \longrightarrow T_p^*M$

For each $X_p \in T_pM$; $\quad (df)_pX_p = X_p(f)$.

Definition 3.2 If $\varphi : M \longrightarrow N$ is a smooth map between manifolds, if $p \in M$, and if $w \in T_p^*M$, then $\varphi_p^*(w) \in T_p^*(M)$ is defined by

$$\varphi_p^*(w)X_p = w(\varphi_{*p}(X_p)).$$

The linear map φ_p^* is called the adjoint of φ_{*p}.

Lemma 3.3 Relative to local coordinates $x^1, \ldots, x^n$ about $p \in M$, The differentials $dx^1, \ldots, dx^n$ at p from a basis of T_p^*M and

$$(dx^i)_p(\frac{\partial}{\partial x^j})_p = \frac{\partial x^i}{\partial x^j}(p) = \delta_{ij}$$

hence dim$T_p^*M = m$ [1;160]

Let $\varphi : M \longrightarrow N$ be smooth map between smooth manifolds. If $x^1, \ldots, x^m$ are coordinates about $p \in M$ and $y^1, \ldots, y^n$ coordinates about $\varphi(p) \in N$, we have basis $(\frac{\partial}{\partial x^i})_p \quad 1 \leq i \leq m$ for T_pM and $(\frac{\partial}{\partial y^j})_{\varphi(p)}$ for $T_{\varphi(p)}N$ and the dual basis $(dx^i)_p$ for T_p^*M and $(dy^j)_{\varphi(p)}$for T_φ^*N. The relations are as follows.

$$(\frac{\partial}{\partial y^j})_{\varphi(p)} = (\frac{\partial x^i}{\partial y^j})_p(\frac{\partial}{\partial x^i})_p \quad 1 \leq i \leq m \quad 1 \leq j \leq n$$

$$(dy^i)_{\varphi(p)} = (\frac{\partial y^j}{\partial x^i})_p(dx^i)_p \quad 1 \leq i \leq m \quad 1 \leq j \leq n$$

Now let $A = \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ be a maximal smooth atlas on M. As a set let

$$T^*M = \bigcup_{p \in M} T_p^*M$$

. We topologies each $T^*(U_\alpha)$ via the bijection

$$\psi_\alpha : T^\alpha(U_\alpha) \longrightarrow U_\alpha \times \mathbb{R}^n$$

defined by

$$\psi_\alpha(w_i(dx^i)_p) = (p, \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix})$$

repeating the case for tangent bundle, T^*M becomes a $2m$ dimensional smooth manifold, and for any $w \in T_p^*M$ we have $w = w_i dx^i$ for every tangent vector $X_p = (X^i(\frac{\partial}{\partial x^i})_p \in T_pM$ and 1-form $w = w_i(dx^i)_p \in T_p^*M$ we have

$$w_p(X_p) = (w_i(dx^i)_p)(X^j(\frac{\partial}{\partial x^j})_p) = w_i X^j(dx^i(\frac{\partial}{\partial x^j}))_p$$

$$= w_i X^j \partial_{ij} = w_i X^i. \qquad 1 \leq i, j \leq n$$

Lie bracket is defined as

$$[\ ,\ ] : D^1(M) \times D^1(M) \longrightarrow D^1(M)$$

$$[\ ,\ ](X, Y) = XY - YX$$

and $(D^1(M), [\ ,\ ])$ is called the Lie algebra of vectorfields on M, or simply Lie algebra of M. In general let $(U; x^1, \ldots, x^n)$ be a coordinate chart on M, then every 1-form w at $p \in U$ is written as $w = \xi_i dx^i$. Then (ξ_i) are the system of Cartesian coordinate in any open subset, π_1^{-1}/U

on T^*M, we can introduce local coordinate (x^i, ξ_i) $1 \leq i \leq n$, which are called induced coordinates. We denote

$$\partial_i = \frac{\partial}{\partial x^i} \quad , \quad \partial^i = \frac{\partial}{\partial \xi_i}$$

.

4. Lie groups and Lie algebra

A Lie group G is a smooth manifold without boundary which is also a group such that the group operation $\varphi : G \times G \longrightarrow G, \quad \varphi(x,y) = xy$ and the inversion $\psi : G \longrightarrow G; \quad \psi(x) = x^{-1}$ are boat smooth.

For example every finite dimensional vector space over $I\!R$ or $\mathbb{C}$ is a Lie group under vector addition.

$$(M(n,\mathbb{C})+), (Gl(n,\mathbb{C}), \times)(M(n,R),+),$$
$$(Gl(n,R),\times).U(n), O(n), Sl(n,R), \dots, S^7, S^3, S^1$$

are all Lie groups.

Definition 4.1 Let G be a Lie group, $g \in G$. Let translation by g is a smooth map $L_g : G \longrightarrow G$ defined by $L_a(x) = ax \quad \forall x \in G$. A vectorfield $X \in D^1(G)$ is left invariant if, for each $g \in G \quad (dl_g)X = X$. The set of all left invariant vectorfields on G is denoted by $L(G)$.

Proposition 4.2 The subset $L(G) \subset D^1(G)$ is the Lie algebra of the Lie group G and the evaluation map $E : L(G) \longrightarrow T_eG$ by $E(X) = X_e$ is an isomorphism of vector spaces (e is the unit of G). Hence, $dimL(G) = dimG$.[1,130]

Let A be $m \times m$ matrix and $\alpha(t) = \exp A \quad t \in R$ such that $\alpha(0) = I$. Then $\alpha(t+s) = \alpha(t)\alpha(s) \quad t,s \in R$ and $\alpha^{-1}(t) = \exp(-tA)$ that is $\alpha(t)$ is a 1-parameter subgroup of $Gl(n,R)$ of G is a Lie group then the exponential map $\exp : L(G) \equiv T_e G \longrightarrow G$ is defined by $\exp(X) = \alpha_X(1)$ where $\alpha_X(t) = \exp tX$.

5. Homogeneous space

Definition 5.1 Let M be a smooth manifold and G a Lie group. A smooth map $\varphi : G \times M \longrightarrow M$ written $\varphi(g, x) = gx$, is said to be an action of G on M, and G is called a Lie transformation group on M if

1. $\varphi(g_2, \varphi(g_1, x) = \varphi(g_1 g_2, x)$ or $g_1(g_2 x) = (g_1 g_2)x$ $\forall g_1, g_2 \in G$ and $\forall x \in M$

2. $\varphi(e, x) = x \; e \in G \forall x \in M$

Definition 5.2 An orbit of the action $G \times M \longrightarrow M$ is a set of points of the form $\{gx_0; g \in G\}$ where $x_0 \in M$. The action is transitive if M itself is an orbit , in which case M is said to be homogeneous space of G.

The orbits of a group action are equivalence classes, two points $x, y \in M$ being equivalent under the action if $\exists g \in G$ such that $gx = y$.

Definition 5.3 Let M be a homogeneous space of G and let $x_0 \in M$. The isotropy of x_0 is the set

$$G_{x_0} = \{g \in G; gx_0 = x_0\}$$

which is a property Lie subgroup of G.

Proposition 5.4 There is a smooth structure on quotient space where G/G_{x_0} and $M = G/G_{x_0}$, is a homogeneous space where $G \times M \longrightarrow M$ is given by $g(hG_{x_0}) = (gh)G_{x_0}$ [1,144]

Corollary 5.5 If G is a Lie group and $H \subseteq G$ is a closed normal subgroup, then the group G/H has a smooth structure in which it is a Lie group.[1,148]

For example $S^{n-1} = o(n)/o(n-1)$ and $S^{2n-1} = U(n)/U(n-1)$.

Definition 5.6 Let $\varphi : I\!R \times n \longrightarrow M$ be a smooth map such that $\varphi(o,p) = p$ and $\varphi(s, \varphi(t,p)) = \varphi(s+t,p)$, That is the additive group $I\!R$ is a transformation group on M. For each $t \in I\!R$ we define $\varphi_t M \longrightarrow M$ such that $\varphi_t(p) = \varphi(t,p)$ then it is easy to show that $\varphi_0 = Id$ and $\varphi_{s+t} = \varphi_s o \varphi_t$ also $\varphi_t^{-1} = \varphi_{-t}$.

$\{\varphi_t; t \in I\!R\}$ is a group. Called 1-parameter group of transformations on M.

6. Connections

Let $U \subseteq I\!R^n$ be open. Given $X, Y \in D^1(U)$, define $D_X(Y) \in D^1(M)$ as follows.

Write $X = X^i \frac{\partial}{\partial x^j}$ $\quad Y = Y^j \frac{\partial}{\partial x^j}$ and define

$$D_X(Y) = X(Y^j)\frac{\partial}{\partial x^j} = X^i \frac{\partial x^j}{\partial x^i}\frac{\partial}{\partial x^i}$$

We can view D as an $I\!R$-bilinear map

$$D : D^1(U) \times D^1(U) \longrightarrow D^1(U)$$

It has the following properties:

1. $D_{fX}(Y) = fD_X(Y)$. $\quad \forall f \in D(U) \forall \quad X, Y \in D^1(U)$

2. $D_X(fY) = X(f)Y + fD_X(Y)$

$D_X(Y)$ is called derivative of the vectorfield Y in X direction. The operation D is called Euclidean connection.

Definition 6.1 Let M be a smooth manifold. An Affine connection or M is an $\mathbb{R}$-bilinear map

$$\nabla : D^1(M) \times D^1(M) \longrightarrow D^1(M)$$

written as $\nabla(X,Y) = \nabla_X(Y)$ with the following properties, for each $X \in D^1(M); \nabla_X : D^1(M) \longrightarrow D^1(M)$ is a linear mapping , that is

1. $\nabla_{fX+gY}(Z) = f(\nabla_X Z)+g(\nabla_Y Z) \quad f, g \in c^\infty(M), Y, Z \in D^1(M)$
2. $\nabla_X(fY) = X(f)Y + f(\nabla_X Y)$

The operator ∇_X is called covariant differentiation with respect to X.

If ∇ is an offine connection on M and $(U_1 x^1, \ldots, x^n)$ is a coordinate chart, set $X = \frac{\partial}{\partial x^i}$ then

$$\nabla_{\frac{\partial}{\partial x^i}}(\frac{\partial}{\partial x^j}) = \Gamma^k_{ij}(\frac{\partial}{\partial x^i})$$

The smooth functions Γ^k_{ij} are called the Christoffel symbols.[3,27]

Definition 6.2 Let M be a manifold with an offine connection ∇. Let $X \in D^1(M)$ and $\alpha : (-\epsilon, \epsilon) \longrightarrow M$ be a differentiable curve on M. X is called parallel along α if $\nabla_{\dot{\alpha(t)}} X = 0$

Definition 6.3 Every manifold M has a connection.[1,301]

Let V be a vector space over $\mathbb{R}$. A bilinear form on V is defined to be a map $\varphi : V \times V \longrightarrow \mathbb{R}$ which is linear in each variable.

Definition 6.4 Suppose ∇ is an offine connection on M and $\varphi : M \longrightarrow M$ be smooth and

$$d\varphi(\nabla_X Y) = \nabla_{d\varphi X}(d\varphi Y) \quad \forall X, Y \in D^1(M)$$

Then φ is called offine transformation of (M, ∇).

Definition 6.5 Let (M,∇) be an offine manifolds, The curve $\alpha:(-\epsilon,\epsilon)\longrightarrow\alpha(t)$ is called a geodesic if the family of tangent vector $\dot{\alpha}(t)$ are parallel along α. That is $\nabla^{\dot{\alpha}(t)}_{\alpha(t)}=0$.

In a coordinate neighborhood $(U,x^1,\ldots,x^n)$; $\alpha(t)=(x^i(t))$is geodesic if it satisfies the system of differential equation

$$\frac{d^2x^k}{dt^2}+\Gamma^k_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt}=0 \qquad [3,30]$$

φ is called symmetric if $\varphi(x,y)=\varphi(y,x)$. a symmetric form is called positive definite if $\varphi(x,y)\geq 0$ and $\varphi(x,x)=0\Longleftrightarrow x=0$ φ is also called an inner product on V or φ is a tensor of type $(0,2)$.

Definition 6.6 A tensorfield g of smooth bilinear forms on a smooth manifold M which assigns to each $p\in M$ a symmetric, positive definite, bilinear form g_p of type $(0,2)$ at T_pM, that is

$$g_p:T_pM\times T_pM\longrightarrow R \qquad \text{(inner product)}$$

is called a Riemannian metric on M and (M,g) is a Riemannian manifold. If g_p fails to be positive definite then g is called pseudo-Riemannian and (M,g) is pseudo-Riemannian. [2,155]

For any coordinate neighborhood $(U,x^1,\ldots,x^n)$,

$$g(\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j})=g_{ij}$$

are called the components of g, that is $g=(g_{ij})$

For $M=I\!R^n$ $\qquad (\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j})=\delta_{ij}$.

Definition 6.7 A connection ∇ on the Riemannian manifold (M,g) is a Riemannian connection if for all $X,Y,Z\in D^1(M)$.

$$Xg(X,Y)=g(\nabla_XY,Z)+g(Y,\nabla_XZ).$$

If ∇ is symmetric then it is called Livi-Civita connection on (M,g).

Proposition 6.8 A Riemannian manifold M has a unique Livi-Civita connection [1,305]

In any local coordinate chart, the matrix (g_{ij}) of metric coefficients is nonsingular, so we can define $(g^{kl}) = (g_{ij})^{-1}$. The coefficient g^{kl} are rational functions of the metric coefficients g_{ij} and satisfy $g_{ik}g^{kl} = \delta_i^l$.

Definition 6.9 Let (M,g) and (N,g') be two smooth Riemannian or (pseudo-Riemannian) manifolds and $\varphi : M \longrightarrow N$ be smooth. φ is called a local isometry if for each $p \in M$, there exist open neighborhoods U of p and V of $\varphi(p)$ such that $\varphi_p^*(h) = g$. In other words φ is an tangent space at p and $\varphi(p)$. That is $g_p(X_p, Y_p) = h_{\varphi(p)}(d\varphi X, d\varphi Y)$

Definition 6.10 A vector X on (M,g) is called an isometry or killing vectorfield if local 1-parameter group of local transformations generated by X in a neighborhood of each $p \in M$, consists of local isometrics . In the same manner, a vectorfield X on (M,∇) is called an infinitesimal offine transformation of M, if for each $p \in M$ a local 1-parameter group of local transformation φ_t of a neighborhood U of p into M, preserves the connection ∇, more precisely if $\varphi_t : U \longrightarrow M$ is an offine mapping, where U is provided with offine connection.

An infinitesimal isometry is necessarily an infinitesimal offine transformation.

The Riemannian manifolds (M,g) is geodesically complete if $\exp_p(X_p)$ is defined for all $p \in M$ and for all $X_p \in T_p(M)$. Equivalently, every local geodesic extends (uniquely) to a geodesic $\alpha(t)$ for $-\infty < t < \infty$.

Proposition 6.11 (Hopt-Rinow) if M is geodesically complete then $\exp_p : T_pM \longrightarrow M$ is surjective.[1,318].

7. Riemannian homogeneous space

Let (M,g) be a connected Riemannian manifold. The group of all isometries $\varphi : M \longrightarrow M$, will be denoted by $I(M)$. The action of $I(M)$ on M preserves all intrinsic properties.

proposition 7.1 (Myers, Steenrod). If (M,g) is a Riemannian manifold, the group $I(M)$, with compact -open topology, is isomorphic, as a topological group, to a Lie group such that the natural action $I(M) \times M \longrightarrow M$ denoted by $(\varphi, x) \longrightarrow \varphi(x)$ is smooth and $M = I(M)/k$, where k is a closed subgroup, hence M is homogeneous manifold.[1,345]

Proposition 7.2 If M is a homogeneous manifold, then it is a complete Riemannian manifold.

Definition 7.3 A Riemannian symmetric space is a Riemannian manifold (M,g) with the property that, for each $p \in M$, is $\varphi_p \in I(M)$ such that $\varphi_p(p) = p$ and $d\varphi_p = -I$, where I is the identity transformation on $T_p(M)$.

Remark that, φ_p reserves every geodesic $s(t)$ through x. That is , if $s(0) = x$, then $\varphi_p(s(t)) = s(-t)$.

Proposition 7.4 If (M,g) is a symmetric space, then M is complete. If (M,g) is connected symmetric space, then M is s homogeneous Riemannian manifold.

8. Lifts

Let M be a manifold of dimension n. For any local coordinates $(U, x^1, \ldots, x^n)$ on M, we have $X = X^i \frac{\partial}{\partial x^i}, w = w_i dx^i$, and local coordinates $(\pi^{-1}(U), x^i, X^i) 1 \leq i \leq n$ for TM, $\pi_1^{-1}(U); x^i, w_i)$ for T^*M. Also

$\partial_i = \frac{\partial}{\partial x^i}, \bar{\partial}_i = \frac{\partial}{\partial X^i}$ basis for tangent spaces of TM and $\partial_i = \frac{\partial}{\partial x^i}, \partial^i = \frac{\partial}{\partial w^i}$ basis for cotangent spaces of T^*M.

Let $w \in D_1(M)$ be 1-form on M, we may regard w as a function on TM and denoted by τw.

For a function $\varphi \in c^\infty(M)$ we denote $\varphi c = \tau(d\varphi) \in TM$ and called complete lift of φ to TM. For any vectorfield X on M we define the complete lift $X^c \in D^1(TM)$ by $X^c\varphi^c = (X\varphi)^c \quad \forall \varphi \in c^\infty(M)$.

For any 1-form $w \in D_1(M)$ we define 1-form $w^c \in D_1(TM)$ by $w^c(X^c) = (w(X))^c \quad \forall X \in D^1(M)$, and w^c is called the complete lift of w to TM.

For $X = X^i\partial_i \in D^1(M)$, $X^\nu = X^i\frac{\partial}{\partial X^i} = \xi^i\bar{\partial}_i$ is called the vertical lift of X to TM. Also $X^h = X^i\partial_i - X^iX^j\Gamma^k_{ij}\bar{\partial}_k$ is called the horizontal lift of X on TM.

For a metric tensorfield $g = (g_{ij})$ on M we define

$$\begin{bmatrix} X^k\partial_k g_{ij} & g_{ij} \\ g_{ij} & 0 \end{bmatrix}$$

as the complete lift of g to TM. In local coordinates if $g = g_{ij}dx^idx^j$ then $g^c = \Gamma g_{ij}\delta\xi^i dx^j$, where $\delta\xi^i = dx^i + \Gamma^k_{jk}dx^j$.

For a vectorfield $X \in D^1(M)$, the vertical lift X^ν on T^*M is a function on T^*M defined by $X^\nu(\quad) = w(X_p)$.

For lift to T^*M we have vertical lift a 1-form w to be a vertical vectorfield w^ν given by

$$w^\nu(X^\nu) = (w(X))^\nu \qquad X \in D^1(M)$$

For a function $X^\nu \in c^\infty(T^*M)$, dx^ν is a 1-form given by

$$dx^\nu = (w_i\partial_j X^i)dx^j + X^i dw_i$$

For a vectorfield $X \in D^1(M)$ we define a vectorfield $X^c \in D^1(T^*M)$ by $X^c = (dx^\nu)$

in, where $\in$ is a tensorfield of type $(2,0)$ on T^*M with components

$$\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$$

where, I is the Kronrcker tensorfield.

For a vectorfield $X \in D^1(M)$ the horizontal lift is defined by $hX(Y^\nu) = (\nabla_X Y)^\nu$.

In coordinates:

$$\begin{array}{ll} X^\nu = w_i X^i & X = X^i \partial_i \\ w^\nu = w_i \frac{\partial}{\partial w_i} = w_i \partial^i & w = w_i dx^i \\ X^c = X^i \partial_i - w_i(\partial_j X^i)\partial^j & X = X^i \partial_i \\ hX = X^i \partial_i + w_i X^j \Gamma^i_{kj} \partial^k & X = X^i \partial_i \end{array}$$

Proposition 8.1 Let g be a pseudo-Riemannian metric on M. Then g^c is a pseudo-Riemannian metric on TM. (Yauo, Kobayashi).

Proposition 8.2 If X is a killing vectorfield of pseudo-Riemannian manifold (M, g), then X^ν and X^c are killing vectorfields of pseudo-Riemannian manifold (TM, g^c).

Let ∇ be a offine connection on M, there exists a unique offine connection ∇^c on TM, defined by

$$\nabla^c_{X^c} y^c = (\nabla_X Y)^c. \qquad \forall X, Y \in D^1(M).$$

If ∇ is a Riemannian connection on M, with respect to the pseudo-Riemannian metric g, then ∇^c is the Riemannian connection of TM with respect to the pseudo-Riemannian metric g^c. Also if X is an infinitesimal offine transformation of M, then both X^ν and X^c are infinitesimal offine transformation of TM with respect to ∇^c.[3]

Proposition 8.3 If the group of offine transformation of M with respect to ∇^c is transitive on TM. Also if M is a pseudo-Riemannian (offine) symmetric space with metric g (connection ∇), then TM is also a pseudo-Riemannian (offine) symmetric space with metric g^c (connection ∇^c).

9. Riemann extension

Define a canonical 1-form θ on T^*M via the following diagram

$$\begin{tikzcd}
& T(T^*M) \arrow[dl, "\pi_2"'] \arrow[dr, "\pi"] & \\
T^*M \arrow[dr, "\pi"'] & & TM \arrow[dl, "\pi_1"] \\
& M &
\end{tikzcd}$$

that is, if $\bar{X}$ is a vectorfield on T^*M, define $\theta(\bar{X}) = \pi_2(\bar{X})(\pi_*(\bar{X}))$.

In terms of coordinates.

$$\theta(x^1, \ldots, x^n, w_1, \ldots, w_n) = w_i dx^i$$

we call θ, the canonical 1-form on T^*M. Denote $\Omega = d\theta$. Ω is called a canonical 2-form on T^*M. In terms of coordinates:

$$\Omega(x^1, \ldots, x^n, w_1, \ldots, w_n) = dw_i \wedge dx^i$$

In matrix form

$$\Omega = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

This implies that Ω has maximal rank, and consequently defines a bundle isomorphism between $T(T^*M)$ and $T^*(T^*M)$.

Let H be the horizontal distribution on T^*M determined by the connection ∇ on M. Then π_* defines an isomorphism of $H_{\bar{p}}$ and the tangent space at $\pi(\bar{p}) \in M$. That is

$$\pi_*(hX)_{\bar{p}} = X_{\pi(\bar{p})} \qquad \bar{p} \in T^*M.$$

The tangent space $T_{\bar{p}}(T^*M)$ splits into horizontal and vertical subspaces,

$$T_{\bar{p}}(T^*M) = H_{\bar{p}} + V_{\bar{p}}$$

where $V_{\bar{p}} = \{\bar{X} \in T_{\bar{p}}(T^*M); \pi_*\bar{X} = 0\}$ and

$$H_{\bar{p}} = \{hX \in T_{\bar{p}}(T^*M); X \in T_pM\}.$$

In general every vectorfield $\bar{X} \in D^1(T^*M)$ can be written as:

$$\bar{X} = \bar{X}^i\partial_i + \bar{w}_i\partial^i = (\bar{X}\partial_i + \Gamma^l_{ik})w_cX^{-k}\partial^i) + (\bar{w}_i - \Gamma^l_{ik}wx^{-k})\partial^i$$

$$= \bar{X}^ie_i + \bar{V}_i\partial^i = \bar{X}^h + \bar{X}^v.$$

where $\bar{X}^h \in H$, $\quad \bar{X}^v \in V$. $\quad \bar{X}^i$ and V_i are generally functions of x and w.

We define the Riemann extension metric $\bar{g}$ on T^*M by

$$\bar{g}(\bar{X},\bar{Y}) = \Omega(\bar{X}^v,\bar{Y}^h) + \Omega(\bar{Y}^v,\bar{X}^h) \quad \bar{X},\bar{Y} \in T(T^*M)$$

In terms of coordinates $\bar{X} = \bar{X}^ie_i + \bar{Y}_i\partial^j, \bar{Y} = \bar{Y}^ie_i + \bar{U}_i\partial^i$

$$\bar{g}(\bar{X},\bar{Y}) = (\Gamma^l_{ki} + \Gamma^l_{ik})w_l\bar{X}^iY^k + \bar{X}^i\bar{U}_i + \bar{Y}^i\bar{V}_i$$

or

$$g : \begin{bmatrix} -(\Gamma^l_{ik} + \Gamma^l_{ki})w_l & I \\ I & 0 \end{bmatrix}$$

This shows that $\bar{g}$ depends only on the symmetric part of the connection ∇ on M. Hence we shall consider M to be an offine manifold with symmetric connection ∇. $(T^*M, \bar{g})$ is called the Riemann extension of (M, ∇).

We have calculated the Lie algebra of vectorfields on $(T^*M, \bar{g})$ and the connection $\bar{\nabla}$ arising from $\bar{g}$ on T^*M.

We have proved that the covariant derivative of $\bar{g}$ is zero that is $\bar{\nabla}\bar{g} = 0$, which is the case for any Riemannian connection ∇ on a Riemannian manifold (M, g). i.e. $\nabla g = o$.

We have modified the metric $\bar{g}$ for a pseudo-Riemannian manifold (M, g) so that T^*M becomes a natural Riemann extension. That is the metric defined on T^*M on the zero section coinsides with the original metric of M.

Let g be the pseudo-Riemannian metric on M and define

$$\bar{g}_c(\bar{X}, \bar{Y}) = \Omega(\bar{X}^v, \bar{Y}^h) + \Omega(\bar{Y}^v, \bar{X}^h) + cg(\pi_* \bar{X}, \pi_* \bar{Y})$$

where $\bar{X}, \bar{Y} \in D^1(T^*M)$ and $c \in I\!R$. For $c = 0$ it is the metric $\bar{g}$, and for $c = 1$ the additional term is just the horizontal lift of g, and for $c > 0$ the additional terms is hypothetic to the horizontal lift of g.

We have proved that $\nabla g = 0 \Longrightarrow \bar{\nabla}\bar{g}_c = 0$.

Definition 9.1 $(T^*M, \bar{g}_c)$ is called a generalized Riemann extension for the pseudo-Riemannian manifold (M, g).

We have calculated the covariant derivatives of all vectorfields on T^*M and hence get the Riemann extension connection $\bar{\nabla}$ on T^*M.

We have calculated the curvature tensorfield $\bar{R}$ on T^*M and prove the following theorem:

Theorem 9.2 Riemann extension $(T^*M, \bar{g})$ of a manifold (M, ∇) is locally symmetric if and only if (M, ∇) is locally symmetric.[4]

Now we show that complete lift and Riemann extension of a pseudo-Riemannian manifolds are isomorphic but for offinly connected manifolds they are not necessary isomorphic.

Let (M,g) be a pseudo-Riemannian manifold. For any local coordinate $(x^i), 1 \leq i \leq n$ in M we have associated the local coordinates (x^i, X^i) and (x^i, w_i) to TM and T^*M respectively, also g^c and $\bar{g}$ the complete lifts and the generalized Riemann extension metrics:

Let $\varphi : TM \longrightarrow T^*M$ be mapping defined by

$$\varphi(x^i, X^i) = (x^i, g_{ij}X^j) \quad , \quad g = g_{ij}$$

It is obviously differentiable and its inverse $\varphi^{-1} = (x^i, w_i) = (x^i, g^{ij}w_j)$ is also differentiable, hence φ is a diffeomorphism. Furthermore,

$$g^c = A^t \bar{g} A$$

where

$$A = \left(\frac{\partial(x^i, w_i)}{\partial(x^j, w_j)} \right) = \begin{pmatrix} \delta_{ij} & 0 \\ X^k \partial_j g_{ik} & g_{ij} \end{pmatrix}.$$

then φ is an isometry. This together with the proposition 8.3 gives the following theorem.

Theorem 9.3 If M is a pseudo-Riemannian symmetric space, then its Riemann extension is a pseudo-Riemannian symmetric space.

Now let $M = I\!R^2$ with coordinates (x^1, x^2) and ∇ with coordinates $\Gamma^2_{11} = x^2 \qquad \Gamma^k_{ij} = 0 \qquad i,j,k = 1,2$ we have proved that $T = 0, \nabla R = 0$ and M is an offine locally symmetric space.

We have constract the Riemann extension $(T^*M, \bar{g})$ and proved that it is locally symmetric pseudo-Riemannian manifold.

We also constract the complete lift (TM, ∇^c) and proved that ∇^c does not arise from any pseudo-Riemannian metric.

We have proved that following theorem:

Theorem 9.4 The generalized Riemann-extension of a pseudo-Riemannian homogeneous manifold (M, g) is a pseudo-Riemannian homogeneous manifold $(T^*M, \bar{g}_e)$.[4]

We have calculated general killing vector fields, curvature tensorfield, general infinitesimal vectorfields and their Lie algebra of $(T^*M, \bar{g}_c)$.[5]

References

[1] L.Conlon; *differentiable manifolds* Birkhäuser advanced text. (1993)

[2] S.Kobayashi and K.Nomizu, *foundations of differential geometry*Vol. Interscience publisher.

[3] Prolongation of tensorfields on tangent bundle J.Math.Soc. No 18.

[4] M.Toomanian, Riemannian s-symettric spaces which are not locally symmetric. Tensor M.S. Vol 44 (1987) Japan. transformation

[5] M.Toomanian. Killing vectorfields and infinitesimal offine on generalized Riemann extension. Tensor M.S. Vol 32 Japan.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 227-270 (2004)

The Stable Iterative Algorithms with the CADNA Library for Solving Sparse Linear Systems

Faezeh Toutounian

Department of Mathematics, Ferdowsi University of Mashhad,
Mashhad, Iran
toutouni@math.um.ac.ir

Abstract: This paper shows the efficiency of iterative methods for solving large sparse systems of linear algebraic equations when they are performed with an efficient round-off error analysis method. In a first part the most robust iterative methods such as GMRES, hybrid GMRES, A^TA-orthogonal s-step Orthomin(k), and QMR algorithms are briefly presented and some of their properties are recalled. Particularly, it is shown that in the floating-point arithmetic there exist some cases in which the properties of these algorithms are lost, e.g ., the result is false, or the coefficients of the GMRES residual polynomial are non significant and lead to serious round-off errors. The subject of this

MSC (2000): 65F10, 65F25, 65Go5
Keywords: Iterative methods, Hybrid methods, CESTAC method, stochastic arithmetic, CADNA library.

paper is to show how by using the CADNA library, it is possible during the run of the codes of these algorithms to detect the numerical instabilities, to stop correctly the process, and to evaluate the accuracy of the results provided by the computer. Numerical examples are used to show the good numerical properties.

1. Introduction

In recent years there has been significant progress in development of iterative methods for solving sparse real linear systems of the form

$$Ax = b, \tag{53}$$

where A is a nonsymmetric matrix of order n. The GMRES, hybrid GMRES, A^TA-orthogonal s-step Orthomin(k), and QMR algorithms are the most popular iterative methods for solving such systems. From the mathematical standpoint, these methods, based on a given initial point $x^{(0)}$ considered as an approximation of the solution to the problem to be solved, consist in computing a sequence $x^{(0)}, x^{(1)}, \ldots, x^{(n)}$ such that, if the method converge, $x^{(n)}$ tends towards a limit x_s, when $n \to \infty$. Obviously it is impossible to reach this limit, and consequently termination criteria are proposed to stop the iterative process, such as

$$\text{if } \|x^{(q)} - x^{(q-1)}\| \leq \epsilon \text{ then stop,}$$

$$\text{if } \|x^{(q)} - x^{(q-1)}\| \leq \epsilon \|x^{(q)}\| \text{ then stop,}$$

where ϵ is an arbitrary positive value. It is clear that these termination are not satisfactory because

$$\|x^{(q)} - x^{(q-1)}\| \leq \epsilon \not\Longleftrightarrow \|x^{(q)} - x_s\| \leq \epsilon,$$

$$\|x^{(q)} - x^{(q-1)}\| \leq \epsilon \|x^{(q)}\| \not\Longleftrightarrow \|x^{(q)} - x_s\| \leq \epsilon \|x_s\|.$$

From the informatical standpoint, the situation is even more serious. In fact, if an iterative method is implemented on a computer, each $X^{(q)} \in F$ only has a certain number of decimal significant digits. If the ϵ selected is less than accuracy of $X^{(q)}$, these termination criteria are no longer meaningful. Thus two problems are raised.

•How can the iteration process be stopped correctly?

•What is the accuracy of the informatical solution given by the machine?

In addition, in the floating-point arithmetic there exist some cases in which the properties of these algorithms are lost, e.g., the coefficients of the GMRES residual polynomial are non-significant and lead to serious round-off errors, or the A^TA-orthogonal method has slow convergence, because of round-off error propagation.

In section 4, we observe that, the stochastic arithmetic allows the development of a termination criterion, called the "optimal termination criterion", which stops the iterative process as soon as a satisfactory informatical solution is reached, and the use of other appropriate criteria for overcoming the instabilities which exist in the algorithms.

In section 2 we briefly describe the GMRES, hybrid GMRES, A^TA-orthogonal s-step Orthomin(k) and QMR algorithms and discuss about the problems which exist in the implementation of these algorithms on a computer.

In section 3 we give a brief description of stochastic round-off analysis, the CESTAC method, and the CADNA software [25, 4]. Section 4 is devoted to the use of the CESTAC method and CADNA library for overcoming the problems which exist in the implementation of the mentioned algorithms on a computer using the floating-point arithmetic. Moreover, we will observe that by using the CADNA library and introducing the appropriate stopping criteria, it is possible, during the run of the code of these algorithms to detect the numerical instabilities and to stop correctly the iterative process, and to restart it in order to improve the computed solution. Some numerical results are given to show the good numerical properties.

2. Hybrid GMRES, A^TA-orthogonal s-step Orthomin(k), and QMR algorithms

2.1 GMRES and Hybrid GMRES algorithms

The GMRES method by Saad and Schultz [22], is one of the most popular methods for solving linear systems with a nonsymmetric matrix. The idea of GMRES is to construct an approximate solution of the form $x_m = x_0 + z_m$ where z_m is an element of the Krylov subspace $K^m(A; r_0) =$ span$\{r_o, Ar_0, \dots, A^{m-1}r_0\}$ with the following property:

$$\|r_m\|_2 = \|b - Ax_m\|_2 = \min_{z \in K^m(A; r_o)} \|r_0 - Az\|_2, \tag{54}$$

where $r_0 = b - Ax_0$. The basic structure of the GMRES algorithm is as follows.

Algorithm 2.1. GMRES

```
1. Compute $r_0 = b - Ax_0, \beta = \|r_0\|_2$, and $v_1 = r_0/\beta$
2. Define the $(m+1) \times m$ matrix $\bar{H}_m = \{h_{ij}\}_{1 \le i \le m+1, 1 \le j \le m}$. Set
   $\bar{H}_m = 0$
3. For $j = 1, 2, ..., m$ Do:
4.        Compute $w_j = Av_j$
5.        For $i = 1, 2, ..., j$ Do:
6.               $h_{ij} = (w_j, v_i)$
```

```
7.              $w_j = w_j - h_{ij} v_i$
8.          EndDo
9.          $h_{j+1,j} = \|w_j\|_2$. If $h_{j+1,j} = 0$ set $m = j$ and goto 12
10.         $v_{j+1} = w_j / h_{j+1,j}$
11.  EndDo
12.  Compute $y_m$ the minimizer of $\|\beta e_1 - \bar{H}_m y\|_2$ and $x_m = x_0 + V_m y$
```

The GMRES iteration constructs a sequence of residual polynomials that minimize the norm of the residual

$$\|r_m\|_2 = \|p_m(A)r_0\|_2 = \min_{\substack{p \in P_m \\ p(0)=1}} \|p(A)r_0\|_2, \; m = 1, 2, \ldots .$$

With these GMRES polynomials the following hybrid GMRES is proposed in [20]:

Start with a random initial guess x_0 .

Phase I: Run GMRES until $\|r_m\|_2$ drops by a suitable amount. Set $\nu = m$.

Phase II: Re-apply the GMRES residual polynomial $p_\nu(z)$cyclically until convergence.

The main idea of this algorithm is to suppose that at the νth GMRES step the relation

$$\frac{\|r_\nu\|_2}{\|r_0\|_2} = \frac{\|p_\nu(A)r_0\|_2}{\|r_0\|_2} = \tau$$

holds for some $\tau < 1$, and moreover we have

$$\|p_\nu(A)\|_2 \simeq \tau.$$

So that by re-applying the GMRES polynomial $p_\nu(z)$ cyclically we can reduce the residual norm. Of course, these assumptions do not always hold and it must modify the algorithm in order to cope with failure. By assumption that storage is not limited, we propose the following safeguarding procedure which differs from Nachtigal, Reichel, and Trefethen algorithm [20] in process 2:

1. If any cycle of ν steps of phase II reduces $\|r_m\|_2$ by a factor less than $\sqrt{\tau}$ - that is, if the convergence is more than twice as slow as expected - return to phase I.

2. Carry out additional GMRES steps $\nu+1, \nu+2, \ldots, \nu'$ of phase I until $\|r_{\nu'}\|_2 / \, |r_\nu\|_2 < \frac{1}{2}$, and calculate a new polynomial $p_{\nu'}(z)$.

3. Begin a new phase II iteration with the new polynomial $p_{\nu'}(z)$, starting from the previous best value x_m , which will come either from the previous phase II if the convergence there was slow but positive, or from the new phase I if there was actual divergence in the previous phase II.

The residual polynomial $p_\nu(z)$ can be implicitly constructed by GMRES. The details of calculating the coefficients of $p_\nu(z)$ can be found in [20].

When the hybrid GMRES is implemented on a computer, at each iteration q of this algorithm we have one approximate solution from phase I, denoted $x_{0\nu}$, and some approximate solutions from phase II, denoted $x_{k\nu}, k = 1, 2, \ldots$. During the run of the program if any of these approximate solution is a satisfactory solution, in the sense that its residual norm is reduced by a factor ϵ (where ϵ is an arbitrary positive value), then the program can be stopped. So, for stopping the process at iteration q with $\nu < n, k = 0, 1, \ldots$ we can use the following termination criterion:

$$\text{if } \|r_{k\nu}\|_2 \leq \epsilon \|r_0\|_2 \ \text{ then stop.}$$

In the case $\nu = n$, the program must be stopped because the GMRES method converges, in the absence of rounding errors, in at most n iterations.

Note that in the above termination criterion ϵ is an arbitrary value, the results of example 1 section 4 (Table 2, in which only the decimal significant digits are printed) show that when ϵ is chosen too large ($\epsilon = 10^{-6}$) the process is broken-off too early and consequently the solution obtained has a poor accuracy. On the contrary when ϵ is chosen too small ($\epsilon = 10^{-16}$) the iterative process is stopped too late and many useless iterations are performed, without improving the accuracy of the solution obtained with $\epsilon = 10^{-15}$. In practice it is absolutely impossible to choose correctly the value of the convergence tolerance ϵ . But as explained in section 4 with CADNA library it is possible to break-off the iteration of an iterative process as soon as a satisfactory computed solution is reached, and this without using any arbitrary ϵ .

Let us now consider the linear system

$$\begin{bmatrix} 21 & 130 & 0 & 2.1 \\ 13 & 80 & 4.74+E8 & 752 \\ 0 & -0.4 & 3.9816+E8 & 4.2 \\ 0 & 0 & 1.7 & 9E-9 \end{bmatrix} .x = \begin{bmatrix} 153.1 \\ 849.74 \\ 7.7816 \\ 2.6E-8 \end{bmatrix}, \qquad (55)$$

which was described in [25]. The exact solution is $x = [1.0,\ 1.0,\ 10^{-8},\ 1.0]^T$. The approximate solution obtained with the initial guess $x_0 = [0, 0, \ldots, 0]^T$ and the FORTRAN code of the hybrid GMRES algorithm, performed on a SUN4 computer, in double precision is as follows

$$x(1) = -89.8760751972095591, \qquad x(2) = 15.6799813793030225,$$
$$x(3) = 2.4747823346160658E-08, \qquad x(4) = 1.0000000003115719.$$

It is necessary to say that this approximate solution has been obtained by phase I with $\nu = 4$. This solution is false and hybrid GMRES algorithm is not able, because of the propagation of the round-off errors, to provide a satisfactory solution for this example. However, nothing in the software nor in the solution allows the user to be aware that the computed solution is false. In section 3 we will show that the CADNA library is able to estimate the propagation of round-off errors during the run of the code and to furnish the accuracy of the computed solution.

Let us consider another linear system with the $n \times n$ block diagonal matrix

$$A = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_{n/2} \end{bmatrix}, \quad \text{where } M_j = \begin{bmatrix} 1 & j-1+\alpha \\ 0 & -1 \end{bmatrix}, \tag{56}$$

which with $\alpha = 0$ corresponds to the example $B_{\pm 1}$ of [21], and the second member $b = [5, -3, 4, -4, 1, \dots, 1]^T$. The exact solution is given by $x_{2j-1} = b_{2j-1} + (j - 1 + \alpha)b_{2j}$ and $x_{2j} = -b_{2j}$. With $\alpha = 1.11$, $n = 150$, and the initial guess $x_0 = [0, 0, \dots, 0]^T$ the computed coefficients α_i of the GMRES polynomial $p_6(z)$ with single and double floating-point arithmetic rounded to nearest mode are presented in Table 1. In section 4, thanks to the CADNA library, we consider that these coefficients are non significants. It is clear that, phase II is not able here to improve the approximate solution which has been obtained by phase I. In general, when ν has a large value, it is possible to have an unstable solution $x_{k\nu}, k = 1, 2, \dots$ and also a GMRES polynomial with unsignificant coefficients. In this situation the use of phase II has obviously no sense. So, in order to avoid the performation of many useless operations of phase II and to prevent an overflow which may occur, it is better the program perform only phase I as soon as such an instability occurs during the run. We now face the question of how can detect these kind of instabilities? The CADNA library is a precious tool for obtaining an answer to this question. In section 4 we show that with CADNA library it is possible, by including a simple test to detect it.

	single precision	double precision
α_0	-2.9381578E+05	1.6409978460272019E+14
α_1	3.0272866E+05	-1.6409978460272156E+14
α_2	5.8763362E+05	-3.2819956920543981E+14
α_3	-6.0545688E+05	3.2819956920544550E+14
α_4	-2.9381784E+05	1.6409978460271962E+14
α_5	3.0272922E+05	-1.6409978460272297E+14

Table 1

2.2 A^TA-orthogonal s-step Orthomin(k) algorithm

In [7], Chronopoulos developes A^TA-orthogonal s-step Orthomin(k) algorithm for nonsymmetric matrices with symmetric part $M = (A + A^T)/2$ positive definite or indefinite. In this method the s directions $\{r_i, \ldots, A^{s-1}r_i\}$ are formed and are A^TA-orthogonalized simultaneously to k of the preceding directions $\{p_j^1, \ldots, p_j^s\}, j = j_i, \ldots, i$ where $j_i = max(0, i-k+1)$. The norm of residual $\|r_{i+1}\|_2$ is minimized simultaneously in all s new directions in order to obtain x_{i+1}. More details of the s–step Orthomin(k) algorithm can be found in [7]. The following notation facilitates the description of the algorithm.

$$W_i = [(Ap_i^j, Ap_i^l)], \text{ where } 1 \leq j, l \leq s$$

$$\underline{a_i} = [a_i^1, \ldots, a_i^s]^T \text{ (the steplengths in updating } x_i \text{)}$$

$$\underline{m_i} = [(r_i, Ap_i^1), \ldots, (r_i, Ap_i^s)]^T$$

$$\underline{c_j^l} = [(A^l r_{i+1}, Ap_j^1), \ldots, (A^l r_{i+1}, Ap_j^s)]^T$$

$\underline{b}_j^l = \{b_j^{l,m}\}_{m=1}^s$ for $j = j_i, \ldots, i$ and $l = 1, \ldots, s$, where $j_i = max(0, i - k + 1)$
(the coefficients to $A^T A$-orthogonalize to the previous directions)

$P_i = [p_i^1, \ldots, p_i^s]$ (the direction vectors)

$R_i = [r_i, Ar_i, \ldots, A^{s-1}r_i]$(the residuals).

A description of s-step Orthomin(k) method can be given as follows:

Algorithm 2.2. s-step Orthomin(k)

```
Select $x_0$
$P_0 = [r_0 = b - Ax_0, Ar_0, \ldots, A^{s-1}r_0]$
**For** $i = 0$ **Until** Convergence **Do**
        Compute $\underline{m}_i, W_i$
        Call Scalar1
        $x_{i+1} = x_i + P_i\underline{a}_i$
        $r_{i+1} = r_i - AP_i\underline{a}_i$
        Compute $\underline{c}_j^i, j = j_i, \ldots, i$
        Call Scalar2
        Compute $R_{i+1} = [r_{i+1}, Ar_{i+1}, \ldots, A^{s-1}r_{i+1}]$
        $P_{i+1} = R_{i+1} + \sum_{j=j_i}^{i} P_j[\underline{b}_j^l]_{l=1}^s$
        Compute $AP_{i+1}$ or,
        $AP_{i+1} = AR_{i+1} + \sum_{j=j_i}^{i} AP_j[\underline{b}_j^l]_{l=1}^s$
```

```
EndFor

Scalar1: Decomposes $W_i$ and solves $W_i \underline{a}_i = \underline{m}_i$.

Scalar2: Solves $W_j \underline{b}_j^l = -\underline{c}_j^l$ for $j = j_i, \dots, i$ and $l = 1, \dots, s$, where
        $j_i = max(0, i - k + 1)$
```

The solution of the linear systems may cause a quick loss of orthogonality of the s-dimensional direction subspaces P_i because the matrix W_i may have a very large condition number. Numerical tests [8, 9, 10] have shown that the condition number of W_i is small for $s \leq 5$. One way to alleviate the orthogonality loss which can occur for large $s > 5$ is to $A^T A$-orthogonalize the s direction vectors in each iteration. In [23], $A^T A$-orthogonal s-step Orthomin(k) was developed and shown to be stable for large values of s(up to $s = 16$). In this method the direction vectors within each subspace P_i are $A^T A$-orthogonalized using the Modified Gram-Schmidt method. The linear systems need not be solved at each iteration since the W_i matrix is the identity matrix if P_i is perfectly $A^T A$-orthogonalized. By using the notation $j_i = max(0, i - k + 1)$ the algorithm can be described as follows:

Algorithm 2.3. $A^T A$-orthogonal s-step Orthomin(k)

```
Select $x_0$

Compute $r_0 = b - Ax_0$

For $i = 0$ Until Convergence Do

    Compute $AP_i = [Ar_i, A^2 r_i, \dots, A^s r_i]$,
    and set $P_i = [r_i, Ar_i, \dots, A^{s-1} r_i]$

    If $(0 < i)$ Then
```

```
            Compute $\underline{b}_j^l = [-(A^l r_i, Ap_j^1), \ldots, -(A^l r_i, Ap_j^s)]^T$, for
            $l = 1, \ldots, s$ and $j = j_{i-1}, \ldots, i-1$
            Compute $P_i = P_i + \sum_{j=j(i-1)}^{i-1} P_j[\underline{b}_j^l]_{l=1}^s$
            Compute $AP_i = AP_i + \sum_{j=j(i-1)}^{i-1} AP_j[\underline{b}_j^l]_{l=1}^s$
        EndIf
        Apply the Modified Gram-Schmidt method to the matrix $AP_i$
        to obtain final $AP_i$ and $P_i$
        Compute $\underline{a}_i = [(r_i, Ap_i^1), \ldots, (r_i, Ap_i^s)]^T$
        $x_{i+1} = x_i + P_i\underline{a}_i$
        $r_{i+1} = r_i - AP_i\underline{a}_i$
    EndFor.
```

The main problem in the use of $A^T A$-orthogonal s-step Orthomin(k) method, with floating-point arithmetic, is the choice of s. Let us consider the results of this method with different values of s and k for the examples 4-6 of section 4.2 (Tables 6, 8, and 10, in which the number of iterations to convergence are printed). These results clearly show that when s has a small or large value the method has slow convergence, and for each problem and each k there exists an s which minimize the number of iterations to convergence. However, as above mentioned, the slow convergence of the method with large value of s is due to the round-off errors propagation. Hence, it is not possible to determine a good value of s without estimating the round-off errors propagation. In section 4, it is shown that by using the CADNA library, which is an efficient tool for doing so, we will be able to determine a good value of s.

Here, another problem is also the choice of the value ϵ for stopping criterion $\|r_i\|_2 \leq \epsilon$. When ϵ is chosen too large, the iterative process is stopped too soon, and consequently the solution obtained has a poor accuracy. On the contrary, when ϵ is chosen small, it is possible, due to the numerical instabilities, many useless iterations are performed without improving the accuracy of the solution. How can the iterative processs be stopped correctly, and restarted in order to improve the computed solution? The CADNA library is a precious tool for obtaining an answer to this question. In section 4 we will show that with CADNA library, it is possible, by including the simple tests to stop and to restart correctly the iterative process.

2.3 The QMR algorithm

The quasi-minimal residual method (QMR) is based on the look-ahead Lanczos algorithm proposed in [13]. In the following, $A \in \mathbb{C}^{N\times N}$ is always assumed to be a given $N \times N$ matrix. Let $v_1, w_1 \in \mathbb{C}^N$ be any two vectors different from the zero vector. Starting with v_1, w_1, the look-ahead Lanczos algorithm generate two sequences of vectors $v_1, v_2, \cdots, v_n$ and $w_1, w_2, \cdots, w_n$, $n = 1, 2, \cdots$, that satisfy

$$\begin{aligned} \text{span}\{v_1, v_2, \cdots, v_n\} &= K_n(v_1, A), \\ \text{span}\{w_1, w_2, \cdots, w_n\} &= K_n(w_1, A^T), \end{aligned} \tag{57}$$

and can be grouped into $k = k(n)$ blocks

$$V_l = [v_{n_l} v_{n_l+1} \cdots v_{n_{l+1}-1}], \qquad W_l = [w_{n_l} w_{n_l+1} \cdots w_{n_{l+1}-1}],$$

$$l = 1, 2, \cdots, k-1,$$

$$V_k = [v_{n_k} v_{n_k+1} \cdots v_n], \qquad W_k = [w_{n_k} w_{n_k+1} \cdots w_n],$$

where

$$1 = n_1 < n_2 < \cdots < n_l < \cdots < n_k \leq n < n_{k+1}.$$

The blocks are constructed such that we have

$$W_j^T V_l = \begin{cases} 0 & \text{if } j \neq l, \\ D_l & \text{if } j = l, \end{cases} \qquad j, l = 1, 2, \cdots, k,$$

where

$$D_l \text{ is nonsingular, } l = 1, 2, \cdots, k-1, \text{ and}$$

$$D_k \text{ is nonsingular if } n = n_{k+1} - 1.$$

The first vectors v_{n_l} and w_{n_l} in each block are called *regular*, the remaining vectors are called *inner*. The kth block is called *complete* if $n = n_{k+1} - 1$; in this case, at the next step $n+1$, a new block is started with the regular vectors $v_{n_{k+1}}$ and $w_{n_{k+1}}$. Otherwise, if $n < n_{k+1} - 1$, the kth block is *incomplete* and at the next step, the Lanczos vectors v_{n+1} and w_{n+1} are added to the kth block as inner vectors.

With these preliminaries, the basic structure of the look-ahead Lanczos algorithm is as follows.

Algorithm 2.4 sketch of the look-ahead Lanczos algorithm [13].

```
0) Choose $v_1, w_1 \in \mathbb{C}^N$ with $\|v_1\| = \|w_1\| = 1$;
    Set $V_1 = v_1, W_1 = w_1, D_1 = W_1^T V_1$;
    Set $n_1 = 1, k = 1, v_0 = w_0 = 0, V_0 = W_0 = \emptyset, \rho_1 = \xi_1 = 1$;
For $n = 1, 2, \cdots$ do:
1) Decide whether to construct $v_{n+1}$ and $w_{n+1}$ as regular or inner
    vectors and go to 2) or 3), respectively;
2) (Regular step.) Compute
    $\tilde{v}_{n+1} = Av_n - V_k D_k^{-1} W_k^T A v_n - V_{k-1} D_{k-1}^{-1} W_{k-1}^T A v_n,$
    $\tilde{w}_{n+1} = A^T w_n - W_k D_k^{-T} V_k^T A^T w_n - W_{k-1} D_{k-1}^{-T} V_{k-1}^T A^T w_n,$   (58)
```

set $n_{k+1} = n+1, k = k+1, V_k = W_k = \emptyset$, *and go to* 4);

3)*(Inner step.) Compute*

$$\begin{aligned}\tilde{v}_{n+1} &= Av_n - \zeta_{n-n_k}v_n - (\eta_{n-n_k}/\rho_n)v_{n-1} - V_{k-1}D_{k-1}^{-1}W_{k-1}^T Av_n,\\ \tilde{w}_{n+1} &= A^T w_n - \zeta_{n-n_k}w_n - (\eta_{n-n_k}/\xi_n)w_{n-1} - W_{k-1}D_{k-1}^{-T}V_{k-1}^T A^T w_n;\end{aligned} \tag{59}$$

4)*Compute* $\rho_{n+1} = \| \tilde{v}_{n+1} \|$ *and* $\xi_{n+1} = \| \tilde{w}_{n+1} \|$;

If $\rho_{n+1} = 0$ *or* $\xi_{n+1} = 0$ *,stop;*

Otherwise, set

$$\begin{aligned}&v_{n+1} = \tilde{v}_{n+1}/\rho_{n+1}, \qquad w_{n+1} = \tilde{w}_{n+1}/\xi_{n+1},\\ &V_k = [V_k \ v_{n+1}], \quad W_k = [W_k \ w_{n+1}], \quad D_k = W_k^T V_k.\end{aligned} \tag{60}$$

Now, we list some properties of Algorithm 2.4 which will be used in the sequel. First, in view of (8), we have

$$\|v_n\| = \|w_n\| = 1, \qquad n = 1, 2, \dots . \tag{61}$$

It is convenient to introduce the notation

$$\begin{aligned}V^{(n)} &= [v_1 \ v_2 \ \dots \ v_n] \quad (= [V_1 \ V_2 \ \dots \ V_k]),\\ W^{(n)} &= [w_1 w_2 \dots w_n] \quad (= [W_1 W_2 \dots W_k]).\end{aligned} \tag{62}$$

Hence, by (5),

$$\begin{aligned}K_n(v_1, A) &= \{V^{(n)}z \mid z \in \mathbb{C}^N\},\\ K_n(w_1, A^T) &= \{W^{(n)}z \mid z \in \mathbb{C}^N\}.\end{aligned} \tag{63}$$

Moreover, the recursions for the v's in (6) and (7) can be rewritten in matrix formulation as follows:

$$AV^{(n)} = V^{(n)}H_n + [0 \ \dots \ 0 \ \tilde{v}_{n+1}]. \tag{64}$$

Here,

$$H_n := \begin{bmatrix} \alpha_1 & \beta_2 & 0 & \cdots & 0 \\ \gamma_2 & \alpha_2 & \beta_3 & \ddots & \vdots \\ 0 & \gamma_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_k \\ 0 & \cdots & 0 & \gamma_k & \alpha_k \end{bmatrix} \tag{65}$$

is a block-tridiagonal matrix of size $n \times n$ that also is upper Hessenberg. Furthermore, the diagonal blocks $\alpha_1, \alpha_2, \cdots, \alpha_k$ of the matrix H_n are all squares, and their sizes are just the lengths of the look-ahead steps.

Let $x_0 \in \mathbb{C}^N$ be an arbitrary initial guess for the solution of (1), and set $r_0 = b - Ax_0$. At the n^{th} iteration QMR computes an approximate solutions of (1) of the form

$$x_n \in x_0 + K_n(r_0, A). \tag{66}$$

If we choose $v_1 = r_0/\|\rho_0\|$ and any $w_1 \in \mathbb{C}^N, \|w_1\|_2 = 1$, as the starting vectors for the look-ahead Lanczos algorithm, then, by (11), the right Lanczos vectors $v_1, v_2, \cdots, v_n$ span Krylov subspace $K_n(r_0, A)$ in (14). Therefore, any iterate (14) can be presented in the form

$$x_n = x_0 + V^{(n)} z_n, \quad z_n \in \mathbb{C}^n, \tag{67}$$

and $V^{(n)}$ is the matrix defined in (10). For this approximate solution x_n the residual satisfies

$$r_n = V^{(n+1)}(f_{n+1} - H_n^{(e)} z_n). \tag{68}$$

where $f_{n+1} = [\|r_0\|_2 \ \ 0 \ \cdots \ 0]^T \in I\!R^{n+1}$, and $H_n^{(e)}$, defined by

$$H_n^{(e)} = \begin{bmatrix} H_n \\ \rho_{n+1}(e_n^{(n)})^T \end{bmatrix}, \quad (e_n^{(n)})^T = [0 \ \ldots \ 0 \ 1]^T \in I\!R^n. \tag{69}$$

Freund and Nachigal [14] suggested to choose z_n as the solution of the least-squares problem

$$\| f_{n+1} - H_n^{(e)} z_n \|_2 = min_{z \in \mathbb{C}^n} \| f_{n+1} - H_n^{(e)} z \|_2 \, . \tag{70}$$

The motivation of this choice of z_n is as follows. The vector z which minimizes (18) can be found with considerably less work than would be needed to minimize the residual norm, since the matrix V^{n+1} will not usually be unitary. More details of the QMR method can be found in [14].

We observe that in each step of a look-ahead Lanczos process, it is necessary to decide whether to construct the Lanczos vectors v_{n+1} and w_{n+1} as regular or inner vectors. As we know [14], for a regular step it is necessary that $D_k = W_k^T V_k$ is nonsingular. Therefore, in implementation of QMR in the floating-point arithmetic, the smallest singular value of matrix D_k is computed and the criterion

$$\sigma_{min}(D_k) \leq Tol$$

is used to check whether this matrix is singular or close to singular, and to decide whether to construct the Lanczos vectors v_{n+1} and w_{n+1} as regular or inner vectors. Here Tol is a suitable chosen tolerance. The efficiency of the algorithm depends on a good choice of the Tol and to construct correctly the Lanczos vectors v_{n+1} and w_{n+1}. In addition, if the quantity $\sigma_{min}(D_k)$ is badly computed, propagation of round-off errors will affect drastically all the computation. In section 4, we show that QMR algorithm with the CADNA library, and using the appropriate test, and optimal terminations, it is possible in constructing the Lanczos vectors v_{n+1} and w_{n+1} to decide correctly, to restart the QMR algorithm if it is necessary, to stop the program as soon as a satisfactory solution is reached, to estimate the accuracy of the solution, and to save computer time, because many useless operations and iterations are not performed.

In the following section we give a brief description of CESTAC method which is an efficient method for solving the numerical problems such as those described above.

3. The CESTAC method

3.1 Basic ideas of the CESTAC method

Any result R provided by a computer always contains an error resulting from round-off errors propagation. It has been proved [2] that a computed result R is modelized to the first order in 2^{-p} as the equation

$$R = r + \Sigma_{i=1}^{n} u_i(d) 2^{-p} \alpha_i,$$

where r is the exact result, α_i is the round-off error, and $u_i(d)$ are quantities depending exclusively on the data. The integer n is the number of arithmetical operations involved in the computation of R, and the integer p is the number of bits in the mantissa.

The CESTAC method (Contrôle et Estimation Stochastique des Arrondis de Calculs) was developed by La Porte and Vignes, and was then generalized by the latter. It is based on a probabilistic approach of the round-off errors propagation, it has been presented in [12, 16, 17], this method allows to estimate the round-off error on each result and consequently provides the accuracy of this result.

The basic idea of this method consists in performing the same code several times in order to propagate the round-off error differently each time. Several samples of R containing different round-off error are then obtained. The first digits common to all the samples are significant and the others are not significant and represent the round-off error propagation. The aim is then to obtain these samples of R. They are obtained by the use of random arithmetic.

Indeed, each result r of any floating-point (FP) arithmetical operator is always bounded by two consecutive FP values R^- and R^+. The random arithmetic consists in randomly choosing either R^- or R^+ with a probability 0.5. Then when a same code is executed N times with a computer using this random arithmetic, for each result of any floating-point arithmetic, N different results $R_i, i = 1, \dots, N$ will be provided. It has been proved [2, 5] that, under certain hypothesis, this N results belong to a quasi-Gaussian distribution centered on the exact result r. So, in practice, the use of the CESTAC method consists in :

(i) Running in parallel N times ($N = 2$ or 3) the program with this new arithmetic. Consequently, for each result R of any floating-point arithmetic operation, a set of N computed results $R_i, i = 1, \dots, N$ is obtained.

(ii) Taking the mean value $\bar{R} = \frac{1}{N}\Sigma_{i=1}^{N} R_i$ of the R_i as the computed result.

(iii) Using Student distribution to estimate a confidence interval for R, and then compute the number $C_{\bar{R}}$ of significant digits of $\bar{R}$, defined by $C_{\bar{R}} = \log_{10}(\sqrt{N}|\bar{R}|/\tau_\beta \sigma)$, with $\sigma^2 = \frac{1}{N-1}\Sigma_{i=1}^{N}(R_i - \bar{R})^2, \tau_\beta$ is the value of the Student distribution for $N-1$ degrees of freedom and a probability level $1 - \beta$.

3.2 Stochastic arithmetic

By using the CESTAC method so that the N runs of the computer program take place in parallel, the N results of each arithmetic operation can be considered as realisations of Gaussian random variable centered on the exact result. We can therefore define a new number, called stochastic number, and a new arithmetic, called stochastic arithmetic, applied to these numbers. We present below the main definitions and properties of this arithmetic. For more details see [6]

Definition 1. We define the set S of stochastic numbers as the set of Gaussian random variables. We denote an element $X \in S$ by $X = (\mu, \sigma^2)$, where μ is the mean value of X and σ its standard deviation. If $X \in S$ and $X = (\mu, \sigma^2)$, there exists λ_β, depending only on β, such that :

$$P(X \in [\mu - \lambda_\beta.\sigma,\ \mu + \lambda_\beta.\sigma]) = 1 - \beta.$$

$I_{\beta,X} = [\mu - \lambda_\beta.\sigma,\ \mu + \lambda_\beta.\sigma]$ is a confidence interval of μ at $(1-\beta)$. An upper bound to the number of significant digits common to μ and each element of $I_{\beta,X}$ is

$$C_{\beta,X} = \log_{10}(\frac{|\mu|}{\lambda_\beta \cdot \sigma}).$$

The following definition is the modelling of the concept of informatical zero proposed in [24].

Definition 2. $X \in S$ is a stochastic zero, denoted $\underline{0}$, if and only if :

$$C_{\beta,X} \leq 0 \quad or \quad X = (0,0).$$

Definition 3. Let $X_1 = (\mu_1, \sigma_1^2)$, $X_2 = (\mu_2, \sigma_2^2)$ be two elements of S. We define the four elementary stochastic operations denoted $(s+, s-, s*, s/)$ on the stochastic numbers by

$$\begin{aligned}
X_1\ s+\ X_2 &\overset{def}{=} (\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2),\\
X_1\ s-\ X_2 &\overset{def}{=} (\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2),\\
X_1\ s*\ X_2 &\overset{def}{=} (\mu_1 * \mu_2,\ \mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2),\\
X_1\ \ s/\ X_2 &\overset{def}{=} (\mu_1/\mu_2,\ (\tfrac{\sigma_1}{\mu_2})^2 + (\tfrac{\mu_1\sigma_2}{\mu_2^2})^2), \qquad \text{with } \mu_2 \neq 0.
\end{aligned}$$

Definition 4. Let X_1 and X_2 be two elements of S, X_1 is stochastically equal to X_2, denoted $X_1\ s=\ X_2$, if and only if

$$X_1\ s-\ X_2 = \underline{0}.$$

Definition 5. Let $X_1 = (\mu_1, \sigma_1^2)$ and $X_2 = (\mu_2, \sigma_2^2)$ be two elements of S. X_1 is stochastically strictly greater than X_2, denoted $X_1\ s>\ X_2$, if and only if :

$$\mu_1 - \mu_2 > \lambda_\beta \sqrt{\sigma_1^2 + \sigma_2^2}$$

Definition 6. Let $X_1 = (\mu_1, \sigma_1^2)$, $X_2 = (\mu_2, \sigma_2^2)$ be elements of S. X_1 is stochastically greater than X_2, denoted $X_1\ s\geq\ X_2$, if and only if :

$$X_1\ \ s>\ X_2 \quad or \quad X_1\ \ s=\ X_2$$

Based on these definitions, the following properties of stochastic arithmetic have been proved:

- $s=$ is reflexive and symmetric but is not transitive,
- $s>$ is transitive
- $s\geq$ is reflexive, anti-symmetric, but is not transitive;

$$a\ s\leq\ b \Rightarrow a\ s=\ b \ or\ a\ s<\ b$$

$$a\ s=\ b \text{ and } b\ s<\ c \Rightarrow a\ s\leq\ c$$

$$a\ s\leq\ b \text{ and } b\ s<\ c \Rightarrow a\ s\leq\ c$$

$\underline{0}$ is absorbent for operation $s*$ and is the neutral element for operation $s+$. Let $x, y \in I\!R$ and $X, Y \in S$ respectively be their representative. If $X\ s<\ Y \Rightarrow x < y$. Thus stochastic arithmetic retrieves properties of exact arithmetic, lost by usual floating-point arithmetic such as associativity, distributivity, the concept of remarkable identities.

On computer, by using the synchronous implemetation of the CESTAC method and by identifying the notions of informatical zero and stochastic zero, it is possible to use the stochastic arithmetic with its definitions. The use of stochastic arithmetic in scientific codes allows:

1. during the run of a scientific code, to estimate the accuracy of any numerical result, to detect the numerical instabilities, and to check the branchings;

2. to eliminate the programming expedients that are absolutely unfounded, such as those used, for example, in termination criteria of iterative methods, and replace them by criteria that directly reflect the mathematical condition that must be satisfied at the solution.

3.3 The CADNA library

CADNA (Control of Accuracy and Debugging for Numerical Applications) is a library for programs written in FORTRAN 77, FORTRAN 90, or in ADA which allows the computation using stochastic arithmetic by automatically implementing the CESTAC method [3, 4]. CADNA is able to estimate the accuracy of the computed results, and to detect numerical instabilities occuring during the run.To use the CADNA library, it suffices to place the instruction USE CADNA at the top of the initial FORTRAN or ADA source code and to replace the declarations of the real type by the stochastic type and to change some statements as printing statements.

During the run, as soon as a numerical anomaly (for example, appearance of informat ical zero in a computation or a criterion) occurs, a message is written in a special file called Cadna_stability_f90.lst. The user must consult this file after the program has run. If it is empty, this means the program has been run without any problem, that it has accordingly been validated, and that the results have been given with their associated accuracy. If it contains messages, the user, using the debugger associated with the compiler, will find the instructions that are the cause of these numerical anomalies, and must reflect in order to correct them if necessary. The program execution time using the CADNA library is only multiplied by a factor 3, which is perfectly acceptable in view of the major advantage offered, i.e., the validation of programs. CADNA is also able to estimate the influence of data errors on the result provided by the computer.

4. Using the CADNA library in iterative methods

4.1 Using the CADNA library in hybrid GMRES method

As we have seen in section 2.1, in implementation of hybrid GMRES method three difficulties arise. Let us first consider the third problem which concerns to the numerical instabilities which may occur in phase II. By remarking that a GMRES polynomial with unsignificant coefficients and also an unstable solution $x_{k\nu}, k = 1, 2, \ldots$ always provide an unsignificant residual norm with large magnitude, we can define with CADNA library the test

$$\text{if } \|r_{k\nu}\|_2 = \underline{0} \quad \text{and} \quad \|r_{k\nu}\|_2 \geq 1.0 \quad \text{then } index = 1, \qquad (71)$$

with initial value $index = 0$, which allows us by checking the value of parameter *index* to detect the presence of an unstable solution. As consequence, the program, during the run, can decide to perform the both phase I and II, or only the phase I according to the value of this parameter *index*.

Now, we consider the first problem which concerns to choose a stopping criterion. As explained in [24], from mathematical point of view, once we know a solution x_m, we can validate its validity by checking the value of the residual norm

$$\|r_m\|_2 = \|b - Ax_m\|_2 = 0.$$

Obviously with usual floating arithmetic this equality is never satisfied even when the solution x_m is the exact solution, because of the round-off errors propagation. However, with CADNA library in view of its properties, the result will be

$$\|r_m\|_2 = \|b - Ax_m\|_2 = \underline{0}.$$

So, as soon as the residual norm of a stable computed solution is equal to the informatical zero, a satisfactory informatical solution is reached and the iterative process must be stopped. Now, by noting that it is possible, in phase II, to have an unstable solution $x_{k\nu}, k = 1, 2, \ldots$ and we can detect it by the test (19), we define the termination criterion

$$\text{if } \|r_{k\nu}\|_2 = \underline{0} \quad \text{then stop,} \tag{72}$$

for checking the value of the residual norm of a stable solution $x_{k\nu}, k = 0, 1, \ldots$. This termination criterion stops the iterative hybrid GMRES process as soon as a satisfactory solution is reached either by phase I or by phase II. It is necessary to say that the instability of the solution $x_{k\nu}, k = 1, 2, \ldots$ must be checked before using the termination criterion (20).

Finally the second problem, which concerns the accuracy of the computed solution, will be solved by using the CADNA library. Since, as it was explained in section 3.3, the CADNA library is able to estimate the accuracy of the computed results and to furnish the results with their exact decimal figures.

Let us now, to present the examples and the results which we obtained by the FORTRAN code of hybrid GMRES method, combined with CADNA library and the above tests. Computation have been performed on a SUN4 computer in double or simple precision with the stochastic arithmetic using the CADNA library.

Example 1. We consider the linear system with

$$A = \begin{bmatrix} 2 & 1 & & & & \\ 0 & 2 & 1 & & & \\ 1 & 0 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 0 & 2 & 1 \\ & & & 1 & 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 3 \\ 4 \\ \vdots \\ 4 \\ 3 \end{bmatrix},$$

which was described in [15], and the dimension equal to 400. The exact solution is given by $x = [1, 1, \dots, 1]^T$. The results obtained by using floating-point arithmetic with $\epsilon = 10^{-6}, 10^{-15}, 10^{-16}$, and CADNA library are presented in Table 2. Only the decimal significant digits are printed. At the last lines of this table we can find the corresponding value of indices ν, k of the presented solution $x_{k\nu}$.

For this example the initial guess has been $x_0 = [0, 0, \dots, 0]^T$. From the results of Table 2, it appears that when ϵ has too large value, the iterative process is stopped too soon (at $\nu = 4, k = 11$), and the solution furnished is not the best that the computer may provide. When ϵ is chosen too small ($\epsilon = 10^{-16}$) the iterative process is stopped too late (at $\nu = 6, k = 1$), many useless iterations are performed without improving the accuracy of the solution with $\epsilon = 10^{-15}$. By using the CADNA library, the optimal termination criterion (20) has stopped the iterative process at $\nu = 4, k = 48$, and solution is reached with about 15 exact significant digits on all the elements.

	$\epsilon = 10^{-6}$	$\epsilon = 10^{-15}$	$\epsilon = 10^{-16}$	CADNA library
$x(1)$	1.0000000	1.0000000000000000	1.0000000000000000	0.100000000000000E+1
$x(2)$	0.999999	1.0000000000000000	1.0000000000000000	0.999999999999999E+0
$x(3)$	1.000000	1.000000000000000	1.000000000000000	0.100000000000000E+1
$x(4)$	1.000000	0.9999999999999999	0.9999999999999999	0.100000000000000E+1
$x(5)$	0.999999	1.0000000000000000	1.0000000000000000	0.999999999999999E+0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x(396)$	1.000000	1.0000000000000000	1.0000000000000000	0.999999999999999E+0
$x(397)$	0.9999999	1.0000000000000000	1.0000000000000000	0.100000000000000E+1
$x(398)$	0.9999999	1.000000000000000	1.000000000000000	0.100000000000000E+1
$x(399)$	1.000000	1.0000000000000000	1.0000000000000000	0.999999999999999E+0
$x(400)$	0.9999999	1.0000000000000000	1.0000000000000000	0.100000000000000E+1
	$\nu = 4$ $k = 11$	$\nu = 4$ $k = 46$	$\nu = 6$ $k = 1$	$\nu = 4$ $k = 48$

Table 2

Remark: For controling the quality of a computed solution X of a linear system $Ax = b$ we can use the normalized residuals test. As explained in [18, 19], this test consists in computing the normalized residuals

$$\rho_i^* = \frac{|\rho_i|}{2^{-p}\sqrt{m_i^q(\Sigma_{j=1}^n(AijX_j)^2 + B_i^2)}}, \qquad i = 1, \dots, n,$$

where

$$\rho_i = B_i - \Sigma_{j=1}^n A_{ij} X_j$$

and A_{ij}, B_i are the normalized floating-point representations of a_{ij} and b_i, respectively, X_j is the jth element of computed solution X, and $q = 1$ for rounding to the nearest mode, and $q = 2$ for other rounding modes. The integer m_i is the number of nonzero elements of row i, and the integer p is the number of bits in the mantissa.

The three following cases can occur:

case I: All the n normalized residuals are of the order of magnitude 1:

$$\rho_i^* \sim 1, \quad \forall i \in [1, 2, \dots, n],$$

thus the computed solution X is a satisfactory informatical solution.

case II: At least one of the normalized residuals is strictly strictly greater than one, but strictly strictly smaller than 2^p:

$$1 \ll \rho_i^* \ll 2^p,$$

thus the computed solution X is not a satisfactory informatical solution, but it is possible to improve it by an increment vector ΔX which may be obtained by solving the linear system $A\Delta X = R$, where R is the residual vector with ith element ρ_i.

case III: At least one of the normalized residuals is of the order of magnitude 2^p:

$$\rho_i^* \sim 2^p,$$

thus X is a bad solution and in general, we can not improve it. If this situation occurs, this means that the using method is not adapted to the proposed system.

By applying this test to the computed solutions of the above example, we discover that all the above computed solutions are the satisfactory informatical solutions except the one which obtained by floating-point arithmetic with $\epsilon = 10^{-6}$. By improving this solution two times, we could obtain a satisfactory informatical solution. We observe that without using CADNA library, it is very difficult to obtain a satisfactory informatical solution.

Example 2. Let us again consider the linear system (3). The solution obtained with CADNA library is as follows

$$x(1) = \underline{0}, \qquad x(2) = \underline{0},$$
$$x(3) = \underline{0}, \qquad x(4) = 0.9999999E + 000.$$

These results show that the first three elements of computed solution are non significant and the last one has 7 significant digits. We observe that only by using the CADNA library it is possible to conclude that the results obtained for the first three elements of computed solution must be due to the round-off errors propagation.

Example 3. Let us again consider the linear system (4). The solutions furnished by using single floating-point arithmetic with $\epsilon = 10^{-5}, \epsilon = 10^{-6}$, and CADNA library are presented in the Table 3. With $\epsilon = 10^{-7}$ any solution has not been obtained, because an overflow occured during the run of code. The CADNA library detected the numerical instabilities and showed that all the coefficients of GMRES polynomial presented in Table 1 are $\underline{0}$, i.e., they have no significant digit. The test of normalized residuals showed that the solution obtained by CADNA library is a satisfactory informatical solution, but those which obtained by floating-point arithmetic are not. By solving the corresponding linear systems $A\Delta X = R$ we could improve the solutions obtained with $\epsilon = 10^{-5}, \epsilon = 10^{-6}$ and obtain the satisfactory informatical solutions which are presented in Table 4. Finally, with many difficulties, we obtained the satisfactory informatical solutions, but what is the accuracy of each element of these solutions? As we observe that the CADNA library not only has obtained a satisfactory informatical solution, but also has furnished the elements of the computed solution with their exact significant digits.

	$\epsilon = 10^{-5}$	$\epsilon = 10^{-6}$	CADNA library
$x(1)$	1.6700622	1.6700689	0.167001E+01
$x(2)$	2.9999490	2.9998803	0.299997E+01
$x(3)$	-4.4399438	-4.4398289	-0.443998E+01
$x(4)$	3.9999347	3.9998538	0.399999E+01
$x(5)$	4.1099434	4.1098661	0.41099E+01
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x(146)$	-0.9999837	-0.9999635	-0.99999E+00
$x(147)$	75.1088562	75.1074753	0.75109E+02
$x(148)$	-0.9999837	-0.9999635	-0.99999E+00
$x(149)$	76.1087952	76.1073151	0.76109E+02
$x(150)$	-0.9999837	-0.9999635	-0.99999E+00
	$\nu = 7$ $k = 0$	$\nu = 11$ $k = 0$	$\nu = 12$ $k = 0$

Table 3

	$\epsilon = 10^{-5}$	$\epsilon = 10^{-6}$
$x(1)$	1.6699998	1.6700000
$x(2)$	3.0000000	3.0000000
$x(3)$	-4.4400001	-4.4400010
$x(4)$	4.0000000	4.0000000
$x(5)$	4.1100001	4.1100001
$\vdots$	$\vdots$	$\vdots$
$x(146)$	-1.0000000	-1.0000000
$x(147)$	75.1100006	75.1100006
$x(148)$	-1.0000000	-1.0000000
$x(149)$	76.1100006	76.1100006
$x(150)$	-1.0000000	-1.0000000
	$\nu = 9$ $k = 0$	$\nu = 5$ $k = 0$

Table 4

4.2 Using the CADNA library in A^TA -orthogonal s-step Orthomin(k) method

As we have seen in section 2.2, in implementation of A^TA-orthogonal s-step Orthomin(k) method two problems arise. The first one is :

- How to determine a good value of s?

Let us first consider Tables 5, 7, and 9 of the examples 4-6, respectively. These Tables present the minimum number of significant digits of the norm of orthogonal direction vectors of P_0(the first s-dimensional subspace) which are furnished by the CADNA library for differents values of s. It emerges from these results that, for each problem, this number begins to decrease from certain s. When it has a small value for some s, the large error exists at the beginning of the iterative process and can lead to the serious round-off errors, and then the slow convergence (see the results of Tables 6, 8, and 10 of the mentioned examples which represent the number of iterations to convergence for different values of s). By noting this remark, it has been observed in experiments that, for double precision, we can obtain a good value of s by taking the highest value of s for which all the orthogonal direction vectors of P_0 have the norm with at least 10 significant digits. By using the CADNA library and increasing the value of s (for example, 4 by 4), it is very easy to determine a such value of s, because the number of significant digits of the norm of orthogonal direction vectors of P_0, for each s, can be furnished by the CESTAC function which exists in this library, and returns the number of significant digits of every stochastic variable.

Now, we consider the second problem which is :

- How can the iterative process be stopped correctly?

As we mentioned in section 2.2, when we use the stopping criterion

$$\|r_i\|_2 \leq \epsilon, \tag{73}$$

it is possible, due to numerical instabilities or/and the stationarity, this stopping criterion is never satisfied. So, we need to use the additional termination criteria for stopping the process in the cases:

(i) The algorithm is stationary and can not converge.

(ii) The computer is not able to distinguish the vector r_i from the null vector and to improve the computed solution, because of the round-off errors propagation.

As explained in [24, 25], the stochastic arithmetic allows the development of two termination criteria for these cases.

In stochastic arithmetic, when the iterative process becomes stationary, that is, the difference between two iterates is nonsignificant, the components of the vector $x_i - x_{i-1}$ are stochastic zeros. So, with CADNA library which allows using stochastic arithmetic by automatically implementing the CESTAC method, and using the stopping criterion

$$\|x_i - x_{i-1}\|_1 = \underline{0}. \tag{74}$$

it is possible to stop the iterative process as soon as it becomes stationary.

In stochastic arithmetic, when the computer is not able to distinguish the vector r_i from the null vector and to improve the computed solution, because of the round-off errors propagation, the components of r_i are stochastic zeros and a satisfactory informatical solution is available. So, with CADNA library, and using the stopping criterion

$$\|r_i\|_2 = \underline{0}. \tag{75}$$

it is possible to stop the iterative process as soon as the case (ii) occurs and a satisfactory informatical solution is reached.

It is clear that, in the above cases which the iterative process is stopped by the criterion (22) or (23) before the criterion (21) is satisfied, the computed solution will not be a solution with desired accuracy ($\|r_i\|_2 \leq \epsilon$) and it is necessary to improve it by an increment vector Δx_i. For doing this, we need the classical type value of the residual vector r_i of the computed solution x_i for solving the linear system $A\Delta x_i = r_i$ by restarting the iterative process. Fortunately, with CADNA library, it suffices for obtaining the classical type value of r_i to use the **old_type** function which exists in this library, and returns the corresponding classical type value of every stochastic variable.

We observe that, with CADNA library, the criteria (22) and (23) stop the iterative process as soon as the cases (i) and (ii) occur, and make it possible to save computation time, because many useless iterations are avoided, to restart the iterative process in order to improve the satisfactory informatical solution which is furnished, and to obtain the solution with the desired accuracy. Consequently, with CADNA library and using the termination criteria (21)-(23), and including the test for restarting the process in the cases in which the process is stopped by the stopping criterion (22) or (23), we can have a stable and efficient A^TA-orthogonal s-step Orthomin(k) algorithm with the value of s furnished by the method discussed above for solving the linear system and obtaining the desired approximate solution (with $\|r_i\|_2 \leq \epsilon$).

Let us now, to present the examples and the results which we obtained by the FORTRAN code of A^TA-orthogonal s-step Orthomin(k) method, with floating-point arithmetic for different values of s, and this code with the CADNA library, and the above tests, for the value of

s furnished by the computer. Computation have been performed on a SUN4 computer in double precision. For floating-point arithmetic the stopping criterion was $\|r_i\|_2 \leq \epsilon$ and the maximum number of iterations allowed set to 1000.

Example 4.

We consider the constant-coefficient elliptic equation

$$-\Delta u + 2P_1 u_x + 2P_2 u_y = f, \tag{76}$$

which was described in [11], on the unit square $\Omega = \{(x,y) | 0 \leq x, y \leq 1\}$ with Dirichlet boundary conditions. Discretizing (24) on $n_1 \times n_1$ grid gives rise to a sparse linear system of equations of order $n = n_1^2$. By using the second order centered differences for the first derivatives and the Laplacian, the coefficient matrix has the form

$$A = \begin{bmatrix} a & d & & & e & & \\ b & a & d & & & \ddots & \\ & b & a & & & & e \\ & & & & \ddots & & \\ c & & & \ddots & \ddots & & \\ & \ddots & & & & & d \\ & & & c & & b & a \end{bmatrix}.$$

After scaling the matrix and right-hand side by $h^2 (h = 1/(n_1 + 1))$, the matrix entries are given by

$$a = 4,\ b = -(1 + p_1),\ c = -(1 + p_2),$$

$$d = -1 + p_1,\ e = -1 + p_2,$$

where $p_1 = P_1 h, p_2 = P_2 h$. In our test we take $P_1 = 0, P_2 = 50$. The grid size is $h = 1/21$, leading to a problem of size 400. The right hand side is determined so that the solution x to the discrete system is 1 everywhere. This allows an easy verification of the results. With $\epsilon = 10^{-9}$ and $x_0 = [0, \dots, 0]^T$ the results obtained are presented in Tables 5 and 6. Table 5 contains the minimum number of significant digits of the norm of orthogonal direction vectors of P_0 for different value of s. Table 6 contains the number of iterations needed to satisfy the stopping criterion (21). For CADNA library there are two numbers in this Table, the first one, denoted by TN, presents the total number of iterations needed in the different runs of the iterative process. The second number, denoted by NR, presents the number of restarting of the iterative process.

s	4	8	12	16	20	24	28	32	36	40
MIN	14	14	14	13	11	9	7	5	3	3

Table 5 : The minimum number of significant digits of the norm of orthogonal direction vectors of P_0.

	double floating-point arithmetic										CADNA library	
s	4	8	12	16	20	24	28	32	36	40	TN	NR
$k = 1$	35	27	19	11	7	6	7	9	18	22	5	1
$k = 2$	46	23	10	7	6	17	16	19	27	34	5	1
$k = 4$	52	12	8	8	8	53	57	65	88	90	5	1

Table 6 : The number of iterations to convergence

The results presented in Table 5 show that the highest value of s for which all the orthogonal direction vectors of P_0 have the norm with at least 10 significant digits is $s = 20$. Table 6 shows that the value $s = 20$ is a good value for this example. With computed value $s = 20$, and the CADNA library, the solution was reached with only 5 iterations for all the values $k = 1, 2, 4$ which is less than those needed with floating-point arithmetic for different values of s. It must be noted that the process was stopped by the stopping criterion (23) at 3th iteration and restarted for improving the computed solution for which the norm of residual was $\|r_3\|_2 = 0.484E-5$, $\|r_3\|_2 = 0.755E-6$, $\|r_3\|_2 = 0.755E-6$, for $k = 1, 2, 4$, respectively. We observe that, for this example, by using CADNA library we could determine a good value of s (s=20), and obtain the desired approximate solution (with $\|r_i\|_2 \leq 10^{-9}$) with the minimum number of iterations. So, for this example, the algorithm using CADNA library is more efficient than that using floating-point arithmetic.

Example 5.

We consider the linear system with

$$A = \begin{bmatrix} \alpha & 1 & & & \\ -1 & \alpha & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & & 1 \\ & & & -1 & \alpha \end{bmatrix}, \quad b = \begin{bmatrix} 1+\alpha \\ \alpha \\ \vdots \\ \alpha \\ \alpha - 1 \end{bmatrix},$$

which was described in [1], and the dimension equal 400. With $\alpha = 10^{-8}$, $\epsilon = 10^{-5}$ and the initial vector $x_0 = [0, \ldots, 0]^T$ the results are given in the Tables 7 and 8.

s	4	8	12	16	20	24	28	32	36	40
MIN	14	14	14	14	14	13	10	7	4	2

Table 7 : The minimum number of significant digits of the norm of orthogonal direction vectors of P_0.

	double floating-point arithmetic										CADNA library	
s	4	8	12	16	20	24	28	32	36	40	TN	NR
k=1	100	50	36	25	20	19	17	15	24	29	18	1
k=2	100	50	37	25	20	20	18	17	23	39	17	1
k=4	100	50	39	25	20	22	20	21	37	58	17	1

Table 8: The number of iterations to convergence

For this example the computed value of s is 28. From the results presented in Table 8 it clearly appears that $s = 28$ is a good value for this example. With the CADNA library, the solution was reached with 18 iterations for $k = 1$, and 17 iterations for $k = 2, 4$. The iterative process was stopped by the stopping citerion (22) at 16th iteration, and restarted in order to improve the computed solution for which the norm of residual was $\|r_{16}\|_2 = 0.285E-3$, $\|r_{16}\|_2 = 0.163E-3$, $\|r_{16}\|_2 = 0.124E-3$, for $k = 1, 2, 4$, respectively. We observe that, with CADNA library for $k = 1$ the total number of iterations to convergence is slightly greater than the smallest which corresponds to $s = 32$, and for $k = 2, 4$ this number which is equal to 17 is less than or equal to those needed for different value of s. So, for this example, with CADNA library the program is able to determine a good value of s ($s = 28$) and to furnish the desired approximate solution (with $\|r_i\|_2 \leq 10^{-5}$) with a reasonable number of iterations. Consequently, $A^T A$-orthogonal s-step Orthomin(k) performed with CADNA library is an efficient tool for solving the linear system of this example.

Example 6.

We consider the linear system with

$$A = \begin{bmatrix} 1 & & & & \alpha \\ & 2 & & & \\ & & \ddots & & \\ & & & n-1 & \\ & & & & n \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

which was described in [23, 26] and the dimension equal 100. With $\alpha = 2 \times 10^6, \epsilon = 10^{-10}$ and the initial vector $x_0 = [0. \dots, 0]^T$ the results are listed in the Tables 9, 10.

s	4	8	12	16	20	24	28	32	36	40
MIN	15	11	5	2	2	1	0	0	0	0

Table 9 : The minimum number of significant digits of the norm of orthogonal direction vectors of P_0.

	double floating-point arithmetic					CADNA library	
s	4	8	12	16	20	TN	NR
k=1	29	14	10	92	*	13	2
k=2	27	15	19	85	*	14	3
k=4	22	74	68	131	*	15	3

Table 10 : The number of iterations to convergence. $*$ = problem reached iteration count limit.

For this example, the highest value of s for which all the orthogonal direction vectors of P_0 have the norm with at least 10 significat digits is $s = 8$. The results of Table 10 show that, with CADNA library, the behavior is similar as in the example 2. It must be noted that the process has been stopped by the stopping criterion (23) two times for $k = 1$, and three times for $k = 2, 4$.

4.3 Using the CADNA library in the QMR method

As we have seen in section 2.3, in implementation of QMR in the floating-point arithmetic, the smallest singular value of matrix D_k is computed and the criterion

$$\text{if } (\sigma_{min}(D_k) \leq Tol) \text{ then go to step 3)}$$

is used to check whether this matrix is singular or close to singular, and to decide whether to construct the Lanczos vectors v_{n+1} and w_{n+1} as regular or inner vectors. Here Tol is a suitable chosen tolerance. The efficiency of the algorithm depends on a good choice of the Tol and to construct correctly the Lanczos vectors v_{n+1} and w_{n+1}, and if the quantity $\sigma_{min}(D_k)$ is badly computed, propagation of round-off errors will affect drastically all the computation. In addition, It has been shown that [14], in exact arithmetic, the stopping criterion in step 4) of algorithm 2.4 will be satisfied after at most N step -except in a very special situation. If $\rho_{n+1} = 0$ or $\xi_{n+1} = 0$ then $K_n(v_1, A)$ is A-invariant subspace or $K_n(w_1, A^T)$ is A^T-invariant subspace, respectively. In the first case x_n is the exact solution and the QMR algorithm must be stopped. In the second case, by restarting the QMR method and using the last available QMR iterate x_{n-1} (which is a good choice [14]) as the new initial guess, it is possible to improve the approximate solution. So, in floating-point arithmetic, we have to use the following criterions for stopping or restarting the QMR algorithm,

$$\text{if } (\rho_{n+1} \leq \epsilon_1 \ ) \text{ then stop,} \tag{77}$$

$$\text{if } \ (\xi_{n+1} \leq \epsilon_1) \text{ then restart,} \tag{78}$$

where ϵ_1 is a suitable chosen tolerance. It is necessary to mention that, it is possible, due to round-off errors propagation, ρ_{n+1} becomes non-significant and in the same time it has a large value $(\rho_{n+1} > \epsilon_1)$. In this situation, it is clear that, the iterations of QMR method are not able to improve the approximate solution and the QMR method must be restarted with the last available QMR iterate x_{n-1}. Finally, for QMR Algorithm a convergence criterion is needed. We can stop the QMR algorithm as soon as the criterion

$$||r_n||_2 \leq \epsilon_2 ||r_0||_2, \tag{79}$$

is satisfied, where ϵ_2 is also a suitable chosen tolerance. As we observe another problem is how to choose ϵ_1 and ϵ_2. We now face the following questions:

- How can we detect the informatical singularity of the matrix D_k and the numerical instabilities which may occur in the program?
- How can the iterative process be stopped or restarted correctly?

In order to overcome these drawbacks, we propose to introduce the stochastic arithmetic in the QMR algorithm, which is able to estimate the round-off error propagation, and to detect the informatical singularity of the matrix D_k by means the following simple test,

$$\text{if } (\sigma_{min}(D_k) = \underline{0}) \text{ then go to step 3)} \tag{80}$$

and to restart or to stop the process by the following tests,

$$\text{if } (\rho_{n+1} \leq \epsilon_1 \text{ or } C_{\rho_{n+1}} < 1\ ) \text{ then (if } ||r_n|| = \underline{0}) \text{ then stop else restart,} \tag{81}$$

$$\text{if } (\xi_{n+1} \leq \epsilon_1 \text{ or } C_{\xi_{n+1}} < 1) \text{ then (if } ||r_n|| = \underline{0}) \text{ then stop else restart,} \tag{82}$$

$$\text{if } (||r_n||_2 = \underline{0}) \text{ then stop.} \tag{83}$$

It has been observed in experiment that, for double precision, $\epsilon_1 = 10^{-8}$ is a good choice for tests (29)-(30) and we will have a stable algorithm.

Let us now, to present an example and the results which we obtained by the FORTRAN code of the QMR algorithm, with floating-point arithmetic for $Tol = 10^{-4}$ and difference value of ϵ_2; and this code with stochastic arithmetic and the test (28) and the termination criteria (29)-(31). Computation have been performed in double precision and the number of iterations allowed set to 200000.

Example 7. We consider the ill-conditioned linear system $Hx = b$ with the Hilbert matrix H $(h_{ij} = 1/(i+j-1))$, and dimension equal to 100. The right hand side is determined so that the exact solution x is 1 everywhere. This allows an easy verification of the results. With $x_0 = [0,\ 0, \cdots,\ 0]^T$, $\epsilon = 10^{-7}$, $\epsilon = 10^{-8}$, and stochastic arithmetic the results obtained are presented in Table 11. These results show that the results obtained with $\epsilon = 10^{-8}$ have not more significant digits than those obtained with $\epsilon = 10^{-7}$, but the iterative process is stopped too late (at iteration 138312). So many useless iterations are performed without improving the accuracy of the solution. It is necessary to mention that with $\epsilon = 10^{-9}$ the iterative process has not been converged. As we observe, in stochastic arithmetic, the better solution is furnished with only 6841 iterations which is very smaller than those of floating-point arithmetic with $\epsilon = 10^{-7}$ and $\epsilon = 10^{-8}$. So, we can conclude that the use of criterion (28) which detect the informatical singularity of matrix D_k allows to stabilize the algorithm, and the use of the optimal criteria (29)-(31) allow to continue the iterations, to restart QMR method if it is necessary, and to stop the program as soon as a satisfactory solution is reached.

	$\epsilon = 10^{-7}$	$\epsilon = 10^{-8}$	Stochastic arithmetic
$x(1)$	0.9999542	1.0000064	1.0000
$x(2)$	1.0007666	0.9998362	1.00
$x(3)$	0.9976219	1.0008133	0.999
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x(98)$	0.9968086	0.9986009	0.995
$x(99)$	0.9963570	0.9983446	0.995
$x(100)$	0.9958868	0.9980741	0.994
iter	19579	138312	6841

Table 11

5. Conclusion

In this paper we have seen in the implementation of iterative methods the following problems arise

- How can the iterative process be stopped correctly?
- What is the accuracy of the computed solution given by computer?
- How can detect the numerial instabilities which may occur in the program?

We observed that the use of CADNA library allows us to solve these problems. It has been shown that the CADNA library with the optimal termination criterion and the appropriate test is able to stop the program as soon as a satisfactory solution is reached, to estimate the accuracy of the solution, to detect the numerical instabilities, to prevent an overflow which may occur, and to save computer time, because many useless operations and iterations are not performed. Consequently, the mentioned iterative methods with CADNA library are the robust and efficient tools for solving large nonsymmetric systems of linear equations.

References

[1] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Compt. 12(1991)58-78.

[2] J. M. Chesneaux, *Study of the computing accuracy by using probabilistic approach,* Contribution to Computer Arithmetic and Self Validating Numerical methods, C. Ulrich ed., IMACS, New Brunswick, NJ, (1990) 19-30.

[3] J. M. Chesneaux , *CADNA, An ADA tool for round-off error analysis and for numerical debugging,* Proc. Congress on ADA in Aerospace, Barcelona,1990.

[4] J. M. Chesneaux, *Descriptif d'utilisation du logiciel CADNA_ F,* MASI Report, No 92-32 (1992).

[5] J. M. Chesneaux and J. Vignes, *Sur la robustesse de la méthode CESTAC,* C.R. Acad. Sci. Paris, Sér. I, Math. 307(1988) 855-860.

[6] J. M. Chesneaux and J. Vignes, *Les fondements de l'arithmétique stochastique,* C.R. Acad. Sci. Paris, Sér. I, Math. 315(1992) 1435-1440.

[7] A. T. Chronopoulos, *s-Step iterative methods for (non)symmetrics (in)definite linear systems*, SIAM J. Numer. Anal. 28(6)(1991) 1776-1789.

[8] A. T. Chronopoulos, and C. W. Gear, *s-Step iterative methods for symmetric linear systems*, J. comput. Appl. Math, 25(1989)153-168.

[9] A. T. Chronopoulos and C. W. Gear, *Implementation of preconditioned s-step conjugate gradient methods on a multi processor system with memory hierarchy*, Parallel comput, 11(1989)37-53.

[10] A. T. Chronopoulos, and S. K. Kim, *The s-step Orthomin and s-step GMRES implemented on parallel computers,* SIAM Conference On Iterative Methods, April 1-5, 1990, The Copper Mountain, CO; University of Minnesota, Dept of Computer Science, Tech. Report 90-15, Minneapolis, MN, 1990.

[11] H. C. Elman, *A stability analysis of incomplete LU factorizations,* J. Math Comp. 47(175)(1986) 191-217.

[12] A. Feldstein and R. Goodman, *Convergence estimates for the distribution of trailing digits,* Journal A.C.M. 23(1976).

[13] R. W. Freund, M. H. Gutknecht and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices,* SIAM J. Sci. Comput. 14(1993)137-158.

[14] R. W. Freund and N. M. Nachtigal, QMR:*a quasi-minimal residual method for non-Heritian linear systems,* Numer. Math. 60 (1991) 315-339.

[15] M. H. Gutknecht, *Variants of BICGSTAB for matrices with complex spectrum,* to appear.

[16] R. W. Hamming, *On the distribution of numbers,* The Bell System Technical Journal (1970).

[17] T. E. Hull and J. R. Swenson, *Test of probabilistic models for propagation of round- off errors,* Communication of A.C.M. 9(2) (1966).

[18] M. La Porte and J. Vignes, *Evaluation statistique des erreurs numèriques dans les calculs sur ordinateur,* Numer-Math, 23(1974) 63-72.

[19] M. La Porte and J. Vignes, *Algorithmes numériques - Analyse et mise en œuvre. vol 1. arithmétique des ordinateurs - Systèmes linéaires.* Editions Technip, Paris (1974).

[20] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?,* SIAM J. Matrix Anal. Appl. 13(3)(1992) 778-795.

[21] N. M. Nachtigal, L. Reichel, and L. N. Trefethen, *A hybrid GMRES algorithm for nonsymmetric linear systems,* SIAM J. Matrix Anal. Appl. 13(3)(1992) 796-825.

[22] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Statist. Comput. 7(3)(1986) 856-869.

[23] C. D. Swanson, and A. T. Chronopoulos , *Orthogonal s-step methods for nonsymmetric linear systems of equations,* ACM Int. Conference On Super computing, July 19-23, 1992, pp 456-464

[24] J. Vignes, *Zéro mathématique et zéro informatique,* C.R.A.S., Paris, t 303, sér 1,(1986) 997-1000 and La Vie des Sciences. 4(1)(1987) 1-13.

[25] J. Vignes, *A stochastic arithmetic for reliable scientific computation,* Math. Comp. Simul, 35(1993) 233-261.

[26] H. Walker, *Implementation of the GMRES method using Householder transformations,* SIAM J. Sci. Stat. Comp. 9(1988)152-163.

Proceedings of the Monthly Colloquium
of the Iranian Mathematical Society
Vol. 2, pp 271-281 (2004)

Local Defining Ideals of Ordinary Singularities

Rahim Zaare-Nahandi

Department of Mathematics and Computer Science
Faculty of Science, University of Tehran, Tehran, Iran
rahimzn@khayam.ut.ac.ir

Dedicated to the late Professor Karim Seddighi

Abstract: We review some basics on the theory of generic singularities in algebraic geometry. We state certain approaches towards the explicit local equations of such singularities. The main point of view is the constructive approach in commutative algebra and algebraic geometry.

1. Introduction

One of the goals in algebraic geometry is the "local classification" of algebraic varieties. Namely, if X is an algebraic variety over an algebraically closed field k and $\mathcal{O}_{X,x}$ is the local ring of regular functions at x (i.e., the ring of "germs" of rational functions on a neighborhood

MSC (2000): 14B05, 14B25
Keywords: Generic singularities, Gröbner basis, local defining ideal.

of x), how can one classify these rings up to isomorphism? This is a far-reaching problem. The "local analytic classification" is the classification of $\widehat{\mathcal{O}}_{X,x}$, the completion of $\mathcal{O}_{X,x}$. The Cohen structure theorem states that if X is smooth at x, then $\widehat{\mathcal{O}}_{X,x} = k[[x_1, \cdots, x_r]]$, the ring of formal power series in r variables where r =dim$X$. Such a classification for $\mathcal{O}_{X,x}$ with X smooth at x, does not exist.

The theory of ordinary singularities is an approach toward the classification of $\widehat{\mathcal{O}}_{X,x}$ when X is singular at x. Ordinary singularities arise from "generic" projection of smooth varieties. More precisely, let $Y \subset \mathbb{P}^n$ be a projective smooth variety of dimension r in the n-dimensional projective space. Let E be a point outside Y. Let $\mathbb{P}^{n-1} \subset \mathbb{P}^n$ be the projective space of dimension $n-1$ and let $\pi : Y \longrightarrow \mathbb{P}^{n-1}$ be the projection from E. If E is in general position, then $Y_1 = \pi(Y)$ has "nice" singularities. We may repeat this procedure of projecting for Y_1, and continue as long as the projection is a birational map onto its image. The singularities of the resulting variety X are called "ordinary singularities". A "node" is the only ordinary singularity of curves, and in this case

$$\widehat{\mathcal{O}}_{X,x} = k[[x_1, x_2]]/(x_1 x_2).$$

The ordinary singularities of surfaces are also fairly simple. They consist of ordinary double points, normal (triple) crossings, and pinch points, where the completion of the local rings are isomorphic to

$$k[[x,y,z]]/(xy), \quad k[[x,y,z]]/(xyz), \quad k[[x,y,z]]/(x^2 z - y^2),$$

respectively. However, the ordinary singularities of higher dimensional varieties are more complicated. Intuitively, ordinary singularities are "stable", and hence they are appropriate models for approximations of critical points of natural phenomena.

Many authors such as K. Mount, O. Villamayor, J. Roberts, J. Lluis and A. Holme have given substantial contributions to the subject. A reasonably complete "parametric" representation of ordinary singularities is now available. However, to apply results from "the ideal theory of the ring of polynomials over a field", one needs the explicit equations of these ordinary singularities. For ordinary "double points", this is rather simple. Indeed, if x is an ordinary double point, then

$$\widehat{\mathcal{O}}_{X,x} = k[[x_1, \cdots, x_n]]/(x_1, \cdots, x_p)(x_{p+1}, \cdots, x_{2p}),$$

or,

$$\widehat{\mathcal{O}}_{X,x} = k[[z, z_1, \cdots, z_n, u_1, \cdots, u_n]]/I,$$

where,

$$I = \text{the ideal of } 2\times 2\text{-minors of } \begin{pmatrix} z_1 & u_1 & z_2 & u_2 & . & . & . & z_n & u_n \\ zu_1 & z_1 & zu_2 & z_2 & . & . & . & zu_n & z_n \end{pmatrix},$$

(See [Z] and [ZZ]).

In the first case, the normalization is the product of two complete regular local rings generalizing a node, while in the second case, the normalization is itself a complete regular local ring generalizing a pinch point.

Using the above classification of the ordinary double points, the quotient forms for the classification of triple points follow from [R1] and [SZ]. Here the goal is to see some picture of the progress on the quotient forms of the completions of local rings at ordinary singularities of higher "multiplicities". We will omit the proofs.

2. Ordinary Singularities with Higher Multiplicities

For some time, in collaboration with P. Salmon, we have been trying to provide a nice presentation for ordinary singularities of higher multiplicities as a quotient of the polynomial ring by an explicit ideal. Let A be the completion of the local ring at a point x on a variety X. Let B be the normalization of A. It is well-known that B is a finite A-module (e.g., see[S]). Let q be the length of B as an A-module. The positive integer q is called the multiplicity of X at x or simply the multiplicity of x. Roughly speaking, a point of multiplicity q is a point at which q "simple branches" of the variety intersect, or, it is a "limiting position" of such points. If B is a domain, the point x is called a "unibranched" point of multiplicity q. Using these notions, we concentrate on unibranched ordinary singularities of multiplicity q. The parametric equations of these points are given in [R2] as:

$$z = yt + xt^2 + \cdots + vt^{q-2} + t^q,$$

$$z_i = y_it + x_it^2 + \cdots + u_it^{q-1}; i = 1, \cdots, m.$$

In order to obtain the local defining ideal of the variety at such a point, we need all relations obtained by "eliminating" t in the above equations. Consider the "resultant matrix":

$$N = \begin{bmatrix} N_0 \\ N_1 \\ \cdot \\ \cdot \\ \cdot \\ N_n \end{bmatrix},$$

where,

$$N_0 = \begin{bmatrix} 1 & 0 & v & \cdot & \cdot & \cdot & y & -z & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & v & \cdot & \cdot & \cdot & y & -z & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & 0 & v & \cdot & \cdot & \cdot & y & -z \end{bmatrix},$$

is a $(q-1) \times (2q-1)$ matrix, and

$$N_i = \begin{bmatrix} u_i & v_i & \cdot & \cdot & \cdot & y_i & -z_i & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & u_i & v_i & \cdot & \cdot & \cdot & y_i & -z_i & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & u_i & v_i & \cdot & \cdot & \cdot & y_i & -z_i \end{bmatrix}.$$

is a $q \times (2q-1)$ matrix.

Conjecture 1. (January 1989) The local defining ideal of the unibranched ordinary singularity with multiplicity q is generated by the maximal minors of N.

Lemma 1. To prove Conjecture 1, it is enough to prove it for the case $y = x = \cdots v = 0$. In this case the maximal minors of N are the same as the maximal minors of

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \cdot \\ \cdot \\ \cdot \\ M_m \end{bmatrix}$$

where

$$M_i = \begin{bmatrix} z_i & zu_i & zv_i & . & . & . & zy_i \\ y_i & z_i & zu_i & . & . & . & zxi \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ u_i & v_i & . & . & . & y_i & z_i \end{bmatrix}.$$

for $i = 1, \cdots, m$ (we are using z_i for $-z_i$ for simplicity).

Remark 1. The concept of resultant matrix first appeared in [SZ]. This generalizes the usual Sylvester resultant of two polynomials in one variable and is in general a larger ideal compared to the classical "u-resultant" (e.g., see [W1]), and hence is more useful from the scheme theoretical point of view.

Remark 2. The set of maximal minors of M is not a minimal generating set for this ideal. Indeed, the number of maximal minors of M is $\begin{pmatrix} mq \\ q \end{pmatrix}$, which is by far larger than m^q, the expected number for a minimal generating set.

Remark 3. Putting $z = 1$, the matrix M_i becomes a circulant matrix (see [D]). In the sense that each row is a permutation of the first row by a cycle of length q. Assigning certain degrees to the variables as in the next section, M_i will be the homogenization of the this circulant matrix.

3. An Approach via Gröbner Algebra.

Gröbner bases are certain tools for computations in the ring of polynomials. To distinguish the "initial monomial" of a polynomial as the largest monomial appearing in the polynomial, a total order is given on the monomials. A generating set of an ideal whose initials generate the ideal of initials of the given ideal, is called a Gröbner basis of the ideal. The initials of a Grobner basis is a set of monomials which can best reflect the given ideal.

Let's reorder the matrix M as

$$M = \begin{bmatrix} z_m & zu_m & zv_m & \cdot & \cdot & \cdot & zy_m \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ z_1 & zu_1 & zv_1 & \cdot & \cdot & \cdot & zy_1 \\ y_m & z_m & zu_m & \cdot & \cdot & \cdot & zx_m \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ y_1 & z_1 & zu_1 & \cdot & \cdot & \cdot & zx_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_m & v_m & \cdot & \cdot & \cdot & y_m & z_m \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_1 & v_1 & \cdot & \cdot & \cdot & y_1 & z_1 \end{bmatrix}.$$

Let $\deg u_i = 1$, $\deg v_i = 2, \cdots, \deg y_i = q-1$, $\deg z_i = \deg z = q$. Consider the "degree reverse lexicographic" order on the monomials induced by

$$z_m > z_{m-1} > \cdots > z_1 > z > y_m > \cdots > y_1 > x_m > \cdots > x_1 > u_m > \cdots > u_1.$$

Definition 1. A maximal minor of M is called *principal* if no entry on its main diagonal is a multiple of z.

Lemma 2. The ideal of maximal minors of M is generated by the set of principal maximal minors of M.

Lemma 3. The initial monomial of a principal maximal minor of M with respect o the above monomila order, is the product of the entries on the main diagonal of the minor.

Conjecture 2. (August 1997) The set of principal minors of M forms a Gröbner basis for the ideal of maximal minors of M.

Theorem 1. Conjecture 2 implies Conjecture 1.

Definition 2. A principal minor is called a *distinguished principal minor* (DP), if the entries on the main diagonal are non-decreasing from the top to the bottom.

Conjecture 3. The set of DP minors of M generates the ideal of maximal minors of M.

Theorem 2. Conjecture 3 implies Conjecture 2.

Lemma 4. The number of DP minors of M is m^q.

Conjecture 3 has lead to certain basic determinantal identities which do not seem to be well-known. Let's state just one of such identities.

Theorem 3. Let

$$D = \begin{bmatrix} L_1 \\ L_2 \\ \cdot \\ \cdot \\ \cdot \\ L_n \end{bmatrix}$$

be a $n \times n$ matrix, $\theta \in S_n$ any permutation with r fixed points. Then

$$\begin{vmatrix} \theta(L_1) \\ L_2 \\ \cdot \\ \cdot \\ \cdot \\ L_n \end{vmatrix} + \begin{vmatrix} L_1 \\ \theta(L_2) \\ \cdot \\ \cdot \\ \cdot \\ L_n \end{vmatrix} + \cdots + \begin{vmatrix} L_1 \\ L_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta(L_n) \end{vmatrix} = r\,|D|\,.$$

4. The “episode” of Elimination Theory!

The concept of resultant has been a central one in the elimination theory heavily studied in the 18th and the 19th centuries mostly with constructive methods. In the early part of the 20th century, many of the leading mathematicians in algebraic geometry and commutative algebra, became less interested in elimination theory and its classical methods. A. Weil in his influential book “Foundation of Algebraic Geometry, 1946” ([W], page 31), right before quoting an important result from C. Chevalley, says, “The device that follows $\cdots$ it may be hoped, finally eliminates from algebraic geometry the last traces of Elimination Theory $\cdots$”. A section of the famous book by B.L. van der Wearden, “Modern Algebra 1953” [W1], is devoted to elimination theory. In the later editions (1959-) [W2], the name “Algebra” is adopted for the book, and van der Wearden eliminates the section on elimination theory! As a result, a simple constructive proof of the existence of a resultant system essentially due to Kronecker, is replaced by a non-constructive proof using the nontrivial Hilbert's Nullstellensatz in the second volume of the book. In the seventies of this century, the tide starts to turn. This is reflected in S. Abhyankar's famous poem of 1970's:

Eliminate, eliminate, eliminate,
Eliminate the eliminators of elimination theory
⋮

Today, with the enormous influence of computers in mathematics, the use of constructive methods has been inevitable [E]. In commutative algebra and algebraic geometry, the basic notion for such constructive methods, has been the notion of "Gröbner basis". We have been trying to put some of the constructive methods of elimination theory in the frame work of the theory of Gröbner bases.

References

[D] P.J. Davis, *Circulant Matrices*, John Wiley & Sons, New York, 1979.

[E] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Springer-Verlag, 1994.

[H] A. Holme, Formal embedding and projection theorems, *Amer. J. Math.*, **95** (1972).

[L] E. Lluis, De las singularidades que aparecen al proyectar variedades algebricas, *Boletin de la Sociedad Matematica Mexicana*, ser. 2, vol. 1 (1956), 1-9.

[MV] K. Mount and O. Villamayor, An algebraic construction of the generic singularities of Boardman-Thom, *Inst. Hautes Etudes Sci. Publ. Math.* No **43**, (1974), 205-244.

[R1] J. Roberts, Generic projections of algebraic geometry, *Amer. J. Math.*, **93** (1971).

[R2] J. Roberts, Singularity subschemes and generic projections, *Trans. Amer. Math. Soc.***212**, (1975), 229-268.

[SZ] P. Salmon and R. Zaare-Nahandi, Algebraic properties of some analytically irreducible triple points, *Rend. Sem. Mat. Univ. Politec. Torino*,**49**,1 (1991), 41-70.

[S] I. Shafarevich, *Basic Algebraic Geometry*, Springer-Verlag, 1977.

[W] A. Weil, *Foundation of Algebraic Geometry*, A.M.S. Colloq. Pub. 29, 1946.

[W1] B.L. van der Wearden, *Modern Algebra*, Fred. Ungar Co., New York, 1953.

[W2] B.L. van der Wearden, *Algebra*, Fred. Ungar Co., New York, 1971.

[Z] R. Zaare-Nahandi, Seminormality of certain generic projections, *Comp. Math.* **52**, (1984), 245-274.

[ZZ] R. Zaare-Nahandi and R. Zaare-Nahandi Jr., Gröbner basis and free resolution of the ideal of 2-minors of a $2 \times n$ matrix of linear forms, *Comm. Algebra* **28**(9), (2000), 4433-4453.

پیشگفتار

زندگی علمی و اجتماعی انجمن ریاضی ایران از همایش ریاضیدانان در سال ۱۳۴۹ آغاز شده و رشد و تکامل آن با تشکیل و ادامهٔ همایش‌های گوناگون دمساز بوده است: کنفرانس سالانهٔ ریاضی ایران که سی و چهارمین آن شهریورماه ۱۳۸۲ در شاهرود برگزار شد، سمینار هفتگی انجمن که چند سال نخست دایر بود و سمینارهای هفتگی دانشگاه‌ها جای آن را گرفت، چندین سمینار تخصصی که با سمینار جبر اهواز در دی‌ماه ۱۳۵۶ آغاز شد و هر از چند گاهی یک سمینار تخصصی نو ظاهر گشت. بدون شک با مشارکت در همایش‌ها امکان گفت و شنود و تبادل نظر در زمینه‌های مختلف فراهم و در همین اثنا بذرهایی کاشته می‌شوند که امید می‌رود در آینده درختان تنومندی در زمینهٔ اندیشهٔ ریاضی به بار آورند. از این نظر می‌توان صرف نیروی انسانی و منابع مالی در راه برگزاری همایش‌ها را توجیه کرد. اما اگر این افکار مدوّن و منتشر نشوند، دامنهٔ برد آنها از حیث زمان و مکان بسیار محدود خواهند ماند. متأسفانه راهی برای تدوین و نشر گفت و شنودهای جنبی کنفرانس متصور نیست مگر قسمت‌های بسیار اندکی که ممکن است در خاطرات بازتاب یابند. اما مباحث اصلی کنفرانس یعنی مقالات ارائه شده را نه تنها می‌توان، بلکه باید به هر قیمت و زحمتی پخش کرد.

به طور کلی انتشارات ادواری و غیر ادواری انجمن‌های علمی نیز جزء ارکان مهم فعالیت‌های آنها هستند. انجمن ریاضی ایران خوشبختانه علاوه بر گزارش کنفرانس‌های سالانه و سمینارهای تخصصی، چندین نشریهٔ ادواری دارد که هر یک دارای منشور و رسالت ویژهٔ خود است: بولتن انجمن ریاضی ایران ویژهٔ مقالات پژوهشی در سطح بین‌المللی به زبان انگلیسی، فرهنگ و اندیشهٔ ریاضی ویژهٔ مقالات مروری به زبان فارسی، خبرنامه و گزارش ویژهٔ اطلاع‌رسانی و اظهار نظرها.

از سال ۱۳۷۸ همایش‌های ماهانهٔ انجمن ریاضی ایران تشکیل گردید که هدف آن آشنا ساختن پژوهشگران جوان با رشته‌های پژوهشی گوناگون است. شیوهٔ کار این همایش آن است که در کمیتهٔ همایش نسبت به دعوت از استادان مقیم داخل و خارج که در زمینه‌ای پیشرفت قابل ملاحظه‌ای داشته و کارهای پژوهشی را هدایت نموده باشند دعوت به عمل آید که مقاله‌ای مروری ارائه دهند و آن را به زبان فارسی یا انگلیسی در اختیار انجمن قرار دهند تا به شکل شایسته منتشر شود. کمیتهٔ اول در سال ۱۳۷۸ عمدتاً در شهر کتاب هفت همایش جالب برگزار کرد. کمیتهٔ دوم در سال ۱۳۷۹ همایش را در دانشگاه‌های مختلف،

در سال ۱۳۸۰ آن را در تالار فرهنگسرای دانشجو و دانشگاه تهران، و در سال ۱۳۸۱ همه را در دانشگاه تهران برگزار نمود.

از آنجا که ویرایش مقالات رسیدهٔ سال اول بیش از حد طول کشید و آخر سر هم امر ویرایش ناقص ماند، کمیته تصمیم گرفت در مسألهٔ ویرایش تجدید نظر کند و به اتکای آن که مقالات را استادان زبردست می‌نویسند، از ویرایش علمی و ادبی آنها به کلی صرف نظر نماید. بدین ترتیب، نویسندگانِ مقالات نهایت جدیت را مبذول فرمودند که مقاله به شکل نهایی و آمادهٔ تکثیر باشد. به علاوه، کمال مطلوب آن است که مقاله چندین روز پیش از برگزاری همایش به دست کمیته برسد تا به تعداد اندکی تکثیر شود و در اختیار شرکت کنندگان قرار گیرد. اگر این کمال مطلوب حاصل شود، در پایان هر سال مجلدی مشتمل بر مقالاتِ همان سال برای چاپ و انتشار آماده می‌شود و در اختیار عموم قرار می‌گیرد.

جلد اول هشت مقاله را در بر گرفت که مشتمل بر مقالات فارسی دریافت شده بود. جلد دوم، یازده مقالهٔ رسیده به زبان انگلیسی را در بر می‌گیرد که آن نیز حاوی مقالاتی از سال ۱۳۷۸ تا بهار ۱۳۸۱ است. برخی از مقالات ارائه شده در طول ۴ سال گذشته به دست کمیته نرسیده است.

از طرف خود و همکارانم در کمیتهٔ همایش، از استادان بزرگواری که دعوت ما را پذیرفتند و به نگارش و تحویل مقالهٔ خود پرداختند متواضعانه سپاسگزارم. همچنین لازم می‌دانم از حمایت‌های جناب آقای دکتر مهدی بهزاد رئیس سابق و جناب آقای دکتر سیدعبادالله محمودیان رئیس فعلی انجمن ریاضی ایران و همچنین از اعضای محترم شورای اجرایی و همکارانم در کمیتهٔ همایش ماهانهٔ انجمن خانم‌ها دکتر: شیوا زمانی و مژگان محمودی و آقایان دکتر: علی آبکار، غلامحسین اسلام‌زاده، علی ایرانمنش، حسین حاج‌ابوالحسن و سیامک یاسمی سپاسگزاری کنم.

اجرای مراحل مختلف همایش مدیون زحمات اعضای فعّال دبیرخانهٔ انجمن است. بدین وسیله از همهٔ این عزیزان به ویژه از تلاش‌های خانم افسانه بختیاری و آقای مرتضی عبدی‌زاده که ظرف چند سال گذشته در برگزاریِ همایش‌ها کمیته را یاری نمودند و خانم فریده صمدیان به خاطرِ حروف‌چینیِ بعضی مقالات و پردازش نهاییِ متن و یکنواختی آنها تشکر می‌کنم. سازماندهی و نظم کارهای دبیرخانه از زمان تصدی آقای منصور شکوهی به ریاست دبیرخانه در تحقّق این انتشار مفید و مؤثر بوده است. از حسن ادارهٔ ایشان قدرشناسی می‌کنم.

در خاتمه، پیروزی اعضای جدید کمیتهٔ همایش ماهانه آقای دکتر علی آبکار، خانم دکتر لیلا خاتمی، آقای دکتر مهدی دهقان (رئیس کمیته)، آقای دکتر عمید رسولیان، آقای دکتر بهروز طایفه رضایی و خانم دکتر مژگان محمودی را خواستارم.

ارسلان شادمان
تهران، بهار ۱۳۸۳

بسم‌الله الرحمن الرحیم

گزارش همایش ماهانه انجمن ریاضی ایران

جلد دوم، بهار ۸۳

اعضای کمیتهٔ همایش ماهانه:

علی آبکار (۱۳۸۰_۸۱)
غلامحسین اسلام‌زاده (۱۳۷۹_۸۱)
مسعود امینی (۱۳۷۸)
علی ایرانمنش (۱۳۷۸_۸۱)
حسین حاج‌ابوالحسن (۱۳۷۸)
آرش رستگار (۱۳۷۸)
شیوا زمانی (۱۳۷۸)
ارسلان شادمان (۱۳۷۹_۸۱)
مژگان محمودی (۱۳۷۹_۸۱)
سیامک یاسمی (۱۳۷۸)

حروفچینی: LaTeX ـ فریده صمدیان
چاپ و صحافی: چاپ بهمن ـ قم
ناشر: بهمن‌آرا
شابک: ۸_۳_۹۵۲۶۸_۹۶۴
نشانی: تهران، صندوق پستی ۴۱۸_۱۳۱۴۵
انجمن ریاضی ایران
تلفن و نمابر: ۸۸۰۸۸۵۵، ۸۸۰۷۷۹۵، ۸۸۰۷۷۷۵
نشانی الکترونیک: iranmath@ims.ir

همایش ماهانهٔ انجمن ریاضی ایران به‌عنوان یکی از فعالیتهای مستمر انجمن از ۱۳۷۸ با اهداف زیر شروع شده است:

نقل از آیین‌نامهٔ مصوب ۷۸/۱/۲۶

- آشنایی جامعه ریاضی ایران با شاخه‌های مختلف ریاضیات.
- آشنایی دانشجویان تحصیلات تکمیلی علوم ریاضی با مسائل و زمینه‌های مختلف تحقیقاتی ریاضی در ایران.
- ایجاد ارتباط بین محققان و دانشجویان ریاضی کشور، به‌ویژه دانشجویان تحصیلات تکمیلی رشته‌های مختلف علوم ریاضی.
- کمک به شناخت توان تحقیقاتی کشور در ریاضیات.
- کمک به ایجاد جوّ مساعد برای رشد ریاضیات در ایران و باروری استعدادهای ریاضی.

مقاله‌های همایش معمولاً مروری تحقیقی هستند و به معرفی یک رشتهٔ فعال پژوهشی می‌پردازند. این مقالات توسط یکی از استادان مدعو داخل یا مقیم خارج ارائه می‌شوند. جهت تسریع در انتشار مقالات، کمیتهٔ همایش از ویرایش آنها معذور است. از این رو، درخواست می‌شود نویسنده مقاله را به شکل نهایی تحویل دهد. چاپ مقاله‌ها در مجلات علمی‌ـترویجی به همان شکل یا با اندکی تغییر بلامانع است. در این صورت توصیه می‌شود نویسندهٔ محترم به جلد و صفحات گزارش همایش ماهانه که مقاله در آن مندرج است اشاره فرماید.

از هر مقاله ۲۵ نسخه به مؤلف تقدیم می‌شود.

زمان برگزاری همایش معمولاً آخرین چهارشنبه هر ماه جز مرداد و اسفند و با رعایت ماه مبارک رمضان است.

Proceedings of the Monthly Colloquium

of the Iranian Mathematical Society

Volume 2 (2004)

Edited by: Arsalan Chademan

CONTENTS

ISBN 964-95268-3-8