TECHNICAL UNIVERSITY OF CIVIL ENGINEERING BUCHAREST

PROCEEDINGS OF THE 15TH WORKSHOP ON MATHEMATICS, COMPUTER SCIENCE AND TECHNICAL EDUCATION

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

VOLUME 1/2018

Bucharest, June 9, 2018

ISSN 2601-9299
ISSN-L 2601-9299

CONTENT

ABSTRACTS SECTION

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 1-4

ON THE EXTENSION OF ABEL DIRICHLET CRITERION

Ileana Bucur
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: bucurileana@yahoo.com

Abstract: We give a variant of Abel – Dirichlet convergence criterion for the case of complex numbers or more generally for the case where the terms of series are functions on some set.
Mathematics Subject Classification (2010): 30B10, 30B50,
Key words: series, uniform convergence, boundary behaviour.

Let us consider two bounded sequences of complex numbers $(a_n)_n$, $(b_n)_n$. We are interesting in the convergence of the series $\sum_n a_n b_n$. It is well known the Abel – Dirichlet criterion where the sequence $(b_n)_n$ is a monotone sequence of real numbers converging to zero.

Theorem 1. *Let $(a_n)_n$, $(b_n)_n$ be two sequences of complex numbers and for any $n \in \mathbb{N}$ let us denote by A_p the partial sum of order p of the series $\sum a_n$*

$$A_p = a_1 + a_2 + ... + a_p .$$

If the sequence $(A_p)_p$ is bounded, if the series $\sum_i |b_{i+1} - b_i|$ is convergent and the sequence $(A_n b_n)_n$ is convergent then the series $\sum a_n b_n$ is convergent.

Proof. It is easy to verify that for any $p, q \in \mathbb{N}$ we have

$$\sum_{i=p}^{p+q} a_i b_i = \sum_{n=p}^{p+q-1} A_n (b_n - b_{n+1}) + A_{p+q} \cdot b_{p+q} - A_{p-1} \cdot b_p =$$

$$= \sum_{n=p-1}^{p+q-1} A_n (b_n - b_{n+1}) + A_{p+q} \cdot b_{p+q} - A_{p-1} \cdot b_{p-1} .$$

Let $M \in \mathbb{R}_+$ such that $|A_n| \le M$ for any $n \in \mathbb{N}$, let $\varepsilon \in \mathbb{R}$ $\varepsilon > 0$ and let $\mathbb{N}_\varepsilon \in \mathbb{N}$ be such that

$$\sum_{n=p-1}^{p+q-1} |b_n - b_{n+1}| < \frac{\varepsilon}{2M}, \ |A_{p+q} \cdot b_{p+q} - A_{p-1} \cdot b_{p-1}| < \frac{\varepsilon}{2}, \quad \forall p - 1 \ge N_\varepsilon, \ \forall q \in \mathbb{N} .$$

We deduce

$$\left|\sum_{i=p}^{p+z} a_i b_i\right| \le \sum_{n=p-1}^{p+q-1} |A_n||b_n - b_{n+1}| + \left|A_{p+q}\cdot b_{p+q} - A_{p-1}\cdot b_{p-1}\right| \le \varepsilon$$

for all $p \ge N_\varepsilon + 1$ and all $q \in \mathbb{N}$, i.e. the series $\sum_i a_i b_i$ is convergent.

Further we shall do some consideration on the necessity of conditions from hypothesis in order to have the same conclusion.

Proposition 2. *Let $(a_n)_n$ be a sequence of complex numbers such that for any sequence $(b_n)_n$ of complex numbers converging to zero and such that the series $\sum_n |b_{n+1} - b_n|$ is convergent, the series $\sum_n a_n b_n$ is convergent. Then the sequence $(A_n)_n$ of partial sums $A_n := \sum_{i\le n} a_i$ is bounded.*

Proof. We suppose the contrary and we consider a subsequence $(A_{n_k})_k$ of the sequence $(A_n)_n$ such that

$$\left|A_{n_{k+1}} - A_{n_k}\right| = \left|\sum_{i=n_k+1}^{n_{k+1}} a_i\right| > k$$

(The construction of the increasing sequence $(n_k)_{k\in\mathbb{N}}$ of natural numbers with the above properties may by done inductively).

We consider now the decreasing sequence $(b_n)_n$ of real number given by

$$b_1 = b_2 = b_3 = \ldots = b_{n_1} = 1,\ b_{n_1+1} = b_{n_1+2} = \ldots = b_{n_2} = \frac{1}{2}, \ldots$$

$$b_{n_k+1} = b_{n_k+2} = \ldots = b_{n_{k+1}} = \frac{1}{k+1}, \ldots .$$

Obviously the sequence $(b_n)_n$ converges to zero and the series

$\sum |b_{n+1} - b_n| = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \ldots + \left(\frac{1}{k} - \frac{1}{k+1}\right) + \ldots$ is convergent. On the other hand we have

$$\left|\sum_{i=n_k+1}^{n_{k+1}} a_i b_i\right| = \frac{1}{k+1}\cdot\left|\sum_{i=n_k+1}^{n_{k+1}} a_i\right| > \frac{k}{k+1} \quad \forall k$$

and therefore the series $\sum_i a_i b_i$ is divergent (which contradicts) fact that does not verify the hypothesis.

Proposition 3. *Let $(a_n)_n$ be a sequence of complex numbers such that for any sequence $(b_n)_n$ of complex numbers for which the series $\sum_n |(b_{n+1} - b_n)|$ is convergent, the series $\sum_n a_n b_n$ is convergent. Then the series $\sum_i a_i$ is convergent.*

Proof. The sequence $\left(b_n^{'}\right)_n$, $\left(b_n^{''}\right)_n$ given by $b_n^{''}=\frac{1}{n}$, $b_n^{'}=1+\frac{1}{n}$ are convergent to zero, respectively to 1 and the series $\sum_n\left|b_{n+1}^{'}-b_n^{'}\right|$, $\sum_n\left|b_{n+1}^{''}-b_n^{''}\right|$ are convergent. Using the hypothesis we get the convergence of the series $\sum_n a_n b_n^{''}$, respectively $\sum_n a_n b_n^{'}=\sum_n a_n\left(1+b_n^{''}\right)$. Hence the series $\sum_n a_n$ is also convergent.

Theorem 2. *Let $\left(a_n\right)_n$ be a sequence of complex numbers such that the series $\sum a_n$ is convergent and let $\left(f_n\right)_n$ be a sequence of complex, bounded functions on a set such that the function $t\to\sum_n\left|f_{n+1}(t)-f_n(t)\right|$ is bounded on T. Then the series $\sum_n a_n f_n$ of functions on T is uniformly convergent on T.*

Proof. For any $n\in\mathbb{N}$ we denote $r_n:=\sum_{i\geq n}a_i$. We have

$$\sum_{n=p}^{p+q}a_n f_n=\sum_{n=p}^{p+q}(r_n-r_{n+1})f_n=\sum_{n=p+1}^{p+q}r_n\left(f_n-f_{n-1}\right)+r_p f_p-r_{p+q+1}f_{p+q}\,.$$

Using the hypotheses we deduce the fact that the sequence $\left(f_n\right)_n$ of complex functions on T is uniformly bounded i.e. there exists $M\in\mathbb{R}_+$ such that

$$\|f_n\|:=\sup\left\{\left|f_n(t)\right|;t\in T\right\}\leq M\,,\ \forall n\in\mathbb{N},\sum_n\left|f_n(t)-f_{n+1}(t)\right|\leq M\,,\ \forall t\in T\,.$$

Since the series $\sum_n a_n$ is convergent we deduce that choosing a real number $\varepsilon,\varepsilon>0$ there exists a natural number N_ε such that $|r_n|<\frac{\varepsilon}{3M}$, for all $n\in\mathbb{N}$, $n\geq N_\varepsilon$. Hence we have $p\geq\mathbb{N}_\varepsilon$,

$$t\in T\Rightarrow\left|\sum_{n=P}^{P+q}a_n f_n(t)\right|\leq\sum_{n=p+1}^{p+q}|r_n|\cdot\left|f_n(t)-f_{n-1}(t)\right|+\left|r_p\right|\cdot\left|f_p(t)\right|+\left|r_{p+q+1}\cdot f_{p+q}(t)\right|<\varepsilon$$

for all $q\in\mathbb{N}$ i.e. the series of functions $\sum_n a_n f_n$ is uniformly convergent.

Applications

1. *Let $\sum_n a_n z^n$ a power series with $a_n\in\mathbb{C}$ and let R be the radius of convergence of this series*

$$R=\frac{1}{\overline{\lim_{n\to\infty}}\sqrt[n]{|a_n|}}\,.$$

We suppose that for some $z_0 = \mathrm{Re}^{i\theta_0}$ *the series* $\sum_n a_n z_0^n$ *is convergent and for any positive number* $r, r < R$ *we denote by* $\Omega_{Z_0 r}$ *the convex covering of the set*

$$\left\{ z \in \mathbb{C} \middle| |z| \le r \right\} \cup \{ z_0 \}.$$

In these conditions the series $\sum_n a_n z^n$ *is uniformly convergent on the set* $\Omega_{Z_0 r}$.

2. *Abel – continuity theorem. If the conditions of the preceding statement 1 are fulfilled then the sum functions* $s(z) = \sum_{n=0}^{\infty} a_n z^n$ *is continuous on the set* $\Omega_{Z_0 r}$.

3. *Let* $(a_n)_n$ *be a sequence of complex numbers such that the series* $\sum a_n$ *is convergent.*

For any number $r < 1$ *we consider the set* $\Omega_r(B)$ *of all linear bounded operators on the Banach space* B *given by*

$$\Omega_r(B) = \left\{ \alpha I + (1-\alpha) A \middle| \alpha \in [0,1], I \text{ the identity on } B, A \text{ linear}, \|A\| \le r \right\}.$$

Then the series $\sum_n a_n T^n$, $T \in \Omega_r(B)$ *is uniformely convergent on the set* $\Omega_r(B)$.

References

[1] Boboc, N.: *Analiza matematica I*, Fundamentum Editura Univ. Bucureşti, Bucuresti, 1999.
[2] Bourbaki, N.: *Fonctions d'une variable réele (Théorie élementaire),* Paris, Herman, 1961.
[3] Dieudonné, J.: *Eléments d'analyse I Gautier – Vilard*, Paris, 1969.
[4] Nicolescu, M., Dinculeanu, N., Marcus, S.: *Analiza Matematica I, II*, Editura Didactica si Pedagogica, Bucuresti, 1971.
[5] Rudin, W.: *Principles of Mathematical Analysis,* Mc. Graw – Hill, New York, 1964.

A NOTE ON GIVENS ROTATIONS, QR DECOMPOSITIONS AND SOLVING OF LINEAR SYSTEMS

Dan Caragheorgheopol
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
and
"Ilie Murgulescu" Institute of Physical Chemistry of the Romanian Academy
Splaiul Independentei 202, RO-060021 Bucharest, Romania
E-mail: dancaraghe@gmail.com

Abstract In numerical linear algebra, Givens rotations, just like Householder reflections, are essential building blocks, used for solving a variety of problems ranging from least the square solution for overdetermined systems of equations to the eigenvalue problem and more. Many of these problems solutions also involve a QR factorization. While Householder reflections can effectively annihilate all the components of a vector, except for the first one, Givens rotations can be used to zero elements in a more selective manner. Hence their advantage in cases of matrices possessing a special structure, like tridiagonal matrices, Hessenberg matrices, etc.
We discuss here possible implementations of Givens rotations with their advantages and disadvantages. We also discuss their use in a number of scenarios and focus on the problem of solving linear systems using the QR decomposition based on Givens rotations . We propose optimized algorithms pertaining to this problem. We also include the Mathcad code for the proposed algorithms.
Mathematics Subject Classification (2010): 97N40, 65F05.
Key words: QR decomposition, QR algorithm, Givens rotations, orthogonal transformations, Mathcad.

1. Introduction

In linear algebra and numerical linear algebra, Givens rotations are a well-known family of orthogonal transformations in the *n*- dimensional space $\mathbb{R}^n$ [3,5]. As the name suggests, they are plane rotations. More specifically, the Givens transformation:

$$G(i,j,\theta)=\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & \iddots & \\ & & \cos\theta & \cdots & \sin\theta & & \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ & & -\sin\theta & \cdots & \cos\theta & & \\ & \iddots & & & & \ddots & \\ 0 & & & \cdots & & & 1 \end{pmatrix}$$

is a plane rotation of angle θ in the (i, j) plane (where the values $\cos\theta$ and $\sin\theta$ are placed on the *i*-th and *j*-th lines and columns of the above matrix). When applied to a vector $f \in \mathbb{R}^n$, such a rotation will only modify the *i*-th and *j*-th elements of *f*. By choosing an appropriate angle θ, we can annihilate one of these components of *f*, as we shall see in the following section.

Furthermore, by successively applying such rotations $G_1, G_2, ..., G_n$ to a $n \times n$ matrix A (i.e., to all of its columns) in a carefully chosen order, one can annihilate, e.g., all the sub-diagonal elements of the matrix *A*. We then have $G \cdot A = R$, where *R* is an upper triangular

matrix and $G = G_n \cdot \ldots G_2 \cdot G_1$ is an orthogonal matrix. If we denote $G^T = Q$, we find that $A = Q \cdot R$, with Q orthogonal and R upper triangular, thus we have a so-called QR decomposition of *A* [3,5].

The QR decomposition of a matrix is useful in many ways. Two important scenarios where such a decomposition is used are:

(1) for *solving linear systems* of large dimension.

If the system to be solved is written in matrix form $Ax = b$, with *A* a $n \times n$ real matrix and $b \in \mathbb{R}^n$ and $A = Q \cdot R$ is a QR decomposition, then we have $Q^T Ax = Q^T b$ and thus $Rx = Q^T b$, which is an upper triangular system that can be solved by back substitution method in $O(n^2)$ floating point operations (*flops*).

(2) for *computing the eigenvalues and eigenvectors* of a matrix, by the QR algorithm [3,5]

In this paper we will analyze the efficient use of Givens rotations for the first case scenario and discuss issues concerning their optimal implementation from a computational point of view. We also show the Mathcad [6] implementation of the proposed algorithms.

2. Givens rotations: mathematical formulas and implementation problems

Givens rotations have been studied and their mathematical properties have been comprehensively discussed, e.g. in [5]. Since they are rotations that operate in a plane, they are best understood at the 2×2 dimension. The conclusions can be easily transferred to the general case.

Let $G(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ be a rotation of angle θ and $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \in \mathbb{R}^2$ be non-zero. Let us also denote by *c* and *s* the cosine and sine of θ, respectively. Then we have:

$$G(\theta) \cdot f = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} f_1' \\ f_2' \end{pmatrix}.$$

It is easy to see that, in order to annihilate f_2', one can choose *c* and *s* according to the following formulas:

$$c = \frac{f_1}{\sqrt{f_1^2 + f_2^2}} \tag{1}$$

$$s = \frac{f_2}{\sqrt{f_1^2 + f_2^2}}, \tag{2}$$

which also ensures that $c^2 + s^2 = 1$. The above formulas in fact mean that we take the angle θ to be the angle of the vector *f* with the *Ox* axis. However, it is worth noticing that we never actually need to compute the θ angle, but we only need the values of *c* and *s*, thus avoiding the computation of inverse trigonometric functions.

As simple as formulas (1) and (2) may seem, when trying to implement an algorithm for a QR decomposition of a $n \times n$ matrix *A* using Givens rotations, we find there are a few problems that must be addressed:

(i) Firstly, we must take care to avoid overflow or underflow which may occur if formulas (1) and (2) are used.

(ii) With every rotation, a square root must be computed. This is a "costly" operation that can be avoided by using the so-called fast Givens rotations. Details on the fast Givens algorithm may be found in, e.g., [3]. This variant of Givens rotations will not be considered here.

(iii) It is not at all convenient, for large values of *n*, to store and update at each step the matrix $G = G_n \cdot \ldots G_2 \cdot G_1 = Q^T$. It is much more efficient to store the two values *c* and *s* that define each rotation and recover *G* (or *Q*) when and *if needed*. However, with each rotation only one position is zeroed in the matrix *A*. It would be therefore very convenient, as pointed out by Stewart [4], to only store *one* number instead of two (*c* and *s*) for each rotation. This number could then be stored in the position that was zeroed by that rotation.

In the following section we discuss possible approaches to *(i)* and *(iii)* as well as their drawbacks and try to find the optimal algorithm for the scenario of using QR-decomposition for solving large linear systems (possibly having a special form, such as Hessenberg form[3]).

3. Using Givens rotations for solving linear systems. An optimized implementation

Let us begin by addressing the overflow/underflow problem. The following algorithm, implemented as a Mathcad function, resemble the few variants that can be found in the literature (see, e.g.,[2,3]).

```
Givens(f) := | if f_2 = 0
             |   c ← 1
             |   s ← 0
             | otherwise
             |   if |f_2| > |f_1|
             |     u ← f_1 / f_2
             |     s ← sign(f_2) / sqrt(1 + u^2)
             |     c ← s·u
             |   otherwise
             |     u ← f_2 / f_1
             |     c ← sign(f_1) / sqrt(1 + u^2)
             |     s ← c·u
             |   v ← (c)
             |       (s)
             | v
```

Figure 3. 1

The **Givens** function in Figure 3.1 first checks if the second component of the vector *f* is not already zero, case in which no further action is needed ($c = 1$ and $s = 0$ means that the resulting rotation is in fact the identity – corresponding to a rotation of angle 0).

Then, we check which component of *f* is larger in absolute value and simplify the fractions in (1) and (2) by it, before computing *c* and *s*. Denoting by *u* the ratio of the smaller component of *f* to the larger one, we have $u \in [-1,1]$ and thus $\sqrt{1+u^2} \in [1,\sqrt{2}]$, which prevents the overflow/underflow.

Moving on to the problem *(iii)* of storing just one number for each rotation, denoted in the following by ρ, we will use the technique proposed by Stewart [4]. The following Table 3.2 shows the two way conversion from (c,s) to ρ and vice-versa.

Compute ρ from known (c,s)	Recover (c,s) from stored ρ
$\rho = 1$ if $c = 0$	$c = 0,\ s = 1$ if $\rho = 1$
$\rho = \frac{1}{2} sign(c) \cdot s$ if $\lvert s\rvert < \lvert c\rvert$	$s = 2\rho,\ c = \sqrt{1-s^2}$ if $\lvert\rho\rvert < 1$
$\rho = \frac{2\,sign(s)}{c}$ if $\lvert c\rvert \le \lvert s\rvert$	$c = \frac{2}{\rho},\ s = \sqrt{1-c^2}$ if $\lvert\rho\rvert > 1$

Table 3.2

The idea is to store the smaller of $c/2$ or $s/2$, but to distinguish the two cases we store $2/c$ instead of $c/2$. This way we have $\rho = \frac{s}{2} < 1$ in one case and $\rho = \frac{2}{c} > 1$ in the other, while $\rho = 1$ is reserved for when $c = 0$ and $\frac{2}{c}$ is not defined. Following are the Mathcad functions for storing *c* and *s* as a single number ρ (Figure 3.2) and for restoring *c* and *s* from ρ (Figure 3.3).

```
econostore(v) := | ρ ← 1  if  v1 = 0
                 | otherwise
                 |     ρ ← (1/2)·sign(v1)·v2  if  |v2| < |v1|
                 |     ρ ← 2·sign(v2)/v1       if  |v1| ≤ |v2|
                 | ρ
```

Figure 3.2

```
econorestore(ρ) := | if  ρ = 1
                   |     c ← 0
                   |     s ← 1
                   | if  |ρ| < 1
                   |     s ← 2ρ
                   |     c ← sqrt(1 − s^2)
                   | if  |ρ| > 1
                   |     c ← 2/ρ
                   |     s ← sqrt(1 − c^2)
                   | (c)
                   | (s)
```

Figure 3.3

A problem, however, remains, when storing just ρ instead of *c* and *s*, namely the problem of *the sign of the vector* $\begin{pmatrix} c \\ s \end{pmatrix}$, which cannot be restored by the above procedure from just the number ρ. Indeed, it takes just a glance at Table 3.2 to notice that, when $|s| < |c|$, the recovered values of *c* and *s* correspond to the cosine and sine of an angle θ in the quadrants I or IV, while in the opposite case, the angle θ in the quadrants I or II. As remarked by Edward [2], this method of storing and then recovering the values of *c* and *s* from ρ creates a discontinuity for the function that maps the pair (f_1, f_2) to (c, s), along the line $f_1 = -f_2$. This discontinuity can lead to situations where a small perturbation of the vector $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$ induce a change of sign in the $\begin{pmatrix} c \\ s \end{pmatrix}$ vector thus leading, e.g., to an eigenvector of opposite sign, when Givens rotations are used in the QR algorithm to compute eigenvalues and eigenvectors (see [2]).

Also, while it is true that the Givens rotation corresponding to $\begin{pmatrix} c \\ s \end{pmatrix}$ creates a zero in the matrix *A* of a linear system just as effectively as the rotation corresponding to $\begin{pmatrix} -c \\ -s \end{pmatrix}$, one must be careful when applying a sequence of such rotations to transform *A* into an upper triangular matrix *R* for the purpose of solving the linear system $Ax = b$. Indeed, the exact same rotations (including the sign of *c* and *s*) must be applied in the same order to the vector *b*, to obtain $Q^T b$, i.e. the right hand side of the transformed system $Q^T Ax = Q^T b$ or, equivalently, $Rx = Q^T b$. This leaves us with two applicable options: *(a)* we apply each rotation to both *A* and *b* and give up the storage of the corresponding ρ (we could store ρ after applying the rotation, but, when trying to use it again, we may end up with $\begin{pmatrix} -c \\ -s \end{pmatrix}$ instead of $\begin{pmatrix} c \\ s \end{pmatrix}$); *(b)* for

each rotation we first compute ρ, then restore $\begin{pmatrix} c \\ s \end{pmatrix}$ (which may differ from the original by its sign) and then apply the rotation to both *A* and *b*. Let us remark that, while option *(a)* requires fewer operations, option *(b)* has the merit of storing for each rotation a copy of the corresponding ρ which, *if used later, will provide the same vector* $\begin{pmatrix} c \\ s \end{pmatrix}$ that was used on *A* and *b*. We show the Mathcad programs for the two options *(a)* and *(b)* in Figures 3.4 and 3.5 respectively.

```
QRSystem(A,b) := | n ← cols(A)
                 | for j ∈ 1..n − 1
                 |   for i ∈ n, n − 1..j + 1
                 |     v ← Givens((Aj,j, Ai,j))
                 |     A ← multiplickcs(v, A, j, i)
                 |     b ← multiplycs(v, b, j, i)
                 | (A  b)
```

Figure 3.4

```
QRSystemStore(A,b) := | n ← cols(A)
                      | for j ∈ 1..n − 1
                      |   for i ∈ n, n − 1..j + 1
                      |     v ← Givens((Aj,j, Ai,j))
                      |     r ← econostore(v)
                      |     v ← econorestore(r)
                      |     A ← multiplickcs(v, A, j, i)
                      |     b ← multiplycs(v, b, j, i)
                      |     Ai,j ← r
                      | (A  b)
```

Figure 3.5

Let us remark, that it is much more efficient to apply each rotation to the vector *b* (which only modifies 2 of its components, at each step) than it would be to compute and store the matrix $Q^T = G = G_n \cdot \ldots G_2 \cdot G_1$ and then compute $Q^T b$. Also, when applying each rotation (which is a rank 2 modification of the identity) to the linear system's matrix A, we in fact modify only two rows of the matrix (namely the rows *i* and *j* when we apply the $G(i, j, \theta)$ rotation). The Mathcad functions **multiplycs** and **multiplickcs** shown in Figures 3.6 and 3.7 perform the update of the matrix A after a $G(i, j, \theta)$ rotation is applied, taking advantage of this fact. The input vector *v* is in fact $v = \begin{pmatrix} c \\ s \end{pmatrix}$, therefore the reader should interpret v_1 as c and v_2 as s.

```
multiplycs(v,A,i,j) := | n ← cols(A)
                       | for k ∈ 1..n
                       |   u ← Ai,k
                       |   Ai,k ← v1·Ai,k + v2·Aj,k
                       |   Aj,k ← −v2·u + v1·Aj,k
                       | A
```

Figure 3.6

```
multiplickcs(v,A,i,j) := | n ← cols(A)
                         | for k ∈ i..n
                         |   u ← Ai,k
                         |   Ai,k ← v1·Ai,k + v2·Aj,k
                         |   Aj,k ← −v2·u + v1·Aj,k
                         | A
```

Figure 3.7

The difference between the two functions **multiplycs** and **multiplickcs** is that the second one performs the update on the elements on rows *i* and *j* only starting from column *i*, instead of the whole rows. This serves two purposes: *i)* if, as in case *(b)* discussed above, we store the

number ρ for each rotation in the place of the annihilated element $A_{i,j}$, then these numbers *must not* be modified by subsequent rotations applied to *A*; *ii)* if, as in case *(a)*, we do not store anything in place of the zeroed elements of *A*, we still have no need to update those elements, because they continue to be equal to zero after applying subsequent rotations; this happens because of the (carefully chosen) order in which we annihilate the elements under the diagonal of *A*, starting from the bottom of the first column and going up, then doing the same with the following columns in order.

Let us show now the result of running the two Mathcad programs **QRSystem** and **QRSystemStore** on an example.

Example 3.1 Let $A=\begin{pmatrix}1&2&3&4\\5&1&6&7\\8&9&1&11\\12&13&14&1\end{pmatrix}$ and $b=\begin{pmatrix}1\\2\\1\\2\end{pmatrix}$. In Figures 3.8 and 3.9 we can see that the upper triangular part of matrices *A1* and *A2* (diagonal elements and above) coincides, except for the fact that in matrix *A2*, that is produced by **QRSystemStore**, the signs of elements on rows 2 and 3 are opposite to the signs of the same elements in *A1*, that is produced by **QRSystem**. The same thing happens to rows 2 and 3 in *b2* compared to the ones in *b1*. This shows that in both cases we will get the same final solution, despite the fact that, when using rotations with the $\begin{pmatrix}c\\s\end{pmatrix}$ vector recovered from a stored value of ρ (i.e., when using **QRSystemStore**), sometimes we get the opposite of the original $\begin{pmatrix}c\\s\end{pmatrix}$ vector.

$$A1 := QRSystem(A,b)_{1,1} = \begin{pmatrix}15.297&15.362&13.663&9.087\\0&4.358&-1.582&-2.89\\0&0&7.268&-6.841\\0&0&0&-7.02\end{pmatrix}$$

$$b1 := QRSystem(A,b)_{1,2} = \begin{pmatrix}2.811\\-0.96\\0.561\\-0.929\end{pmatrix}$$

Figure 3.8

$$A2 := QRSystemStore(A,b)_{1,1} = \begin{pmatrix}15.297&15.362&13.663&9.087\\0.163&-4.358&1.582&2.89\\0.277&-0.027&-7.268&6.841\\24.083&0.105&0.146&-7.02\end{pmatrix}$$

$$b2 := QRSystemStore(A,b)_{1,2} = \begin{pmatrix}2.811\\0.96\\-0.561\\-0.929\end{pmatrix}$$

Figure 3.9

On the other hand, under the diagonal, the two matrices *A1* and *A2* differ, as expected. While *A1* has only zeros under diagonal, the matrix *A2* stores in each under-diagonal position the corresponding number ρ for the rotation that annihilated that element.

Let us now evaluate how well the proposed algorithms perform in terms of floating point operations (*flops*). When computing the number of flops involved in program **QRSystem** and **QRSystemStore**, one finds approximately $n^3+O(n^2)$ flops in both cases. This is similar to the flops number found for solving linear systems using the QR method based on Householder reflections (see [3], [1]). However, the QR method based on Givens rotations has the advantage of flexibility, in the sense that we can use Givens rotations to annihilate any element we choose from the system's matrix *A*, which allows us to take advantage of cases when *A* has a special form, such as, e.g., Hessenberg form, or tridiagonal form. We illustrate this statement by showing, in Figure 3.10, how the function **QRSystem** can be modified for the

case when *A* is Hessenberg. The number of flops for the **QRSystemHess** function is $3n^2 + O(n)$, which represent a significant improvement over the general case.

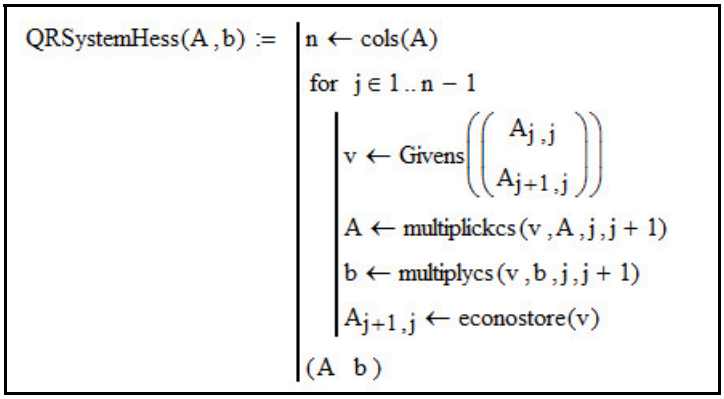


Figure 3.10

4. Conclusion

Givens rotations can be used to obtain the QR decomposition of a matrix, thus solving linear systems. This method, just like the method that uses Householder reflections, has the advantage of performing only orthogonal transformations, which gives both methods excellent numerical stability. Moreover, Givens rotations are simple plane rotations. Every such rotation can be stored as one single number, in the place of the matrix element that it annihilates. For general matrices, the QR method using Givens rotations involves a similar flops number to the one using Householder reflections, but for matrices possessing special structures, such as, Hessenberg form, tridiagonal form or sparse matrices, the flops number can be greatly improved.

References

[1] Caragheorgheopol, D.: A discussion on QR decomposition algorithms, *The 14-th Workshop of Scientific Communications*, Department of Mathematics and Computer Science, T.U.C.E.B., 2017.

[2] Edward Anderson: *Discontinuous Plane Rotations and the Symmetric Eigenvalue Problem*, LAPACK Working Note. University of Tennessee at Knoxville and Oak Ridge National Laboratory, 2000.

[3] Golub, G.H. and Van Loan, C.F.: *Matrix computations (3rd edition)*, The Johns Hopkins University Press, 1996.

[4] Stewart, G.W.: The Economical Storage of Plane Rotations, *Numer. Math.* **25**, 1976.

[5] Wilkinson, J.H.: *The Algebraic Eigenvalue Problem*, Claredon Press, 1965.

[6] Mathcad 14 Help, Parametric Technology Corporation, 2007 (http://www.ptc.com).

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 12-17

BAYESIAN INFERENCE TO DERIVE THE PROBABILITY OF EXCEEDANCE OF THE FLOODS MAXIMUM DISCHARGES

Daniel Ciuiu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
Romanian Institute for Economic Forecasting, Bucharest
E-mail: dciuiu@yahoo.com
Radu Drobot
Department of Hydraulic Constructions,
Technical University of Civil Engineering, Bucharest, Romania
E-mail: drobot@utcb.ro

Abstract The design of hydraulic structures like spillways, dykes or diversions is based on the probability of exceedance of the maximum discharges during flood events. The registred data are usually reflecting current events, which correspond to empirical probabilities of exceedance in the range 2-5%. Exceptional events, corresponding to a return period of 100 years or even more (meaning probabilities of exceedance less than 1%) could occur in the period of registered data. The main problem is the real probability of exceedance is not known for the outliers, and the values of the statistical parameters are influenced by the empirical probability which is assigned to extreme values. In order to derive the "real" return period, the Bayesian inference is used. The proposed distribution are Gamma 2 and Pareto with two parameters. We will also use the Gibbs algorithm in order to simulate the value of one parameter depending on the other parameter and on the sample values, according to the posterior distribution. Together with each simulated parameter, we simulate also a random variable according the above Gamma 2, respectively Pareto distribution, in order to estimate the exceedance probabilities.
Mathematics Subject Classification (2010): 62F15, 62C12.
Key words: Bayesian inference, Pareto, Gamma, discharges, volumes.

1. Introduction

In the case of the Bayesian inference, the current distribution deppending on one or several parameters is considered a conditional distribution: the involved parameters are considered random variables.

Because the (possible vectorial) parameter $\theta \in \Theta$ is a continuous random variable, we consider first [7,5,2] a probability density function for θ, denoted by $g(\theta)$.

This pdf is called the prior pdf of θ. The posterior pdf of θ is [7,5]

$$g\left(\theta_0 \middle| X = x_i\right) = \frac{g(\theta_0) \cdot P\left(X = x_i \middle| \theta = \theta_0\right)}{\int\limits_{\Theta} g(\theta) \cdot P\left(X = x_i \middle| \theta\right) d\theta} \tag{1}$$

if X is a discrete random variable, respectively

$$g\left(\theta_0 \middle| x\right) = \frac{g(\theta_0) \cdot f\left(x \middle| \theta_0\right)}{\int\limits_{\Theta} g(\theta) \cdot f\left(x \middle| \theta\right) d\theta} \tag{1'}$$

if X is a continuous random variable with the conditional pdf $f(x/\theta)$.

The above posterior distribution of the parameter $\theta|X_1,...,X_n$ is used for parameter estimation. The first estimator is the estimator-mode, which is the value of θ such that the posterior pdf has the maximum value. It results that this estimator is the mode of the posterior distribution.

If we consider the case of continuous prior distribution, it results that the log-likelihood is

$$\ln g\left(\theta|X_1,...,X_n\right)=\ln C+\sum_{i=1}^{n}\ln p\left(X_i=x_i|\theta\right)+\ln g\left(\theta\right) \tag{2}$$

in the discrete case, respectively

$$\ln g\left(\theta|X_1,...,X_n\right)=\ln C+\sum_{i=1}^{n}\ln f\left(x_i|\theta\right)+\ln g\left(\theta\right) \tag{2'}$$

in the continuous case. We notice that in the case of non-informative prior distribution, the estimator according the maximum likelihood method is that obtained using classical maximum likelihood method. Otherwise, to the terms (initial, or after derivatives) we add the correction term from the prior distribution.

The other two Bayesian estimators minimize the expected loss according the above posterior distribution [7,5,2]. More exactly, if the loss function is the square of distance between the estimator and the parameter, we obtain the estimator-expectation, and if the loss function is the absolute value of the mentioned distance we obtain the estimator median (i.e. the expectation and the median of the posterior distribution). The estimator-mode is not obtained from an expected loss function as the other two Bayesian estimators, but it is a limit of such Bayesian estimators [7].

Instead of the whole sample we use generally [2,5,7] a statistics, and consequently the posterior θ is conditioned on this statistics instead on the above whole sample. This statistics is in the continuous case of expectation of normal distribution and of exponential distribution the sample mean $\overline{X}$, and in the discrete case of Bernoulli, binomial, geometric and Poisson distributions the sample sum (the sum of values) [5,2].

The Gibbs algorithm consider in the case of vectorial parameter the posterior distribution of each component θ_i conditioned on the sample (or on the above statistics) and the other components [3,1,8]. At each step the parameter θ is simulated by the Monte Carlo methods [11], and for estimations there are considered only the values generated at the last n_2 iterations (the first n_1 iterations are used only to change the posterior distribution - after each simulation of a component, the last posterior distribution becomes prior distribution). An analogous separation of expectation and variance in the case of normal distribution is made in [4] (in the classical approach there is used the mixed bivariate posterior distribution, and from it the marginal distributions).

The Chauvenet test for normal distribution is presented in [10]. First of all we compute

$$z=\frac{0.435-0.862\cdot a}{1-3.604\cdot a+3.213\cdot a^2}\text{, where} \tag{3}$$

$$a=\frac{2\cdot n-1}{4\cdot n} \tag{3'}$$

and n is the size of the sample. The outliers X_i are the values such that

$$\left|X_i-\overline{X}\right|>z\cdot\sigma\text{,} \tag{4}$$

where $\overline{X}$ and σ^2 are the sample expectation and the sample variance.

2. Methodology

For Gamma distribution, in the case of fixed value of *a* we consider, as in the exponential case prior $\Gamma(c,d)$ distribution for $\lambda=\frac{1}{b}$. It results that

$$\lambda|\overline{X},a\sim\Gamma\left(n\cdot a+c,\frac{d}{n\cdot d\cdot\overline{X}+1}\right). \tag{5}$$

Of course, in the noninformative case, we take $c=1$ and $d\to\infty$, and the second parameter of the posterior Gamma distribution becomes $\frac{1}{n\cdot\overline{X}}$. If we consider now fixed value of *b*, it results that $\frac{\overline{X}}{b}$ is distributed $\Gamma\left(n\cdot a,\frac{1}{n}\right)$. This distribution is proportional to $\frac{\overline{X}^{n\cdot a-1}}{\Gamma(n\cdot a)}$. In order to have after each iteration for an average between data and prior information, consider the prior distribution as

$$g(a)\propto\frac{\alpha^{n\cdot a-1}}{\Gamma(n\cdot a)}, \tag{6}$$

and the posterior distribution

$$g\left(a|\overline{X},b\right)\propto\frac{\beta^{n\cdot a-1}}{\Gamma(n\cdot a)}, \tag{6'}$$

where $\beta=\sqrt{\frac{\alpha\cdot\overline{X}}{b}}$.

Therefore, considering first prior values for *a* and *c* one and $d\to\infty$ (noninformative), we generate first λ according the posterior distribution (3), we take $b=\frac{1}{\lambda}$, and finally we generate *a* as follows. First we notice that if $n\cdot a$ is integer in (4) and (4'), the values of pdfs are proportional to the probabilities value of Poisson distribution of parameter α, respectively β. Therefore we generate first such Poisson distribution as in [11]: the first integer value such the sum of corresponding exponential random variables is greater than one. In this way we have obtained the integer part of $n\cdot a$, denoted by k. For the fractionary part, we generate 100 uniform random variables in $[0,1]$ and we order them increasing. Next we estimate by the Monte Carlo methods

$$A=\int_0^1\frac{x^{k+t-1}}{\Gamma(k+t)}dt\,. \tag{7}$$

Finally, we generate an uniform random variable $U\in[0,1]$, and the fractionary part of $n\cdot a$ is the last value in the above increasing sequence such that the sum of $\frac{x^{k+U_i-1}}{\Gamma(k+U_i)}$ is less than $U\cdot A$.

We repeat the generation of *a* and *b* for n_1 times without tacking into account these values for estimations, and for n_2 times for estimations.

In the case of Pareto parameters, we take into account first [9] that the three parameters Pareto cdf is

$$F(x)=1-\left(1-\frac{a(x-c)}{b}\right)^{\frac{1}{a}}. \tag{8}$$

For the limit case $a\to0$ we obtain $X-c\sim\exp\left(\frac{1}{b}\right)$. Consider now $c=0$. For applying the Gibbs algorithm, we take into account that [9] if $Y\sim\exp(1)$ we have

$$X = \frac{b\left(1-e^{-a\cdot Y}\right)}{a} \sim Pareto\left(a,b\right). \tag{9}$$

From here we obtain

$$\frac{\ln b - \ln\left(b - a\cdot X\right)}{a} = Y \sim \exp\left(1\right), \tag{9'}$$

and finally

$$\frac{\ln b}{a} - \frac{\overline{\ln\left(b - a\cdot X\right)}}{a} \sim E_n\left(1\right). \tag{9''}$$

Therefore, for one of the two parameters fixed, we simulate $Y \sim E_n\left(\frac{1}{n}\right)$ as average of n exponential, and we solve in the other parameter the equation

$$\frac{\ln b}{a} - \frac{\overline{\ln\left(b - a\cdot X\right)}}{a} = Y \tag{10}$$

by the Newton method.

For applying the Chauvenet test for non-normal distributions, we take into account that if $X \sim N\left(m,\sigma^2\right)$ then $\frac{X-m}{\sigma}$ is standard normal distributed. Applying the standard normal cdf Φ, it results that the outliers X_i are characterized by

$$1-\Phi\left(z\right) \le \Phi\left(\frac{X_i - \overline{X}}{\sigma}\right) \le \Phi\left(z\right). \tag{11}$$

For a non-normal distribution we replace above $\Phi\left(\frac{X_i - \overline{X}}{\sigma}\right)$ by $F\left(X_i\right)$, where F is the considered non-normal cdf.

3. Applications

Example 1 *Consider the values of maximum discharges and volumes of Prut River at Rădăuţi gauging station. We have yearly data from 1970 to 2008 (39 years). As distributions we consider log-Gamma and log-Pareto distributions (the logarithms of data are Gamma, respectively Pareto).*

The results are presented, for logarithms, in the following table.

Table 1: The estimated parameters *a* and *b* for logarithms of data, Gamma and Pareto distributions, Prut River

Method	Estimator	Gamma		Pareto	
		a	*b*	*a*	*b*
Moments' method		113.20709	0.05963	56.10375	385.4543
Gibbs algorithm	Expectation	113.03972	0.05837	55.30752	386.04413
	Mode	109.78543	0.05892	54.31241	380.12421
	median	112.89132	0.06028	55.32142	385.36123

For outliers we take into account that $n = 38$, resulting $z = 0.49342$, $\Phi\left(z\right) = 0.99268$ and

$1-\Phi(z)=0.00732$, the last two values being both the limits for cdf and for the probabilities of exceedance. In the case of log-Gamma distribution we obtain two outliers: the minimum value 163 (log value 5.09375, exceedance 0.99767 by moments' method, 0.99541 using the estimator-expectation, 0.99123 using estimator-mode and 0.99816 using estimator-median) and maximum value 4240 (log value 8.35232, exceedance 0.00867 by moments' method, 0.00415 using the estimator-expectation, 0.00234 using estimator-mode and 0.01117 using estimator-median).

Example 2 *Consider the values of maximum discharges and volumes of Bistrita River at Frumosu gauging station. We have also yearly data from 1978 to 2015 (38 years). As distributions we consider now Gamma and Pareto distributions.*

The results are presented in the following table.

Table 2: The estimated parameters *a* and *b* for the data, Gamma and Pareto distributions, Bistrita River

Method	Estimator	Gamma		Pareto	
		a	*b*	*a*	*b*
Moments' method		3.70752	178.61279	1.35377	1558.6846
Gibbs algorithm	Expectation	3.93023	180.87314	1.30213	1557.90133
	Mode	4.01452	175.43121	1.32153	1555.97543
	Median	3.75124	177.95642	1.34261	1557.995543

For outliers we take into account that $n=39$, resulting $z=0.49359$, $\Phi(z)=0.99285$ and $1-\Phi(z)=0.00715$. In the case of Pareto distributions, the outliers are the last two values, 1170 and 1980. For 1170 the exceedance probabilities are 0.00535 for estimator-expectation, 0.00512 for estimator-mode, and less than $5\cdot10^{-6}$ for moments' method and estimator-median. For the maximum value, 1980, we obtain exceedance probability less than $5\cdot10^{-6}$ in all cases. In the case of Gamma distribution, only the maximum value, 1980 is outlier: the exceedance probability is 0.00315 by moments' method, 0.00469 using estimator-expectation, 0.00403 using estimator-mode and 0.00324 using estimator-median.

4. Conclusions

The used statistics such that the posterior distribution of a component of vectorial parameter θ conditioned by the other parameters and the sample is in fact the classical sufficient statistics: multiplying the pdfs' product relation from the definition of sufficient statistics by the prior pdf of θ, in the formula of Bayes the conditioned distribution of the sample on the sufficient statistics does not depend on θ. Therefore, this pdf goes out of the integral from denominator, and θ_i conditioned by the sample and the other parameters is identical distributed as θ_i conditioned by the considered statistics and the other parameters.

References

[1] Blake, A. and Mumtaz, H.: *Applied Bayesian econometrics for central banks*, Technical Book, Center for Central Banks Studies, 2012.
[2] Carlin, B. and Louis, T.: *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman

& Hall/ CRC, London, 2000.
[3] Casella, G. and Edward, I.G.: Explaining the Gibbs Sampler, *The American Statistician*, **Vol. 46, No. 3** (1992), 167-174.
[4] Ciuiu, D.: Bayes Signification Tests in Linear Regression and Economic applications, *Scientific Journal Mathematical Modelling in Civil Engineering*, **Vol. 9, No. 1** (2013), 16-29.
[5] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B.: *Bayesian Data Analysis*, Chapman & Hall/ CRC, Boca Raton, London, New York, Washington, 2000.
[6] Lindley, D.V. and Smith, A.F.M.: Bayes Estimates for the Linear Model, *Journal of the Royal Statistical Society. Series B (Methodological)*, **Vol. 34, No. 1** (1972), 1-41.
[7] Preda, V.: *Teoria deciziilor statistice*. Ed. Academiei, Bucharest, 1992.
[8] Robert, C. and Casella, G.: *Monte Carlo Statistical Methods*, Springer Verlag, 2004.
[9] Singh, V.P. and Guo, H.: Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy (POME), *Hydrological Sciences-Journal des Sciences Hydrologyques*, **Vol. 40, No. 2** (1995), 165-181.
[10] Trandafir, R. and Iatan, I.F.: *Modelare-simulare. Noţiuni teoretice şi aplicaţii*, Conspress, Bucharest, 2014.
[11] Vaduva, I.: *Modele de simulare*, Bucharest University Printing House, 2004.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 18-23

A SURVEY OF GENERALIZED INVERSE MATRICES

Ştefania Constantinescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: c_aurora32@yahoo.com

Abstract: This paper presents two forms of generalized inverse matrices - Penrose-Moore pseudo-inverse and the generalized inverse Drazin - and also some fields where they found applicability. Some properties of these matrices, algorithms of obtaining and examples of applications are showed.
Mathematics Subject Classification (2010): 15A09, 65F05
Key words: generalized inverse matrix, Drazin inverse.

1. Introduction

It is well-known that a quadratic matrix of complex numbers is invertible if it is nonsingular. In different domains of applied mathematics, there was a need of having a type of "invertible matrix" also for rectangular matrix or singular quadratic.

- Fredholm's method of solving a certain integral in 1903, was probably the first script about generalized inverse matrices.
- In 1906, Moore expressed the generalized inverse of a matrix in an algebraic **framework**, which was published in 1920 and in the thirties Neumann used generalized inverses in his studies of continuous geometries and regular rings.
- In 1955, Kaplansky and Penrose independently proved that the Moore inverse matrix could be represented in four equations, nowadays known as Penrose-Moore equations.
- In 1958, Drazin [5] introduced a new type of generalized inverse matrix with applications in: finite Markov Chains theory, studies of differential singular and difference equations, cryptography, iterative methods in numerical analysis, dynamics of multibody systems and others.

Thus, the notion of inverse matrix to right R or to left L for a rectangular matrix $A \in M_{m,n}(C)$ appeared:

$$AR = I_m,\ LA = I_n .$$

Starting from these definitions we reached to notion of generalized inverse, the most used being the generalized inverse matrix introduced by mathematicians E.H. Moore in 1920 and Roger Penrose in 1955, named also "pseudo-inverse of a matrix", and the Drazin generalized inverse matrix introduced by Michael Drazin in 1958.
The generalized inverse matrices made themselves necessary and found their applicability in solving unsubstantial systems appeared in the smallest squares method, in determining eigenvalues and eigenvectors or in recognition forms [7] (as application of the least squares method), in statistics, differential equations, Markov chains, population models, cryptography, control theory. More domains of applying the generalized inverse are emphasized in [8].

2. The generalized inverse Penrose-Moore: definitions, properties

2.1. Definitions. Properties

Penrose proved [8] that for any matrix $m \times n$ of complex numbers there is an unique matrix $X \in M_{m,n}(C)$ which satisfies the following relations:

$$\begin{aligned} &AXA = A \\ &XAX = A \\ &(AX)^H = AX \\ &(XA)^H = XA \end{aligned} \tag{1}$$

where A^H is the hermitian matrix of A (transpose and conjugate of A).
These relations are known as *Penrose Lemmas.*

Definition 1. *Matrix X is the generalized inverse matrix of A, noted with* A^I *also named the Penrose-Moore generalized inverse matrix (pseudo-inverse matrix of A).*

The Penrose generalized inverse matrix is useful in solving problems of minimizing distance

$$\rho(x) = \|b - Ax\| \tag{3}$$

where b is a m-dimensional vector, and A is a $m \times n$ matrix with $m \geq n$ and has the columns linear independent ($rank(A)=n$). The problem of determining vector x of minimum length is in fact the least squares problems.
If $n \geq m$ and the lines A are linear independent then [6]

$$A^I = (A^H A)^{-1} A^H . \tag{4}$$

Theorem 1. [3] $x_0 = A^I \cdot b$ *is the best approximate solution of system* $b = Ax$.

If $rank(A) = r < \min(m,n)$ then A^I is determined by singular-value decomposition method presented forwards.

2.2. Singular-value decomposition method

Definition 2. [6]. *For a matrix* $A \in M_{m,n}(C)$ *with rank r there is the decomposition*

$$A = UDV^H \tag{5}$$

where :
$U \in M_{m,m}(C)$ *and* $V \in M_{n,n}(C)$ *are unitary matrix* ($U^H = U^{-1}$, $V^H = V^{-1}$)
$D \in M_{m,n}(C)$ *matrix which has on diagonal the square root* s_i, $i = \overline{1,r}$ *of eigenvalues of quadratic matrix* $A^* A^H$, *in number equal to r,* $D = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$, $S \in M_{r,r}(C)$ *being the singular values matrix. Lines* $\{m-r,...,m\}$ *and columns* $\{n-r,...,n\}$ *of D have null elements.*
In the following algorithm the decomposition of matrix A was made with function MathCAD *svd2*.

Theorem 2. [6]. *With notations above*

$$AI = V \cdot D^+ \cdot U^H \tag{6}$$

where $D^+ = \begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix}$, *and elements of matrix* S^{-1} *are* $s_i^{-1} = \frac{1}{s_i}$, $i = \overline{1,r}$.

3. Algorithm of obtaining

Algorithm of generating the Penrose-Moore pseudo-inverse matrix

```
Penrose_Moore (A) :=
    m ← last(A^⟨1⟩)
    n ← last[(A^T)^⟨1⟩]
    r ← rank(A)
    return AI ← (A^T · A)^-1 · A^T   if r = n
    return AI ← A^T · (A · A^T)^-1   if r = m
    otherwise
        A ← A^T   if m < n
        (s; U; V) ← svd2(A)
        V ← V^T
        for i ∈ 1 .. r
            s_i ← (s_i)^-1   if s_i ≠ 0
        AI ← V · diag(s) · U^T
        return AI^T
```

1. Case of large linear equations systems

$$A := \begin{pmatrix} 5 & 1 & -1 \\ 1 & 6 & 0 \\ 1 & -1 & 4 \\ 1 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix} \qquad \operatorname{rank}(A) = 3 \qquad AI := \text{Penrose\_Moore}(A)$$

$$AI = \begin{pmatrix} 0.16045 & -0.01302 & 0.04256 & 0.04954 & 0.05934 \\ -0.02247 & 0.1544 & 0.00743 & -0.05129 & 0.0009 \\ -0.0452 & 0.06364 & 0.21813 & 0.03625 & -0.04601 \end{pmatrix} \quad AI \cdot A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

2. Case of small linear equations system

$$A := \begin{pmatrix} 6 & -1 & 1 & 2 & 3 \\ 1 & 7 & 0 & -3 & 2 \\ 0 & 1 & 4 & -1 & 1 \end{pmatrix} \quad \operatorname{rank}(A) = 3 \quad AI := \text{Penrose\_Moore}(A)$$

$$AI = \begin{pmatrix} 0.12143 & 0.02577 & -0.04184 \\ -0.01613 & 0.11388 & -0.0159 \\ -0.00006 & -0.04559 & 0.23933 \\ 0.04125 & -0.04011 & -0.03598 \\ 0.05761 & 0.02836 & 0.02259 \end{pmatrix} \quad A \cdot AI = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

3. Case when rank(A)<min(m,n)

$$A := \begin{pmatrix} 3 & -2 & -1 & 0 & 0 \\ -6 & 4 & 2 & 1 & 0 \\ 1.5 & -1 & -0.5 & 0 & 1 \\ 3 & -2 & -1 & 0 & -1 \end{pmatrix} \quad AI := \text{Penrose\_Moore}(A) \quad \text{rank}(A) = 3$$

$$AI = \begin{pmatrix} 0.10084 & 0 & 0.07563 & 0.07563 \\ -0.06723 & 0 & -0.05042 & -0.05042 \\ -0.03361 & 0 & -0.02521 & -0.02521 \\ 0.94118 & 1 & 0.70588 & 0.70588 \\ 0.11765 & 0 & 0.58824 & -0.41176 \end{pmatrix}$$

$$AI \cdot A = \begin{pmatrix} 0.64286 & -0.42857 & -0.21429 & 0 & 0 \\ -0.42857 & 0.28571 & 0.14286 & 0 & 0 \\ -0.21429 & 0.14286 & 0.07143 & 0 & 0 \\ 0 & -0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad A \cdot AI = \begin{pmatrix} 0.47059 & 0 & 0.35294 & 0.35294 \\ 0 & 1 & 0 & 0 \\ 0.35294 & 0 & 0.76471 & -0.23529 \\ 0.35294 & 0 & -0.23529 & 0.76471 \end{pmatrix}$$

4. Examples of applications

4.1. Recognition of forms [4]

The linear degradation of an image can be represented by a such relation

$$x_{out}(i, j) = \iint_D x_{in}(u, v) h(i, j; u, v) du dv \tag{7}$$

where *:*

- $x_{in}(u, v)$ represents the original image in point $(u, v) \in D$, D - the set of points of image to be reconstituted,
- $x_{out}(i, j)$ data measured in pixel (i,j) of image to be reconstituted,
- $h(i, j; u, v)$ is a function of deformation supposed to be known.

After meshing relation (7) we obtain

$$x_{out}(i, j) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{in}(u, v) h(i, j; u, v) \tag{8}$$

or

$$H \cdot x_{in} = x_{out} \quad , \quad H = \left(h(i, j; u, v)\right)_{\substack{1 \le i \le m \\ 1 \le j \le n}} \tag{9}$$

H being the matrix of degradation of original image, supposed to be known from the measurement method (with X rays).

The reconstruction criterion of deformed image will be given by the minim distance between x_{out} *si* $\tilde{x}_{in}$ similarly to relation (2). According to Theorem 1, the best solution of system (9) is obtained using the Penrose-Moore pseudo-inverse matrix.

4.2. The hunting out of defects of a concrete beam

For discovering the gaps of a concrete beam, using the methodology of computer tomograph, we can proceed thus:

- We gather a series of projections scanned under different angles of the beam .
- We build the mathematical model of the beam's interior using an iterative process.
- We obtain a model of form $Ax=b$

Mathematical formulation of the problem

-We note with $x \in R^n$ the image of beam's interior to be analyzed.

- We note with $b \in R^m$ the vector of measurements, $b_i = A_i x,\ i = 1,2,...,m$

- The image of beam's interior represents solution x of the small linear system $Ax=b$, because there are less measurements m than dimension n of the image

- The problem becomes $\min_{x \in R^n \ Ax=b} \|x\|_\alpha$ whose solution returns to the smallest squares method, which can be solved with the Penrose generalized inverse matrix.

Regarding the norm of x usually, from counters reasons, we consider $\alpha = 0 \vee 1 \vee 2$, e.g. :

$\|x\|_0$ =*number of non-null elements of x*

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$ and

$\|x\|_2 = (\sum_{i=1}^{n} |x_i|)^{\frac{1}{\alpha}}$. We choose $\alpha = 2$.

According to Theorem 1 $x_0 = A^I b$ is the image of beam's interior to be analyzed, the most appropriate with respect to the gathered information.

5. The Drazin inverse

Definition 3.[2] *Let* $A \in C^{m \times m}$. *The smallest nonnegative integer k such that*

$$rank(A^k) = rank(A^{k+1}) \tag{10}$$

is named the ***index*** *of A, and it is noted with k=Ind(A).*

Definition 4. [2] *Let* $A \in C^{m \times m}$ *with Ind(A)=k and* $X \in C^{m \times m}$ *such that*

$$AX=XA,\ A^{k+1}X=A^k,\ XAX=X, \tag{11}$$

then X is named the ***inverse Drazin matrix*** *or* ***Drazin pseudo-inverse*** *of A, and it noted with* $X=A^D$.

5.1. The iterative Newton-Raphson method of approximating the Drazin pseudo-inverse matrix [10]

In [10] is proved that by taking $\frac{1}{x}$ as root of function $s(y) = \frac{1}{y} - x$, and using the Newton-Raphson method one can obtain a sequence

$$y_{n+1} = y_n - \frac{s(y)}{s'(y)} = y_n(2 - xy_n)\ ,\ n = 1,2,... \tag{12}$$

for a suitable y_0 .

We put relation (12) as form

$$A_0 = \beta A^l\ ,\ A_{n+1} = A_n(2I - A \cdot A_n)\ \ n = 1,2,... \tag{14}$$

by taking $0 < \beta < \frac{2}{\|A\|_2^{\,l+1}}$, $l \geq k$ and one can demonstrate that $(A_n)_{n \geq 1}$ converges to A^D , where I is the unitary matrix of m order.

5.2. The algorithm for determination the Drazin inverse with the Newton-Raphson method

Determination of the index of a quadratic matrix

$$\begin{aligned}
&\text{Index}(A) := \\
&\quad k \leftarrow 1 \\
&\quad r1 \leftarrow \text{rank}\left(A^{k}\right) \\
&\quad r2 \leftarrow \text{rank}\left[A^{(k+1)}\right] \\
&\quad \text{while } r1 \neq r2 \\
&\quad\quad r1 \leftarrow r2 \\
&\quad\quad k \leftarrow k+1 \\
&\quad\quad r2 \leftarrow \text{rank}\left(A^{k}\right) \\
&\quad \text{return } k
\end{aligned}$$

6. Example of application [9, 10]

The solution of a linear equations system with the singular quadratic coefficients' matrix $Ax=b$, where:

$$A := \begin{pmatrix} 2 & 4 & 6 & 5 \\ 1 & 4 & 5 & 4 \\ 0 & -1 & -1 & 0 \\ -1 & -2 & -3 & -3 \end{pmatrix} \quad b = \begin{pmatrix} 8 \\ 7 \\ 3 \\ -3 \end{pmatrix}$$

has the approximate solution given by relation

$$x = A^{D}b + \left(I - A^{D}A\right)z \ , \ z \in R(A) + N(A) \tag{15}$$

where $R(A) = \left\{ y \in R^{m} \middle| (\exists) u \in R^{m} \ \ a.i. \ y = Au \right\}$, $N(A) = \left\{ u \in R^{m} \ \ s.t. Au = \mathbf{0} \right\}$

7. Conclusions

In this paper we have emphasized few properties of generalized inverse matrices, some ways of obtaining and few domains where they have found their applicability. The references represent a selection of articles and books titles that deal with this subject.

References

[1] Baksalary, J.K., Baksalary, O.M., Trenkler, G.: A revisitation of formulas for the Moore–Penrose inverse of modified matrices, *Linear Algebra and its Applications*, 372 (2003) 207–224.

[2] Ben-Israel and T.N.E. Greville: *Generalized Inverses: Theory and Applicatio*ns, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, no. 15, Springer, New York, NY, USA, 2nd edition, 2003.

[3] Chountasis, S., Katsikis, V.N., Pappas, D.: Applications of the Moore-Penrose Inverse in Digital Image Restoration, *Mathematical Problems in Engineering*, Hindawi, Volume 2009, Article ID 170724, 12 pages http://dx.doi.org/10.1155/2009/170724.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 24-29

COMMON EXTENSIONS FOR A FAMILY OF LATTICE OPERATORS

Rodica-Mihaela Dăneţ
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: rodica.danet@gmail.com
Marian-Valentin Popescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: popescu.marianvalentin@gmail.com
Nicoleta Popescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: popescu.snicoleta@gmail.com

Abstract: In the *first* part of this paper we study sufficient conditions for the existence of a lattice operator which extends an arbitrary family of lattice operators. The obtained results generalize our previous results devoted to the simultaneous extension of two lattice operators and also the results of Z. Lipecki (1980, 1985) concerning the extension of a single lattice operator *T*. In 1982, Lipecki proved that all extensions of *T* which are lattice operators are precisely the extreme points of the set of all positive extensions of *T*. In the *second* part of this paper, with the aim of generalizing this result in a future work, we will go through the first step, inspired by a result of Z. Lipecki, D. Plachky and W. Thomsen (1979). So we will characterize the extreme points of the set of all common positive extensions of a family of lattice operators.
Mathematics Subject Classification (2010): 46A22, 47B60, 47B65.
Key words: lattice operator, extension of a lattice operator, sublinear operator.

1. Introduction

This work has two parts. The first part generalizes some results obtained by the authors in [3]. In the second part, we are doing the necessary preparations as that in a future work we will be able to connect our problem with the notion of extreme points.

So, following an idea of Z. Lipecki, D. Plachky and W. Thomsen [7, Theorem 3] and actually an idea of [4, Theorem 1], we give a characterization of the extreme points of the set of all common positive extensions of a given family $\left(T_\delta\right)_{\delta\in\Delta}$ of operators defined on vector subspaces $G_\delta\left(\delta\in\Delta\right)$ of a vector lattice E. Notice that, in his paper [5], which followed to [7], Z. Lipecki used [7, Theorem 3] to prove that *"An extreme positive extension of a lattice operator is again a lattice operator and a converse to this assertion also holds."* - see [5, Theorem 2].

As a consequence (see [5, Corollary 2]), Z. Lipecki proved that any lattice operator $T:G\to F$, defined on a *majorizing* vector sublattice G of a given real vector lattice E, extends to a lattice operator. (Here, F is an arbitrary Dedekind complete real vector lattice.) In the next work we intent to generalize [5, Theorem 2] to a family of lattice operators.

2. Preliminaries

For a better understanding we begin, as in [3], to remember some basic notions that lead to the concept of lattice operator.

Firstly we suppose that X and Y are two ***vector spaces*** and let $T:X\to Y$ and $S:X\to Y$ be two mappings. We say (see, for example, [1]) that

a) T is a *linear operator* if it has the following properties for all x,y in X and $\alpha\in\mathbb{R}$

1) $T(x+y)=T(x)+T(y)$ (i.e., T is *additive*),

2) $T(\alpha x)=\alpha T(x)$ (i.e., T is *homogeneous*).

b) S is a *sublinear operator* if it has the following properties for all x,y in X and $\alpha\geq 0$ in $\mathbb{R}$

1) $S(x+y)\leq S(x)+S(y)$ (i.e., S is *subadditive*),

2) $S(\alpha x)=\alpha S(x)$ (i.e., S is *positively homogeneous*).

Secondly we assume that E and F are two ***ordered vector spaces***, $T:E\to F$ is a linear operator and $S:E\to F$ is a mapping. We say that

a) T is a *positive operator* if $T(x)\geq 0$ for all $x\geq 0$ in E,

b) S is *monotone* (or, equivalently, *increasing*) if $S(x)\leq S(y)$ for all $x\leq y$ in E.

Now we suppose that E and F are two ***vector lattices*** and $T:E\to F$ is a linear operator. Obviously T preserves the algebraic operations. A lattice operator is a linear operator that preserves the lattice operations, too. More precisely we say that T is a *lattice operator* (or, equivalently, *lattice homomorphism*) if for all x,y in E,

$$T(x\vee y)=T(x)\vee T(y) \tag{1}$$

or, equivalently,

$$T(x\wedge y)=T(x)\wedge T(y). \tag{2}$$

In [2] are given other properties equivalent with (1). Also are given other definitions and few examples of lattice operators, some of them known from the literature.

For the second part ("3.B.") of this paper we need the following notions.

1) We suppose that X is a (real) vector space. We recall that a set $C\subset X$ is called *convex* if for all $x,y\in C$ and $\alpha\in[0,1]$, it follows that $\alpha x+(1-\alpha)y\in C$.

2) Also we recall that an *extreme point* for C is a point of C which does not lie in any open line segment joining two points of C. Intuitively, an extreme point is a "corner" or a "vertex" of C. We will denote by $extr(C)$ the set of all extreme points of C.

3) We recall that a matrix $A=(a_{ij})_{i=\overline{1,n},\ j=\overline{1,n}}$ is called *positive* (and we denote this by $A\geq 0$) if $a_{ij}\geq 0$ for all $i=\overline{1,n}$ and $j=\overline{1,n}$.

4) Also, a positive matrix A is called *stochastic* (more precisely *row stochastic*), if each row sum of A is 1, or equivalently, $Ae=e$, where $e=(1,1,...,1)$ is the unit element of C^n.

5) A matrix $A=(a_{ij})_{i=\overline{1,n},\ j=\overline{1,n}}$ is called a $(0,1)$*-matrix* if for each pair (i,j) either $a_{ij}=0$ or $a_{ij}=1$.

6) A matrix $A=(a_{ij})_{i=\overline{1,n},\ j=\overline{1,n}}$ is called a *permutation matrix* if $Ae_i=e_k$, where $i\mapsto k$ is a permutation of $\{1,...,n\}$ and $\{e_1,e_2,...,e_n\}$ is the canonical basis of $\mathbb{C}^n$.

7) Let E be an ordered vector space and $G\subseteq E$ a vector subspace. We say that G is a *majorizing subspace* if for all $x\in E$ there exists $v\in G$ such that $x\leq v$ (or, equivalently, there exists $w\in G$ with $w\leq x$).

3. Main results

3.A. Extending a family of lattice operators

3.A.1. Description of the sublattice generated by a family of vector sublattices and an additional vector

The following result generalizes [3, Proposition 2].

Proposition 1. *Let E be a vector lattice, $(G_\delta)_{\delta\in\Delta}$ a family of vector sublattices in E and $x_0 \in E \setminus span\left(\bigcup_{\delta\in\Delta} G_\delta\right)$. Denote $S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup\{x_0\}\right)$ the vector sublattice of E, generated by $\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup\{x_0\}$. Then $S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup\{x_0\}\right) = M - M$, with*

$$M = \left\{ \bigvee_{i=\overline{1,n}} \left(\sum_{\delta\in\Delta} v_{\delta i} + \alpha_i x_0 \right) \middle| n \in \mathbb{N}^*,\ \forall i = \overline{1,n},\ (v_{\delta i})_{\delta\in\Delta} \in \Phi\left((G_\delta)_{\delta\in\Delta}\right) \text{ and } \alpha_i \in \mathbb{R}_+ \right\},$$

where $\Phi\left((G_\delta)_{\delta\in\Delta}\right)$ is the collection of all families $\{v_\delta \in G_\delta \,|\, \delta\in\Delta\}$ such that $v_\delta \neq 0$ for at most finitely $\delta\in\Delta$.

(The above notation $\Phi\left((G_\delta)_{\delta\in\Delta}\right)$ was used for the first time by D. Maharam in his theorem concerning the common positive extension of a family of positive linear functionals – see [9] or [11, Theorem 3].)

Proof. Obviously, M is closed under finite suprema and multiplication by positive scalars. We can also prove that $S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup\{x_0\}\right) = M - M$. (To this end we start by observing first that $0\in M$, $M \subset M - M$, $x_0 \in M - M$, $G_\delta \subset M - M$ for all $\delta\in\Delta$. We can prove that $M - M$ is a sublattice of E. Also for any sublattice H of E, such that $H \supseteq \{x_0\}$ and $H \supseteq G_\delta$, for all $\delta\in\Delta$, it follows that $H \supseteq M - M$.) ■

3.A.2 Common extension for a family of lattice operators by using an additional set

The following result generalizes [3, Theorem 6].

Theorem 2. *Let E_0 and F be two vector lattices, $(G_\delta)_{\delta\in\Delta}$ a family of vector sublattices in E_0 and $M \subseteq E_0$ a wedge, closed under finite suprema, such that $M \supseteq G_\delta$, for all $\delta\in\Delta$. Consider $E = S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup M\right)$ $(= M - M)$ the sublattice generated by $(G_\delta)_{\delta\in\Delta}$ and M. Let $T_\delta : G_\delta \to F$ $(\delta\in\Delta)$ be a lattice operator and $P : M \to F$ a mapping such that:*

1) *$P(z_1 \vee z_2) = P(z_1) \vee P(z_2)$, $\forall z_1, z_2 \in M$;*
2) *$P = T_\delta$ on G_δ, for all $\delta\in\Delta$;*
3) *P is positively homogeneous.*

Then the following are equivalent.

i) *There exists a lattice operator $L : E \to F$, such that $L = P$ on M (and consequently $L = T_\delta$ on G_δ for each $\delta\in\Delta$);*

ii) *P is additive on M.*

Proof. i)$\Rightarrow$ii). It is obvious.

ii) $\Rightarrow$i) The operator L is defined by $L(z_1 - z_2) = P(z_1) - P(z_2)$ for all $z_1, z_2 \in M$. This operator L has the following properties:

a) L is well-defined (according to ii));

b) L is positive and linear;
c) L is a lattice operator.

To prove "c)" we use the following identity:

$$(z_1 - z_2) \vee (z_1' - z_2') = (z_1 + z_2') \vee (z_1' + z_2) - (z_2 + z_2')$$

see, for example [8, Theorem 11.5(V)]. ■

Remark. (see also [3, the Remark after Theorem 6]) If F is an Archimedean vector lattice, we can delete the hypothesis 3) in Theorem 2, because we use it only for ii)$\Rightarrow$i), to prove that L is homogeneous. But for F Archimedean, this follows from the additivity and the positivity of L.

3.A.3. Common Extension for a family of lattice operators

Obviously, a lattice operator is a positive operator, but the investigation of the proofs of extension theorems for positive linear operators does not immediately yield extension theorems for lattice operators. As far as we know, this problem was *first* solved (for lattice-group homomorphisms and for special case where the range space is the reals) by A. Hayes (1962). The general case was examined by several authors: D.H. Fremlin (1974), C.D. Aliprantis and O.Burkinshaw (1985), W.A.J. Luxemburg and A.R. Schep (1974), Z. Lipecki (1979, 1980, 1982, 1985), A.W. Wickstead (1980), B. de Pagter (1981), E.R. Aron, A. Hayer, J.J. Madden (1982), G. Buskes (1983, 1987), R-M. Dăneţ (1987, 1993, 2001). Next we apply the result from the previous section to obtain a common extension for a family of lattice operators. Notice that the idea of the proof of this result belongs to Z. Lipecki [6, Corollary]. But our next result is a generalization of the Lipecki's result. It is also a generalization of [3, Theorem 7].

Theorem 3. *Let E be a vector lattice, F a Dedekind complete vector lattice and $(G_\delta)_{\delta\in\Delta}$ a family of vector sublattices in E. Let $T_\delta : G_\delta \to F$ $(\delta \in \Delta)$ be a lattice operator and $P : E \to F$ a mapping such that:*

1) $P(x_1 \vee x_2) = P(x_1) \vee P(x_2)$, $\forall x_1, x_2 \in E$;

2) $P\left(\sum_{\delta\in\Delta} v_\delta + x\right) = \sum_{\delta\in\Delta} T_\delta(v_\delta) + P(x)$ *for all* $(v_\delta)_{\delta\in\Delta} \in \Phi\left((G_\delta)_{\delta\in\Delta}\right)$, *and* $x \in E$, *where* $\Phi\left((G_\delta)_{\delta\in\Delta}\right)$ *is the collection of all families* $\{v_\delta \in G_\delta \mid \delta \in \Delta\}$ *such that* $v_\delta = 0$ *for cofinitely many* δ *(or, equivalently,* $v_\delta \neq 0$, *for at most finitely many* $\delta \in \Delta$*).*

Then there exists a lattice operator $L : E \to F$, such that $L = T_\delta$ on G_δ, for each $\delta \in \Delta$.

Proof. Step 1. Let $x_0 \in E \setminus span\left(\bigcup_{\delta\in\Delta} G_\delta\right)$. Denote $E_1 = S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right) \cup \{x_0\}\right)$ the sublattice generated by the family $(G_\delta)_{\delta\in\Delta}$ and the vector x_0. Then $E_1 = M - M$, where

$$M = \left\{ z \in E \;\middle|\; z = \bigvee_{i=\overline{1,n}} \left(\sum_{\delta\in\Delta} v_{\delta i} + \alpha_i x_0 \right), \text{ with } n \in \mathbb{N}^*, \text{ and for each } i = \overline{1,n}, \right.$$
$$\left. (v_{\delta i})_{\delta\in\Delta} \in \Phi\left((G_\delta)_{\delta\in\Delta}\right) \text{ and } \alpha_i \in \mathbb{R} \right\}.$$

Properties of M and P:

I) M is a wedge;
II) M is closed under finite suprema;
III) $G_\delta \subseteq M$ for each $\delta \in \Delta$;
IV) P additive on M.

Therefore, from the Theorem 2 and the remark that follows it we infer that there exists $T : S\left(\left(\bigcup_{\delta\in\Delta} G_\delta\right)\cup\{x_0\}\right) = M - M \to F$, a lattice operator such that $T = P$ on M and so $T = T_\delta$ on G_δ, for each $\delta \in \Delta$. (Remember that T is defined by $T(z_1 - z_2) = P(z_1) - P(z_2)$.)

Step 2 By applying Zorn lemma, we obtain a lattice operator $L : E \to F$ which extends T_δ, for each $\delta \in \Delta$. ■

3.B. A characterization of the extreme points of the common positive extensions for a family of positive linear operators.

3.B.1. Comments

In this part of our paper we will study the connection between the common extension of a family of lattice operators and the notion of extreme points. This idea is suggested by the following result concerning the characterization of a lattice operator on $\mathbb{C}^n$ by using its matrix in the canonical basis of $\mathbb{C}^n$.

Proposition 4. [10, Proposition 4.4] *A stochastic matrix A defines a lattice operator on $\mathbb{C}^n$ if and only if A is an extreme point in the set S_n of all stochastic matrices of order n.*

Remark. Notice that "*The matrix A is an extreme point of S_n if and only if A is a $(0,1)$-matrix with exactly one* 1 *in each row.*" [10, Proposition 4.3], and that "*A stochastic matrix A defines a lattice isomorphism if and only if A is a permutation matrix.*" [10, Corollary to the Proposition 4.4]. The method used in [10] to prove the last result serves to determine all lattice operators on $\mathbb{C}^n$. In fact if A defines a lattice operator, then $A \geq 0$ (since $|Ax| = Ax$, whenever $x \in \mathbb{C}^n$, $x \geq 0$) and each row of A contains at most one element $y > 0$. The converse is obvious. Similarly, A defines a lattice operator if $A \geq 0$ and the replacement of each non-zero entry by 1 transforms A in a permutation matrix.

Returning to the purpose of this part of our paper, we will begin by analyzing the connection between the problem of the common extension of a family of positive linear operators and the notion of extreme points.

3.B.2. Positive common extensions for a family of positive linear operators and extreme points

Note that, if $G \subset E$ is a vector subspace and $T : G \to F$ is a positive linear operator, we denote by $E_+(T)$ the set of all positive linear extensions of T to the whole E. The following result (see Theorem 5 below) characterizes the extreme common positive linear extensions for a family $(T_\delta)_{\delta\in\Delta}$ of positive linear operators defined on vector subspaces G_δ, $\delta \in \Delta$, of a vector lattice E, that is, it characterizes any operator $L \in extr\left(\bigcap_{\delta\in\Delta} E_+(T_\delta)\right)$. This result is in the line of the following statement (which generalizes [4, Theorem 1]) and even generalizes it: *"Let E be a vector lattice, $G \subseteq E$ a vector subspace and F a Dedekind complete vector lattice. Let also $T : G \to F$ be a positive linear operator. Then $S \in extrE_+(T)$ if and only if $\inf_{v\in G} S(|x - v|) = 0$ for all $x \in E$."* - see [7, Theorem 3].

Theorem 5. *Let E be a vector lattice and F a Dedekind complete vector lattice. Let $(G_\delta)_{\delta\in\Delta}$ be a family of vector subspaces of E and for all $\delta \in \Delta$, consider $T_\delta : G_\delta \to F$ be a positive*

linear operator. Let also $L:E\to F$ *be a common positive linear extension of the family* $\left(T_\delta\right)_{\delta\in\Delta}$, *that is,* $L\in\bigcap_{\delta\in\Delta}E_+\left(T_\delta\right)$. *Then, the following are equivalent:*

i) $L\in extr\left(\bigcap_{\delta\in\Delta}E_+\left(T_\delta\right)\right)$;

ii) $\inf_{v\in G}\left\{L\left(\left|x-\sum_{\delta\in\Delta}v_\delta\right|\right)\middle|\ \left(v_\delta\right)_{\delta\in\Delta}\in\Phi\left(\left(G_\delta\right)_{\delta\in\Delta}\right)\right\}=0$ *for all* $x\in E$, *where* $\Phi\left(\left(G_\delta\right)_{\delta\in\Delta}\right)$ *denotes the collection of all families* $\left\{v_\delta\in G_\delta\middle|\delta\in\Delta\right\}$ *such that* $v_\delta\neq 0$, *for at most finitely* $\delta\in\Delta$.

References

[1] Cristescu, R.: *Ordered Vector Spaces and Linear Operators,* Ed. Academiei, Bucharest, Romania , Abacus Press, Tunbridge Wells, Kent, England, 1976.

[2] Dăneț, R-M.: *Riesz Homomorphisms. Quasi-Riesz Homomorphisms*, in *Order Structures in Functional Analysis,* ed. R. Cristescu, Vol. **4**, pp. 45-89, Publishing House of the Romanian Academy, Bucharest, 2001.

[3] Dăneț, R-M., Popescu, M-V. and Popescu, N.: *On the simultaneous extensions of two lattice operators,* Proceedings of the 14th Workshop of Scientific Communications, Department of Mathematics and Computer Science, Dedicated to Professor Gavriil Păltineanu at 75th anniversary, Technical University of Civil Engineering Bucharest, May 27, 2017, 25-30.

[4] Douglas, R.G.: *On extremal measures and subspace density*, Michigan Math. J. **10**(1964), 243-246.

[5] Lipecki, Z.: *Extension of positive operators and extreme points* II, Colloq. Math. **42**(1979), 285-289.

[6] Lipecki, Z.: *Extension of vector lattice-homomorphisms revisited*, Indag. Math. (Proceedings), **88**(2) (1985), 229-233.

[7] Lipecki, Z., Plachky, D., Thomsen, W.: *Extension of positive operators and extreme points* I, Colloq. Math. **42**(1979), 279-284.

[8] Luxemburg, W.A.J., Zaanen, A.C.: *Riesz Space* I, North Holand, Amsterdam-London, 1971.

[9] Maharam, D.: *Consistent extensions of linear functionals and of probability measures*, in Proc. Sixth Berkley Symp. on Math. Stat. and Probab. (Berkley, 1970/71), Univ. California Press, 1972 vol. II, 127-147.

[10] Schaefer, H.H.: *Banach lattices and positive operators*, Springer Verlag, Berlin, Heidelberg, New York, 1974.

[11] Schmidt, K.D.: *Decomposition and extension of abstract measures in Riesz spaces*, Rend. Istit. Math. Univ. Trieste **29**(1998), Suppl., 135-213.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 30-36

USING R FOR SOIL PARAMETER ESTIMATION

Gabriela-Roxana Dobre
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: roxana.dobre@utcb.ro
Alina Elisabeta Sandu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: alina.sandu@utcb.ro

Abstract: An accurate estimation of soil-water retention curve (SWRC) is required for a detailed knowledge on water percolation into the soil. The model parameters are related to the soil properties. Because the experimental determination of the model parameters in the field or in laboratory is tedious, time consuming and involves uncertainty for practical applications, inverse modeling methods are used. The soil parameters are estimated by fitting retention models to the observed data using optimization techniques. The paper compares two freely accessible computer programs for estimating the hydraulic parameters using gradient methods for least-squares curve-fitting problem: HydroMe, an add-on R-package and SWRC Fit.

Mathematics Subject Classification (2010): 97M50, 93E24, 49M15

Key words: optimization, R programming language, inverse modeling, hydraulic parameters, unsaturated zone

1. Introduction

The dynamics of the water in vadose (unsaturated) zone, between the land surface and the saturated zone, it is affected by human activity, including agriculture, mining, construction, and waste disposal. The mathematical models used to predict the mobility of the water and pollutants in unsaturated zone are based on Richards' equation. The solution of Richards' equation requires well-defined hydraulic properties of the soil. Soil hydraulic properties are essential in irrigation and drainage studies, for predicting leaching of nutrients and for other agronomical and environmental applications.

Soil Water Retention Curves (SWRC) define the relationship between the amount of water in a soil and soil suction. SWRC has been successfully used to estimate all unsaturated soil property functions. The unsaturated soil hydraulic properties, are in general highly nonlinear functions of the pressure head.

The shape of soil water retention curves is given by different analytical models: Brooks and Corey, 1964 [1]; van Genuchten, 1980 [10]; Durner, 1994 [3]; Fredlund and Xing 1994 [4]; Kosugi, 1996 [5]; Seki 2007 [9].

The model parameters for the soil water characteristic curve of a variably saturated soil are generally estimated by fitting various retention models to the observed retention data using inverse analysis [8].

Classical gradient-based methods find the solution in the neighborhood of a starting point. In case of nonlinear least-squares fitting applied for soil parameter estimation Gaus-Newton and Levenberg-Marquardt method are usual technique in hydrological studies. Also, local search methodologies cannot provide a unique solution and may fail to locate the global minimum, so these cases require the use of global optimization methods.

2. Soil hydraulic properties

The equation of the water flow in the unsaturated media can be modeled mathematically by Richards' equation

$$C(h)\cdot\frac{\partial h}{\partial t}=div\big(K(h)\cdot grad(h+z)\big)+S;\ C(h)=\frac{d\theta}{dh} \quad (1)$$

where $X=\mathrm{X}(\mathrm{x},\mathrm{y},\mathrm{z})$ is the spatial variable; t is the time $[T]$; $h=h(X,t)\le 0$ is soil water pressure head $[L]$; $\theta=\theta(h)$ is the volumetric water content $[L^3L^{-3}]$; z is the elevation head $[L]$; $K=K(h)$ is the unsaturated hydraulic conductivity $[LT^{-1}]$; $S=S(X,t)$ is the volumetric sink term representing the sources of water $[L^3L^{-3}T^{-1}]$, $C=\mathrm{C}(h)$ is the soil specific moisture capacity function $[L^{-1}]$.

The relationship between the amount of water in a soil and soil suction are given by parametric equations described in Table 1 [7,9].

Model	Equation	Parameters	Software Package	
			R (HydroMe)	SWRC Fit
Gardner (1958)	$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{(1+\lvert\alpha h\rvert^{\mathrm{n}})}$	$\theta_s,\theta_r,\alpha,n$	✓	
Brooks and Corey (1964)	$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{\lvert\alpha h\rvert^{n}};\ \theta(h)=\theta_r+\frac{\theta_s-\theta_r}{\lvert h/hb\rvert^{\lambda}}$	$\theta_s,\theta_r,\alpha,n$ $\theta_s,\theta_r,\lambda,hb$	✓	✓
Campbell (1974)	$\theta(h)=\theta_s\lvert\alpha h\rvert^{\mathrm{n}}$	θ_s,α,n	✓	
van Genuchten (1980)	$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{(1+\lvert\alpha h\rvert^{n})^{1-1/n}}$	$\theta_s,\theta_r,\alpha,n$	✓	✓
Fredlund and Xing (1994)	$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{(1+\lvert h/a\rvert^{n})^{m}}C^*(h)$	θ_s,θ_r,a,n,m	✓	✓
Durner (1994)	$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{(1+\lvert\alpha 1h\rvert^{n1})^{1-1/n1}}w1+\frac{\theta_s-\theta_r}{(1+\lvert\alpha 2h\rvert^{n2})^{1-1/n2}}w2$	$\theta_s,\theta_r,\alpha1,n1,\alpha2,n2$		✓
Kosugi (1996)	$\theta(h)=\theta_r+\frac{(\theta_s-\theta_r)\ln(\frac{h}{hm})}{\sigma}\mathrm{Q}$	$\theta_s,\theta_r,hm,\sigma$	✓	✓
Seki (2007)	$\theta(h)=\theta_r+\mathrm{w1Q}\frac{(\theta_s-\theta_r)\ln\left(\frac{h}{hm1}\right)}{\sigma1}+(1-w1)\mathrm{Q}\frac{(\theta_s-\theta_r)\ln(\frac{h}{hm2})}{\sigma2}$	$\theta_s,\theta_r,w1,hm1,\sigma1,hm2,\sigma2$		✓

Table 1: Overview of different soil hydraulic models

Water retention hydraulic parameters contained in the water retention models are: θ_s is the saturated water content- it is the moisture content when suction potential is very high (almost at the drying point); θ_r is the residual water content- it is the moisture content when suction potential is very low (almost at the saturation point); α- it is the inverse of air-entry potential or bubbling pressure; n, $\mathrm{n}>0$ -it is a parameter or index for the pore-size distribution; $C^*(h)$-it is a correction factor; $a,\sigma,hm,hm1,hm2,n1,n2,\ m1,m2,w1,\alpha1,\alpha2$ are curves shape empirical parameters.

For the determination of SWRC, laboratory measurement of water content versus pressure head are made using pressure plate apparatus. Pressure plate extractors of 15 bar limits are used to determine the water-holding characteristics of a soil sample layer between 0-19 cm;

There were carried out more repetitions of the experiment: for each value of pressure applied to a disturbed soil core (from 1 hPa to −15.500 hPa), water content was measured.

3. Model parameter estimation

We outline the advantages and disadvantages of the different optimization techniques, used for parameter estimation in the unsaturated media. If the classical gradient methods predicted multiple local optimum solutions for different initial solution, the global optimization algorithms help in estimating the exact solution (global optimum solution) from the feasible solutions and do not require the initial solution and gradient.

The direct problem represents a simulation model where the pressure heads and/or volumetric water content are calculated, while the inverse problem consists of determining the parameters of the SWRC for which the modeling error between the calculated and observed values is minimal.

The objective function includes data that can be of different types, magnitudes and accuracy so each residual should be weighted according to the relationship:

$$f(p) = \sum_{i=1}^{ns} \sum_{j=1}^{nt} w_i^j \left[\left(v_i^j(p)\right)^{calc} - \left(v_i^j\right)^{obs}\right]^2 \quad (2)$$

where: p is the trial vector of unknown parameter values; $(v_i^j)^{obs}$) is the measured value for v (pressure head and /or water content); $(v_i^j(p))^{calc}$ is the corresponding predicted model for v (h and/or θ) using the soil hydraulic parameters of the optimized trial vector p, ns, nt are the number for all measurements of v in space and time; $w = \frac{1}{k\sigma^2}$ where σ is the standard deviation and k is the number of samples [8].

The calibration is done by determining the optimal set of parameters p^*such as

$$f(p^*) = minf(p) \quad (3)$$

From the classical gradient methods for parameter estimation, Gaus-Newton and Levenberg–Marquardt [6] are preferred by the soil scientist.

The Gauss–Newton algorithm for parameter estimation is an iterative algorithm for least-squares curve-fitting of continuous and continuously differentiable mathematical models. The inputs for this algorithm include a mathematical model with parameters to be estimated, starting values for the parameters, first-order partial derivatives of the model with respect to the parameters, and experimental data to facilitate the estimation process.

Since the Gauss–Newton algorithm requires starting values to be as close to the exact solutions as possible, the choice for starting values is critical to guaranteeing success. In the HydroMe package [7], the guidelines for choosing the starting values are based on the recommendations from the original references for the hydraulic models from Table 1.

The parameter values of the nonlinear constitutive model are searched for by using the Levenberg–Marquardt approximation can provide a fast convergence. The parameter identification results illustrate that the proposed parameter inversion procedure has not only higher computing efficiency but also better identification accuracy. The Levenberg–Marquardt algorithm is an iterative technique that locates the minimum of a multivariate function that is expressed as the sum of squares of nonlinear real-valued functions

The Levenberg–Marquardt algorithm approximates the normal gradient descent method, while if it is small, the expression transforms into the Gauss–Newton method form.

The iterative process from algorithm has the form

$$p^{k+1} = p^k - (J^{k^T} J^k + \lambda_M diag\left(J^{k^T} J^k\right))^{-1} J^{k^T} J^k \quad (4)$$

where J is the Jacobian matrix of f, $J^k = J(p^k) = \left(\frac{\partial f_i}{\partial p_j}\right)_{\substack{i=\overline{1,n} \\ j=\overline{1,m}}}$ is calculated at the current point p^k ; λ_M is Marquardt parameter. If λ_M tends to infinity we have the steepest descent method and if $\lambda_M = 0$ we have the Gauss-Newton method and its convergence is slow but safe [6].

4. Result and discussion

The soil hydraulic parameters for analyzing water movement in variably saturated soil can be determined by fitting soil hydraulic model to a soil water retention curve.

SWRC Fit performs nonlinear fitting of six soil hydraulic models to measured soil water retention curve; the relationship between the soil water potential and volumetric water content, using Levenberg-Marquardt method [9]. Considering the ordinary least squares regression that minimizes the sum of squared error to find the best fit for the data set, the starting points are selected at random and then clustered together to provide a reasonable degree of coverage. If the start values are very far from the optimum, the algorithm may not converge.

SWRC Fit automatically determines all the necessary conditions for the nonlinear fitting, such as the initial estimate of the parameters, and, therefore, users can simply input the soil water retention data to obtain the necessary parameters. SWRC Fit uses six models: the Brooks and Corey model, the van Genuchten model, Kosugi's lognormal, pore-size distribution model, Durner's bimodal pore-size distribution model, and Seki bimodal log-normal pore-size distribution model. For soils with a heterogeneous pore structure, Durner developed a multimodal retention function, constructed by a linear superposition of curves of the van Genuchten model.

By looking at the results, the accuracy of the fit with different models can be compared in both R^2 values and the fitting curves with measured data points are also shown in a graph.

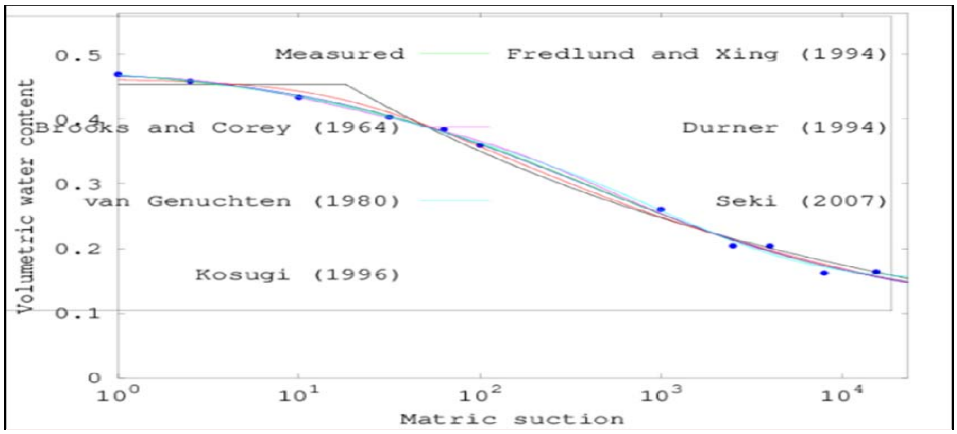


Figure 1:Soil water retention curves

Also, the package HydroMe is executable in the freely downloadable R programming software [11,12]. HydroMe is a freely accessible package for fast, efficient, and accurate estimation of soil hydraulic parameters in some commonly used water retention models [9].

Soil hydraulic models can be called by the corresponding R name

- Brooks-Corey water retention model- $Brook(x, thr, ths, alp, nscal)$
- Campbell water retention model- $Campbel(x, ths, alp, nscal)$
- Fredlund-Xing water retention model-$SSfredlund(x, thr, ths, alp, nscal, mscal)$
- Gardner water retention model- $SSgardner(x, thr, ths, alp, nscal)$
- Kosugi water retention model- $SSkosugi(x, thr, ths, alp, nscal)$

- van Genuchten water retention model- $SSvgm(x, thr, ths, alp, nscal, mscal)$

according to hydraulic function from table 1.

In HydroMe hydraulic parameter estimation is achieved using the nls (Nonlinear Least Squares) function based on Gauss-Newton algorithm. In order to provide a faster convergence, $nlsLM$ function provide a modification of the Levenberg-Marquardt algorithm, with support for lower and upper parameter bounds. The $summary$ function is used to produce result summaries of the results of various model fitting functions. For nonlinear models, nls or $nlsLM$ model fit in R does not calculate R-squared. For general nls models the residual standard error, which is usually called ss or root mean squared error (RMSE) it's a measure of how close the fit is to the points. RMSE divide the sum of the squared residuals by the degrees of freedom ($d.f.$) (i.e the difference between the number of observations and the number of variables):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{calc} - y_i^{obs}\right)^2}{d.f.}} \quad (5)$$

where n is the number of observation.

We are using RStudio because provide the most widely used open source and enterprise-ready professional software for the R statistical computing environment [13].

Model	Software Package Parameters			
	R (HydroMe)-GN	$RMSE$	SWRC Fit-LM	R^2
Gardner (1958)	$\theta s = 0.48351136$ $\theta r = 0.10417396$ $\alpha = 0.04444743$ $n = 0.51360760$	0.007774		
Brooks and Corey (1964)	$\theta s = 0.519582785$ $\theta r = -23.680125779$ $\alpha = 1.548785507$ $n = 0.001487437$	0.02443	$\theta s = 0.45388$ $\theta r = 1.0106e-06$ $hb = 18.022$ $\lambda = 0.15126$	0.98844
Campbell (1974)	$\theta s = 3.660390\text{e}-01$ $\alpha = 1.084433\text{e}-26$ $n = 2.621248\text{e}-03$	0.1385		
van Genuchten (1980)	$\theta r = 1.397300e-01$ $\theta s = 4.917393e-01$ $\alpha = 2.697493e-05$ $n = 4.177069e-01$	0.01298	$\theta s = 0.46178$ $\theta r = 4.7374e-06$ $\alpha = 0.039232$ $n = 1.1682$	0.99491
Fredlund and Xing (1994)	$\theta s = 0.4780593$ $\theta r = -0.2609676$ $\alpha = 41.7475678$ $n = 0.6222185$ $m = 0.4360705$	0.009107	$\theta s = 0.48720$ $\theta r = 0.11994$ $a = 1107.9$ $m = 3.8589$ $n = 0.46270$	0.99729
Durner (1994)			$\theta s = 0.47111$ $\theta r = 0.092595$ $w1 = 0.35267$ $\alpha 1 = 0.19301$ $n1 = 1.3561$ $\alpha 2 = 0.0046025$ $n2 = 1.3417$	0.99710
Kosugi (1996)	$\theta s = 6.040\text{e}-03$ $\theta r = 5.592\text{e-}01$ $\alpha = 4.802\text{e}-06$ $n = 1.191\text{e}+01$	0.01514	$\theta s = 0.47419$ $\theta r = 0.11619$ $hm = 413.11$ $\sigma = 2.9711$	0.99703
Seki (2007)			$\theta s = 0.47725$ $\theta r = 0.15069$ $w1 = 0.68417$ $hm1 = 97.297$ $\sigma 1 = 2.6192$ $hm2 = 1572.8$ $\sigma 2 = 1.2482$	0.99771

Table 2 Simulation results

5. Conclusion

To overcome the excessive costs associated with the measurement of unsaturated soil properties, indirect estimation procedures have been developed to obtain unsaturated soil property functions based on SWRC. Both computer programs, R (HydroMe) and SWRC Fit are freely accessible for estimating the hydraulic parameters. The default algorithm used by HydroMe for parameter estimation is Gauss–Newton that is very sensitive to the choice of starting values which is critical for convergence. To provide a faster convergence Levenberg–Marquardt approximation is used by both computer programs to estimate hydraulic parameters in some commonly used water retention models. To establish the accuracy of nonlinear models, HydroMe offers data about the residual standard error or residual sum-of-squares and does not calculate R-squared. R-squared is based on the underlying assumption that you are fitting a linear model. For nonlinear models, the research literature shows that it

is an invalid goodness-of-fit. Also, SWRC Fit uses R^2values for accuracy of fitting parameters of six models and draw the fitting curves. The range of parameter are not chosen by default so error of calculation can appear if lower and upper bounds are not defined. The users can choose the model to be used for further analysis by comparing the fitting results.

References

[1] Brooks, R. H. and Corey, A. T.: *Hydraulic properties of porous media*, Hydrol. Paper 3, Colorado State Univ., Fort Collins, CO, USA, 1964.
[2] Campbell, G. S.: *A simple method for determining unsaturated conductivity from moisture 25 retention data*, Soil Sci., 117(6), 311–317, 1974.
[3] Durner,W.: *Hydraulic conductivity estimation for soils with heterogeneous pore structure*, Water Resour. Res., 30(2), 211–223, 1994.
[4] Fredlund DG and Xing A.*: Equations for the soil-water characteristic curve*. Canadian Geotechnical Journal 31, 1994.
[5] Kosugi, K.: *Lognormal distribution model for unsaturated soil hydraulic properties*, Water Resour.Res. 32(9), 2697–2703, 1996.
[6] Marquardt, D.: *An algorithm for least-squares estimation of nonlinear parameters*, J. Soc. Indust. Appl. Math., 11, 431–441, 1963.
[7] Omuto, C.T., Gumbe, L.O.: *Estimating water infiltration and retention characteristics using a computer program in R*, Computers & Geosciences, Volume 35, Issue 3, March, Pages 579-585, 2009, https://doi.org/10.1016/j.cageo.2008.08.011
[8] Ritter, A., Hupet, F., Muñoz-Carpena, F., Lambot, S., Vanclooster, M.: *Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods*, Agricultural Water Management 59(2):77-96, 2003.
[9] Seki, K.: *SWRC fit - a nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure.* Hydrol. Earth Syst. Sci. Discuss., 4: 407-437, 2007, doi:10.5194/hessd-4-407
[10] van Genuchten, M. T.: *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils,* Soil Sci. Soc. Am. J., 44, 892–898, 1980.
[11] R Development Core Team: R: *A Language and Environment for Statistical Computing*, 2014, http://www.R-project.org
[12] The Comprehensive R Archive Network:
http://cran.r-project.org/ web/packages/HydroMe/
[13] Rstudio
https://www.rstudio.com/

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 37-42

APPLYING MOMENTS OF ORTHOGONAL POLYNOMIALS TO PATTERN RECOGNITION

Corina Grosu
Politehnica University of Bucharest,
Bucharest, Romania
E-mail: cgr90@yahoo.com
Marta Grosu
Politehnica University of Bucharest,
Bucharest, Romania
E-mail: marta_grosu@yahoo.com

Abstract: One of the principal applications of orthogonal moments lies in their use to image reconstruction in pattern recognition. We present here some properties related to 2D transformations of moments corresponding to a class of bivariate orthogonal polynomials. These polynomials have already been introduced in the literature as complex forms of the bivariate Hermite polynomials [3], [4], [9], but their use in connection with image processing is new. A possible application is related to the construction of depth mappings for video games (in Autodesk 3ds max or Unity) by creating shaders based on such moments.
Mathematics Subject Classification (2010): 33C45, **33C70**, **33D50**
Key words: orthogonal complex polynomials, image moments , invariants of orthogonal moments

1. Introduction

The introduction of orthogonal moments for image analysis origins with Teague [8], in 1980. The main types of orthogonal polynomials considered in his paper were the Legendre and the Zernicke polynomials, orthogonal on the unit disk. Although their use in pattern recognition [2], [10], [11] was especially due to their behavior relative to rigid geometric transformations (translations, rotations and scale), other types of orthogonal moments, invariant to geometric transformations, were also considered. While the image function for which the moments are computed is real valued, the necessity for fast, stable and accurate computations imposed the use of both real and complex orthogonal polynomials, notably Hermite polynomials [3], [4], [9]. Moreover, the study of 2D transformations was mostly made in a product basis of the type $\{H_p(x)H_q(y)\}$.The analysis concerning the transformation formulae of different types of moments was in principal related to the following requirement: from the action of the transformation on the coordinates (either planar or spatial) find an algorithm to obtain the transformation of the moments. The simplest such case was for geometric moments of an image function

$$m_{p,q}(f) = \iint_{R^2} x^p y^q f(x,y)dxdy$$

but unfortunately, computational instability and errors imposed the use of other types of moments.

Our present paper deals with the moments associated to a class of bivariate orthogonal polynomials which includes, as special cases, the bivariate Hermite polynomials [3]. Such a basis is more suitable for computing plane orthogonal moments and for recovering the initial data from the moment sequence [2], [10], [11]. Although the applications considered in this paper are mainly related to pattern recognition through moment' invariants, a more

specialized type of application can also be considered. Thus the construction of depth mappings for video games (in Autodesk 3ds max or Unity) can be improved by creating shaders based on such moments.

The paper is structured as follows. In chapter 2 we recall the main properties of the bivariate complex Hermite and generalized Laguerre polynomials. Then, in chapter 3 we introduce the $\{z_{p,q}^{(\beta)}(f)\}$ moments of an image function f and show that previous results concerning 2D moments are particular cases of the results obtained for $\{z_{p,q}^{(\beta)}(f)\}$ moments. More general results based on the generating function for the corresponding orthogonal polynomials will form the contents of a forthcoming paper.

2. Bivariate polynomials and their connection to complex Hermite and generalized Laguerre polynomials

In our present paper we are mainly concerned with linear transformations of the coordinates, hence the bivariate polynomials will take as arguments the variables $z = x + iy, \bar{z} = x - iy$, respectively $z = |z|e^{i\varphi} = re^{i\varphi}, \bar{z} = |z|e^{-i\varphi} = re^{-i\varphi}$. Nevertheless, for certain inversion formulae and generating functions we shall require that the arguments are complex variables $z_1, z_2 \in C$ (see proposition 2.2) not necessarily satisfying $z_2 = \bar{z}_1$.

Bivariate complex orthogonal polynomials $\{H_{p,q}(z,\bar{z})\}_{p,q\in N}$ are polynomials characterized by the following definition [3], [4], for $z = x + iy, \bar{z} = x - iy$:

$$H_{p,q}(z,\bar{z}) = p!q!(\frac{i}{2})^{p+q}\sum_{j=0}^{p}\sum_{k=0}^{q}(-1)^{q+j}i^{j+k}\frac{H_{j+k}(x)}{j!k!}\frac{H_{p+q-j-k}(y)}{(p-j)!(q-k)!},\quad p,q\in N.$$

We are interested in the main properties of these polynomials, summarized in the following propositions.

Proposition 2.1 [3], [4] The bivariate complex Hermite polynomials have the following properties:

i) $\{H_{p,q}(z,\bar{z})\}_{p,q\in N}$ form an orthogonal basis in the space $L^2(R^2, e^{-(x^2+y^2)}dxdy)$,

ii) the orthogonality relation is expressed by

$$\iint_{R^2} H_{p,q}(x+iy,x-iy)\overline{H_{s,t}(x+iy,x-iy)}e^{-(x^2+y^2)}dxdy = \pi p!q!\delta_{p,s}\delta_{q,t},\ p,q\in N.$$

iii) $\{H_{p,q}(z,\bar{z})\}_{p,q\in N}$ can be connected with the generalized Laguerre polynomials

$$H_{p,q}(z,\bar{z}) = (-1)^{\min(p,q)}(\min(p,q))!|z|^{|p-q|}e^{i(p-q)\varphi}L_{\min(p,q)}^{|p-q|}(z\bar{z}).$$

The following proposition contains the transformation formulae of the bivariate complex Hermite polynomials in terms of other types of monomials or polynomials. It will be used to express the complex (CGZ) moments in terms of other types of moments.

Proposition 2.2 [3], [4] For complex variables z_1, z_2

i) the explicit form of the bivariate Hermite polynomials $\{H_{p,q}(z_1,z_2)\}_{p,q\in N}$ is given by

$$H_{p,q}(z_1,z_2) = p!q!\sum_{k=0}^{\min(p,q)}\frac{(-1)^k}{k!}\frac{z_1^{p-k}}{(p-k)!}\frac{z_2^{q-k}}{(q-k)!} = \sum_{k=0}^{\min(p,q)}(-1)^k\binom{p}{k}\binom{q}{k}k!z_1^{p-k}z_2^{q-k}$$

ii) the relation of the bivariate complex Hermite polynomials $\{H_{p,q}(z,\bar{z})\}_{p,q\in N}$ with products of Hermite polynomials

$$H_{p,q}(z,\bar{z}) = H_{p,q}(x+iy, x-iy) = (\frac{i}{2})^{p+q} \sum_{k=0}^{p} \sum_{j=0}^{q} (-1)^{q+k} i^{j+k} \binom{p}{k} \binom{q}{j} H_{j+k}(x) H_{p+q-j-k}(y)$$

with the inverse relation

$$H_p(x)H_q(y) = \sum_{k=0}^{p+q} \sum_{m=0} \frac{p!q!i^q(-1)^{m-k}}{k!(p-k)!(m-k)!(p+q-m)!} H_{m,p+q-m}(z,\bar{z})$$

iii) $\{H_{p,q}(z,\bar{z})\}_{p,q\in N}$ can be connected with the generalized Laguerre polynomials

$$H_{p,q}(z,\bar{z}) = (-1)^{\min(p,q)}(\min(p,q))!|z|^{|p-q|} e^{i(p-q)\varphi} L_{\min(p,q)}^{|p-q|}(z\bar{z}).$$

The introduction of the complex (CGZ) moments will require the use of generalized Laguerre polynomials of argument $z\bar{z}$, therefore we shall need some properties of $\{L_q^{(\alpha)}(x^2)\}_{q\in N}$.

Proposition 2.4 The generalized Laguerre polynomials $\{L_q^{(\alpha)}(x^2)\}_{q\in N}$ possess the following properties

i) $\{L_q^{(\alpha)}(x^2)\}_{q\in N}$ form an orthogonal basis in the space $L^2((0,\infty), x^{2\alpha+1}e^{-x^2}dx)$

ii) the orthogonality relation is expressed by

$$\int_0^{\infty} L_p^{(\alpha)}(x^2) L_q^{(\alpha)}(x^2) x^{2\alpha+1} e^{-x^2} dx = \frac{\Gamma(\alpha+1)}{2} \binom{p+\alpha}{p} \delta_{pq}$$

iii) the explicit form of $\{L_q^{(\alpha)}(x^2)\}_{q\in N}$ is given by

$$L_q^{(\alpha)}(x^2) = \frac{(\alpha+1)_q}{q!} {}_1F_1(-q, \alpha+1, x^2) = \sum_{k=0}^{q} (-1)^k \binom{q+\alpha}{q-k} \frac{x^{2k}}{k!}.$$

Proof. Statement i) follows from [7], while statements ii) and iii) are direct consequences of the explicit form of the generalized Laguerre polynomials $\{L_q^{(\alpha)}(x)\}_{q\in N}$ given in [5].

Since we shall also use the Hermite and Laguerre functions, we recall their main properties in the following propositions.

3. Complex Laguerre moments

A generalization of the bivariate polynomials considered in the previous chapter was introduced in [4] as $\{Z_{p,q}^{(\beta)}(z_1, z_2)\}_{p,q\in N}$. Accordingly, for a parameter $\beta > -1$ to be specified in connection with the applications,

$$Z_{p,q}^{(\beta)}(z_1, z_2) = \begin{cases} \frac{1}{q!} \sum_{k=0}^{q} (-1)^{q-k} \binom{q}{k} \frac{(\beta+1)_p}{(\beta+1)_{p-k}} z_1^{p-k} z_2^{q-k}, p \geq q \\ \frac{1}{p!} \sum_{k=0}^{p} (-1)^{p-k} \binom{p}{k} \frac{(\beta+1)_q}{(\beta+1)_{q-k}} z_1^{q-k} z_2^{p-k}, p \leq q \end{cases}, \; p,q \in N.$$

Proposition 3.1 [4] For $p,q \in N, p \geq q$

i) $Z_{p,q}^{(\beta)}(z,\bar{z}) = \begin{cases} z^{p-q} L_q^{(\beta+p-q)}(z\bar{z}), p \geq q \\ \bar{z}^{q-p} L_p^{(\beta+q-p)}(z\bar{z}), p \leq q \end{cases}$

ii) if $\beta = 0$, then $Z_{p,q}^{(0)}(z,\bar{z}) = H_{p,q}(z,\bar{z})$.

Proof. The statements follow from direct calculations and the fact that the Laguerre polynomial can be alternatively expressed as

$$L_q^{(\alpha)}(x^2)=\frac{(\alpha+1)_q}{q!}\sum_{k=0}^{q}(-1)^k\binom{q}{k}\frac{x^k}{(\alpha+1)_k}=\frac{(\alpha+1)_q}{q!}\sum_{k=0}^{q}(-1)^k\binom{q}{k}\frac{x^{q-k}}{(\alpha+1)_{q-k}}.$$

Remark. The previous proposition is valid for general arguments $z_1, z_2 \in C$.

The following proposition records the form of the above mentioned polynomials corresponding to a specified change of coordinates.

Proposition 3.2 Let $z=re^{i\varphi}, \bar{z}=re^{-i\varphi}$ and $p,q\in N, p\geq q$. Then

$$H_{p,q}(r,\varphi)=\sum_{k=0}^{q}(-1)^k\binom{p}{k}\binom{q}{k}k!r^{p+q-2k}e^{i(p-q)\varphi}$$

$$Z_{p,q}^{(\beta)}(r,\varphi)=\frac{1}{q!}\sum_{k=0}^{q}(-1)^{q-k}\binom{q}{k}\frac{(\beta+1)_p}{(\beta+1)_{p-k}}r^{p+q-2k}e^{i(p-q)\varphi}$$

$$Z_{p,q}^{(\beta)}(r,\varphi)=r^{p-q}e^{i(p-q)\varphi}L_q^{(\beta+p-q)}(r^2).$$

Proof. The statements follow from direct calculations.

The main properties of the polynomials $\{Z_{p,q}^{(\beta)}(z,\bar{z})\}_{p,q\in N}$, with the notations from proposition 3.2, are recorded in the following proposition. They are in fact direct consequences of the established connection between the polynomials $\{Z_{p,q}^{(\beta)}(z,\bar{z})\}_{p,q\in N}$ and the generalized Laguerre polynomials (see proposition 2.3).

Proposition 3.3 [4]

i) $\{Z_{p,q}^{(\beta)}(z,\bar{z})\}_{p,q\in N}$ are orthogonal with the orthogonality relation

$$\iint_{R^2} Z_{m,n}^{(\beta)}(r,\varphi)\overline{Z_{p,q}^{(\beta)}(r,\varphi)}r^{2\beta+1}e^{-r^2}drd\varphi=\pi\frac{\Gamma(\beta+1+\max(m,n))}{(\min(m,n))!}\delta_{m,p}\delta_{n,q}$$

ii) for $p,q\in N, p\geq q$, $q!Z_{p,q}^{(\beta)}(z,\bar{z})=\sum_{k=0}^{q}\binom{q}{k}(\beta)_k(-1)^k H_{p-k,q-k}(z,\bar{z})$ and

$$H_{p,q}^{(\beta)}(z,\bar{z})=q!\sum_{k=0}^{q}\frac{(-\beta)_k}{k!}(-1)^k Z_{p-k,q-k}^{(\beta)}(z,\bar{z}).$$

Definition 3.4 An image function is a piecewise continuous function $f: D\subset R^2\to R$ with compact support and finite integral [10].

In order to achieve translation invariance we shall consider, like in [11], a modular constant depending on the image function.

Definition 3.5 Let $f: D\subset R^2\to R$ be an image function. For any image function f define the modular constant $\sigma(f)$ by $\sigma(f)=\sigma_0[\iint_{R^2} f(x,y)dxdy]^{1/2}$.

Definition 3.6 For a parameter $\beta>-1$ define the complex (CGZ) moments of f by

$$z_{p,q}^{(\beta)}(f)=\iint_{R^2} Z_{p,q}^{(\beta)}\left(\frac{z}{\sigma(f)},\frac{\bar{z}}{\sigma(f)}\right)\left(\frac{z\bar{z}}{\sigma^2(f)}\right)^{\beta}\exp\left(-\left(\frac{z\bar{z}}{2\sigma^2(f)}\right)\right)f(x,y)dxdy,$$

with $z=x+iy, \bar{z}=x-iy$.

Proposition 3.7 The relation between the (CGZ) moments and the Gauss-Hermite (GH) moments $d_{p,q}$ from [10], obtained for the case $\beta=0$ and the constant σ replaced by $\sigma(f)$, is

$$z_{p,q}^{(0)} = \frac{1}{2^{p+q}} d_{p,q}.$$

Proof. According to [4], for $\beta = 0$, $Z_{p,q}^{(0)}(\frac{z}{\sigma(f)}, \frac{\bar{z}}{\sigma(f)}) = H_{p,q}(\frac{z}{\sigma(f)}, \frac{\bar{z}}{\sigma(f)})$, therefore it suffices to prove the relation for the bivariate Hermite polynomials. But

$$z_{p,q}^{(0)}(f) = \iint_{R^2} p!q!(\frac{i}{2})^{p+q} \sum_{j=0}^{p}\sum_{k=0}^{q} (-1)^{q+j} i^{j+k} \frac{H_{j+k}(\frac{x}{\sigma(f)})}{j!k!} \frac{H_{p+q-j-k}(\frac{y}{\sigma(f)})}{(p-j)!(q-k)!} e^{-\frac{x^2+y^2}{2\sigma^2(f)}} f(x,y)dxdy =$$

$$= \frac{1}{2^{p+q}} \sum_{j=0}^{p}\sum_{k=0}^{q} (-1)^{q-j} i^{p+q-j-k} \binom{p}{j}\binom{q}{k} k! \iint_{R^2} H_{j+k}(\frac{x}{\sigma(f)}) H_{p+q-j-k}(\frac{y}{\sigma(f)}) e^{-\frac{x^2+y^2}{2\sigma^2(f)}} f(x,y)dxd = \frac{1}{2^{p+q}} d_{p,q}$$

with [10] $d_{p,q} = \sum_{j=0}^{p}\sum_{k=0}^{q} (-1)^{q-j} i^{p+q-j-k} \binom{p}{j}\binom{q}{k} \iint_{R^2} H_{j+k}(\frac{x}{\sigma(f)}) H_{p+q-j-k}(\frac{y}{\sigma(f)}) e^{-\frac{x^2+y^2}{2\sigma^2(f)}} f(x,y)dxdy$.

Since we are interested in pattern recognition, we assume that certain transformations do not deform the image, i.e. the function $f : D \subset R^2 \to R$ is translation and rotation invariant. In fact, when analyzing the image, translation invariance is achieved through a change of the origin.

Proposition 3.8 For any image function $f : D \subset R^2 \to R$, a plane rotation by an angle α acts on the GH moments as $\hat{d}_{p,q} = \exp(-i(p-q)\alpha))d_{p,q}$, with

$$\hat{d}_{p,q} = \sum_{j=0}^{p}\sum_{k=0}^{q} (-1)^{q-j} i^{p+q-j-k} \binom{p}{j}\binom{q}{k} \iint_{R^2} H_{j+k}(\frac{\hat{x}}{\sigma(f)}) H_{p+q-j-k}(\frac{\hat{y}}{\sigma(f)}) e^{-\frac{\hat{x}^2+\hat{y}^2}{2\sigma^2}} f(\hat{x},\hat{y})d\hat{x}d\hat{y}$$

and $\hat{z} = \hat{x} + i\hat{y} = (x\cos\alpha - y\sin\alpha) + i(x\sin\alpha + y\cos\alpha)$.

The proof is easily obtained by noticing that $\hat{f}(x,y) = f(\hat{x},\hat{y}) = f(x,y)$, hence $\sigma(\hat{f}) = \sigma(f)$ and

$$H_{p,q}(\frac{\hat{z}}{\sigma(f)}, \frac{\hat{\bar{z}}}{\sigma(f)}) \exp(-(\frac{\hat{z}\hat{\bar{z}}}{2\sigma^2(f)})) = \exp(-i(p-q)\alpha) H_{p,q}(\frac{z}{\sigma(f)}, \frac{\bar{z}}{\sigma(f)}) \exp(-(\frac{z\bar{z}}{2\sigma^2(f)})).$$

Proposition 3.9 For any image function $f : D \subset R^2 \to R$, a plane rotation by an angle α acts on the $\{z_{p,q}^{(\beta)}(f)\}$ moments as $\hat{z}_{p,q}^{(\beta)}(f) = \exp(-i(p-q)\alpha))z_{p,q}^{(\beta)}(f)$, with $\hat{z}_{p,q}^{(\beta)}(f) = z_{p,q}^{(\beta)}(\hat{f})$.

Proof. From propositions 3.1 and 3.2 and $\hat{z}\hat{\bar{z}} = z\bar{z} = r^2$ we obtain that, for $p, q \in N, p \geq q$

$$\hat{z}_{p,q}^{(\beta)}(f) = \iint_{R^2} Z_{p,q}^{(\beta)}(\frac{\hat{z}}{\sigma(f)}, \frac{\hat{\bar{z}}}{\sigma(f)}) (\frac{\hat{z}\hat{\bar{z}}}{\sigma^2(f)})^{\beta} \exp(-(\frac{\hat{z}\hat{\bar{z}}}{2\sigma^2(f)})) f(\hat{x},\hat{y})d\hat{x}d\hat{y} =$$

$$= \int_0^{\infty}\int_0^{2\pi} r(\frac{r}{\sigma(f)})^{p-q+2\beta} L_q^{(\beta+p-q)}(\frac{r^2}{\sigma^2(f)}) e^{-\frac{r^2}{2\sigma^2(f)}} e^{i(p-q)(\varphi-\alpha)} drd\varphi = e^{-i(p-q)\alpha} z_{p.q}^{(\beta)}(f)$$

while for $p, q \in N, p \leq q$.

$$\hat{z}_{p,q}^{(\beta)}(f)=\iint_{R^2} Z_{p,q}^{(\beta)}\left(\frac{\hat{z}}{\sigma(f)},\frac{\bar{\hat{z}}}{\sigma(f)}\right)\left(\frac{\hat{z}\bar{\hat{z}}}{\sigma^2(f)}\right)^{\beta}\exp\left(-\left(\frac{\hat{z}\bar{\hat{z}}}{2\sigma^2(f)}\right)\right)f(\hat{x},\hat{y})d\hat{x}d\hat{y}=$$

$$=\int_0^{\infty}\int_0^{2\pi} r\left(\frac{r}{\sigma(f)}\right)^{q-p+2\beta} L_p^{(\beta+q-p)}\left(\frac{r^2}{\sigma^2(f)}\right)e^{-\frac{r^2}{2\sigma^2(f)}}e^{-i(q-p)(\varphi-\alpha)}drd\varphi=e^{-i(p-q)\alpha}z_{p.q}^{(\beta)}(f)$$

Corollary 3.10 For any image function $f:D\subset R^2\to R$, all the GH rotation invariants given in [10] are also CGH rotation invariants.

Proposition 3.11 For any image function $f:D\subset R^2\to R$, an uniform scaling by a factor $s>1$acts on the $\{z_{p,q}^{(\beta)}(f)\}$ moments as $\hat{z}_{p,q}^{(\beta)}(f)=s^2 z_{p,q}^{(\beta)}(f)$, with

$\hat{z}_{p,q}^{(\beta)}(f)=z_{p,q}^{(\beta)}(\hat{f})$ and $\hat{f}(x,y)=f\left(\frac{x}{s},\frac{y}{s}\right)$.

Proof. For $p,q\in N, p\geq q$ we remark that $\sigma(\hat{f})=s\sigma(f)$ hence a change of variable gives

$$\hat{z}_{p,q}^{(\beta)}(f)=\iint_{R^2} Z_{p,q}^{(\beta)}\left(\frac{z}{\sigma(\hat{f})},\frac{\bar{z}}{\sigma(\hat{f})}\right)\left(\frac{z\bar{z}}{\sigma^2(\hat{f})}\right)^{\beta}\exp\left(-\left(\frac{z\bar{z}}{2\sigma^2(\hat{f})}\right)\right)f\left(\frac{x}{s},\frac{y}{s}\right)dxdy=$$

$$=s^2\iint_{R^2} Z_{p,q}^{(\beta)}\left(\frac{z}{\sigma(f)},\frac{\bar{z}}{\sigma(f)}\right)\left(\frac{z\bar{z}}{\sigma^2(f)}\right)^{\beta}\exp\left(-\left(\frac{z\bar{z}}{2\sigma^2(f)}\right)\right)f(x,y)dxdy$$

and a similar relation holds for $p,q\in N, p\leq q$.

References

[1] Chihara, T.: *An introduction to orthogonal polynomials*, Gordon and Breach Science Publishers, 1978.

[2] Flusser, J., Suk, T. and Zitová, B.: *Moments and Moment Invariants in Pattern Recognition*, Wiley, Chichester, 2009.

[3] Ghanmi, A.: Operational formulae for the complex Hermite polynomials $H_{p,q}(z,\bar{z})$, *Integral Transform. Spec. Funct.* 2**4** (2013), 884 - 895.

[4] Ismail, M.E.H., Zhang, R.: Classes of bivariate orthogonal polynomials, *SIGMA* **12** (2016), 37.

[5] Rainville, E.D.: *Special functions*, The Macmillan Company, 1960.

[6] Srivastava, H.M. and Manocha, H.L.: *A treatise on generating functions*, John Wiley and Sons, 1984.

[7] Szegő, G.: *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1959.

[8] Teague, M.R.: Image Analysis via the General Theory of Moments, *Journal of Optical Society of America*, **70** (1980), 920 – 930.

[9] Wünsche, A.: Hermite and Laguerre 2D polynomials, *Journal of Computational and Applied Mathematics*, **133** (2001), 665 – 678.

[10] Yang, B., Suk, T., Dai, M., Flusser, J.: *2D and 3D Image-Analysis by Gaussian-Hermite Moments*, Wiley, 2016.

[11] Yang, B., Kostkova, J., Flusser, J., Suk, T.: Scale invariants from Gaussian-Hermite moments, *Signal Processing* **132** (2017), 77 - 84.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 43-48

ON THE CONVERGENCE OF CERTAIN SERIES

Ghiocel Groza
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: grozag@utcb.ro
Marilena Jianu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: marilena_jianu@yahoo.com

Abstract: The convergence of a series whose terms are defined by a recurrent sequence of real numbers is proved. Two numerical examples are presented.
Mathematics Subject Classification (2010): 40A05, 40A25
Key words: recurrent sequence, convergent series

1. Introduction

Recurrent sequences are a useful notion into pure or applied mathematics. Thus, for example, in the study of difference equations (see [1]) they are the central object. General notions used in this paper follow those from [2] and [6].

In this paper we consider a recurrent sequence $\{x_n\}_{n\geq 0}$ which generalizes a series of sequences which were used to approximate the solution of differential equations by using either Taylor or Newton series (see [3]-[5]). The main result shows that the series $\sum_{n=0}^{\infty} x_n$ converges absolutely and gives an estimation when the sum of the series is approximated by a partial sum. Two numerical examples are presented.

2. Main result

Let m be a positive integer. Suppose $\{a_{i,n}\}_{n\geq 0}$, $i=1,2,...,m$, $\{b_n\}_{n\geq 0}$ are m+1 sequences of real numbers, ρ_1, ρ_2, with $\rho_1 < \rho_2 < 1$, are two positive constants and N_1 is a positive integer such that, for all $n \geq N_1$ and $i=1,2,...,m$,

$$|a_{i,n}| \leq \rho_1^n, \ |b_n| \leq \rho_2^n. \tag{1}$$

Theorem 1 *Let $\{f_n\}_{n\geq 0}$ be a sequence of positive real numbers such that the sequence $\{g_n\}_{n\geq 1}$, $g_n := \frac{f_{n-1}}{f_n} \leq 1$, is a decreasing sequence which converges to zero. Consider $\{x_n\}_{n\geq 0}$ a recurrent sequence of real numbers such that $x_0, x_1, ..., x_{m-1}$ are arbitrary real numbers and, for every $n \geq m$,*

$$x_n = \frac{f_{n-m}}{f_n}\left(b_{n-m} - \sum_{i=1}^{m}\sum_{j=0}^{n-m} \frac{f_{n-i-j}}{f_{n-m-j}} a_{i,j} x_{n-i-j} \right). \tag{2}$$

Then the series

$$\sum_{n=0}^{\infty} x_n \tag{3}$$

converges absolutely.

Proof. By (1) we may choose a constant $K \geq 1$ such that, for all n and $i = 1,2,...,m$,

$$\left| a_{i,n} \right| \leq K\rho_1^n. \tag{4}$$

For $n \geq m$, denote $\beta_{i,n} = \frac{f_{n-i}}{f_n}$, $i = 1,2,...,m$, $\gamma_n = \sum_{i=1}^{m} \frac{\beta_{i,n-m+1}}{\rho_2^i}$ and, for $n \geq 2m-1$, $\alpha_n = n+1-2m$, $S_{1,n} = \sum_{r=0}^{\alpha_n} \left(\frac{\rho_1}{\rho_2} \right)^r \gamma_{n-r}$. Since, for $i = 1,2,...,m$, $\beta_{i,n} = \prod_{j=n-i+1}^{n} g_j$ it follows that $\{\gamma_n\}_{n \geq m}$ is a decreasing sequence which converges to zero. Hence, for every fixed $n \geq 2m-1$, and $r \in \{0,1,\ldots,\alpha_n\}$, $\{\gamma_{n-r}\}_r$ is an increasing sequence and, because $\rho_1 < \rho_2$, $\left\{ \left(\frac{\rho_1}{\rho_2} \right)^r \right\}_r$ is a decreasing sequence. Then, by Chebyshev's sum inequality, we get

$$\frac{1}{\alpha_n+1} \sum_{r=0}^{\alpha_n} \left(\frac{\rho_1}{\rho_2} \right)^r \gamma_{n-r} \leq \left(\frac{1}{\alpha_n+1} \sum_{r=0}^{\alpha_n} \left(\frac{\rho_1}{\rho_2} \right)^r \right) \cdot \left(\frac{1}{\alpha_n+1} \sum_{r=0}^{\alpha_n} \gamma_{n-r} \right),$$

which implies

$$S_{1,n} \leq \frac{1}{1-\frac{\rho_1}{\rho_2}} \cdot \frac{1}{n+2-2m} \sum_{s=2m-1}^{n} \gamma_s .$$

Since $\lim_{n\to\infty} \gamma_n = 0$, by Stolz-Cesàro theorem, we get $\lim_{n\to\infty} \frac{\sum_{s=2m-1}^{n} \gamma_s}{n+2-2m} = 0$ and consequently $\lim_{n\to\infty} S_{1,n} = 0.$

Let $\varepsilon_i \in (0,1)$, $i = 1,2,3$, such that $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 \leq 1$. Then there exists a positive integer $N_2 \geq \max\{N_1 + m, 2m-1\}$ such that, for all $n \geq N_2$,

$$S_{1,n} \leq \frac{\varepsilon_1}{K}. \tag{5}$$

Consider

$$S_{2,n} := \sum_{j=\alpha_n+1}^{n+1-m}\left(\frac{\rho_1}{\rho_2}\right)^j \sum_{i=1}^{m}\frac{f_{n-i-j-m+1}}{f_{n-j-m+1}\rho_1^i} = \sum_{j=\alpha_n+1}^{n+1-m}\left(\frac{\rho_1}{\rho_2}\right)^j \sum_{i=1}^{m}\frac{\beta_{i,n-j-m+1}}{\rho_1^i},$$

where $f_r = 1, \text{if } r < 0$. Let $t = \min_{0\le r\le m-1}\{f_r\}$ and $T = \max_{-1\le r\le m-2}\{f_r\}$. Then

$$S_{2,n} \le m\left(\frac{\rho_1}{\rho_2}\right)^{\alpha_n+1}\cdot\frac{T}{t}\sum_{i=1}^{m}\rho_1^{-i}.$$

Hence $\lim_{n\to\infty} S_{2,n} = 0$ and there exists a positive integer $N_3 \ge N_2$ such that, for all $n \ge N_3$,

$$S_{2,n} \le \frac{\varepsilon_2}{K}. \tag{6}$$

Then we choose a constant $C \ge \frac{1}{\varepsilon_3}$ such that, for all $n \le N_3$,

$$|x_n| \le C\frac{f_{n-m}\rho_2^{n-m}}{f_n}. \tag{7}$$

We shall prove by induction that (7) holds for every n. It is true for all $n \le N_3$. Assume it is true for all $n' \le n$ and we'll show that

$$|x_{n+1}| \le C\frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}. \tag{8}$$

By (1), (2) and (4) - (7) it follows that

$$\begin{aligned}
|x_{n+1}| &\le \frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}\left(1+\sum_{i=1}^{m}\sum_{j=0}^{n+1-m}\frac{K\rho_1^j f_{n+1-i-j}}{f_{n+1-m-j}}\cdot\frac{Cf_{n+1-m-i-j}\rho_2^{-i-j}}{f_{n+1-i-j}}\right)\\
&= C\frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}\cdot\left(\frac{1}{C}+K\sum_{j=0}^{n+1-m}\left(\frac{\rho_1}{\rho_2}\right)^j\sum_{i=1}^{m}\frac{f_{n+1-m-i-j}}{\rho_2^i f_{n+1-m-j}}\right)\\
&\le C\frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}\left(\frac{1}{C}+K(S_{1,n}+S_{2,n})\right)\\
&\le C\frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}\cdot(\varepsilon_1+\varepsilon_2+\varepsilon_3) \le C\frac{f_{n+1-m}\rho_2^{n+1-m}}{f_{n+1}}.
\end{aligned}$$

This proves (8) and, consequently, (7) holds for every n. Since, by (7), there exists a positive constant C' such that $|x_n| \le C'\rho_2^{n-m}$, the theorem follows. □

The following result gives an estimation of the error when the sum of the series (3) is approximated by a suitable partial sum.

Corollary 1 *Under the hypotheses of Theorem 1 and by using the same notations, suppose S is the sum of the series (3). If* $S_N = \sum_{k=0}^{N} x_k$ *, then, for every* N,

$$|S - S_N| \le C \sum_{k=N+1}^{\infty} \frac{f_{k-m}\rho_2^{k-m}}{f_k}. \tag{9}$$

Moreover, if there exists a positive real function g such that, for every positive integer n, $g(n) = f_n$ *and* $h(x) := \frac{g(x-1)}{g(x)}$, $x \ge N_3 + 1$, *is a decreasing function, then, for every* $N > N_3$,

$$|S - S_N| \le C \int_N^{\infty} \frac{g(x-m)\rho_2^{x-m}}{g(x)} dx. \tag{10}$$

Proof. By (7) it follows (9). Since $\frac{g(x-m)}{g(x)} = \prod_{k=0}^{m-1} h(x-k)$ is a decreasing function, by (9) we get (10). □

Similarly, the result which follows can be proved.

Corollary 2 *Under the hypotheses of Theorem 1 and by using the same notations, the sequence* $\{x_n\}_{n\ge 0}$ *belongs to the space* l_1. *If* $\sigma = \|\{x_n\}_{n\ge 0}\|_{l_1} = \sum_{k=0}^{\infty} |x_k|$ *and* $\sigma_N = \sum_{k=1}^{N} |x_k|$, *then, for every* N,

$$|\sigma - \sigma_N| \le C \sum_{k=N+1}^{\infty} \frac{f_{k-m}\rho_2^{k-m}}{f_k}.$$

Moreover, if there exists a positive real function g such that, for every positive integer n, $g(n) = f_n$ *and* $h(x) := \frac{g(x-1)}{g(x)}$ *is a decreasing function, then, for every* N,

$$|\sigma - \sigma_N| \le C \int_N^{\infty} \frac{g(x-m)\rho_2^{x-m}}{g(x)} dx.$$

3. Numerical examples

Example 1. Consider the recurrent sequence $\{x_n\}_{n\ge 0}$ defined by

$$x_n = -\frac{1}{(n-1)n}\big((n-1)x_{n-1} + x_{n-2}\big),\ n \ge 2,\ x_0 = x_1 = 1. \tag{11}$$

In this case $m = 2$, $f_n = n!$, $b_n = 0$, $a_{1,0} = a_{2,0} = 1$ and $a_{i,j} = 0$, for every $j = 1,2,\ldots$, $i = 1,2$. Then the hypotheses of Theorem 1 hold. We may choose, for example $\rho_1 = 0.5$, $\rho_2 = 0.8$ and $K = 1$. We'll seek N such that, for every for every $n \ge N$, $|S - S_N| \le 10^{-3}$. In order to find a

suitable constant *C*, used in the estimations (9) and (10), we must find N_i, i = 1,2,3, as in the proof of Theorem 1.

Thus, by (1), we may choose $N_1 = 1$. Then, for $n \geq 3$, $\alpha_n = n-3$, $\beta_{1,n} = \frac{1}{n}$, $\beta_{2,n} = \frac{1}{(n-1)n}$, $\gamma_n = \frac{\beta_{1,n-1}}{0.8} + \frac{\beta_{2,n-1}}{0.8^2} = \frac{1}{0.64(n-2)(n-1)}(0.8n - 0.6)$ and

$$S_{1,n} = \sum_{r=0}^{n-3} \left(\frac{5}{8}\right)^r \frac{0.8(n-r)-0.6}{0.64(n-r-2)(n-r-1)}.$$

Consider $\varepsilon_1 = 0.8$, $\varepsilon_2 = \varepsilon_3 = 0.1$. Hence by choosing $N_2 = 9$, for every $n \geq 9$, it follows (5). Then

$$S_{2,n} = \sum_{j=n-2}^{n-1} \left(\frac{\rho_1}{\rho_2}\right)^j \sum_{i=1}^{2} \frac{\beta_{i,n-j-1}}{\rho_1^i} \leq 2\left(\frac{\rho_1}{\rho_2}\right)^{n-2} \frac{T}{t}\left(\rho_1^{-1} + \rho_1^{-2}\right) = 2\left(\frac{5}{8}\right)^{n-2}(2+4) = 12\left(\frac{5}{8}\right)^{n-2}.$$

Hence $N_3 = 13$ and consequently, for every $n \geq 13$, (6) holds.

Now, by (11) we compute x_i, i=0,1,…,13 and we choose

$$C \geq \max\left\{\varepsilon_3^{-1}, \max_{n\leq 13}\left\{\frac{x_n}{\beta_{m,n}\rho_2^{n-m}}\right\}\right\} = \max\left\{10, \max_{n\leq 13}\left\{\frac{(n-1)nx_n}{0.8^{n-2}}\right\}\right\}.$$

Hence we get that we may use C = 10. Finally, by taking $g(x) = \Gamma(x)$, from (10) with $N = 22$, we get $S_{22} = 1.7982$ and $|S - S_{22}| \leq .00081 < 10^{-3}$.

Example 2. Let $\{x_n\}_{n\geq 0}$ be the recurrent sequence defined by

$$x_n = \frac{n}{2^n}\left(\frac{1}{2^{n+1}} - \frac{n-1}{2^{n-1}} \cdot x_{n-1} - \frac{1}{2} \cdot x_{n-2}\right),\ n \geq 2,\ x_0 = 2,\ x_1 = -1. \tag{12}$$

Here $m = 2$, $f_n = \frac{2^{\frac{n(n+1)}{2}}}{n!}$, $b_n = \frac{1}{2^{n+1}}$, $a_{1,0} = 1$; $a_{2,0} = \frac{1}{2}$ and $a_{i,j} = 0$, for every $j = 1,2,\ldots$, $i = 1,2$. By Theorem 1 it follows that the series (3) converges. We'll choose *N* such that, for every $n \geq N$, $|S - S_N| \leq 10^{-4}$. Consider $\rho_1 = 0.2$, $\rho_2 = 0.9$ and $K = 1$. We'll find a suitable constant *C*, used in the estimations (9) and (10), as in Example 1.

Thus, by (1), we may choose $N_1 = 1$. Then, for $n \geq 3$, $\alpha_n = n-3$, $\beta_{1,n} = \frac{n}{2^n}$, $\beta_{2,n} = \frac{n(n-1)}{2^{2n-1}}$, $\gamma_n = \frac{\beta_{1,n-1}}{0.9} + \frac{\beta_{2,n-1}}{0.9^2} = \frac{n-1}{0.81}\left(\frac{0.9}{2^{n-1}} + \frac{n-2}{2^{2n-3}}\right) = \frac{n-1}{0.81 \cdot 2^{2n-3}}\left(0.9 \cdot 2^{n-2} + n - 2\right)$ and

$$S_{1,n}=\sum_{r=0}^{n-3}\left(\frac{2}{9}\right)^r\frac{n-r-1}{0.81\cdot 2^{2(n-r)-3}}\left(0.9\cdot 2^{n-r-2}+n-r-2\right).$$

Consider $\varepsilon_1=0.7$, $\varepsilon_2=0.2$ and $\varepsilon_3=0.1$. Hence by choosing $N_2=5$, for every $n\geq 5$, it follows (5). Then

$$S_{2,n}=\sum_{j=n-2}^{n-1}\left(\frac{\rho_1}{\rho_2}\right)^j\sum_{i=1}^{2}\frac{\beta_{i,n-i-j-1}}{\rho_1^i}\leq 2\left(\frac{\rho_1}{\rho_2}\right)^{n-2}\frac{T}{t}\left(\rho_1^{-1}+\rho_1^{-2}\right)=2\left(\frac{2}{9}\right)^{n-2}(5+25)=60\cdot\left(\frac{2}{9}\right)^{n-2}.$$

Hence $N_3=7$ and consequently, for every $n\geq 7$, (6) holds.

Now, by (12) we find x_i, i=0,1,…,8 and we choose

$$C\geq\max\left\{\varepsilon_3^{-1},\max_{n\leq 7}\left\{\frac{x_n}{\beta_{m,n}\rho_2^{n-m}}\right\}\right\}=\max\left\{10,\max_{n\leq 7}\left\{\frac{(n-1)nx_n}{0.9^{n-2}}\right\}\right\}.$$

Hence for $C=10$ by taking $g(x)=\dfrac{2^{0.5x(x+1)}}{\Gamma(x)}$, from (10) with $N=11$, we get $S_{11}=0.05101$ and $\left|S-S_{11}\right|\leq .00004<10^{-4}$.

References

[1] Elaydi, S.: *An introduction to difference equations*, Springer, 2005.
[2] Fihtenholt, G.M.: *A course of differential and integral calculus*, vol. II, Ed. Tehnică, Bucharest, 1964 (Romanian).
[3] Groza, G. and Pop, N.: A numerical method for solving of the boundary value problems for ordinary differential equations, *Result. Math.* **53** (2009), No. 3-4, 295-302.
[4] Groza, G., Jianu, M. and Pop, N.: Infinitely differentiable functions represented into Newton interpolating series, *Carpathian J. Math.*, **30** (2014), No. 3, 309-316.
[5] Groza, G. and Jianu, M.: Polynomial approximations of solutions of boundary value problems for ODEs which arise from engineering, *Proc. of the International Conference-RIGA,* Bucharest, May 19-21, 2014, 131-143.
[6] Kolmogorov, A. and Fomine, S.: *Eléments de la théorie des fonctions et de l'analyse fonctionnelle*, Editions Mir, Moscou, 1974.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 49-54

SOME PROPERTIES OF RELIABILITY POLYNOMIAL OF A HAMMOCK NETWORK

Marilena Jianu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: marilena_jianu@yahoo.com
Leonard Dăuş
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: daus@utcb.ro

Abstract: Motivated by the study of hammock (aka brick-wall) networks, in this paper we introduce the notion of X-path. Using the Jordan Curve Theorem for piecewise smooth curves, we prove duality properties for hammock networks. Consequences for reliability polynomials are given.
Mathematics Subject Classification (2010): 05C31, 05A99, 94Cxx, 68Rxx
Key words: networks, reliability polynomial, lattice paths

1. Introduction

In order to improve the reliability of a network, Moore and Shannon [5] introduced a new type of circuit called brick-wall (or hammock) network. A network of this type is formed by $w \times l$ identical devices disposed in w lines, each line consisting of l devices connected in series. Besides the horizontal connections, there exist also vertical connections. Out of all $(l-1)(w-1)$ possible vertical connections, half are present and the other half are absent. The vertical connections are arranged regularly in an alternate way which gives rise to the "brick-wall" pattern shown in Fig.1. These networks have a large area of applicability in various domains of science, from electronics to medicine.

Each device of the network can be closed with probability p and open with probability $q = 1 - p$ (all the devices work independently). The *reliability* of the network is the probability that the network is closed (i.e. there exist a path made of closed devices connecting the terminals S and T).

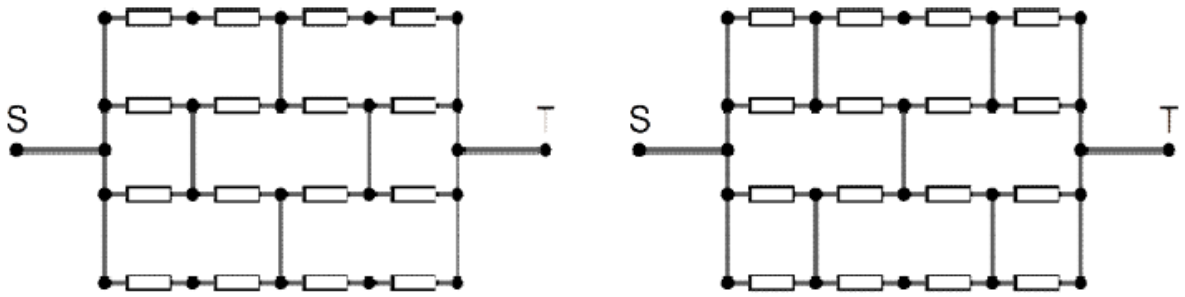


Fig. 1

2. The reliability polynomial of a network

A network is a probabilistic graph [1], $N=(V,E)$, where V is the set of nodes (vertices) and E is the set of (undirected) edges. The edges can be represented as independent identically distributed random variables: each edge operates (is closed) with probability p and fails (is open) with probability $q=1-p$. We assume that nodes do not fail; the fail of the network is always a consequence of edge failures.

Let $K\subset V$ be a subset of some special nodes (terminals). The K - reliability of the network N is the probability that there exists a path (a sequence of adjacent edges) made of operational (closed) edges between any pair of nodes in K. It is a polynomial function in p denoted by $h_K(p)$. If $K=V$ then $h_K(p)$ is called the all-terminal – reliability of the network. If the subset K contains two nodes: S (source / input) and T (target / output) then $h_K(p)$ is called two-terminal reliability. We denote $h_K(p)=h(p)$.

A *pathset* in the network N is a subset of E which contains a path between the nodes S and T. A minimal pathset (*minpath*) is a pathset P such that, if any edge e of P is removed, then $P-\{e\}$ is no longer a pathset (the nodes S and T are disconnected). We denote by $\mathcal{P}$ the set of all the pathsets of N.

A *cutset* in the network N is a subset of edges, $C\subset E$, such that the complementary set, $E-C$, contains no path between S and T ($E-C$ is not a pathset). A minimal cutset (*mincut*) is a cutset C such that, if any edge e of C is removed, then $C-\{e\}$ is no longer a cutset ($E-C\cup\{e\}$ is a pathset). We denote by $\mathcal{C}$ the set of all the cutsets of N.

If $n=|E|$ is the size of the graph, N_i is the number of pathsets with exactly i edges and C_i, the number of cutsets with exactly i edges, then the reliability of the network can be expressed as

$$h(p)=\sum_{P\in\mathcal{P}} p^{|P|}q^{n-|P|}=\sum_{i=1}^{n} N_i p^i (1-p)^{n-i}, \tag{1}$$

or, in terms of cutsets, as

$$h(p)=1-\sum_{C\in\mathcal{C}} q^{|C|}p^{n-|C|}=1-\sum_{i=1}^{n} C_i (1-p)^i p^{n-i}. \tag{2}$$

3. Hammock networks

Brick-wall networks were also named by Moore and Shannon *hammock networks*, from their appearance when nodes S and T are pulled apart and every vertical connection collapses into a node, as rectangular "bricks" deforms into rhombs. As can be seen, the "hammock" representation fits the above definition of the probabilistic graph, unlike the "brick-wall" representation, where the vertical edges have no probability assigned (it is assumed they are always closed). In Fig. 2 a brick-wall network of dimensions $w=3$, $l=7$ and the corresponding hammock network are represented. Notice that, in order to preserve the regularity of the hammock network, the terminal nodes S and T can be replaced by some "fictive" terminal nodes, $S_1,S_2,\ldots$, respectively, $T_1,T_2,\ldots$.

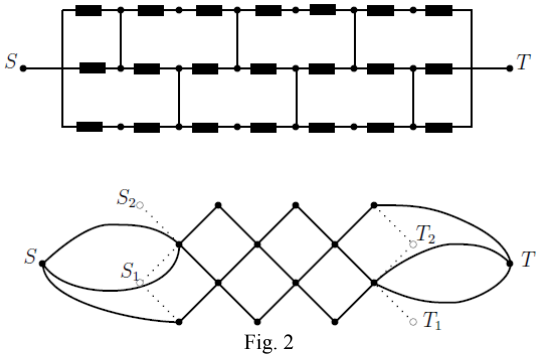


Fig. 2

Definition 1 *Let* $S \subset \mathbb{Z}^2$. *A lattice path with steps in S is a sequence of lattice points,* $v_0, v_1, \ldots, v_k \in \mathbb{Z}^2$, *such that* $v_i - v_{i-1} \in S$ *for all* $i = 0,1,\ldots,k$.

Definition 2 *An* $\boldsymbol{X}$ *– path is a lattice path* $v_0, v_1, \ldots, v_k$ *with steps in the set*

$$S = \{(1,1),(-1,1),(1,-1),(-1,-1)\},$$

such that $v_i \neq v_j$, $\forall i \neq j$.

As can be seen in Fig. 3 (a), from a lattice point (x, y) it is allowed to move in 4 directions and reach one of the 4 neighbor points (if (x, y) is a starting point then any direction may be chosen, if not, we must take into account that $v_i \neq v_j$, $\forall i \neq j$). The neighbor points are:

$$(x+1, y+1), \quad (x-1, y+1), \quad (x+1, y-1), \quad (x-1, y-1).$$

We can notice that the sum of coordinates of any neighbor point has the same parity as $x + y$. We say that a lattice point (x, y) is even (odd) if $x + y$ is even (odd). An **X** – path is even (odd) if it contains even (respectively, odd) points. For example, the **X** – path represented in Fig. 3 (b) contains only odd points.

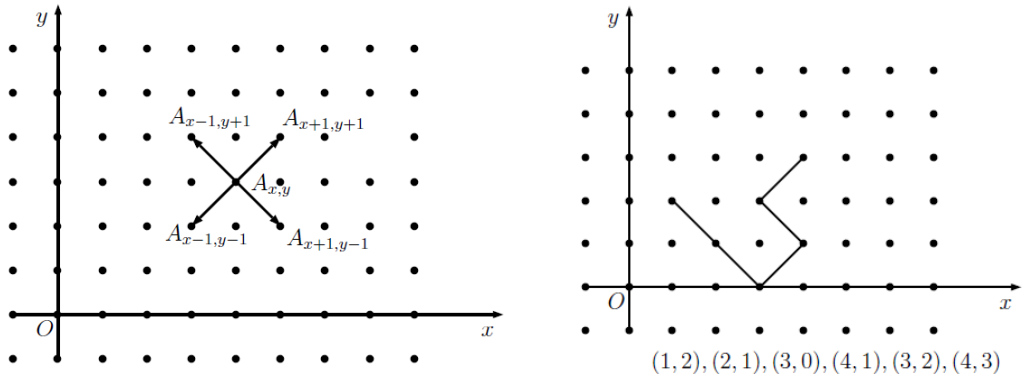


Fig. 3

Let $\mathcal{V}_{l,w} = \{A_{x,y} = (x, y) \in \mathbb{Z}^2 : 0 \leq x \leq l, 0 \leq y \leq w\}$ be the set of all lattice points in the rectangle $[0, l] \times [0, w]$ and $V_{l,w}^{(1)} = \{A_{x,y} \in \mathcal{V}_{l,w} : x + y = \text{even}\}$, $V_{l,w}^{(2)} = \{A_{x,y} \in \mathcal{V}_{l,w} : x + y = \text{odd}\}$ be the subsets of even (respectively, odd) points in $[0, l] \times [0, w]$.

We denote by $\mathcal{E}_{l,w}=\left\{A_{x,y}A_{x',y'}: A_{x,y},A_{x',y'}\in\mathcal{V}_{l,w},\left|x-x'\right|=\left|y-y'\right|=1\right\}$ the set of all the line segments of length $\sqrt{2}$ connecting points of $\mathcal{V}_{l,w}$. Let $E_{l,w}^{(1)}=\left\{A_{x,y}A_{x',y'}\in\mathcal{E}_{l,w}: x+y=\text{even}\right\}$ be the subset of even edges and $E_{l,w}^{(2)}=\left\{A_{x,y}A_{x',y'}\in\mathcal{E}_{l,w}: x+y=\text{odd}\right\}$, the set of odd edges (the two disjoint subsets form a partition of $\mathcal{E}_{l,w}$).

A *hammock network of the first kind* of dimensions (l,w) is the probabilistic graph $H_{l,w}^{(1)}=\left(V_{l,w}^{(1)},E_{l,w}^{(1)}\right)$, while a *hammock network of the second kind* is $H_{l,w}^{(2)}=\left(V_{l,w}^{(2)},E_{l,w}^{(2)}\right)$. We assume that each edge is closed with probability p and open with probability $1-p$. The input (source) nodes are $S_j=A_{0,y}$ (with y = even for the first kind and y = odd for the second kind), and the output (target) nodes are $T_k=A_{l,z}$ (with $l+z$ = even, respectively, odd).

A subset of even (respectively, odd) edges $P\subset E_{l,w}^{(i)}$ is a *pathset* in $H_{l,w}^{(i)}$ if it contains an **X** – path connecting a source node S_j with a target node T_k. Let $\mathcal{P}_{l,w}^{(i)}$ be the set of all pathsets in $H_{l,w}^{(i)}$. A subset $C\subset E_{l,w}^{(i)}$ is a *cutset* in $H_{l,w}^{(i)}$ if $E_{l,w}^{(i)}-C$ contains no **X** – path connecting a source node S_j with a target node T_k. Let $\mathcal{C}_{l,w}^{(i)}$ be the set of all cutsets in $H_{l,w}^{(i)}$. By using these notations in formulas (1) and (2), the reliability polynomials of hammock networks of the first and of the second kind, $h_{l,w}^{(1)}(p)$ and $h_{l,w}^{(2)}(p)$, are written:

$$h_{l,w}^{(i)}(p)=\sum_{P\in\mathcal{P}_{l,w}^{(i)}}p^{|P|}(1-p)^{lw-|P|}=1-\sum_{C\in\mathcal{C}_{l,w}^{(i)}}(1-p)^{|C|}p^{lw-|C|},\ i=1,2 \tag{3}$$

Remark 1 If l = odd or w = odd, then the hammock networks $H_{l,w}^{(1)}$ and $H_{l,w}^{(2)}$ are isomorphic and the reliability polynomials are identical: $h_{l,w}^{(1)}=h_{l,w}^{(2)}$. If l and w are both even numbers, then we have two different networks of dimensions (l,w): $h_{l,w}^{(1)}\neq h_{l,w}^{(2)}$. Fig. 4 represents the two hammock network of the first and second kind for $l=2$ and $w=2$.

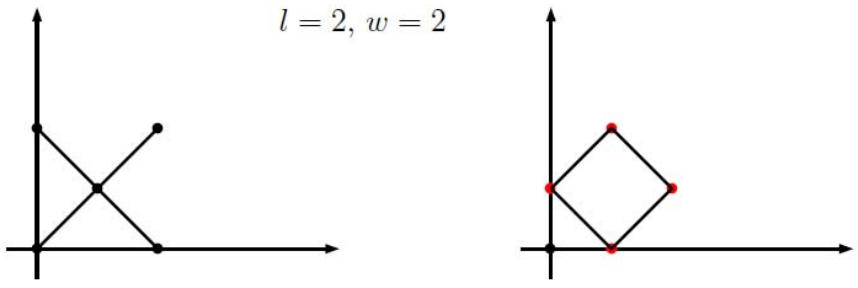


Fig. 4

4. Dual network

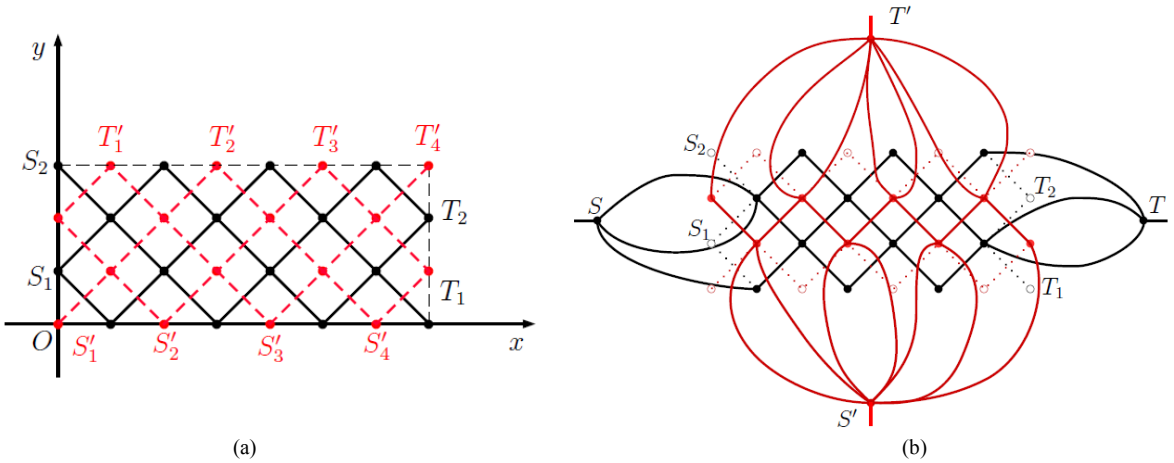
For every edge $e\in\mathcal{E}_{l,w}$, $e=A_{x,y}A_{x+1,y\pm1}$, we denote by $\overline{e}=A_{x+1,y}A_{x,y\pm1}$ its complementary edge (the edge that *cuts* e). It can be seen that the complementary edge of an even edge is odd and the complementary edge of an odd edge is even. Thus, if $e\in E_{l.w}^{(i)}$, then $\overline{e}\in\overline{E_{l,w}^{(i)}}=\mathcal{E}_{l,w}-E_{l,w}^{(i)}=E_{l,w}^{(2/i)}$. By using the notation $\overline{V_{l,w}^{(i)}}=\mathcal{V}_{l,w}-V_{l,w}^{(i)}=V_{l,w}^{(2/i)}$, the dual network of $H_{l,w}^{(i)}=\left(V_{l,w}^{(i)},E_{l,w}^{(i)}\right)$ is $\overline{H_{l,w}^{(i)}}=\left(\overline{V_{l,w}^{(i)}},\overline{E_{l,w}^{(i)}}\right)$ with the source nodes $S_j'=A_{x,0}\in\overline{V_{l,w}^{(i)}}$ and the target

nodes $T'_k = A_{z,w} \in \overline{V_{l,w}^{(i)}}$ (see Fig. 5 (a)). The probability of an edge $\overline{e} \in \overline{E_{l,w}^{(i)}}$ to be closed is the probability of the edge $e \in E_{l.w}^{(i)}$ to be open ("cut"): $q = 1 - p$.

Remark 2 The networks $\overline{H_{l,w}^{(i)}}$ and $H_{w,l}^{(2/i)}$ are isomorphic (since they are symmetric with respect to the first bisectrix) and the reliability of the dual network can be written

$$\overline{h_{l,w}^{(i)}}(q) = h_{w,l}^{(2/i)}(p). \tag{4}$$

Let $G_{l,w}^{(i)}$ be the graph obtained from $H_{l,w}^{(i)}$ by replacing back the "fictive" nodes $S_1, S_2, \ldots$ and $T_1, T_2, \ldots$ by the terminal nodes S and T, respectively, and let $\overline{G_{l,w}^{(i)}}$ be the graph obtained from $\overline{H_{l,w}^{(i)}}$ by the same operation (the terminal nodes, in this case, are S' and T'). We can notice that, if we consider the terminal nodes S and T as being placed to $\pm\infty$, then $\overline{G_{l,w}^{(i)}}$ is the dual graph of $G_{l,w}^{(i)}$, as can be seen in Fig. 5 (b).



(a) (b)

Fig. 5

5. The reliability polynomial of a hammock network

The main result of the paper is **Theorem 1** whose corollaries make the connection between the reliability polynomials of a hammock network and its dual network. The proof of this theorem relies on the **Jordan Curve Theorem** which asserts that *every simple closed plane curve divides the plane into an "interior" region bounded by the curve and an "exterior" region, so that every continuous path connecting a point of one region to a point of the other intersects with that curve somewhere.*

Theorem 1 *Let* $\Sigma = \{e_1, e_2, \ldots, e_n\} \subset E_{l,w}^{(i)}$ *be a subset of edges of the network* $H_{l,w}^{(i)}$ *and let* $\overline{\Sigma} = \{\overline{e}_1, \overline{e}_2, \ldots, \overline{e}_n\} \subset \overline{E_{l,w}^{(i)}}$ *be the set of complementary edges. The following statements hold:*

i) If Σ *is a mincut in* $H_{l,w}^{(i)}$ *then* $\overline{\Sigma}$ *is a minpath in* $\overline{H_{l,w}^{(i)}}$.

ii) If Σ *is a minpath in* $H_{l,w}^{(i)}$ *then* $\overline{\Sigma}$ *is a mincut in* $\overline{H_{l,w}^{(i)}}$.

Corollary 1 *Let* $\Sigma = \{e_1, e_2, \ldots, e_n\} \subset E_{l,w}^{(i)}$ *be a subset of edges of the network* $H_{l,w}^{(i)}$ *and let* $\overline{\Sigma} = \{\overline{e}_1, \overline{e}_2, \ldots, \overline{e}_n\} \subset \overline{E_{l,w}^{(i)}}$ *be the set of complementary edges. The following statements hold:*

i) Σ *is a pathset in* $H_{l,w}^{(i)}$ *if and only if* $\overline{\Sigma}$ *is a cutset in* $\overline{H_{l,w}^{(i)}}$.

ii) Σ *is a cutset in* $H_{l,w}^{(i)}$ *if and only if* $\overline{\Sigma}$ *is a pathset in* $\overline{H_{l,w}^{(i)}}$.

As a consequence, by using the equation (3) and Remark 2, it follows Corollary 2.

Corollary 2 *For any* $l, w \geq 1$ *and* $i = 1, 2$ *the following relation is true for all* $p \in [0,1]$*:*

$$h_{l,w}^{(i)}(p) = 1 - h_{w,l}^{(2/i)}(1-p). \tag{5}$$

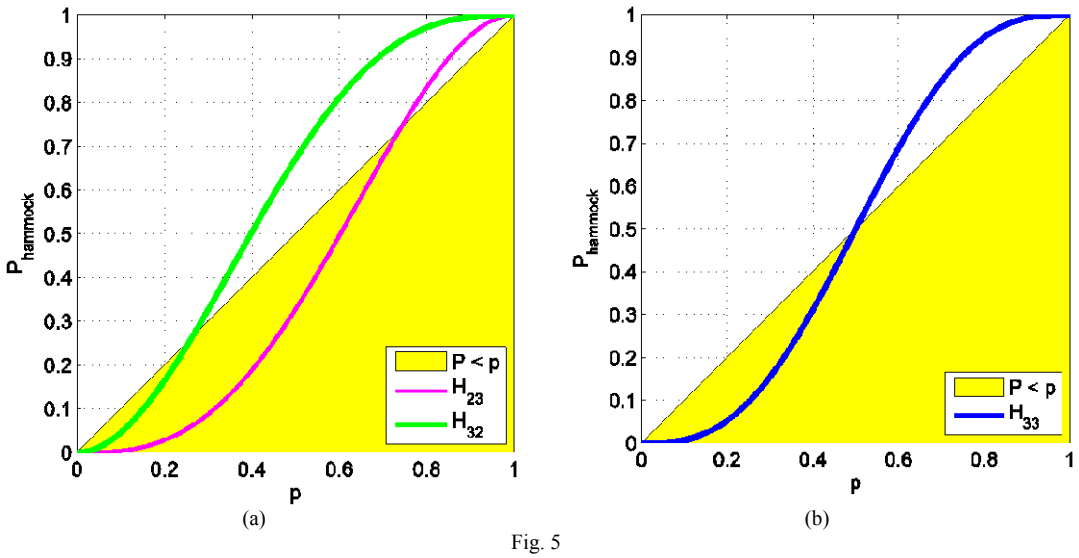
As we saw in Remark 1, if at least one of the numbers l and w is odd then $h_{l,w}^{(1)} = h_{l,w}^{(2)} = h_{l,w}$ and next corollary follows.

Corollary 3 *If l = odd or w = odd then the following relation is true for all* $p \in [0,1]$*:*

$$h_{l,w}(p) = 1 - h_{w,l}(1-p). \tag{6}$$

For $l \neq w$ this means that the graphics of the polynomials $h_{l,w}(p)$ and $h_{w,l}(p)$ are symmetric one to each other with respect to the point $\left(\frac{1}{2}, \frac{1}{2}\right)$, as can be seen in Fig. 6 (a).

For $l = w = 2k+1$ it means that the point $\left(\frac{1}{2}, \frac{1}{2}\right)$ is a center of symmetry for the graphic of the polynomial $h_{2k+1,2k+1}(p)$ (see Fig. 6 (b)).



(a) (b)

Fig. 5

References

[1] Colbourn, C.J.: *The Combinatorics of Network Reliability*, Oxford University Press, Oxford, 1987.

[2] Cowell, S.R., Beiu, V., Dăuş, L. and Poulin, P.: On the exact reliability enhancements of small hammock networks, *IEEE Access*, accepted, (2018).

[3] Dăuş, L., Beiu, V., Cowell, S.R. and Poulin, P.: Brick-wall lattice paths and applications, *Tech. Rep. arXiv*, 17 Apr. 2018, http://arxiv.org/abs/222.7960

[4] Groza, G.: Analiză combinatorică şi algoritmica grafurilor, Conspress, Bucureşti, 2015.

[5] Moore, E.F. and Shannon, C.E.: Reliable circuits using less reliable relays Part I, J. Frankl. Inst. 262(3): 191-208, 1956.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 55-58

WEB-BASED MATHEMATICS EDUCATION - FUTUREMATH

Ion Mierluş Mazilu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: ion.mierlusmazilu@utcb.ro

Abstract: Future Mathematics strives to meet the current needs of teachers and learners alike. Today's students are accustomed to learning with the aid of state-of-the-art technologies. Mathematics educators who wish to implement these technologies in their classroom teaching are often deterred in doing so due to time restraints. The Mathematics Learning Platform (MLP) is designed for acquiring or sharing digital teaching and learning resources in an Open Educational Resource setting, thus promoting a supportive community of mathematics educators involved in the digitization process.
The FutureMath project aims to respond to the requirements of modern society and to make mathematics' learning and teaching more digitalized, effective and accessible. Additionally, the aim is to explore and develop the most motivational, learner centered methods, techniques and resources for engineering mathematics learning and teaching with the help of technology. All the learning resources developed in the project will be made available for free under the idea of Open Source or Open Educational Resource (OER).
Mathematics Subject Classification (2010): 97D40, 97U50
Key words: Mathematics education, learning platform, web-based education, online exercises

1. Introduction

To be a higher educator now is a quite demanding task. Technological developments bring constantly different kind of possibilities to be utilized also in education. Also the new generation is comfortable with almost any technology and they prefer using e.g. videos, online contents and online assessment for studying purposes. The rapid development of technology and computer programs is one of them. There is a wide variety of mathematical computer software (Maple, Mathematica, Mathcad, Matlab etc.) that can offer quick and easy solutions to different kind of mathematical or engineering problems. Spreadsheet programs, such as Excel, also support complex calculations through their complex library of functions. These tools are important in solving different kind of problems, but also may be one cause of low students' motivation on acquiring substantial mathematical knowledge. [2]
A meaningful and efficient process of teaching mathematics has to take into account all the factors that have a substantial impact on the process. The new ways of teaching must rely on technology to create flexible and friendly learning environments in which students can easily acquire and understand new information. Also, the existing technology must be used to its fullest potential to solve a wide variety of problems, make connection between math problems and real life situations and enhance the quality of learning. The teachers must encourage deeper knowledge through geometric exploration programs where learners can visualize functions and data, and experiment with mathematical formulas.
Mathematics is one of the key subjects for any career in engineering, science or business and a good mathematical knowledge is required for understanding and mastering different engineering disciplines. Unfortunately, in the recent decades, the mathematical competences of students have weakened and all those involved in teaching mathematics are facing more and more challenges. There are several factors that had a substantial impact on the process of teaching mathematics. [1]

2. Web-based learning mathematics – FutureMath project

Aside from the creation of a suitable web-based repository (Mathematics Learning Platform - MLP), the project is dedicated to providing a basis for digital mathematics teaching, learning and assessment materials in particular for courses in Mathematics for engineering students. In doing so, pedagogical issues are considered, for example, individualized learning possibilities that take into account learner types. Furthermore, key approaches such as collective thinking, collaboration and shared problem solving skills are taken into account while planning the resources. These skills are not only beneficial for successful studies but also for success in the working world.

With this foundational work and the future contributions from mathematics educators throughout Europe, the accessibility of ready-made digitized materials will increase. These flexible alternatives, which can either replace or be combined with traditional teaching methods, are attractive to students. They can help to engage students more intensely in the learning process, thereby increasing their chances of successfully learning mathematics. The individual learning solutions and differentiated feedback are key to students' self-motivation, so that students are also supported in the self-study phase.

Mathematics plays one of the most important roles in developments of our modern and technology-centered society. Additionally, it lays the basis for technical studies, but is also needed e.g. in economics and life science. In fact, good mathematical skills are crucial for science and economy. The FutureMath project aims not only to develop mathematical competence in Europe but also to pay attention to the quality of mathematics' education. In fact, based on studies (i.e. Hanushek and Wossman, 2007) the quality of education has a strong positive effect on economic growth.

Unfortunately, various studies have shown that mathematical competence in Europe has weakened in recent decades. The lack of mathematical proficiency is already causing problems in engineering mathematics' and other courses in European HEIs. In fact, this seems to be a global problem, and e.g. the learning outcomes of Eastern European countries have been weaker than expected, especially in mathematics, after they moved to the Western European model of education (e.g. SEFI 2002). Compounding the issues, the resources allocated to teaching have been decreased so that there are fewer resources for teaching and the development of teaching.

Additionally, in recent years the study groups have been increasing and becoming even more heterogeneous. This naturally causes problems for organization of mathematics' teaching as for example the entry level of competence in mathematics varies greatly depending on the background studies. Under these circumstances, taking into account individual needs or organizing dynamic and creative activities becomes almost impossible during the classroom sessions. As a sum of many factors, it has been reported that the drop-out rates are high in the field of technology.

However, mathematical skills are a prerequisite in technical studies and mathematics lay the basis for understanding different engineering disciplines. Thus, the students' poor skills in mathematics slow down or even prevent their studies. In principle, an engineer must be able to think analytically and to be capable of logical reasoning. In addition, an engineer needs to understand mathematics, which allows them to deal with and understand technical problems. Overall, mathematics penetrates deep into the engineering professional field, affecting the opportunities to absorb and learn engineering subjects. Thus, for example to be able to make new technological innovations, the understanding and skills of mathematics are crucial.

Unfortunately, the lack of basic skills and knowledge of mathematics among the European engineering students complicates and in the worst case, even prevents future technological development in Europe. In order to maintain the competitiveness of Europe, the basic level of

mathematical proficiency needs urgently to be increased on a large scale. Based on above described situation, the proposed project aims to improve the mathematical proficiency of European engineering students by developing methods and best practices to learn, teach and assess mathematics effectively. Since the objectives of the project are international, the best results can be achieved with transnational co-operation. Based on the results of a survey collected at TAMK in 2014, students expect more digital learning possibilities and utilizations of ubiquitous technology in mathematics' studies. This is very natural as the whole of society is changing. Big data, open data, cloud services, digitalization, IoT etc. affect society and social activities on a large scale. As working life is constantly changing, its expectations and requirements have become more diverse. The 21st century skills, such as collective thinking, collaboration, creativity and shared problem solving skills are key components in modern working life and therefore the university teaching and learning should also train these skills.

The FutureMath project aims to respond to the requirements of modern society and to make mathematics' learning and teaching more digitalized, effective and accessible. Additionally, the aim is to explore and develop the most motivational, learner centered methods, techniques and resources for engineering mathematics learning and teaching with the help of technology. All the learning resources developed in the project will be made available for free under the idea of Open Source or Open Educational Resource (OER).

Overall, the project respects and enables i.e. collective thinking, collaboration and shared problem solving skills. The project aims to develop and improve technology-enhanced methods and resources to teach, learn and study engineering mathematics under the themes such as collaboration, peer instruction and assessment, mostly based on approaches of e-learning 2.0 and 21st century skills.

Furthermore, the objectives are e.g. to pay attention to the different learner types, individual learning solutions, flexibility, effective feedback and assessment.

Future Mathematics strives to meet the current needs of teachers and learners alike. Today’s students are accustomed to learning with the aid of state-of-the-art technologies. Mathematics educators who wish to implement these technologies in their classroom teaching are often deterred in doing so due to time restraints. The Mathematics Learning Platform (MLP) is designed for acquiring or sharing digital teaching and learning resources in an Open Educational Resource setting, thus promoting a supportive community of mathematics educators involved in the digitization process.

The FutureMath project aims to respond to the requirements of modern society and to make mathematics' learning and teaching more digitalized, effective and accessible. Additionally, the aim is to explore and develop the most motivational, learner centered methods, techniques and resources for engineering mathematics learning and teaching with the help of technology. All the learning resources developed in the project will be made available for free under the idea of Open Source or Open Educational Resource (OER).

The FutureMath project develops pedagogical methods and resources to teach and learn mathematics more effectively by providing personalized learning possibilities with the help of ubiquitous technology. The underlying notion is to support digitalization of European engineering mathematics education in a large scale. By these means, it is supposed to improve the efficiency, accessibility and quality of mathematics teaching and learning on European level which, in fact, is one of the four common objectives of EU's Strategic Framework of Education and Training 2020. Additionally, as an impact of the project, improving of transversal and basic skills (ET2020), such as digital skills and mathematical skills, will be a central focus. With these actions, it is expected not only to develop innovative learning approaches but also to enrich the teaching, support personalized learning and increase the flexibility and attractiveness.

In addition to the mathematics learning platform (MLP), the proposed project aims to develop innovative pedagogical methods, techniques, materials and resources not only to teach and learn mathematics but also to assess mathematics' learning. The key approaches while planning the resources are i.e. collective thinking, collaboration and shared problem solving skills - the skills that are necessary for success in working life. Furthermore project resources will respect individual learning solutions. Therefore, different learning types will be taken into account in the project's material production. In this way, it is also possible to decrease the inequality among different kinds of learners.
Overall, the one main objective of this project is to increase the global large-scale awareness about the possibilities ubiquitous technology offers for mathematics learning throughout MLP. Our aim is to make mathematics learning more motivational, interesting and increase accessibility and the alternative modern methods for mathematics learning. [1]

3. Conclusions

Using the experience from the FutureMath project, we can say that still, more important than if technology is used, is the way that computers are used in teaching and learning mathematics [1]. Therefore, teachers' ability to select appropriate software and materials, or to create their own materials, plays an essential role in the process of effectively and successfully integrate technologies into classroom teaching.
Current research reveals the positive effects of technology assisted teaching and learning, when this is done at its fullest potential. Therefore it is important for teachers to find a way to integrate technology into their classrooms [2].

References

[1] http://www.futuremath.eu/index.php/en/
[2] Kinnari-Korpela, H., Korpela, A.: Enhancing learning in engineering studies: experiences on short video lecturing. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2014 (1), pp. 2207–2216. Chesapeake, VA: AACE.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 59-61

ABOUT THE EXISTENCE, UNICITY AND NUMERICAL SOLVING FOR A NONLINEAR DIFFERENTIAL EQUATIONS SYSTEM

Lucian Niţă
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: lucian.s.nita@gmail.com
Iuliana Popescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail:iulianapopescu1@gmail.com

Abstract: In this paper we consider a nonlinear differential equations system of first order. One can see that the hypothesis for the (classical) theorem of existence and unicity (for first order differential equations system) are not fulfilled. In [1] is given a more general theorem for existence and unicity. We prove that this theorem can be used to get the existence and unicity for our system. Then, we use a numerical method to approximate the solution of the system in some point.
Mathematics Subject Classification (2010): 34A34, 65L06, 34A45
Key words: existence, unicity, Lipschitz condition, numerical solving.

1. Introduction

Theorem 1. Let

$$\begin{cases} y_1' = f_1(x, y_1, y_2) \\ y_2' = f_2(x, y_1, y_2) \end{cases} \tag{1}$$

be a differential equations system of first order, which fulfills the following conditions:

a) Let $a, b, c \in R_+^*$ fixed, $(x_0, y_{10}, y_{20}) \in R^3$ and $D \subseteq R^3$ defined by

$$D = \{(x, y, z) \in R^3 \mid |x - x_0| \le a, |y_1 - y_{10}| \le b, |y_2 - y_{20}| \le c\}. \tag{2}$$

The functions $f_1,\ f_2 : D \to R$ are continuous on D.

b) There exist $A > 0, B > 0$ such that for any $(x, y_{11}, y_{21}) \in D, (x, y_{12}, y_{22}) \in D$ we have

$$\begin{aligned} |f_1(x, y_{11}, y_{21}) - f_1(x, y_{12}, y_{22})| &\le A|y_{12} - y_{11}| + B|y_{22} - y_{21}| \\ |f_2(x, y_{11}, y_{21}) - f_2(x, y_{12}, y_{22})| &\le A|y_{12} - y_{11}| + B|y_{22} - y_{21}| \end{aligned} \tag{3}$$

(the Lipschitz condition).

Then, there exists a unique solution of the system: $y_1 = \varphi(x),\ y_2 = \psi(x), \varphi$ and ψ being derivable on an interval $|x - x_0| \le h,\ (h \le a)$ such that $y_{10} = \varphi(x_0),\ y_{20} = \psi(x_0)$.

Theorem 2. Let be a Banach space

$$y' = f(x, y) \tag{4}$$

a differential equation of first order which fulfills the following conditions:

a) Let $I \subseteq R$ be an interval. The function $f : I \times E \to E$ is continuous on $I \times E$

There exists $h \in L^1_{loc}(I),\ h \ge 0$ such that $\forall (x, y_1) \in I \times E,\ (x, y_2) \in I \times E$, we have

$$\| f(x, y_1) - f(x, y_2) \| \le h(x) \| y_1 - y_2 \|. \tag{5}$$

Let $(x_0, y_0) \in I \times E$.

Then, the problem: $y' = f(x, y),\ y(x_0) = y_0$ has a unique solution on I, which can be obtained as the uniform convergent limit (on any compact set included in I) of the Picard successive approximations sequence:

$$y(x_0) = y_0,\ y_{n+1}(x) = y_0 + \int_{x_0}^{x} f(s, y_n(s))\mathrm{d}s,\ n \geq 0,\ x \in I. \tag{6}$$

2. Application

Let $E = R^2,\ I = R,\ y = (y_1, y_2),\ f = (f_1, f_2)$,

$$f_1(x, y) = \cos x\sqrt{|y|},\ f_2(x, y) = 2\sqrt{2}\cos x\sqrt{|y|}. \tag{7}$$

Then

$$\|f(x, y_1) - f(x, y_2)\| \leq 3\sqrt{(\cos x\sqrt{|y_1|} - \cos x\sqrt{|y_2|})^2} \tag{8}$$

$$= 3\sqrt{4\sin^2\frac{x(\sqrt{|y_1|} - \sqrt{|y_2|})}{2}\sin^2\frac{x(\sqrt{|y_1|} + \sqrt{|y_2|})}{2}} \leq$$

$$\leq 3\sqrt{4\frac{x^2(\sqrt{|y_1|} - \sqrt{|y_2|})^2}{4}\frac{x^2(\sqrt{|y_1|} + \sqrt{|y_2|})^2}{4}} = 3\frac{x^2}{2}\|\,|y_1| - |y_2|\,\|.$$

If $y_1, y_2 \in E,\ y_1 = (a, b),\ y_2 = (c, d)$ then $|y_1| = \sqrt{a^2 + b^2},\ |y_2| = \sqrt{c^2 + d^2}$,

$$|y_1 - y_2| = \sqrt{(a-c)^2 + (b-d)^2}.$$

We prove that $\|\,|y_1| - |y_2|\,\| \leq |y_1 - y_2|$:

$$|\sqrt{a^2 + b^2} - \sqrt{c^2 + d^2}| \leq \sqrt{(a-c)^2 + (b-d)^2} \tag{9}$$

$$-2\sqrt{(a^2 + b^2)(c^2 + d^2)} \leq -2(ac + bd)$$

$$\sqrt{(a^2 + b^2)(c^2 + d^2)} \geq ac + bd.$$

If $ac + bd < 0$ the inequality holds.

If $ac + bd \geq 0$ the inequality is also true, using the Cauchy-Buniakowsky-Schwarz inequality.

Hence $\|f(x, y_1) - f(x, y_2)\| \leq 3\frac{x^2}{2}\|y_1 - y_2\|$.

We define $h : I \to R,\ h(x) = 3\frac{x^2}{2}$. One can see that has the properties needed in theorem 2.

We prove now that $f = (f_1, f_2)$ does not fulfill the Lipshitz condition, needed in theorem 1: let us suppose that $\exists A, B > 0$ such that

$$|f_1(x, y_1) - f_1(x, y_2)| \leq A|y_{12} - y_{11}| + B|y_{22} - y_{21}| \tag{10}$$

$\forall x \in I\ y_1, y_2 \in E,\ y_1 = (y_{11}, y_{21}),\ y_2 = (y_{12}, y_{22})$.

We suppose $A \geq B$.

Then, we should have:

$$|f_1(x, y_1) - f_1(x, y_2)| \leq A(|y_{12} - y_{11}| + |y_{22} - y_{21}|). \tag{11}$$

We will have

$$|f_1(x, y_1) - f_1(x, y_2)| = \sqrt{(\cos x\sqrt{|y_1|} - \cos x\sqrt{|y_2|})^2} = \tag{12}$$

$$|\cos x\sqrt{|y_1|}-\cos x\sqrt{|y_2|}|=2|\sin x\frac{(\sqrt{|y_1|}-\sqrt{|y_2|})}{2}\sin x\frac{(\sqrt{|y_1|}+\sqrt{|y_2|})}{2}|.$$

We take $x=\frac{2}{\sqrt{|y_2|}-\sqrt{|y_1|}}$, $|y_1|\neq|y_2|$ and (12) become

$$|f_1(x,y_1)-f_1(x,y_2)|=2\left|\sin\frac{(\sqrt{|y_1|}+\sqrt{|y_2|})^2(|y_1|+|y_2|)}{|y_2|^2-|y_1|^2}\sin 1\right|. \tag{13}$$

For $y_1=(1,1)$, $y_2=\left(1+\frac{1}{n},1+\frac{1}{n}\right)$, $n\in N^*$ we get

$$2\left|\sin\frac{(\sqrt{|y_1|}+\sqrt{|y_2|})^2(|y_1|+|y_2|)}{|y_2|^2-|y_1|^2}\sin 1\right|=2\left|\sin\frac{\left(1+\sqrt{1+\frac{1}{n}}\right)^2\left(2+\frac{1}{n}\right)}{\frac{1}{n^2}+\frac{2}{n}}\sin 1\right|= \tag{14}$$

$$=2\left|\sin\left[\frac{n^2}{2n+1}\left(1+\sqrt{1+\frac{1}{n}}\right)^2\left(2+\frac{1}{n}\right)\right]\sin 1\right|.$$

Using (11) we conclude that $\exists A>0$ such that

$$2\left|\sin\left[\frac{n^2}{2n+1}\left(1+\sqrt{1+\frac{1}{n}}\right)^2\left(2+\frac{1}{n}\right)\right]\sin 1\right|\leq A\frac{2}{n}. \tag{15}$$

For a large *n*, there exists $\exists\varepsilon>0$ small enough, such that:

$$2\left|\sin\left[\frac{n^2}{2n+1}\left(1+\sqrt{1+\frac{1}{n}}\right)^2\left(2+\frac{1}{n}\right)\right]\sin 1\right|>\varepsilon. \text{ and } A\frac{2}{n}<\varepsilon. \tag{16}$$

We deduce that f_1 does not fulfills the Lipschitz condition.

Using theorem 2, the differential equation $y'=f(x,y)$ (with a condition $y_0=y(x_0)$) has a unique solution.

3. Numerical application

For $x_0=0$, $y_{10}=1$, $y_{20}=0$ using the Runge-Kutta of the 4th order method, we obtained the results

Step size, h	$y_1(1)$	$y_2(1)$
0.01	1.6632079	1.8758352
0.001	1.6626662	1.8743029
0.0001	1.6626124	1.874151

References

[1] Rădulescu, S., Rădulescu, M.: *Teoreme şi probleme de analiză matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1982

[2] Toma, I., Mosnegutu, V., Constantinescu, S.: *Analyse Mathématique*, Editura Conspress, Bucureşti, 2014.

[3] Saucez, Ph., Wouwer, A.V., Vilas, C.: *Simulation of ODE/PDE Models with MATLAB®, OCTAVE and SCILAB*, Springer International Publishing, 2014

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 62-64

CONVERGENCE PROPERTIES FOR THE ATTRACTORS ASSOCIATED TO A SEQUENCE OF ITERATED FUNCTION SYSTEMS

Lucian Niţă
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: lucian.s.nita@gmail.com
Daniel Tudor
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: danieltudor@gmail.com

Abstract: In this paper we consider a sequence of iterated function systems, defined and taking values in a Banach space. The sequence is built using a system of contractions and a sequence of linear and bounded operators (in the same Banach space), that converges (in operatorial norm) to a linear and bounded operator. For each iterated function system we have an attractor. We obtain a sequence of attractors. The problem that we solve is if this sequences of attractors converge in the Hausdorff-Pompeiu metric.
Mathematics Subject Classification (2010): 28C20, 4GG12
Key words iterated function system, attractor, Hausdoff-Pompeiu metric

1. Preliminary Facts

In this section, we will recall some definitions and results regarding the Hausdorff-Pompeiu metric, iterated function system and the attractor associated to an iterated function system (I.F.S.) on a complete metric space. For more details, one can consult [1].

Let (X,d) be a metric space. We denote by $\mathcal{P}^*(X)$ the class of non-empty and bounded subsets of X. For $x \in X, A \in \mathcal{P}^*(X)$, the distance between x and A is:

$d(x,A) = \inf\{d(x,y) / y \in A\}$. For $A, B \in \mathcal{P}^*(X)$, we denote $d(A,B) = \sup_{x \in A} d(x,B)$, and $d(B,A) = \sup_{y \in B} d(y,A)$. We define $\delta(A,B) := \max\{d(A,B), d(B,A)\}$. Let us denote by $\mathcal{K}(X)$ the class of non-empty and compact subsets of X.

Theorem 1. *a)* $\delta : \mathcal{K}(X) \times \mathcal{K}(X) \to [0,\infty)$ *is a metric on* $\mathcal{K}(X)$;

b) If $\omega : X \to X$ *is a Lipschitz function, L being its Lipschitz constant, then* $\delta(\omega(A), \omega(B)) \le L \cdot \delta(A,B), \forall A, B \in \mathcal{K}(X)$;

c) If $(E_i)_{1 \le i \le n} \subset \mathcal{K}(X)$, $(F_i)_{1 \le i \le n} \subset \mathcal{K}(X)$ *then* $\delta(\bigcup_{i=1}^{n} E_i, \bigcup_{i=1}^{n} F_i) \le \max_{i \in \overline{1,n}} \delta(E_i, F_i)$.

Definition 1. *The metric* δ *from Theorem 1 is called the Hausdorff-Pompeiu metric.*

Theorem 2. *i) If* (X,d) *is complete, then* $(\mathcal{K}(X), \delta)$ *is also complete;*

ii) If (X,d) *is compact, then* $(\mathcal{K}(X), \delta)$ *is also compact.*

Definition 2. *Let* (X,d) *be a complete metric space and* $\omega_i : X \to X, i \in \overline{1,n}$ *contractions of ratio* $r_i \in (0,1)$. *The family* $(\omega_i)_{1 \le i \le n}$ *is called iterated function system (I.F.S.).*

For an I.F.S. $(\omega_i)_{1\le i\le n}$, on a complete metric space (X,d), we define $S:\mathcal{K}(X)\times\mathcal{K}(X)$, $S(A)=\bigcup_{i=1}^{n}\omega_i(A), A\in\mathcal{K}(X)$.

Theorem 3. *S is a contraction (on the complete metric space $(\mathcal{K}(X),\delta)$) of ratio $r\le\max_{1\le i\le n} r_i$. Consequently, there exists an unique set $K\in\mathcal{K}(X)$ such that* $K=S(K)=\bigcup_{i=1}^{n}\omega_i(K)$.

Definition 3. *The set K from Theorem 3 is called the attractor associated to the I.F.S.. $(\omega_i)_{1\le i\le n}$.*

2. Results obtained

Now we consider a Banach space $(X,\|\ \|)$ and we denote: $\mathcal{L}(X)=\{T:X\to X / T \text{ is linear and continuous}\}$.

Lemma 1. *Let $\omega:X\to X$ be a contraction of ratio r and $T\in\mathcal{L}(X)$ such that $\|T\|+r<1$. Then, the application $\omega^T:X\to X, \omega^T=T+\omega$ is a contraction on $(X,\|\ \|)$.*

Proof. For $x,y\in X$, we have: $\|\omega^T(x)-\omega^T(y)\|=\|T(x)-T(y)+\omega(x)-\omega(y)\|\le$
$\le\|T(x-y)\|+\|\omega(x)-\omega(y)\|\le(\|T\|+r)\|x-y\|<\|x-y\|$ (according to $\|T\|+r<1$, the last inequality is true).

Lemma 2. *Let $\omega:X\to X$ be a contraction of ratio r and $(T_j)_{j\ge1}\subset\mathcal{L}(X)$ such that $\sup_{j\ge1}\|T_j\|+r<1$, and $T_0\in\mathcal{L}(X)$ such that $\lim_{j\to\infty}\|T_j-T_0\|=0$. Then, $\forall\varepsilon>0,\forall K\in\mathcal{K}(X)$, $\exists j_0\in\mathbf{N}^*$ such that $\forall j\ge j_0, \delta(\omega^{T_j}(K),\omega^{T_0}(K))<\varepsilon$, that means $\omega^{T_j}(K)\xrightarrow[j\to\infty]{}\omega^{T_0}(K)$.*

Proof. Let $K\in\mathcal{K}(X)$, arbitrary, fixed. $\omega^{T_j}(K)=\{T_j(x)+\omega(x)/x\in K\}$,
$\omega^{T_0}(K)=\{T_0(x)+\omega(x)/x\in K\}$. Let $y\in\omega^{T_j}(K)$; hence, there is $a\in K$ such that
$y=T_j(a)+\omega(a)$; $d(y,\omega^{T_0}(K))=\inf\{\|T_j(a)+\omega(a)-(T_0(b)+\omega(b))\|/b\in K\}\le$
$\le\inf\{\|T_j(a)-T_0(a)\|+\|T_0(a)-T_0(b)\|+\|\omega(a)-\omega(b)\|/b\in K\}\le$
$\le\inf\{\|T_j-T_0\|\cdot M+(\|T_0\|+r)\|a-b\|/b\in K\}$, where $M=\max_{a\in K}\|a\|$. Taking
$b=a\in K$, we deduce that $d(y,\omega^{T_0}(K))\le\|T_j-T_0\|\cdot M$. But $\lim_{j\to\infty}\|T_j-T_0\|=0$, so, for $\varepsilon>0$,
we find $j_0\in\mathbf{N}^*$ such that: $\forall j\ge j_0,\|T_j-T_0\|\le\frac{\varepsilon}{M}$. Then, $d(y,\omega^{T_0}(K))\le\varepsilon$,
$\forall y\in\omega^{T_j}(K),\forall j\ge j_0$. It results: $d(\omega^{T_j}(K),\omega^{T_0}(K))\le\varepsilon$. Similarly,
$d(\omega^{T_0}(K),\omega^{T_j}(K))\le\varepsilon$, so $\delta(\omega^{T_j}(K),\omega^{T_0}(K))<\varepsilon$.

Let us consider (A,d), where $A\in\mathcal{K}(X)$, $d(x,y)=\|x-y\|,\forall x,y\in A$. Hence, (A,d) is complete and, from Theorem 2, (a), $(\mathcal{K}(A),\delta)$ is complete. (we denoted δ the restriction of δ on $\mathcal{K}(A)$). We consider now an I.F.S. $(\omega_i)_{1\le i\le n}$ on (A,d), and let $(T_j)_{j\ge1}\subset\mathcal{L}(X)$ such that

$\max_{1\le i\le n} r_i + \sup_{j\ge 1}\|T_j\| < 1$ (for example, if $r = \max_{1\le i\le n} r_i < 1$, we may take: $T_j(x) = \frac{j}{j+1}\cdot\frac{1-r}{2}\cdot x$, $\forall x \in X$ and we have $\|T_j\| = \frac{j}{j+1}\cdot\frac{1-r}{2}$, $\sup_{j\ge 1}\|T_j\| = \frac{1-r}{2} < 1-r$, so, $r + \sup_{j\ge 1}\|T_j\| < 1$).

As before, we denote $\omega_i^{T_j} = \omega_i + T_j$. Then, (according Lemma 1) $\omega_i^{T_j}$ is a contraction on (A,d); hence, for a fixed j, . $(\omega_i^{T_j})_{1\le i\le n}$ is an I.F.S. on (A,d). Using Theorem 3, $\exists! K^j = \bigcup_{i=1}^{n} \omega_i^{T_j}(K^j)$ (the attractor associated to the I.F.S. $(\omega_i^{T_j})_{1\le i\le n}$). Now, we suppose that $\exists T_0 \in \mathcal{L}(X)$ such that $\lim_{j\to\infty}\|T_j - T_0\| = 0$ and we denote by $\omega_i^{T_0} = \omega_i + T_0$; obviously, $r_{T_0} := \max_{1\le i\le n} r_i + \|T_0\| < 1$. Let $K = \bigcup_{i=1}^{n} \omega_i^{T_0}(K)$ (the attractor associated to the I.F.S. $(\omega_i^{T_0})_{1\le i\le n}$).

Theorem 4. *We have* $\lim_{j\to\infty} \delta(K^j, K) = 0$.

Proof. $\delta(K^j, K) = \delta(\bigcup_{i=1}^{n} \omega_i^{T_j}(K^j), \bigcup_{i=1}^{n} \omega_i^{T_0}(K)) \le \delta(\bigcup_{i=1}^{n} \omega_i^{T_j}(K^j), \bigcup_{i=1}^{n} \omega_i^{T_0}(K^j)) +$

$+ \delta(\bigcup_{i=1}^{n} \omega_i^{T_0}(K^j), \bigcup_{i=1}^{n} \omega_i^{T_0}(K))$. Let $\varepsilon > 0$ arbitrary, fixed. From Theorem 1(c), we have

$\delta(\bigcup_{i=1}^{n} \omega_i^{T_j}(K^j), \bigcup_{i=1}^{n} \omega_i^{T_0}(K^j)) \le \max_{1\le i\le n} \delta(\omega_i^{T_j}(K^j), \omega_i^{T_0}(K^j) < \varepsilon$ (here we use the fact that (A,d) is compact, so, if $M_j = \max_{a\in K^j}\|a\|$, as in the proof of Lemma 2, then $\exists M > 0$ such that $\sup_{j\ge 1} M_j \le M$). On the other hand, $\delta(\bigcup_{i=1}^{n} \omega_i^{T_0}(K^j), \bigcup_{i=1}^{n} \omega_i^{T_0}(K)) \le \max_{1\le i\le n}(\delta(\omega_i^{T_0}(K^j), \omega_i^{T_0}(K)) \le$

$\le \max_{1\le i\le n}[r_i^{T_0}\delta(K^j, K)]$. For the last inequalities we used Theorem 1(b,c). Hence, for a large enough $j \in \boldsymbol{N}^*$, we get: $\delta(K^j, K) \le r^{T_0}(K^j, K) + \varepsilon$, that is: $(1 - r^{T_0})\delta(K^j, K) \le \varepsilon$. We conclude that $\lim_{j\to\infty} \delta(K^j, K) = 0$.

Remark. We supposed that the sequence $(\omega_i^{T_j})_{1\le i\le n}$ on (A,d) has taken values only in *A*.

References

[1] Secelean, N.: *Masura si Fractali,* Editura Universitatii Lucian Blaga din Sibiu, 2002.
[2] Niţă, L., Tudor., D.: *Dependence of the Hutchinson vector measure with respect to a parameter,* The 14th Workshop of Scientific Communications, Department of Mathematics and Computer Science, 2017.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 65-68

THALES THEOREM OVER AN ARBITRARY FIELD

Sever Angel Popescu
Department of Mathematics and Computer Science,
Technical University of Civil Engineering Bucharest, Romania
E-mail: angel.popescu@gmail.com

Abstract: We state and prove a version of the famous Thales theorem over an arbitrary field. Consequently, we give a new proof for the classical geometrical version of the same theorem over the real number field.
Mathematics Subject Classification (2000): 01-01, 12-01, 14P99.
Key words: Geometry over arbitrary fields, Thales theorem, ordered fields, non-Archimedean fields.

1. Some generalities on geometry over an arbitrary field

The idea of doing geometry over an arbitrary field also appears in [2], Ch. 3. Maybe it is living even from the beginning of the XX century. But here I use my own point of view on definitions and notation.

Let K be an arbitrary field and let "∞" be a symbol which is not in K, such that $a/0=\infty, a\cdot\infty=\infty, a\neq 0, a\in K$. Let us denote $(P)=K\times K$ the *plane* associated to K [2]. Thus, a point M in *(P)* is simply an ordered pair $(x,y)\in K\times K$ and we write $M(x,y)$ and say "*M of coordinates x and y*". A *straight line* in *(P)* is a subset of *(P)* of the form: $(d)=\{(x,y)\in(P): ax+by+c=0\}$, where $a,b,c\in K$ are three fixed elements of K such that $a\neq 0, or\ b\neq 0$. We can simply write: $(d): ax+by+c=0$. We say that two straight lines $(d_1): a_1x+b_1y+c_1=0,$ and $(d_2): a_2x+b_2y+c_2=0$ are *parallel lines* if $a_1b_2=a_2b_1$, or, equivalently, $\frac{a_2}{b_2}=\frac{a_1}{b_1}, in\ K\cup\{\infty\}$. Let $\left(M_0(x_0,y_0), M_1(x_1,y_1), M_2(x_2,y_2)\right)$ be three points in (P). Let us define two *ratio functions*, $R_l, R_r: (P)\times(P)\times(P)\to K\cup\{\infty\}$, one "on the left" and one "on the right":

(1) $R_l(M_0,M_1,M_2)=\frac{x_1-x_0}{x_2-x_0}$, the *left ratio function*, and

(2) $R_r(M_0,M_1,M_2)=\frac{y_1-y_0}{y_2-y_0}$, the *right ratio function.*

It is clear that the order in the triplet (M_0,M_1,M_2) is essential. For instance,

$$R_l(M_0,M_1,M_2)=\frac{1}{R_l(M_0,M_2,M_1)}.$$

Remark 1. It is easy to prove that for any two distinct points $H_1(t_1,u_1), H_2(t_2,u_2)\in(P)$, there exists a unique straight line (d) in (P) such that $H_1, H_2\in(d)$ and its "equation" is:

(6) $(d): \frac{x-t_1}{t_2-t_1} = \frac{y-u_1}{u_2-u_1}$ $(in\ K \cup \{\infty\})$. We denote it by (H_1H_2). Here $(d): ax+by+c$, where $a = u_2 - u_1, b = t_1 - t_2, c = u_1t_2 - u_2t_1$. In order to prove the uniqueness it is sufficient to prove that two parallel straight lines with a common point are identical.

It is not difficult to prove the following three results.

Proposition 1. Let $M_0, M_1, M_2 \in (P)$ be three distinct points in the fixed plane *(P)*. Then the following three statements are equivalent.

i) $M_0 \in (M_1M_2)$

ii) $M_1 \in (M_0M_2)$

iii) $M_2 \in (M_0M_1)$.

If one of these statements is true, then we say that M_0, M_1, M_2 are *collinear*.

Proposition 2 M_0, M_1, M_2 are collinear if and only if $R_l(M_0, M_1, M_2) = R_r(M_0, M_1, M_2)$.

Corollary 3 The equality $R_l(M_0, M_1, M_2) = R_r(M_0, M_1, M_2)$ does not depend on the order of the points M_0, M_1, M_2 in spite of the fact that the definitions of R_l, R_r depends on the order of the point M_0, M_1, M_2.

2. The abstract Thales theorem

Let now $(d_1): a_1x + b_1y + c_1 = 0$ and $(d_2): a_2x + b_2y + c_2 = 0$ be two non-parallel straight lines, i.e. $a_1b_2 \neq a_2b_1$ and let $M_0(x_0, y_0)$ be their intersection point. Thus, (x_0, y_0) is the unique solution of the linear system:

(3) $\begin{cases} a_1x + b_1y + c_1 = 0 \\ a_2x + b_2y + c_2 = 0 \end{cases}$, i.e. $\begin{cases} a_1x_0 + b_1y_0 + c_1 = 0 \\ a_2x_0 + b_2y_0 + c_2 = 0 \end{cases}$.

Let $M_1(x_1, y_1), M_2(x_2, y_2)$ be other two distinct points on (d_1) such that $M_1 \neq M_0, M_2 \neq M_0$. Thus,

(4) $\begin{cases} a_1x_1 + b_1y_1 + c_1 = 0 \\ a_1x_2 + b_1y_2 + c_1 = 0 \end{cases}$.

Let $N_1(z_1, w_1), N_2(z_2, w_2)$ be also other two distinct points on (d_2), i.e.

(5) $\begin{cases} a_2z_1 + b_2w_1 + c_2 = 0 \\ a_2z_2 + b_2w_2 + c_2 = 0 \end{cases}$, such that $N_1 \neq M_0, N_2 \neq M_0$.

Theorem 4 Let us consider the above notation and assumptions to be available here. Then the following statements are logically equivalent:

i) The straight lines (M_1N_1) and (M_2N_2) are parallel lines.

ii) $R_l(M_0, M_1, M_2) = R_l(M_0, N_1, N_2)$.

iii) $R_r(M_0, M_1, M_2) = R_r(M_0, N_1, N_2)$.

Proof. First of all, it is easy to see that if M_0, M_1, M_2 belongs to one and the same straight lines, then $R_l(M_0, M_1, M_2) = R_r(M_0, M_1, M_2)$. Thus ii) $\Leftrightarrow$ iii). Let us prove now that i) $\Leftrightarrow$ ii). Let us look at the "equations" of the two straight lines:

(7) $(M_1N_1): \dfrac{x - x_1}{z_1 - x_1} = \dfrac{y - y_1}{w_1 - y_1}$

(8) $(M_2N_2): \dfrac{x - x_2}{z_2 - x_2} = \dfrac{y - y_2}{w_2 - y_2}$.

Thus, (M_1N_1) and (M_2N_2) are parallel lines if and only if

(9) $\dfrac{w_1 - y_1}{z_1 - x_1} = \dfrac{w_2 - y_2}{z_2 - x_2}$.

For convenience, let us assume $b_1 \neq 0, b_2 \neq 0$. We write (9) as:

(10) $\dfrac{(w_1 - y_0) - (y_1 - y_0)}{(z_1 - x_0) - (x_1 - x_0)} = \dfrac{(w_2 - y_0) - (y_2 - y_0)}{(z_2 - x_0) - (x_2 - x_0)}$

By using the obvious equalities (see (3), (4), (5)):

(11) $w_1 - y_0 = -\dfrac{a_2}{b_2}(z_1 - x_0),\ w_2 - y_0 = -\dfrac{a_2}{b_2}(z_2 - x_0)$ and

(12) $y_1 - y_0 = -\dfrac{a_1}{b_1}(x_1 - x_0),\ y_2 - y_0 = -\dfrac{a_1}{b_1}(x_2 - x_0)$, we find:

(13) $\dfrac{-\dfrac{a_2}{b_2}(z_1 - x_0) + \dfrac{a_1}{b_1}(x_1 - x_0)}{(z_1 - x_0) - (x_1 - x_0)} = \dfrac{-\dfrac{a_2}{b_2}(z_2 - x_0) + \dfrac{a_1}{b_1}(x_2 - x_0)}{(z_2 - x_0) - (x_2 - x_0)}$, or

(14) $\left[\dfrac{a_1}{b_1} - \dfrac{a_2}{b_2}\right]\left[(z_1 - x_0)(x_2 - x_0) - (x_1 - x_0)(z_2 - x_0)\right] = 0.$ Since $(d_1), (d_2)$ are not parallel lines, (9) is equivalent to $(z_1 - x_0)(x_2 - x_0) = (x_1 - x_0)(z_2 - x_0)$, or to

(15) $\dfrac{x_1 - x_0}{x_2 - x_0} = \dfrac{z_1 - x_0}{z_2 - x_0}$. Thus, (9) is equivalent to

(16) $R_l(M_0, M_1, M_2) = R_l(M_0, N_1, N_2)$, i.e. with ii) from the statement of the theorem, and the proof is over.

3. The classical version of Thales theorem

Let us recover the classical Thales theorem from Theorem 4. With the above notation and assumptions, we finally have to prove that for *K*=**R**, the real numbers field, the equality (15) is equivalent to the following metric equality:

(17) $\dfrac{\|M_0M_1\|^2}{\|M_0M_2\|^2} = \dfrac{\|M_0N_1\|^2}{\|M_0N_2\|^2}$, where $\|M_0M_1\| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$ is the length of the segment $[M_0M_1]$ in the usual Cartesian plane *xOy*. Let us write (17) in the usual language of the Plane Analytical Geometry:

(18) $\dfrac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{(x_2 - x_0)^2 + (y_2 - y_0)^2} = \dfrac{(z_1 - x_0)^2 + (w_1 - y_0)^2}{(z_2 - x_0)^2 + (w_2 - y_0)^2}$ and use formulas (11) and (12) to find

(19) $\dfrac{(x_1 - x_0)^2}{(x_2 - x_0)^2} = \dfrac{(z_1 - x_0)^2}{(z_2 - x_0)^2}$.

We know that in the usual Plane Analytical Geometry one can consider the Cartesian coordinate plane xOy in an arbitrary position. Let us take it such that $O \equiv M_0$ and the straight line $(d_1) \equiv Ox - axis$, in the direction of $\overrightarrow{OM_1}$. It is easy to see that in this last case we have: $|x_1 - x_0| = x_1 - x_0$, $|x_2 - x_0| = x_2 - x_0$ and the differences $(z_1 - x_0)$, $(z_2 - x_0)$ have the same sign. Thus, the equality (19) is equivalent to the equality (15). Thus, the proof of the classical Thales theorem is over.

4. Some final remarks

Remark 2. In the above note we did not assume that the field *K* is an ordered field because we did not need the axiom of betweeness (see [2]). Even if *K* were an ordered field, we also supplied a version of Thales theorem over *K*, without assuming that *K* is an Archimedean field [we say that an ordered field $(K, \prec)$ is Archimedean if for any $x \in K$, there is n, a natural number, such that $x \prec n$]. Thus, Thales theorem can also belong to non-Archimedean Geometry.

Here is an example of a non-Archimedean ordered field.

Example 1. ([2], p. 159) Let **R** be the ordered field of real numbers and let **F** = **R**(t) be the field of rational functions over **R**. We say that

$$\frac{f(t)}{g(t)} = \frac{a_n t^n + a_{n-1} t^{n-1} + ... + a_0}{b_m t^m + b_{m-1} t^{m-1} + ... + b_0} \succ 0, a_n \neq 0, b_m \neq 0, \text{ if } \frac{a_n}{b_m} > 0 \text{ as a real number.}$$

It is not difficult to prove that $(\mathbf{R}(t), \succ)$ is an ordered field in which one has the following relations:

(20) $0 \prec 1 \prec 2 \prec ... \prec n \prec ... \prec t \prec t+1 \prec t+2 \prec ... \prec t^2 \prec t^3 \prec ...$

Since $n \prec t, \forall n$ a natural number, we see that $(\mathbf{R}(t), \succ)$ is not an Archimedean field. In spite of this one can state and prove a Thales theorem over it.

Remark 3. Over $F_2 = \frac{Z}{2Z}$ we cannot state Thales theorem because in its associated plane we have only four points.

References

[1] Artin, M.: *Algebra*, Prentice-Hall, Inc; New Jersey, 1991.
[2] Hartshorne, R.: *Geometry, Euclid and Beyond*, Springer, 1997.

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 69-74

EIGENVALUES, EIGENVECTORS AND THE DIAGONALIZATION OF A MATRIX WITH MATLAB, MATHCAD AND R SOFTWARE

Alina Elisabeta Sandu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: alina.sandu@utcb.ro
Gabriela-Roxana Dobre
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: roxana.dobre@utcb.ro

Abstract: In the study of the matrices a special place is hold by diagonal square matrices. One of the reasons is that they can be easily power up. So to get a square matrix in to the diagonal shape it is very important. To do such thing an essential part it is to find the eigenvalues and the eigenvectors of a square matrix. The eigenvalues of a matrix can be found by solving the characteristic equation, but this equation can have big orders for matrices with order bigger than 4, so solving the characteristic equation can be very difficult.
So, there are some software packages like **MatLAB**, **R** or **MathCad** which can help finding the eigenvalues, eigenvectors and even the diagonal form for a square matrix. In this paper we will explain how every program find eigenvalues and eigenvectors for a square matrix, we will describe the functions there are used to do that, and there are also be presented some examples for matrices of order bigger than 4 to find the eigenvalues, eigenvectors and the diagonal form, which can be very difficult for students to calculate other than with software packages. A problem from National Mathematical Countest "Traian Lalescu" will be resolved in the last part, using the presented software packages.
Mathematics Subject Classification (2010): 15xx, 65xx,
Key words: eigenvalues, eigenvector, diagonalization of a matrix, R software package, MatLAB, MathCAD.

1. Introduction

In this paper it will be presented how to calculate the eigenvalues, the eigenvectors and the diagonal form for a square matrix, using the software packages **MatLAB**, **R** and **MathCAD**. In the final part it will be resolved a problem from National Mathematical Countest "Traian Lalescu", edition 2012, also using the software packages **MatLAB**, **R** and **MathCAD**.
To calculate the eigenvalues and the eigenvectors for a square matrix A of order n, it is necessary to resolve the characteristic equation: $\det(A-\lambda\cdot I_n)=0$, which can be very difficult for order n bigger then 3. So, the software packages can make it very easy for finding the eigenvalues and the eigenvectors for a matrix A of order n, bigger then 3.

2. Using MatLAB

For calculate the eigenvalues, the eigenvectors and the diagonal form for a square matrix using the software **MatLAB**, it is necessary to make the following steps:

- define the square matrix A;
- use the command *disp(eig(A))* to find the eigenvalues of matrix *A;*
- for determinate the eigenvectors and the diagonal form of matrix A it will be used the comand *[X,L]=eig(A),* where X is the square matrix that conteined the eigenvectors

and L is the diagonal form matrix, a square matrix of the same order as matrix A, which has on his diagonal the eigenvalues of A.

Here is an example of a square matrix A of order 4 and how can be easily find in **MatLAB** his eigenvalues, eigenvectors and the diagonal form:

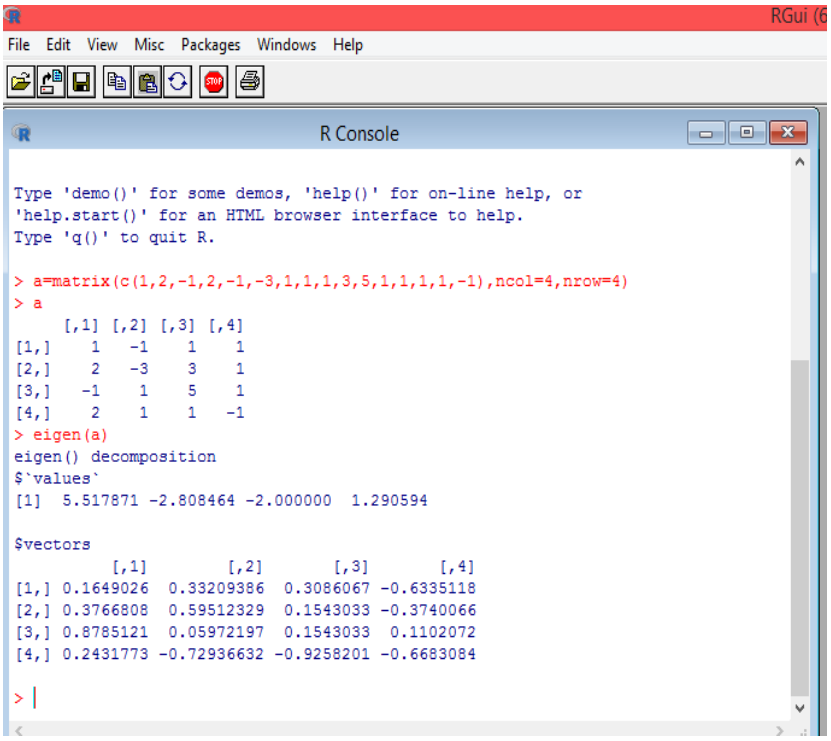
```
>> syms A
>> A=[1 -1 1 1;2 -3 3 1;-1 1 5 1;2 1 1 -1]

A =

     1    -1     1     1
     2    -3     3     1
    -1     1     5     1
     2     1     1    -1

>> disp(eig(A))
   -2.8085
   -2.0000
    1.2906
    5.5179

>> [X,L]=eig(A)

X =

    0.3321    0.3086   -0.6335    0.1649
    0.5951    0.1543   -0.3740    0.3767
    0.0597    0.1543    0.1102    0.8785
   -0.7294   -0.9258   -0.6683    0.2432

L =

   -2.8085         0         0         0
         0   -2.0000         0         0
         0         0    1.2906         0
         0         0         0    5.5179
```

3. Using R software

For calculate the eigenvalues, the eigenvectors for a square matrix using the software **R**, it is necessary to make the following steps:

- define the square matrix A, using the comand matrix( );
- use the command *eigen(A)* to find the eigenvalues and the eigenvectors of matrix A.

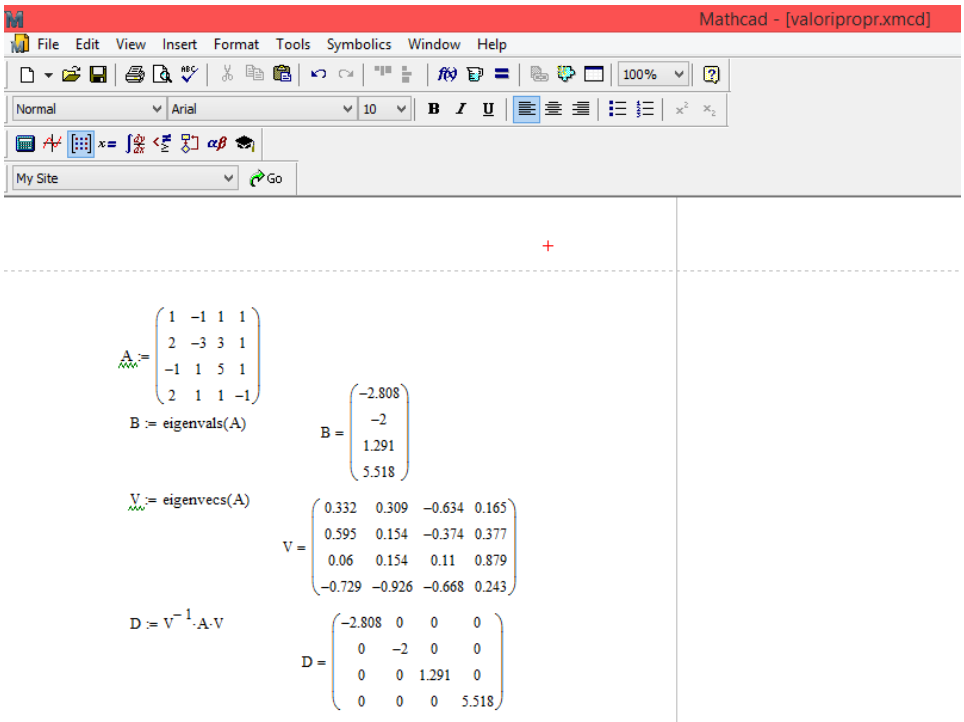
Here is an example of a square matrix A of order 4 and how can be easily find using **R** software his eigenvalues and the eigenvectors:



4. Using MathCAD

To determinate the eigenvalues, the eigenvectors and the diagonal form for a square matrix using the software **MathCAD**, it is necessary to make the following steps:

- define the square matrix A;
- use the comand *eigenvals(A)* to find the eigenvalues of the matrix *A;*
- use the comand *eigenvecs(A)* to find the eigenvectors of the matrix *A;*
- for determinate de diagonal form of the matrix A, first note with V the matrix of the eigenvectors, then make the operation $V^{-1} \cdot A \cdot V$.



5. Resolving a problem from Mathematical Contest "Traian Lalescu" phase local TUCEB, April 2012 using MatLAB, MathCAD and R software

Let be $T:\mathbf{R}^3\to\mathbf{R}^3$ a linear transformation with asociated matrix in canonical base of $\mathbf{R}^3$, the following matrix:

$$A=\begin{pmatrix}\frac{2}{3} & \frac{1}{3} & \frac{2}{3}\\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3}\\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3}\end{pmatrix}.$$

Find the eigenvalues of the linear transformation T and the associeted eigenvectors. Calculate A^{2012}.

(i) Solving the problem using **MatLAB:**

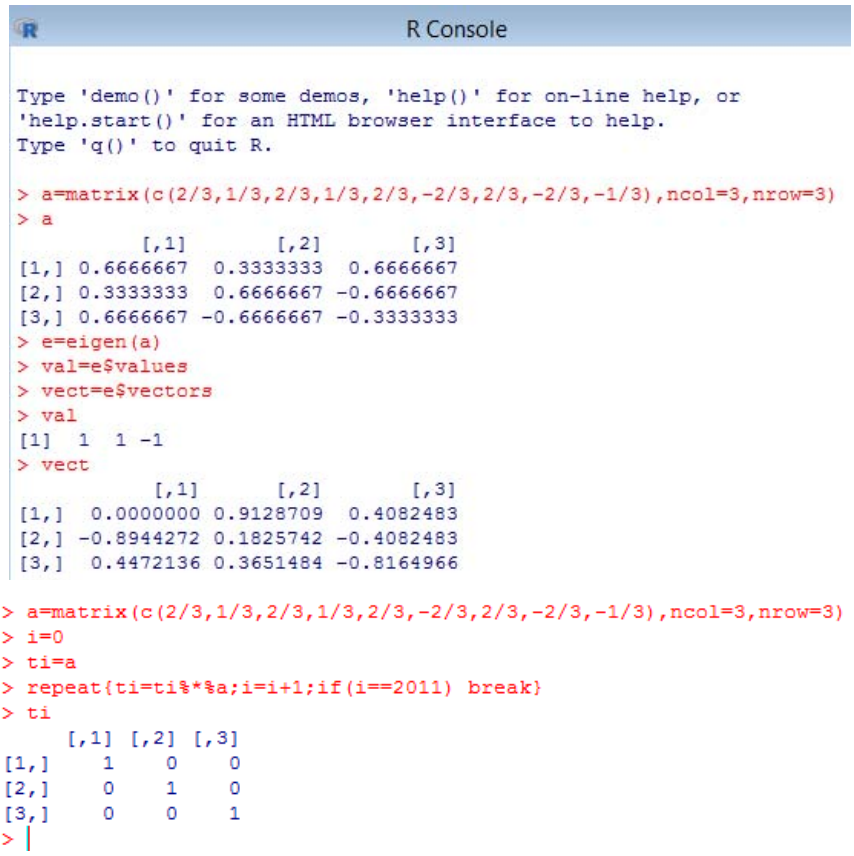
- Define the matrix *A* ;
- Find the eigenvalues of the matrix *A* using the comand *disp(eig(A))*;
- Find the eigenvectors and the diagonal form of the matrix *A* using the comand *[X,L]=eig(A)*; notate with *X* the matrix of the eigenvectors and with *L* the diagonal form of the matrix *A*;
- Using the comand $X*L\text{^}2012*X\text{^}-1$, we will find A^{2012}.

```
>> syms A
>> A=[2/3 1/3 2/3;1/3 2/3 -2/3;2/3 -2/3 -1/3]

A =

    0.6667    0.3333    0.6667
    0.3333    0.6667   -0.6667
    0.6667   -0.6667   -0.3333

>> disp(eig(A))
   -1.0000
    1.0000
    1.0000

>> [X,L]=eig(A)

X =

    0.4082   -0.5774    0.7071
   -0.4082    0.5774    0.7071
   -0.8165   -0.5774         0
```

```
X =

    0.4082   -0.5774    0.7071
   -0.4082    0.5774    0.7071
   -0.8165   -0.5774         0


L =

   -1.0000         0         0
         0    1.0000         0
         0         0    1.0000

>> X*L^2012*X^-1

ans =

    1.0000    0.0000   -0.0000
    0.0000    1.0000    0.0000
   -0.0000    0.0000    1.0000
```

(ii) Solving the problem using **R** software:

- Define the matrix *a* ;
- Find the eigenvalues and the eigenvectors of the matrix *a* using the comand *eigen(a)* ;
- For calculate a^{2012} we will use the programing skills of the **R** software: the alghoritm is: initiate *i=0 and ti=a,* then using the comand *repeat* multiple *ti*a, i=i+1* until *i=2011.* We obtain a^{2012}.



(iii)Solving the problem using **MathCAD**:

- Define the matrix A ;
- Find the eigenvalues of the matrix A using the comand *eigenvals(A)*;
- Find the eigenvectors of the matrix A using the comand *eigenvecs(A)*; notate C the matrix of the eigenvectors;
- Calculate the diagonal form of the matrix A, using $C^{-1} \cdot A \cdot C$
- Use the comand A^{2012} and verify the result with $C \cdot D^{2012} \cdot C^{-1}$.

$$A := \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{-2}{3} \\ \frac{2}{3} & \frac{-2}{3} & \frac{-1}{3} \end{pmatrix}$$

$$\text{eigenvals}(A) = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \qquad \text{eigenvecs}(A) = \begin{pmatrix} 0.913 & 0.408 & -0.044 \\ 0.183 & -0.408 & -0.902 \\ 0.365 & -0.816 & 0.429 \end{pmatrix}$$

$$C := \text{eigenvecs}(A)$$

$$D := C^{-1} \cdot A \cdot C \qquad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$A^{2012} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$C \cdot D^{2012} \cdot C^{-1} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

References

[1] Sandu, A.E.: *Algebra liniara, geometrie analitica si diferentiala*, *Note de curs şi aplicaţii*, Editura Conspress, Bucuresti, ISBN 978-973-100-416-7, 2016.
[2] Breaz, N., Craciun, M., Gaspar, P., Miroiu, M., Paraschiv-Munteanu, I.: *Modelare matematica prin Matlab*, ed. StudIS, ISBN:978-606-624-303-2, 2013.
[3] Miroiu, M., Petrehus, V., Zbaganu, G.: *Initiere in R,* ed. StudIS, ISBN:978-606-624-304-9
[4] Gavrilă, C., Petrehuş, V., Teodorescu, N., Nartea, C., Popescu, I., Sandu, A.E.: *Mathcad, aplicaţii, modelare şi simulare,* Editura Conspress, ISBN 978-973-100-356-6, 2014.
[5] R Development Core Team: R: A Language and Environment for Statistical Computing, 2014, http://www.R-project.org
[6] The Comprehensive R Archive Network:
https://cran.r-project.org/web/packages/hydroPSO/
http://cran.r-project.org/ web/packages/HydroMe/
[7] Rstudio
https://www.rstudio.com/

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 75-81

GIVENS ROTATIONS AND THE QR ALGORITHM FOR THE EIGENVALUE PROBLEM

Daniel Tudor
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: danieltudor@gmail.com
Dan Caragheorgheopol
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
and
"Ilie Murgulescu" Institute of Physical Chemistry of the Romanian Academy
Splaiul Independentei 202, RO-060021 Bucharest, Romania
E-mail: dancaraghe@gmail.com

Abstract Givens rotations are plane rotations that are often used in numerical linear algebra to create zeros in a matrix in a selective manner. This further allows one to obtain the QR decomposition of a matrix *A* and further on, to apply the QR algorithm to find the eigenvalues of *A* and, if *A* is symmetric, its eigenvectors. This method has the advantage of numerical stability, relatively low number of operations, if carefully implemented, and it allows further optimization in case of special matrices such as, e.g., tridiagonal or Hessenberg.

We discuss in this paper possible optimizations for the QR algorithm based on Givens rotations for general and symmetric matrices and provide the Mathcad programs for the optimized implementation.
Mathematics Subject Classification (2010): 97N40, 65F15.
Key words: QR algorithm, Givens rotations, eigenvalues, eigenvectors, orthogonal transformations, Mathcad.

1. Introduction

In numerical linear algebra, orthogonal transformations of the real n – dimensional space such as reflections and rotations are extensively used to perform annihilation of an element or group of elements of a matrix [1,3,4,6]. This techniques are used, e.g., to reduce matrices to triangular or Hessenberg or tridiagonal form and they have the important advantage of being numerically stable, due to the orthogonality of transformations involved.

One type of such transformations are the so-called Givens rotations [3,4,6], which are in fact plane rotations, performed in one of the coordinates plane. For instance, the Givens rotation of angle θ performed in the plane *(i,j)* corresponds to the matrix:

$$G(i,j,\theta)=\begin{pmatrix} 1 & & & \cdots & & & 0 \\ & \ddots & & & & \iddots & \\ & & \cos\theta & \cdots & \sin\theta & & \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ & & -\sin\theta & \cdots & \cos\theta & & \\ & \iddots & & & & \ddots & \\ 0 & & & \cdots & & & 1 \end{pmatrix},$$

where it is understood that the modifications from the identity matrix appear on rows and columns *i* and *j*. As we shall see in the following section, by applying certain Givens rotations successively in a specific order, one can annihilate all the sub-diagonal elements of a matrix

A, which means that there exist rotations $G_1, G_2, ..., G_p$ such that $G_p \cdot \ldots G_2 \cdot G_1 \cdot A = R$, where *R* is an upper triangular matrix. Denoting by $G = G_p \cdot \ldots G_2 \cdot G_1$, it follows that *G* is orthogonal and that $G \cdot A = R$, and therefore $A = Q \cdot R$, if $Q = G^T$. Therefore we can obtain the so-called QR decomposition [2,3,4] of a matrix by means of Givens rotations.

One important reason for computing the QR factorization of a matrix *A* is for applying the QR algorithm [2,3,4,6], which allow us to find the eigenvalues of *A*, and if *A* is symmetric, also the eigenvectors of *A*. This algorithm, however, requires the computation of a significant number of QR factorizations, which is why it is of crucial importance to optimize the number of floating point operations (*flops*), as well as the data storage needed for each QR decomposition.

In this note, we discuss the problems that occur when implementing the QR algorithm based on Givens rotations and propose optimized Mathcad [7] programs for computing the eigenvalues and, in the symmetric case, an orthogonal set of eigenvectors for a matrix *A*. For each of these programs, we analyze the performance in terms of the number of flops they require.

2. Givens rotations, QR decomposition and the QR algorithm

We briefly present, in this section, how Givens rotations can be used to annihilate all the sub-diagonal elements of a matrix *A*, thus allowing us to perform a QR factorization. Then we discuss the QR algorithm and explain how the combination of these techniques allows us to find the eigenvalues of *A*. For more in-depth information we refer the reader to, e.g., [3,4].

Let us first explain how a Givens rotation $G(i, j, \theta)$ acts on a vector $v \in \mathbb{R}^n$ and how it can be made to annihilate one of its components. If $v = \begin{pmatrix} v_1 & \ldots & v_i & \ldots & v_j & \ldots & v_n \end{pmatrix}^T$ and we denote $c = \cos\theta$ and $s = \sin\theta$, then we have

$$G(i,j,\theta)\cdot v = \begin{pmatrix} 1 & & & \cdots & & & 0 \\ & \ddots & & & & \iddots & \\ & & c & \cdots & s & & \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ & & -s & \cdots & c & & \\ & \iddots & & & & \ddots & \\ 0 & & & \cdots & & & 1 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ c\cdot v_i + s\cdot v_j \\ \vdots \\ -s\cdot v_i + c\cdot v_j \\ \vdots \\ v_n \end{pmatrix} \overset{not}{=} \begin{pmatrix} v_1' \\ \vdots \\ v_i' \\ \vdots \\ v_j' \\ \vdots \\ v_n' \end{pmatrix} \overset{not}{=} v'.$$

We can see that the vector v' obtained after applying the $G(i, j, \theta)$ rotation to *v* has all the components of *v* unchanged, except for the ones on rows *i* and *j*, that are modified by the formulas:

$$\begin{cases} v_i' = c\cdot v_i + s\cdot v_j \\ v_j' = -s\cdot v_i + c\cdot v_j \end{cases} \tag{1}$$

By choosing *c* and *s* according to:

$$c = \frac{v_i}{\sqrt{v_i^2 + v_j^2}},\ s = \frac{v_j}{\sqrt{v_i^2 + v_j^2}} \tag{2}$$

we have $c^2 + s^2 = 1$ verified and by substituting (2) into (1), we find $v_j' = 0$, i.e., we have successfully annihilated the *j*-th component of v'.

Let us remark that we in fact do not need, for any practical purposes, to know the value of angle θ, but only the values of c and s, for every Givens rotation. This relieves us of the burden of computing inverse trigonometric functions.

When applying a Givens rotation to a $n \times n$ matrix A, it is as if applying that rotation to every column of A. Therefore, only elements on rows i and j of the matrix A will be changed, according to formulas similar to (1). To be precise, we will have:

$$\begin{cases} A_{i,k}' = c \cdot A_{i,k} + s \cdot A_{j,k} \\ A_{j,k}' = -s \cdot A_{i,k} + c \cdot A_{j,k} \end{cases}, (\forall) k = 1 \ldots n \qquad (3)$$

Thus, we can annihilate an element $A_{j,k}'$ of our choice by using the values of c and s prescribed by (2) with v_i and v_j replaced by $A_{i,k}$ and $A_{j,k}$ respectively.

In order to annihilate all the sub-diagonal elements of a matrix A by using Givens rotations, one should proceed with care, so that an element already zeroed is not modified later, by a subsequent rotation. E.g., we can start by annihilating all the sub-diagonal elements on the first column: first a rotation in the (1,n) plane to annihilate $A_{1,n}$, then one in the (1, n-1) plane for zeroing $A_{1,n-1}$ and so on, with the rotation in (1,i) plane to annihilate $A_{1,i}$, for $i = n, n-1, \ldots 1$. The effects of these rotations only overlap on the first row, while all the other rows are affected by only one rotation each – the rotation that zeroes the element on that row. Now we can move to the second column and annihilate all its sub-diagonal elements from bottom to top. This time we use rotations in the (2,i) plane to annihilate $A_{2,i}$, for $i = n, n-1, \ldots 2$. The careful reader will notice that, by applying these rotations we actually do modify the already zeroed elements on column 1. However, the modifications consist of applying the formulas in (3), with $k = 1$, and since all the elements on column 1 are already zero (except for the first one), the result will be again zero. Furthermore, since we already know we will obtain zero, we can reduce the number of flops by using an algorithm that omits the recalculations for the already zeroed columns. Of course, the procedure continues with the next columns, from bottom to top, in a similar manner, until the initial matrix A is transformed into an upper triangular matrix R. This leads, in a way that has already been described in the previous section, to a so-called QR factorization of the matrix A, with $Q = G^T$ and $G = G_p \cdot \ldots G_2 \cdot G_1$. The transformations $G_1, G_2, ..., G_p$ are the successive rotations that were applied to A until R was obtained. Let us remark that we have;

$$Q = \left(G_p \cdot \ldots G_2 \cdot G_1\right)^T = G_1^T \cdot G_2^T \cdot \ldots G_p^T . \qquad (4)$$

Now let us see how we can use the method just described of obtaining a QR decomposition of a matrix A in a QR algorithm that allows us to find the eigenvalues of A. The QR algorithm [1,2,3,4] consists of the following steps:

Let $A = A_0$. For $k = 0, 1, 2, \cdots$, we repeat:

$$Step1: A_k = Q_k R_k \ (QR \text{ decomposition})$$

$$Step2: A_{k+1} = R_k Q_k \ (\text{define } A_{k+1} \text{ as the product of } Q_k \text{ and } R_k \text{ in the reversed order})$$

If certain conditions are satisfied (see, e.g.,[1]), the sequence $\left(A_k\right)_{k \in N}$ converges to an upper triangular matrix (diagonal matrix, if A is symmetric). As we shall see, the matrices A_k, $k = 0, 1, 2, ...$ are similar, hence they have the same eigenvalues and therefore, the upper triangular or diagonal matrix obtained has the eigenvalues of A on the diagonal. Moreover, if A is symmetric, we will be able to compute an orthogonal basis of eigenvectors as well.

Now let us see why the matrices A_k and A_{k+1} are similar (even unitary similar) for an

arbitrary *k*. Indeed, we have $A_k = Q_k R_k$ and therefore $Q_k^T A_k = R_k$, considering that Q_k is orthogonal. However, $A_{k+1} = R_k Q_k$ and it follows that $A_{k+1} = Q_k^T A_k Q_k$, which proves the unitary similarity of A_k and A_{k+1}. Let now *A* be symmetric. From the fact that $A_{k+1} = Q_k^T A_k Q_k$ for all *k*, it follows by induction that in fact we have:

$$A_{k+1} = Q_k^T \cdots \underbrace{Q_1^T \overbrace{Q_0^T A Q_0}^{A_1} Q_1}_{A_2} \cdots Q_k = \left(Q_0 Q_1 \cdots Q_k\right)^T A \left(Q_0 Q_1 \cdots Q_k\right) \text{ for all } k. \quad (5)$$

Since the sequence $\left(A_k\right)_{k\geq 0}$ converges to the diagonal form of *A*, denoted by Λ, we have $Q^T A Q \approx \Lambda$, for a sufficiently large *k*, where we denoted $Q = Q_0 Q_1 \cdots Q_k$. It follows that *Q* is an orthogonal matrix whose columns are the eigenvectors of *A*.

Let us look at this algorithm now from a computational point of view. Clearly, storing matrices Q_k and multiplying them by A_k is far from desirable. Therefore, we need to organize the algorithm in a more favorable way, before implementing it in a computer program. In the following section we propose a way to accomplish that, for the general case as well as for the symmetric case, when the eigenvectors can also be computed, along with the eigenvalues.

```
GivensStore(f) :=
    if f2 = 0
        c ← 1
        s ← 0
    otherwise
        if |f2| > |f1|
            u ← f1 / f2
            s ← sign(f2) / sqrt(1 + u^2)
            c ← s·u
        otherwise
            u ← f2 / f1
            c ← sign(f1) / sqrt(1 + u^2)
            s ← c·u
    ρ ← 1 if c = 0
    otherwise
        ρ ← (1/2)·sign(c)·s if |s| < |c|
        ρ ← 2·sign(s)/c if |c| ≤ |s|
    ρ
```

Figure 3. 1

3. Using Givens rotations for solving the eigenvalues problem. An optimized implementation

We have discussed in [2] how the QR decomposition of a matrix *A* can be optimally implemented using Givens rotations and we have presented there the Mathcad programs that do it. For reader's convenience, we will present them here briefly as well (with some slight modifications, where needed).

The function **GivensStore** from Figure 3.1 performs first a computation of the *c*, s numbers that define a Givens rotation that annihilates the second component of *f*. In theory we used formulas (2), but for computational purposes they had to be slightly modified to avoid overflow/underflow, by simplifying them by the larger component of *f* (in absolute value) (see [2]). A number ρ is then computed based on the values of *c* and *s*. Later we will be able to restore the values of *c*, s by using the function **econorestore** (Figure 3.2). This way of storing all the information concerning a plane rotation in just one number was suggested by Stewart [5], since we annihilate one element of *A* with each rotation and thus are able to store the number ρ defining the rotation in the emptied position of *A*. This was amply discussed in [2]. It is important to keep in mind that with this technique, we may recover $-(c,s)$ instead of (c,s). In general, this poses no problem, as long as we take care to be consistent throughout.

```
econorestore(ρ) := | if ρ = 1
                   |     c ← 0
                   |     s ← 1
                   | if |ρ| < 1
                   |     s ← 2ρ
                   |     c ← √(1 − s^2)
                   | if |ρ| > 1
                   |     c ← 2/ρ
                   |     s ← √(1 − c^2)
                   | (c)
                   | (s)
```

Figure 3.2

```
multiplycs(v, A, i, j) := | n ← cols(A)
                          | for k ∈ 1..n
                          |     u ← A_i,k
                          |     A_i,k ← v1·A_i,k + v2·A_j,k
                          |     A_j,k ← −v2·u + v1·A_j,k
                          | A
```

Figure 3.3

The **multiplycs** function shown in Figure 3.3 performs the update of the matrix A when a rotation $G(i, j, \theta)$ is applied, taking advantage of the fact that only rows i and j are changed, while the function **multiplickcsl** in Figure 3.4 does the same thing, but updates the elements of rows i and j only from the i-th element to the last one. This last feature is important especially when we want to store the values ρ that define each rotation in the corresponding zeroed position, in order to prevent these values from being modified by subsequent rotations. It also reduces the number of operations needed. On the other hand, the **multiplickcsr** function in Figure 3.5 performs the update of matrix A when $G(i, j, \theta)^T$ is applied to the right, i.e. when we compute $A \cdot G(i, j, \theta)^T$. In this case the elements of columns i and j are modified.

```
multiplickcsl(v, A, i, j) ≡ | n ← cols(A)
                            | for k ∈ i..n
                            |     u ← A_i,k
                            |     A_i,k ← v1·A_i,k + v2·A_j,k
                            |     A_j,k ← −v2·u + v1·A_j,k
                            | A
```

Figure 3.4

```
multiplickcsr(v, A, i, j) := | n ← cols(A)
                             | for k ∈ 1..n
                             |     u ← A_k,i
                             |     A_k,i ← v1·A_k,i + v2·A_k,j
                             |     A_k,j ← −v2·u + v1·A_k,j
                             | A
```

Figure 3.5

In order to perform the steps in the QR algorithm, we define two functions. At iteration k, function **Step1**, shown in Figure 3.6, produces the matrix R_k by successively applying Givens rotations $G_1, G_2, ..., G_p$ to (the left side of) A_k until it becomes upper triangular. For each rotation applied, the ρ number is computed and stored in the corresponding position of a new matrix, denoted by S. This is necessary, because at the next step, by using function **Step2** shown in Figure 3.7, we will multiply the resulting R_k by Q_k to the right, which , according to (4), means we need to successively multiply R_k to the right by $G_1^T, G_2^T, \ldots G_p^T$. Thus, we will recover these rotations from the stored ρ values in S. The matrices S and A are juxtaposed by the Mathcad function **augment** and returned as a single matrix by function **Step1**. Finally, the function **AlgQR** shown in Figure 3.8 just performs k iterations of steps 1 and 2.

```
Step1(A) :=
    n ← cols(A)
    for j ∈ 1..n − 1
        for i ∈ n, n − 1..j + 1
            S_{i,j} ← GivensStore( (A_{j,j}, A_{i,j}) )
            A ← multiplickcsl(econorestore(S_{i,j}), A, j, i)
    augment(S, A)
```

Figure 3.6

```
Step2(M) :=
    n ← rows(M)
    A ← submatrix(M, 1, n, n, 2·n − 1)
    for j ∈ 1..n − 1
        for i ∈ n, n − 1..j + 1
            A ← multiplickcsr(econorestore(M_{i,j}), A, j, i)
    A
```

Figure 3.7

In Figure 3.9 we show the result of running **AlgQR** on a test matrix *A*. The eigenvalues of *A* can be found on the diagonal of the resulting upper triangular matrix.

```
AlgQR(A, k) :=
    for i ∈ 1..k
        A ← Step2(Step1(A))
    A
```

Figure 3.8

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 3 & 2 & 6 & 5 \\ 1 & 5 & 2 & 2 \\ 2 & 1 & 3 & 4 \end{pmatrix} \qquad \text{AlgQR}(A, 50) = \begin{pmatrix} 10.317 & 1.473 & 0.98 & -3.793 \\ 0 & -2.692 & -0.711 & 3.121 \\ 0 & 0 & 1.117 & -1.112 \\ 0 & 0 & 0 & 0.258 \end{pmatrix}$$

Figure 3.9

Let us now take a look at the symmetric case, when we can also compute the eigenvectors of *A*. As it follows from (5), the eigenvectors will be the columns of the matrix $Q = Q_0 Q_1 \cdots Q_k$, while each Q_i is the product of a sequence of Givens rotations, according to (4). Instead of storing these Q_i matrices and then multiply then, which would be very costly, from a computational point of view, we start with the identity matrix and keep multiplying it by Givens rotations (which is much cheaper than normal matrix multiplication, since we only modify two rows or two columns) until we first obtain Q_1, then $Q_1 \cdot Q_2$ and so on, up to $Q = Q_0 Q_1 \cdots Q_k$. Still, compared to the algorithm for finding just the eigenvalues, we need to store and update one more $n \times n$ matrix. In Figures 3.10 and 3.11 we show the functions **Step1Sym** and **Step2Sym** which perform the same tasks as **Step1** and **Step2**. Additionally, function **Step1Sym** computes the updates of *Q*, as explained above and function **Step2Sym** just carries it forward to the next step without modifications. Finally, the function **AlgQRSym**, shown in Figure 3.12 just initializes *Q* to the identity matrix and performs *k* iterations of **Step1Sym** and **Step2Sym** . It returns the diagonal matrix Λ having the eigenvalues of *A* on the diagonal and the Q^T matrix, juxtaposed in a single $n \times 2n$ matrix. The eigenvectors of *A* are

```
Step1Sym(A, Q) :=
    n ← cols(A)
    for j ∈ 1..n − 1
        for i ∈ n, n − 1..j + 1
            S_{i,j} ← GivensStore( (A_{j,j}, A_{i,j}) )
            v ← econorestore(S_{i,j})
            A ← multiplickcsl(v, A, j, i)
            Q ← multiplycs(v, Q, j, i)
    augment(S, A, Q)
```

Figure 3.10

```
Step2Sym(M) :=
    n ← rows(M)
    A ← submatrix(M, 1, n, n, 2·n − 1)
    Q ← submatrix(M, 1, n, 2·n, 3·n − 1)
    for j ∈ 1..n − 1
        for i ∈ n, n − 1..j + 1
            A ← multiplickcsr(econorestore(M_{i,j}), A, j, i)
    (A  Q)
```

Figure 3.11

the columns of *Q*, i.e., the *rows* of Q^T .

```
AlgQRSym(A, k) :=
    n ← cols(A)
    Q ← identity(n)
    for i ∈ 1..k
        A1 ← A
        A ← Step2Sym(Step1Sym(A, Q))1,1
        Q ← Step2Sym(Step1Sym(A1, Q))1,2
    augment(A, Q)
```

Figure 3.12

Below, in Figure 3.13 we present an example of running the **AlgQRSym** function on a symmetric matrix. The first four columns of the resulting 4×8 matrix represent the diagonal matrix with A's eigenvalues on the diagonal. The last four columns represent Q^T, and the eigenvectors of A are its *rows*.

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 5 \\ 3 & 1 & 5 & -1 \\ 4 & 5 & -1 & 7 \end{pmatrix} \quad \text{AlgQRSym}(A, 50) = \begin{pmatrix} 12.26 & 0 & 0 & 0 & 0.393 & 0.507 & 0.128 & 0.756 \\ 0 & 6.274 & 0 & 0 & -0.288 & 4.803 \times 10^{-3} & -0.909 & 0.3 \\ 0 & 0 & -2.514 & 0 & 0.748 & 0.235 & -0.394 & -0.479 \\ 0 & 0 & 0 & -0.021 & -0.45 & 0.829 & 0.038 & -0.329 \end{pmatrix}$$

Figure 3.13

4. Conclusion

An estimate flops count shows that each iteration of the algorithm for computing the eigenvalues for a $n \times n$ matrix requires $4n^3 + O(n^2)$ floating point operations. In the case of symmetric matrices, if the eigenvectors are also required, then each iteration uses an extra $3n^3 + O(n^2)$ flops, which leads us to a total of $7n^3 + O(n^2)$ flops per iteration. This method of computing eigenvalues of a matrix using the QR algorithm based on Givens rotations, for which we have presented here the optimized algorithms and the corresponding Mathcad programs, is numerically stable, due to the fact that we use only unitary transformations. It involves a reasonable number of flops, compared to other options available and it has the advantage that we can considerably reduce this number of flops in cases of matrices having a special structure, such as, e.g., Hessenberg or tridiagonal matrices.

References

[1] Atkinson, K.: *An introduction to numerical analysis (2nd edition)*, Wiley, 1989.
[2] Caragheorgheopol, D.: A note on Givens rotations, QR decompositions and solving of linear systems , *The 15-th Conference "Mathematics, Computer Science and Education"*, Department of Mathematics and Computer Science, T.U.C.E.B., 2018, to appear.
[3] Golub, G.H. and Van Loan, C.F.: *Matrix computations (3rd edition)*, The Johns Hopkins University Press, 1996.
[4] Horn, R.A. and Johnson, C.R.: *Matrix analysis*, Cambridge University Press, 1985.
[5] Stewart, G.W.: The Economical Storage of Plane Rotations, *Numer. Math.* **25**, 1976.
[6] Wilkinson, J.H.: *The Algebraic Eigenvalue Problem*, Claredon Press, 1965.
[7] Mathcad 14 Help, Parametric Technology Corporation, 2007 (http://www.ptc.com).

The 15th Workshop of Scientific Communications,
Department of Mathematics and Computer Science, June 9, 2018, p. 82-87

ABOUT DIRECT SUMS OF DECOMPOSABLE OPERATORS AND DECOMPOSABLE SYSTEMS

Mariana Zamfir
Department of Mathematics and Computer Science
Technical University of Civil Engineering of Bucharest, Romania
E-mail: zamfirvmariana@yahoo.com

Abstract: In the present article some results concerning the direct sums of two decomposable operators (respectively, two decomposable systems) are presented. It is shown that the direct sum of two operators (two operator systems) is a decomposable operator (system) if and only if each operator (system) is decomposable.
Mathematics Subject Classification (2010): 47B47, 47B40.
Key words: spectral maximal space; exact form; direct sum; (weakly) decomposable.

1. Introduction

The main purpose of the this work is to present in a systematic way, several results concerning the behavior of decomposable operators (respectively, decomposable systems) in Banach spaces with respect to the property of direct sum.
The first section is dedicated to the preliminaries. Let us briefly recall the notations and terminology dealt with by the present paper.
In the following discussion, let $\mathbf{B}(X)$ be the Banach algebra of all linear bounded operators on a given complex Banach space X, and let $\mathbf{S}(X)$ be the family of all linear closed subspaces of X. Furthermore, if $\mathbf{C}$ is the complex plane and $\mathbf{C}^n$ is the space of all elements $z=(z_1, z_2, ..., z_n)$, with $z_1, z_2, ..., z_n \in \mathbf{C}$, we denote by $\mathbf{F}(\mathbf{C})$ (respectively, by $\mathbf{F}(\mathbf{C}^n)$) the family of all closed subsets $F \subset \mathbf{C}$ (respectively, $F \subset \mathbf{C}^n$).
As usual, the spectrum of an operator $T \in \mathbf{B}(X)$ is denoted by $\sigma(T)$ and it is defined as the set of all numbers $\lambda \in \mathbf{C}$ for which the operator $\lambda I - T$ is no inversable in $\mathbf{B}(X)$.
For any $Y \in \mathbf{S}(X)$, invariant to an operator $T \in \mathbf{B}(X)$ (respectively, to a commuting operator system $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$), $T|Y$ means the restriction operator of T to Y (respectively, $a|Y=(a_1|Y, a_2|Y, ..., a_n|Y)$ means the restriction system of a to Y).
If p is an integer and $\sigma=(s_1, s_2, ..., s_n)$ is an n-tuple of indeterminates, then $\Lambda^p[\sigma, X]$ will denote the space of all exterior forms of degree p in s, having coefficients in X; any $\psi \in \Lambda^p[\sigma, X]$ can be written as $\psi = \sum_{1 \le j_1 \le ... \le j_p \le n} x_{j_1 ... j_p} s_{j_1} \wedge ... \wedge s_{j_p}$, $x_{j_1 ... j_p} \in X$.
In the remaining sections, we will discuss about direct sums of decomposable operators and decomposable operators; several results were obtained by Colojoară and Foiaş in [4], by Frunză in [6] and [7], by Bacalu in [2], and by Bacalu and Zamfir in [12] and [13].

2. Direct sums of operators with the single-valued extension property

Definition 2.1. ([4], [5]) An operator $T \in \mathbf{B}(X)$ has the *single-valued extension property* if for any analytic function $f : G \to X$ (where $G \subset \mathbf{C}$ open), verifying $(\lambda I - T) f(\lambda) \equiv 0$, it results that $f(\lambda) \equiv 0$.

For $T \in \mathbf{B}(X)$ having the single-valued extension property and for any $x \in X$, we consider the open set $\rho_T(x)$ of all elements $\xi \in \mathbf{C}$ such that there is a unique X-valued analytic function $x_T(\cdot)$ defined on an open neighborhood of ξ, verifying $(\lambda I - T) x_T(\lambda) \equiv x$; $\rho_T(x)$ is the *local resolvent set of x with respect to T* and $\sigma_T(x) = \mathbf{C} \setminus \rho_T(x)$ is the *local spectrum of x with respect to T*. We have $\sigma_T(x) \subset \sigma(T)$ and we denote by

$$X_T(F) = \{x \in X;\ \sigma_T(x) \subset F,\ F \subset \mathbf{C}\}.$$

Theorem 2.1. ([4]) Let $T_1 \in \mathbf{B}(X_1)$ and $T_2 \in \mathbf{B}(X_2)$. Then $T_1 \oplus T_2 \in \mathbf{B}(X_1 \oplus X_2)$ has the single-valued extension property if and only if both T_1 and T_2 have the single-valued extension property. Moreover, for $x_1 \in X_1$ and $x_2 \in X_2$, we have:

$$\sigma_{T_1 \oplus T_2}(x_1 \oplus x_2) = \sigma_{T_1}(x_1) \cup \sigma_{T_2}(x_2).$$

Proposition 2.1. ([4]) Let $T_1 \in \mathbf{B}(X_1)$ and $T_2 \in \mathbf{B}(X_2)$ be two operators having the single-valued extension property. Then the following statements are hold:

1) $X_{1_{T_1}}(F_1) \oplus X_{2_{T_2}}(F_2) \subseteq (X_1 \oplus X_2)_{T_1 \oplus T_2}(F_1 \cup F_2)$, $F_1, F_2 \subseteq \mathbf{C}$.

2) $X_{1_{T_1}}(F) \oplus X_{2_{T_2}}(F) = (X_1 \oplus X_2)_{T_1 \oplus T_2}(F)$, $F \subseteq \mathbf{C}$.

3. Direct sums of decomposable and weakly decomposable operators

Definition 3.1. ([4]) Let $T \in \mathbf{B}(X)$ and let Y be a linear closed subspace of X invariant to T. Y is called *spectral maximal space of T* if it contains any other linear closed subspace Z of X, also invariant to T, such that $\sigma(T|Z) \subseteq \sigma(T|Y)$.

Definition 3.2. ([4]) An operator $T \in \mathbf{B}(X)$ is called *decomposable* if for any finite open covering $\{G_i\}_{i=1}^n$ of the spectrum $\sigma(T)$, there is a system $\{Y_i\}_{i=1}^n$ of spectral maximal spaces of T such that following two conditions are verified:

1) $\sigma(T|Y_i) \subseteq G_i$, for all $i = 1, 2, ..., n$

2) $X = Y_1 + Y_2 + ... + Y_n$.

Theorem 3.1. ([4]) Let $T \in \mathbf{B}(X)$ be decomposable. Then the following assertions are established:

1) T has the single-valued extension property.

2) $X_T(F)$ is a spectral maximal space of T and $\sigma(T|X_T(F)) \subseteq F \cap \sigma(T)$, for any $F \in \mathbf{F}(\mathbf{C})$.

3) If Y is a spectral maximal space of T, then $Y = X_T(\sigma(T|Y))$.

Lemma 3.1. ([4], [3]) Let $T_1 \in \mathbf{B}(X_1)$, $T_2 \in \mathbf{B}(X_2)$, and $T_1 \oplus T_2 \in \mathbf{B}(X_1 \oplus X_2)$. Then:

(i) $\sigma((T_1 \oplus T_2)|(Y_1 \oplus Y_2)) = \sigma(T_1|Y_1) \cup \sigma(T_2|Y_2)$, for $Y_i \subseteq X_i$ linear closed subspace invariant to T_i, $i = 1, 2$.

(ii) If $Y \subseteq X_1 \oplus X_2$ is a spectral maximal space of $T_1 \oplus T_2$, then $Y = Y_1 \oplus Y_2$, where Y_1 is a spectral maximal space of T_1 and Y_2 is a spectral maximal space of T_2.

Theorem 3.2. ([4]) Let $T_1 \in \mathbf{B}(X_1)$ and $T_2 \in \mathbf{B}(X_2)$ be two decomposable operators. Then $T_1 \oplus T_2 \in \mathbf{B}(X_1 \oplus X_2)$ is a decomposable operator. Conversely, if $T_1 \oplus T_2 \in \mathbf{B}(X_1 \oplus X_2)$ is decomposable, then both T_1 and T_2 are decomposable.

Definition 3.3. ([4]) An operator $T \in \mathbf{B}(X)$ is called *weakly decomposable* if:

1) T has the single-valued extension property and the space $X_T(F)$ is closed, for any $F \in \mathbf{F}(\mathbf{C})$

2) for any finite open covering $\{G_i\}_{i=1}^n$ of $\sigma(T)$, there is a system $\{Y_i\}_{i=1}^n$ of spectral maximal spaces of T with $\sigma(T \mid Y_i) \subset G_i$, $i = 1, 2, ..., n$ and $X = \overline{Y_1 + Y_2 + ... + Y_n}$.

Lemma 3.2. ([14]) An operator $T \in \mathbf{B}(X)$ is weakly decomposable if and only if the following conditions are hold:

(i) T has the single-valued extension property and the space $X_T(F)$ is closed, for any $F \in \mathbf{F}(\mathbf{C})$;

(ii) for any finite open covering $\{G_i\}_{i=1}^n$ of $\sigma(T)$, the set Z of all elements $x \in X$, with $x = x_1 + x_2 + ... + x_n$, $\sigma_T(x_i) \subset G_i$, $i = 1, 2, ..., n$, is dense in X.

Theorem 3.3. ([14]) Let $T_1 \in \mathbf{B}(X_1)$ and $T_2 \in \mathbf{B}(X_2)$. Then $T_1 \oplus T_2 \in \mathbf{B}(X_1 \oplus X_2)$ is weakly decomposable if and only if both T_1 and T_2 are weakly decomposable.

4. Direct sums of operator systems with the single-valued extension property

Definition 4.1. ([6], [11]) The commuting operator system $a = (a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$ is said to be *nonsingular* on X if the Koszul complex $E(X, a)$ is exact, where

$$E(X,a): 0 \to X = \Lambda^n[\sigma, X] \xrightarrow{\delta_n} \Lambda^{n-1}[\sigma, X] \xrightarrow{\delta_{n-1}} ...$$
$$... \xrightarrow{\delta_3} \Lambda^2[\sigma, X] \xrightarrow{\delta_2} \Lambda^1[\sigma, X] \xrightarrow{\delta_1} \Lambda^0[\sigma, X] = X \to 0$$

or, equivalent, the complex $F(X, a)$ is exact, where

$$F(X,a): 0 \to X = \Lambda^0[\sigma, X] \xrightarrow{\delta^0} \Lambda^1[\sigma, X] \xrightarrow{\delta^1} \Lambda^2[\sigma, X] \xrightarrow{\delta^2} ...$$
$$... \xrightarrow{\delta^{n-2}} \Lambda^{n-1}[\sigma, X] \xrightarrow{\delta^{n-1}} \Lambda^n[\sigma, X] = X \to 0.$$

For an integer p, the homology modules of $E(X, a)$ can be written as

$$H_p(X,a) = \mathrm{Ker}\left(\delta_{p+1} : \Lambda^{p+1}[\sigma, X] \to \Lambda^p[\sigma, X]\right) / \mathrm{Im}\left(\delta_p : \Lambda^p[\sigma, X] \to \Lambda^{p-1}[\sigma, X]\right)$$

and respectively, the cohomology modules of $F(X, a)$ as

$$H^p(X,a) = \mathrm{Ker}\left(\delta^p : \Lambda^p[\sigma, X] \to \Lambda^{p+1}[\sigma, X]\right) / \mathrm{Im}\left(\delta^{p-1} : \Lambda^{p-1}[\sigma, X] \to \Lambda^p[\sigma, X]\right).$$

If $G \subset \mathbf{C}^n$ is an arbitrary open set, $C^\infty(G, X)$ is the set of all X-valued continuous functions

on G admitting partial derivatives of any order, and $\sigma=(s_1,s_2,...,s_n)$ is a system of indeterminates, then α is the operator that acts on the exterior forms $\psi\in\Lambda^p\left[\sigma,C^\infty(G,X)\right]$ in indeterminates s having coefficients in $C^\infty(G,X)$, defined as

$$(\alpha\psi)(z)=\left[(z_1-a_1)s_1+...+(z_n-a_n)s_n\right]\wedge\psi(z),\ z\in G$$

and $\alpha\oplus\overline{\partial}$ is the operator that acts on the exterior forms $\psi\in\Lambda^p\left[\sigma\cup d\overline{z},C^\infty(G,X)\right]$ in indeterminates s and $d\overline{z}=(d\overline{z}_1,\ d\overline{z}_2,\ ...,\ d\overline{z}_n)$ with coefficients in $C^\infty(G,X)$:

$$\left((\alpha\oplus\overline{\partial})\psi\right)(z)=\left[(z_1-a_1)s_1+...+(z_n-a_n)s_n+\frac{\partial}{\partial\overline{z}_1}d\overline{z}_1+...+\frac{\partial}{\partial\overline{z}_n}d\overline{z}_n\right]\wedge\psi(z),\ z\in G.$$

Definition 4.2. ([6], [11]) The complementary in $\mathbf{C}^n$ of the set of all $z=(z_1,z_2,...,z_n)\in\mathbf{C}^n$ such that the system $z-a=(z_1-a_1,\ z_2-a_2,...,z_n-a_n)$ is nonsingular on X is called *the spectrum of* $a=(a_1,a_2,...,a_n)\subset\mathbf{B}(X)$ *on* X and it is denoted by $\sigma(a,X)$.

Definition 4.3. ([6]) For $a=(a_1,a_2,...,a_n)\subset\mathbf{B}(X)$ and for any $x\in X$, we consider the open set $\rho(a,x)$ of all elements $z=(z_1,z_2,...,z_n)\in\mathbf{C}^n$ such that there are n X-valued analytic functions $f_1,f_2,...,f_n$ defined on an open neighborhood V of z, satisfying the equality $(\zeta_1-a_1)f_1(\zeta)+...+(\zeta_n-a_n)f_n(\zeta)\equiv x$, with $\zeta=(\zeta_1,\zeta_2,...,\zeta_n)\in V$; $\rho(a,x)$ is the *analytic resolvent set of* x *with respect to* a and $\sigma(a,x)=\mathbf{C}^n\setminus\rho(a,x)$ is the *analytic spectrum of* x *with respect to* a.

Definition 4.4. ([6], [7]) The complementary in $\mathbf{C}^n$ of the reunion of all open sets $G\subset\mathbf{C}^n$ such that there is an exterior form $\psi\in\Lambda^{n-1}\left[\sigma\cup d\overline{z},C^\infty(G,X)\right]$ satisfying the equality

$$xs_1\wedge s_2\wedge...\wedge s_n=\left[(z_1-a_1)s_1+...+(z_n-a_n)s_n+\frac{\partial}{\partial\overline{z}_1}d\overline{z}_1+...+\frac{\partial}{\partial\overline{z}_n}d\overline{z}_n\right]\wedge\psi(z)$$

is called *the spectrum* of x *with respect to* $a=(a_1,a_2,...,a_n)$ and is denoted by $sp(a,x)$.

Definition 4.5. ([6]) We say that the system $a\subset\mathbf{B}(X)$ verifies *the cohomology property* (L) or the system a has the *single-valued extension property* if

$$H^{n-1}\left(C^\infty(G,X),\alpha\oplus\overline{\partial}\right)=0,\ \text{for any } G\subset\mathbf{C}^n \text{ open.}$$

In this case, we denote by

$$X_{[a]}(F)=\left\{x\in X;\ sp(a,x)\subset F,\ F\subset\mathbf{C}^n\right\}$$

$$X_a(F)=\left\{x\in X;\ \sigma(a,x)\subset F,\ F\subset\mathbf{C}^n\right\}.$$

Proposition 4.1. ([2], [12]) If $a=(a_1,a_2,...,a_n)\subset B(X_1)$ and $b=(b_1,b_2,...,b_n)\subset B(X_2)$ are two operator systems, then we have

$$H^p\left(C^\infty(G,X_1),\alpha\oplus\overline{\partial}\right)\oplus H^p\left(C^\infty(G,X_2),\beta\oplus\overline{\partial}\right)=$$

$$=H^p\left(C^\infty(G,X_1\oplus X_2),(\alpha\oplus\beta)\oplus(\overline{\partial}\oplus\overline{\partial})\right)$$

for any $G \subset \mathbf{C}^n$ open set and every $p \in \mathbf{Z}$.

Theorem 4.1. ([2], [12]) Let $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X_1)$ and $b=(b_1, b_2, ..., b_n) \subset \mathbf{B}(X_2)$ be two operator systems. The systems a and b verify condition (L) if and only of the system $a \oplus b = (a_1 \oplus b_1, a_2 \oplus b_2, ..., a_n \oplus b_n) \subset \mathbf{B}(X_1 \oplus X_2)$ verifies condition (L).

Proposition 4.2. ([2], [12]) Let $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X_1)$ and $b=(b_1, b_2, ..., b_n) \subset \mathbf{B}(X_2)$ be two operator systems verifying condition (L). Then:

1) $\sigma(a \oplus b, x_1 \oplus x_2) = \sigma(a, x_1) \cup \sigma(b, x_2)$, $x_1 \in X_1$, $x_2 \in X_2$.
2) $sp(a \oplus b, x_1 \oplus x_2) = sp(a, x_1) \cup sp(b, x_2)$, $x_1 \in X_1$, $x_2 \in X_2$.
3) $X_{1_{[a]}}(F) \oplus X_{2_{[b]}}(F) = (X_1 \oplus X_2)_{[a \oplus b]}(F)$, $F \in \mathbf{F}(\mathbf{C}^n)$.
4) $X_{1_a}(F) \oplus X_{2_b}(F) = (X_1 \oplus X_2)_{a \oplus b}(F)$, $F \in \mathbf{F}(\mathbf{C}^n)$.

5. Direct sums of decomposable operator systems

Definition 5.1. ([6]) A linear closed subspace $Y \subset X$ is a *spectral maximal space* of the system $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$ if Y is invariant to a and for any other subspace $Z \subset X$, invariant to a, the inclusion $\sigma(T \mid Z) \subseteq \sigma(T \mid Y)$ implies the inclusion $Z \subset Y$.

Definition 5.2. ([6]) A mapping $E : \mathbf{F}(\mathbf{C}^n) \to \mathbf{S}(X)$ is called *spectral capacity of* $(\mathbf{C}^n, X)$-*type* if the following conditions are verified:

(1) $E(\varnothing) = \{0\}$, $E(\mathbf{C}^n) = X$;

(2) $E\left(\bigcap_{i \in I} F_i\right) = \bigcap_{i \in I} E(F_i)$, for any family $\{F_i\}_{i \in I} \subset \mathbf{F}(\mathbf{C}^n)$

(3) for any open finite covering $\{G_j\}_{j=1}^m$ of $\mathbf{C}^n$ we have

$$X = E(\overline{G}_1) + E(\overline{G}_2) + ... + E(\overline{G}_m).$$

The commuting operator system $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$ is said to be *decomposable* if there is a spectral capacity E of $(\mathbf{C}^n, X)$-type such that:

(4) $a_i E(F) \subset E(F)$, for all $F \in \mathbf{F}(\mathbf{C}^n)$ and for any $i = 1, 2, ..., n$

(5) $\sigma(a, E(F)) \subset F$, for all $F \in \mathbf{F}(\mathbf{C}^n)$.

If a is decomposable, then it admits a unique spectral capacity E ([6]).

Definition 5.3. ([2]) A decomposable system $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$ is called *strongly decomposable* if the restriction system $a \mid Y = (a_1 \mid Y, a_2 \mid Y, ..., a_n \mid Y) \subset \mathbf{B}(Y)$ is decomposable, for any spectral maximal space Y of a.

Theorem 5.1. ([6]) If $a=(a_1, a_2, ..., a_n) \subset \mathbf{B}(X)$ is decomposable, then the system a verifies condition (L).

Proposition 5.1. ([2], [12]) A system $a=(a_1,a_2,...,a_n)\subset B(X)$ is decomposable if and only if the following relations are verified:

1) a satisfies condition (L), $X_{[a]}(F)$ is closed and $\sigma\left(a,X_{[a]}(F)\right)\subset F$, for any $F\in\mathbf{F}\left(\mathbf{C}^n\right)$

2) for any open finite covering $\{G_j\}_{j=1}^m$ of $\mathbf{C}^n$ we have $x=x_1+x_2+...+x_m$, with $sp\left(a,x_j\right)\subset G_j$, $j=1, 2, ..., m$, for all $x\in X$.

Theorem 5.2. ([2], [12]) Let $a=(a_1, a_2, ...,a_n)\subset\mathbf{B}(X_1)$ and $b=(b_1, b_2, ...,b_n)\subset\mathbf{B}(X_2)$ be two systems. Then the system $a\oplus b=(a_1\oplus b_1,a_2\oplus b_2, ..., a_n\oplus b_n)\subset\mathbf{B}(X_1\oplus X_2)$ is decomposable if and only if both systems a and b are decomposable.

Corollary 5.1. ([2], [12]) Let $a=(a_1, a_2, ...,a_n)\subset\mathbf{B}(X_1)$ and $b=(b_1, b_2, ...,b_n)\subset\mathbf{B}(X_2)$. Then $a\oplus b\subset\mathbf{B}(X_1\oplus X_2)$ is strongly decomposable if and only if both a and b are strongly decomposable.

References

[1] Apostol, C.: *Spectral decompositions and functional calculus*, Rev. Roum. Math. Pures et Appl., 13, 1481-1528, 1968.

[2] Bacalu, I.: *Descompuneri spectrale reziduale* (*Residually spectral decompositions*), Stud. Cerc. Mat., I (1980), II (1980), III (1981).

[3] Bacalu, I.: *S -Spectral Decompositions*, Ed. Politehnica Press, Bucharest, 2008.

[4] Colojoară, I., Foiaş, C.: *Theory of generalized spectral operators*, Gordon Breach, Science Publ., New York-London-Paris, 1968.

[5] Dunford, N., Schwartz, J.T.: *Linear Operators*, Interscience Publishers, New-York, Part I: General Theory, 1958; Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space, 1963; Part III: Spectral Operators, 1971.

[6] Frunză, Şt.: *O teorie axiomatică a descompunerilor spectrale pentru sisteme de operatori I* (*An axiomatic theory of spectral decompositions for systems of operators I*), Stud. Cerc. Mat., 27, 655-711, 1975.

[7] Frunză, Şt.: *The Taylor spectrum and spectral decompositions*, J. Func. Anal., 19, 390-421, 1977.

[8] Lange, R., Wang, S.: *New approaches in spectral decomposition*, Amer. Math. Soc, 1992.

[9] Laursen, K.B., Neumann, M.M.: *An introduction to local spectral theory*, London Math. Soc. Monographs New Series, Oxford Univ. Press., New-York, 2000.

[10] Şerbănescu, C., Bacalu, I.: *Generalizing certain properties of decomposable systems*, Recent Advances on Applied Mathematics an Computational Methods in Engineering, ISBN 978-1-61804-292-7, 26-35, 2015.

[11] Taylor, J.L.: *A joint spectrum for several commuting operators*, J. Func. Anal., 6, 172-191, 1970.

[12] Zamfir, M., Bacalu, I.: *Direct sums of decomposable systems*, Scient. J. Math. Modelling in Civil Engineering, **7**, no 2 BIS, 69-81, 2011.

[13] Zamfir, M., Bacalu, I.: *On the local spectral properties of operator systems in Banach spaces*, Applied Sciences, 14, 89-97, 2012.

[14] Zamfir, M.: *Some properties of weakly decomposable operators*, Proc. of 1-st Workshop, Technical University of Civil Engineering, Bucharest, Romania, MatrixRom, ISSN 2392-6317, 152-157, 2014.

THE CONTINUITY OF THE ROOTS OF A POLYNOMIAL OVER Qptilde AS FUNCTIONS OF ITS COEFFICIENTS

Sever Achimescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: sachimescu@ahoo.com
Stelian Corneliu Andronescu
Departament of Matthematics and Computer Science,
University of Pitesti, Romania
E-mail: corneliuandronescu@yahoo.com

Abstract: For any polynomial with coefficients in the ring **Qp**tilde we prove that its roots are continuous functions of the coefficients.

CONNECTION BETWEEN SPECTRA AND (CO)HOMOLOGY THEORIES

Cristian Costinescu
Department of Mathematics and Computer Science
Technical University of Civil Engineering Bucharest, Romania
E-mail: cristiancostinescu13@yahoo.com

Abstract: In the algebraic topology it is useful to introduce (co)homology theories on the category of spectra **S**. In this paper we assign a (co)homology theory on **S** to a spectrum and this situation turns out to be invertible.
We also give some examples of spectra and (co)homology theories.

SEPARATION OF VARIABLES: APPLICATIONS IN HEAT TRANSFER

Oana Dumitru
Undergraduate student 2nd year
Building Services Engineering Faculty
Technical University of Civil Engineering Bucharest, Romania
E-mail: oana.dumitru@efden.org

Abstract: This paper presents some applications of the method of separation of variables [1,2] in the field of heat transfer phenomena, as a special topic [3,4] that extends the link between two disciplines in the Engineering Curricula: Special Mathematics and Heat Transfer.
Besides the usual approach used in the mathematical manuals [1,2] regarding the heat equation (which represents the unsteady heat conduction in a 1D planar geometry or "heat conduction in a rod" [2]), supplementary examples [4] are provided in this presentation, concerning the steady heat conduction in a 2D geometry, described by the Laplace equation, and the laminar heat convection in a pipe flow, described by the PDE:

$$2w_{x,avg}\left(1-\frac{r^2}{r_C^2}\right)\frac{\partial T}{\partial x}=a\left(\frac{\partial^2 T}{\partial r^2}+\frac{1}{r}\frac{\partial T}{\partial r}\right)$$

References

[1] Olariu, V., Prepeliță, V.: *Matematici speciale,* Editura Didactică și Pedagogică, Bucharest, 1985.
[2] Boyce, W.E, DiPrima, R.C., Meade, D.B.: *Elementary Differential Equations and Boundary Value Problems*, John Wiley & Sons, Inc., 2017.
[3] Băltărețu, Fl.: *Transferts thermiques,* Lecture Notes, Technical University of Civil Engineering, Bucharest, 2018.
[4] Băltărețu, Fl.: *Special Topics in Heat Transfer,* Personal communication, 2018.

INTERNET OF THINGS (IO T) TO IMPROVE ICT EDUCATION IN TAMK

Esa Kujansuu
ICT Engineering Department
Tampere University of Applied Sciences
Kuntokatu 3, FI-33520 Tampere, Finland
E-mail: esa.kujansuu@tamk.fi

Tampere University of Applied Sciences is a multidisciplinary and international higher education institution located in the Tampere Region in Finland. TAMK has around 10,000 undergraduate students. TAMK's expertise ranges from engineering to health and social care and creativity, with special emphasis on practically oriented education and R&D activities. TAMK's profile, both as a modern and dynamic education institution as well as an active project actor, allows for genuine, long-lasting and confidential partnerships with companies and SMEs, as well as with public and third sector organizations.

TAMK has five strategic focus areas where educational excellence is combined with practice oriented, user-driven research, development and innovation actions. Multidisciplinary solutions are developed to meet the changing needs of operational environments. TAMK's research and development portfolio includes a wide range of successful local, regional, national and international projects.

TAMK is an active player in the field of IoT in Finland. TAMK organises a yearly IoT seminar, gathering hundreds of participants from education and industry. IoT lab has been developed in TAMK during the last years to serve not only engineering education but also other disciplines like health and business. IoT lab offers possibilities for teaching staff to use IoT technologies to enrich the learning experiences by offering authentic learning experiences for the students in the classrooms. IoT lab in TAMK collaborates with other universities in Finland. This collaboration enables more authentic learning possibilities for the students. The universities develop improved educational possibilities through IoT in a national project. TAMK also offers Master Degrees in Engineering in the field of IoT. The master program of Information Technology offers students courses covering all topics of IoT technologies.

SOME REMARKS ON THE WEIERSRASS APPROXIMATION THEOREMS

Gavriil Păltineanu
Department of Mathematics and Computer Science,
Technical University of Civil Engineering Bucharest, Romania
E-mail: gavriil.paltineanu@gmail.com

Abstract: The first Weierstrass approximation theorem prove the density of algebraic polynomials in the space of real-valued continuous functions on a finite interval in the uniform topology and the second Weierstrass approximation theorem prove the density of trigonometric polynomials in the space of $2\pi-$periodic real-valued continuous functions on $\boldsymbol{R}$ in the same topology. In this paper the equivalence of this two theorems and an elementary proof of the first Weierstrass approximation theorem are presented.

References

[1]. Paltineanu, G., Bucur, I., *Some Density Theorem in the Set of Continuous Functions with Values in the Unit Interval,* Mediterr, J. (2017) 14:44, *published online* March 2, 2017.
[2].Pinkus, A., *The Weierstrass Approximation Theorems,* Surveys in Approximations Theory, Vol.1, 2005, pp 1-37.