

Series on

**Mathematical Modelling  
of  
Environmental and Life Sciences Problems**

**Proceedings of the fourth workshop  
September, 2005, Constanța, Romania**

## Scientific Committee

Lazăr Dragoş

Marius Iosifescu

George Dincă

Dorel Homentcovschi

Alexandru Morega

Ioan Roşca

Silviu Sburlan

Harry Vereecken

Series on

**M**athematical **M**odelling  
of  
**E**nvironmental and  
**L**ife **S**ciences **P**roblems

Proceedings of the fourth workshop  
September, 2005, Constanța, Romania

Edited by

**Stelian Ion**

**Gabriela Marinoschi**

**Constantin Popa**



EDITURA ACADEMIEI ROMÂNE  
București, 2006

© EDITURA ACADEMIEI ROMÂNE, 2006

All rights reserved.

Addres: EDITURA ACADEMIEI ROMÂNE  
Calea 13 Septembrie nr. 13, sector 5,  
050711, Bucharest, Romania,  
Tel.: 4021-318 81 46, 4021-318 81 06,  
Fax: 4021-318 24 44,  
E-mail: [edacad@ear.ro](mailto:edacad@ear.ro),  
Internet: <http://www.ear.ro>

**Descrierea CIP a Bibliotecii Naționale a României**

**SERIES ON MATHEMATICAL MODELLING OF  
ENVIRONMENTAL AND LIFE SCIENCES PROBLEMS.  
WORKSHOP (2005; Constanța**

**Series on Mathematical modelling of environmental and life  
sciences problems: Proceedings of the fourth Workshop:  
Constanța Romania, september 2005/ed.: Stelian Ion, Gabriela  
Marinoschi, Constantin Popa. - București: Editura Academiei  
Române, 2006**

ISBN (10) 973-27-1358-5 ; ISBN (13) 978-973-27-1358-7

I. Ion, Stelian (ed.)

II. Marinoschi, Gabriela (ed.)

III. Popa, Constantin (ed.)

51(063)

**This book was sponsored by SOFTWIN Bucharest**

Editor: Dan-Florin DUMITRESCU

Technical editing: Sofia MORAR

Computer editing: Stelian ION

Cover design: Mariana ȘERBĂNESCU

---

11.07.2006. Format 16/70×100.

UDC: 519.283: 57(028); 51

---

## Table of Contents

<b>Mathematical aspects of the study of the cavitation in liquids</b> <i>Alina Barbulescu and Cristian Stefan Dumitriu</i> .....	7
<b>Evolutionary Algorithms in Image Reconstruction from Limited Data</b> <i>Andrei Băutu, Elena Băutu and Constantin Popa</i> .....	15
<b>The Dynamics of Systems Modeling Acute Inflammation</b> <i>Cristina Bercia</i> .....	27
<b>Self-Propulsion of an Oscillatory Wing Including Ground Effects</b> <i>Adrian Carabineanu</i> .....	39
<b>Thermal Coupling Numerical Models for Boundary Layer Flows over a Finite Thickness Plate Exposed to a Time-Dependent Temperature</b> <i>Emilia Mladin Cerna and Dorin Stanciu</i> .....	55
<b>Mathematical Modeling of the Dynamic Crack Propagation in a Double Cantilever Beam</b> <i>Eduard-Marius Craciun, Tudor Udrescu and George Cîrlig</i> .....	69
<b>Noise prediction model for wind turbines</b> <i>Alexandru Dumitrache and Horia Dumitrescu</i> .....	79
<b>Quasimonotone ODE Approximation of Nonlinear Diffusion Process</b> <i>Stelian Ion</i> .....	87
<b>Minimum Free Energy Configuration of the Planar Lipidic Bilayer. Analytical Solutions</b> <i>Stelian Ion and Dumitru Popescu</i> .....	95
<b>Numerical Study of Axisymmetric Slow Viscous Flow Past Two Spheres</b> <i>Gheorghe Juncu</i> .....	103
<b>Approach to nonstationary (transient) Birth-Death Processes</b> <i>Alexei Leahu</i> .....	113
<b>Statistical simulation and analysis of some software reliability models</b> <i>Alexei Leahu and Elena Carmen Lupu</i> .....	119
<b>A Mathematical Model Describing the Vulnerability to Pollution of Groundwater in the Proximity of Slatina Town</b> <i>Anca Marina Marinov and Victor Moldoveanu</i> .....	123

<b>Analysis of a Preconditioned CG Method for an Inverse Bioelectric Field Problem</b>	
<i>Marcus Mohr, Constantin Popa and Ulrich R�de</i> .....	135
<b>Dosimetric Estimates in Biological Tissue Exposed to Microwave Radiation in the Near Field of an Antenna</b>	
<i>Mihaela Morega, Alina Machedon and Marius Neagu</i> .....	147
<b>Lower bounds on the weak solution of a moving-boundary problem describing the carbonation penetration in concrete</b>	
<i>Adrian Muntean and Michael B�hm</i> .....	161
<b>Discretization Techniques and Numerical Treatment For First Kind Integral Equations</b>	
<i>Elena Pelican and Elena B�utu</i> .....	171
<b>A fast approximation for discrete Laplacian</b>	
<i>Constantin Popa and Tudor Udrescu</i> .....	181
<b>Gibbs regularized tomographic image reconstruction with DW algorithm based on generalized oblique projections</b>	
<i>Constantin Popa and Rafal Zdunek</i> .....	191
<b>Post-Synaptic Nicotinic Currents Triggered by the Acetylcholine Distribution within the Synaptic Cleft</b>	
<i>Anca Popescu and Alexandru Morega</i> .....	201
<b>Effect of tricyclic antidepressants on the frog epithelium</b>	
<i>Corina Prica, Emil Neaga, Beatrice Macri, Dumitru Popescu and Maria Luiza Flonta</i> .....	211
<b>On the Solvability of Navier-Stokes Equations</b>	
<i>Cristina Sburlan</i> .....	223
<b>The Influence of Initial Fields on the Propagation of Attenuated Waves along an Edge of a Cubic Crystal</b>	
<i>Olivian Simionescu-Panait</i> .....	231
<b>Model for molecular dynamics simulation of radiation-induced defect formation in fcc metals</b>	
<i>Daniel �opu, Bogdan Nicolescu and M.A. G�r�u</i> .....	243

## Mathematical aspects of the study of the cavitation in liquids

Alina Barbulescu\* and Cristian Stefan Dumitriu\*\*

In a liquid, an ultrasonic field can carry along small bubbles or can produce cavitation bubbles, whose movements determine drastic effects as: erosion, unpassivation and emulsification, chemical reactions, sonoluminescence, pressure variation, that have as a effect oscillations whose frequencies differ from that of the incident ultrasound wave.

We found out that the frequency of these electrical signals, generated by the cavitation bubbles, at their exterior, corresponds to the frequency of the mechanical waves, generated by the collapse of the cavitation bubbles. We present the mathematical models for the voltage induced by the cavitation bubbles in diesel.

### 1. Experimental set-up

The cavitation is the process of the appearance of one or many gas cavities in a liquid. An ultrasonic field that goes over a liquid can produce or move cavitation bubbles.

To make the study of the ultrasonic cavitation in liquids, we used an ultrasound generator. The frequency of the ultrasound produced by it is constant.

The experimental set-up consists in a core-tank, which contains the studied liquid. Two metallic electrodes are put in the tank. They are connected with an acquisition card, that digitizes analog signals and stores the resulting digital pattern in the on-board memory.

In some papers ([2], [4], [6]) we studied the voltage induced in water by the cavitation bubbles produced by the ultrasonic generator.

Now we shall make a comparative study of the signals captured in diesel and in crude petroleum.

---

\* “Ovidius” University, Constanța, Romania e-mail: [abarbulescu@univ-ovidius.ro](mailto:abarbulescu@univ-ovidius.ro)

\*\* Utilnavorep S.A., Constanța, Romania.

## 2. Definitions

In order to discuss our results we need some notions concerning the time series analysis.

**Definition 1.** A discrete time process is a sequence of random variables  $(X_t; t \in \mathbf{Z})$ .

**Definition 2.** A discrete time process  $(X_t; t \in \mathbf{Z})$  is called stationary if:

$$(\forall) t \in \mathbf{Z}, E(X_t^2) < \infty,$$

$$(\forall) t \in \mathbf{Z}, E(X_t) = \mu, (\forall) t \in \mathbf{Z},$$

$$(\forall) h \in \mathbf{Z}, \text{Cov}(X_t, X_{t+h}) = \gamma(h).$$

where  $E(X)$  is the expectation value of the random variable  $X$  and  $\text{Cov}(X, Y)$  is the correlation of the random variables  $X$  and  $Y$ .

**Definition 3.** The function defined on  $\mathbf{Z}$ , by:

$$\rho(h) = \frac{\text{Cov}(X_t, X_{t+h})}{\sqrt{\sigma^2(X_t)\sigma^2(X_{t+h})}} = \frac{\gamma(h)}{\gamma(0)}$$

is called the autocorrelation function of the process  $(X_t; t \in \mathbf{Z})$ .

$\sigma^2(X_t)$  is the variance of the variable  $X_t$ .

The most used estimator of  $\rho(h)$  is the empiric autocorrelation function, ACF:

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{n-|h|} (x_t - \bar{x})(x_{t+|h|} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2},$$

where  $x_t$  is a realization of  $X_t$ ,  $h$  is the lag,  $n$  is a fixed natural number and  $\bar{x} = \frac{\sum_{t=1}^n x_t}{n}$  is the arithmetic mean of the values  $x_1, \dots, x_n$ .

**Definition 4.** If  $(X_t; t \in \mathbf{Z})$  is a stationary process, the function defined by:

$$\tau(h) = \frac{\text{Cov}(X_t - X_t^*, X_{t-h} - X_{t-h}^*)}{D^2(X_t - X_t^*)}, \quad h \in \mathbf{Z}_+$$

is called the partial autocorrelation function, where  $X_t^*$  ( $X_{t-h}^*$ ) is the affine regression of  $X_t$  ( $X_{t-h}$ ) with respect to  $X_{t-1}, \dots, X_{t-h+1}$ .

The most used estimator of  $\tau(h)$  is the empiric partial autocorrelation function, PACF.

**Definition 5.** Consider

$$\begin{aligned} B(X_t) &= X_{t-1} \\ \Phi(B) &= 1 - \varphi_1 B - \dots - \varphi_p B^p, \quad \varphi_p \neq 0, \\ \Theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_p B^p, \quad \theta_p \neq 0, \\ \Delta^d X_t &= (1 - B)^d X_t. \end{aligned}$$



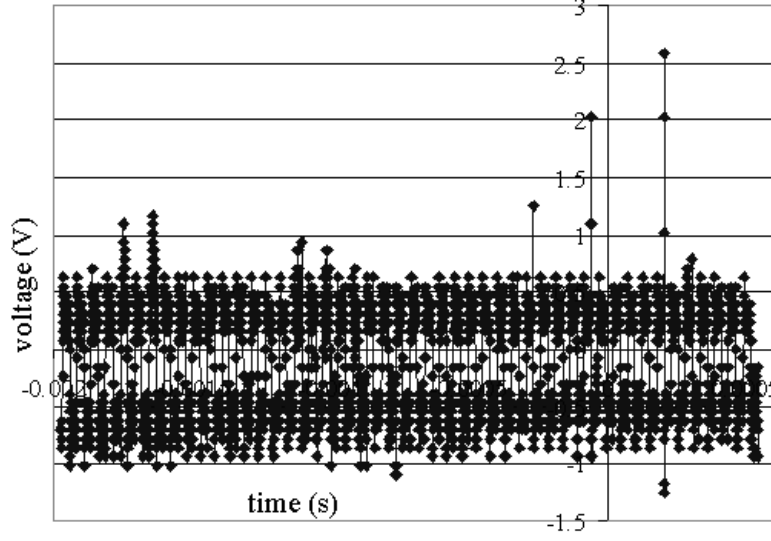


Fig. 1. The voltage induced in diesel.

The process  $(X_t; t \in \mathbf{Z}_+)$  is called an autoregressive integrated moving average process of  $p$ ,  $d$  and  $q$  orders, ARIMA  $(p, d, q)$ , if:

$$\Phi(B)\Delta^d X_t = \Theta(B)\varepsilon_t,$$

where the absolute values of the roots of  $\Phi$  and  $\Theta$  are greater than 1 and  $(\varepsilon_t; t \in \mathbf{Z})$  is a white noise.

If  $d = 0 = q$ , the ARIMA $(p, d, q)$  process is an autoregressive of  $p$  order, AR $(p)$ , process.

If  $p = d = 0$ , the ARIMA $(p, d, q)$  process is an moving average of  $q$  order, MA $(q)$  process.

If  $d = 0$ , the ARIMA $(p, d, q)$  process is an autoregressive moving average of  $p$  and  $q$  orders, ARMA $(p, q)$  process.

**Remarks.** The ARMA $(p, q)$  process is stationary.

### 3. Results

In the Figure 1 we can see the electrical signal induced by the cavitation in diesel (voltage, function of time), captured by the acquisition card and processed by us.

We made the analysis of this signal. It can be seen that there are some aberrant values, that must be removed. After this process, the remained values were studied.

First, the autocorrelation function (ACF) of the voltage, at the lags between 1 and 16, was calculated and the confidence interval, at the confidence level 95%, was determined.

```

Variable:      V
Regressors:    NONE
Non-seasonal differencing: 0
No seasonal component in model.

Parameters:
AR1  AR2  AR3  MA1  MA2  MA3  MA4  MA5  MA6
95.00 percent confidence intervals will be generated.
Split group number: 1  Series length: 5033
No missing data.
Melard's algorithm will be used for estimation.

Termination criteria:
Parameter epsilon: .001
Maximum Marquardt constant: 1.00E+09
SSQ Percentage: .001
Maximum number of iterations: 10

FINAL PARAMETERS:
Number of residuals      5033
Standard error           .13150094
Log likelihood           3071.9924
AIC                     -6125.9849
SBC                     -6067.2709

      Analysis of Variance:
      DF  Adj.  Sum of Squares  Residual Variance
Residuals      5024                86.934397                .01729250

      Variables in the Model:
      B      SEB      T-RATIO      APPROX.  PROB.
AR1      2.3797957      .01630930      145.91647      .00000000
AR2     -2.2495648      .02800818     -80.31813      .00000000
AR3      .8465101      .01384651      61.13528      .00000000
MA1      1.2790430      .02146364      59.59115      .00000000
MA2     -0.8222232      .02608657     -31.51902      .00000000
MA3     -0.0313566      .02580848     -1.81497      .02443314
MA4      .1189112      .02581246      4.60674      .00000419
MA5     -0.0998738      .02364224     -4.22438      .00002438
MA6      .0416114      .01665942      2.49777      .01252949

```

Fig. 2. The coefficients of the model ARIMA(3, 0, 6).

The values of ACF were outside the confidence interval. The form of the ACF was an exponential decreasing and that of PACF was of damped sine wave oscillation. These remarks enable us to think that the process could be of ARIMA type. We also thought at this type of models because the simple models are not convenient point of view of the errors. Point of view of physics, the ARIMA models found by us ([2], [4], [6]) for the signals induced in water satisfy the experimental and theoretical results known from the literature.

Forty models were analyzed. To chose between them, the Schwarz (SBC) and Akaike (AIC) criteria were used. The preferred values were that of the SBC criterion. The selected model – ARIMA(3, 0, 6) – had the least SBC value.

The first step was to test the hypothesis  $H_0$ : *the coefficients of the model are zero*, at the significance level 5%.

The values of the model coefficients and of the  $t$ -ratio are given in the Figure 2.

The values of the  $t$ -ratio (the Figure 2, the last nine rows and the column 4) are greater than the values of the quantila of the Student function with 5 033 liberty degrees, at the significance level 5%.

Also, the probabilities to accept the hypothesis  $H_0$  are practically zero (the last column of the Figure 2), so  $H_0$  is rejected.

In order to prove that the model is a good one, point of view of statistics, the autocorrelation function and the partial autocorrelation function of the residuals were calculated. The graphs of these functions can be seen in the Figures 3 and 4 and their values, in the Tables 1 and 2.

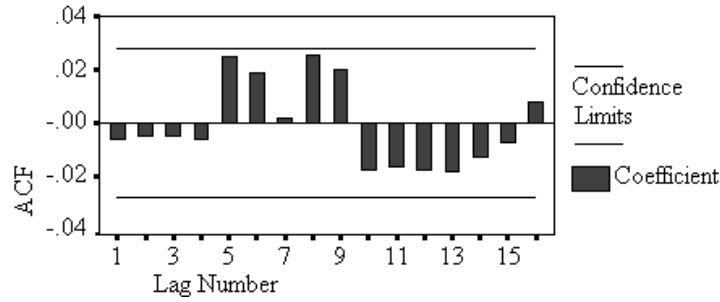


Fig. 3. The ACF of the residuals in the model ARIMA(3, 0, 6).

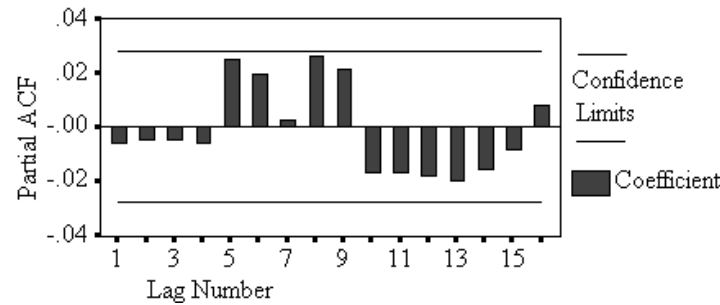


Fig. 4. The PACF of the residuals in the model ARIMA(3, 0, 6).

In the Figures 3 and 4 we see that the values of the autocorrelation function and of the partial autocorrelation function of the residuals are inside the confidence intervals, at 0.95 confidence level.

The following data are provided by the table 1:

- the lags, between 1 and 16 – in the column 1;
- the autocorrelation of the errors – in the column 2;
- in the column 4 – the values of the Box-Ljung statistics, which are in the interval  $[0.184, 18.572]$ , so, less than  $\chi^2(15)$  ;

– in the last column: the probability to accept the hypothesis that the residuals form a white noise, which are between 0.741 and 0.968.

The values of the partial autocorrelation function, at the lags between 1 and 16, are given in the second column of the Table 2. They are very small. Also, the modulus of the standard errors (Tables 1 and 2) is small (0.014).

So, the hypothesis that the residuals form a white noise can be accepted. Therefore, the model is well selected.

*Table 1*  
The values of PACF of the residuals

Lag	Pr-Aut-Corr.	Stand. Err.	-.5	-.25	0	.25	.5
			+-----+-----+-----+-----+				
1	-.006	.014					.*
2	-.005	.014					.*
3	-.005	.014					.*
4	-.006	.014					.*
5	.025	.014					.*
6	.019	.014					.*
7	.002	.014					.*
8	.026	.014					.I*
9	.021	.014					.*
10	-.017	.014					.*
11	-.017	.014					.*
12	-.018	.014					.*
13	-.020	.014					.*
14	-.016	.014					.*
15	-.008	.014					.*
16	.008	.014					.*

Plot Symbols: Autocorrelations \*  
Two Standard Error Limits .

*Table 2*  
The values of ACF of the residuals

Lag	Corr.	Err.	-.5	-.25	0	.25	.5	Box-Ljung Prob.
			+-----+-----+-----+-----+					
1	-.006	.014						.184 .968
2	-.005	.014						.295 .863
3	-.005	.014						.408 .939
4	-.006	.014						.590 .964
5	.025	.014						3.759 .885
6	.019	.014						5.602 .869
7	.002	.014						5.622 .785
8	.026	.014						8.927 .849
9	.020	.014						10.983 .877
10	-.017	.014						12.521 .852
11	-.016	.014						13.865 .741
12	-.018	.014						15.466 .817
13	-.018	.014						17.166 .832
14	-.013	.014						17.994 .907
15	-.007	.014						18.270 .849
16	.008	.014						18.572 .891

Plot Symbols: Autocorrelations \* Two Standard Error Limits .

Total cases: 5033 Computable first lags: 5032

#### 4. Conclusions

In the experiments made we found out that when an ultrasound propagates through a liquid, a potential difference between two points appears. It has both harmonic and subharmonic components.

The equation of the voltage induced in diesel, at 80 W is:

$$V_n - 2.3798V_{n-1} + 2.2495V_{n-2} - 0.8465V_{n-3} = \varepsilon_n - 1.279\varepsilon_{n-1} + \\ + 0.8222\varepsilon_{n-2} + 0.0313\varepsilon_{n-3} - 0.1189\varepsilon_{n-4} + 0.0998\varepsilon_{n-5} - 0.0416\varepsilon_{n-6},$$

where  $n \in \mathbf{N}, n \geq 3$  and  $\{\varepsilon_n, n \in \mathbf{N}\}$  is a white noise.

An analogous study, made for the voltage induced in crude petroleum, in the same condition as for the diesel (Figure 5), conducted us to an ARIMA(3, 1, 4) model, without a constant term.

It was expected that the results don't differ too much, since the chemical compositions of the two liquids were not too different.

The two models differs also from that obtained in water:

– at 80 W, which was ([2]) an AR(2), given by:

$$V_n = 1.5636298V_{n-1} - 0.89193194V_{n-2} + \varepsilon_n,$$

where  $n \in \mathbf{N}, n \geq 3$  and  $\{\varepsilon_n, n \in \mathbf{N}\}$  is a white noise.

– at 120 W, which was of ARMA(2,1) type, given by ([6]):

$$V_n = 1.3006553V_{n-1} - 0.7035790V_{n-2} + \varepsilon_n - 0.6128040\varepsilon_{n-1},$$

where  $n \in \mathbf{N}, n \geq 3$  and  $\{\varepsilon_n, n \in \mathbf{N}\}$  is a white noise.

– at 180 W, which was of ARIMA(2, 1, 0) type, given by ([4]):

$$(1 - 1.2313304B + 0.84409B^2)(1 - B)V_n = \varepsilon_n,$$

where  $n \in \mathbf{N}, n \geq 3$  and  $\{\varepsilon_n, n \in \mathbf{N}\}$  is a white noise.

So, we proved that the voltage induced by the cavitation bubbles in liquids depends on the liquids and on the power of the ultrasonic generator.

#### References

- [1] A. Barbulescu, *Time series, with applications*, Junimea, Iași, 2002 (in Romanian).
- [2] A. Barbulescu, *Some models for the voltage*, Analele Științifice ale Universității Ovidius, Seria Matematica, vol. **10** (2), 2002, pp. 1–7
- [3] A. Barbulescu, V. Marza, *Some results regarding the ultrasonic cavitation*, Acta Universitatis Apulensis, Mathematics–Informatics, Part B, no. 7/2004, pp. 31–38
- [4] A. Barbulescu, V. Marza, *Some models for the voltage induced in a liquid by cavitation*, Proceedings of Conference 2004: Dynamical systems and applications, Antalya, Turkey, 5-10.07.2004, pp. 158–165.

- [5] A. Barbulescu, V. Marza, *Electrical phenomena induced by cavitation in oil*, Scientific Bulletin of the Politehnica University of Timișoara, Transactions on Mechanics, Tome **49**(63).
- [6] A. Barbulescu, V. Marza, *Electrical effect induced at the exterior of the cavitation bubbles*, Acta Physica Polonica (submitted).

## Evolutionary Algorithms in Image Reconstruction from Limited Data

Andrei Băutu<sup>\*‡</sup>, Elena Băutu<sup>\*\*‡</sup> and Constantin Popa<sup>\*\*\*‡</sup>

We consider in this paper two classes of evolutionary methods for improving the ART Kaczmarz procedure in case of data limitation: genetic algorithm and particle swarm optimization, respectively. They are combined in various ways with the classical Kaczmarz projection method, in two classes of hybrid algorithms. Experiments illustrating the efficiency of these new methods are presented for consistent least-squares formulation of some image reconstruction problems.

### 1. Limitation of data in practice and theory

In this section we shall analyse from both practical and theoretical viewpoints the “data limitation” aspect appearing in two very important practical problems: medical computerized tomography (MCT, for short) and electromagnetic geotomography (EGT, for short). The corresponding idealized and simplified (two dimensional) situations are presented in Figures 1 and 2 below. In Figure 1 we supposed that for each position of the CT scanner, only one X-ray is emitted ( $S_i R$ ,  $i = 1, 2, \dots, m$ ).  $S_i$  is the source and  $R$  is the receptor; the body-section which is analysed is “contained” in the rectangular region  $ABCD$ .

Figure 2 describes an EGT problem;  $ABCD$  is the rectangular underground region which is analysed,  $AB$  and  $CD$  are holes in which are introduced electromagnetic waves sources  $S_1, S_2, \dots, S_p$  and receptors  $R_1, R_2, \dots, R_q$ , respectively (see [2]).

In both cases, the rectangular regions  $ABCD$  are uniformly discretized in a number  $n$  of pixels,  $P_1, P_2, \dots, P_n$  (see figure 3).

---

\* “Mircea cel Bătrân” Naval Academy, Constanța, Romania, e-mail: [abautu@anmb.ro](mailto:abautu@anmb.ro)

\*\* “Ovidius” University, Constanța, Romania, e-mail: [erogojina@univ-ovidius.ro](mailto:erogojina@univ-ovidius.ro)

\*\*\* “Ovidius” University, Constanța, Romania, e-mail: [cpopa@univ-ovidius.ro](mailto:cpopa@univ-ovidius.ro)

‡ This paper was supported by the PNCDI INFOSOC Grant 131/2004.

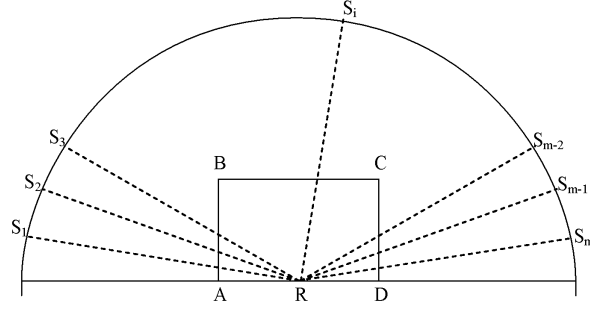


Fig. 1. The MCT problem.

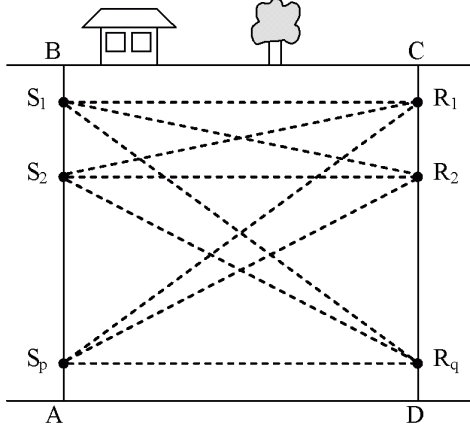


Fig. 2. The EGT problem.

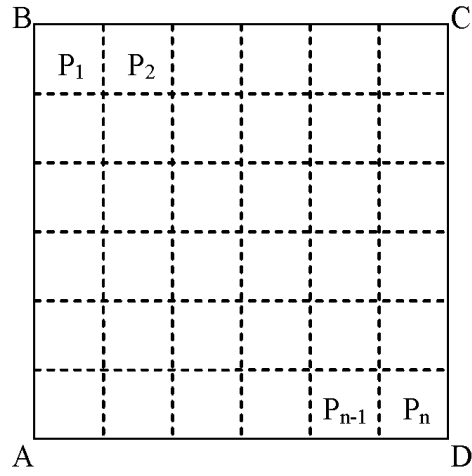


Fig. 3. Pixels discretization.

The mathematical model of the reconstruction procedure, for both EGT and MCT problems is a linear least-squares formulation,

$$\min \|Ax - b\| \quad (1)$$

with  $A$  an  $m \times n$  matrix and  $b \in \mathbb{R}^m$ . The right hand side  $b$  is constructed by measuring the X-rays or electromagnetic waves intensities at sources and receptors (see [3], [5] for details). Concerning the matrix  $A$ , the number  $n$  of its columns is exactly the number of pixels in the discretization from figure 3, whereas the number  $m$  of its rows corresponds to the number of X-rays in the MCT case ( $S_i R$  denoted by  $E_i$ ,  $i = 1, 2, \dots, m$ ) or electromagnetic waves source-receptor combinations in the EGT case ( $S_k R_l$ ,  $k = 1, \dots, p$ ,  $l = 1, \dots, q$  denoted by  $E_i$ ,  $i = 1, 2, \dots, m = pq$ ). The value of the  $(A)_{ij}$  component is the length of the segment intersection between the  $E_i$  ray and the pixel  $P_j$  (see Figure 4). If such an intersection is empty, the



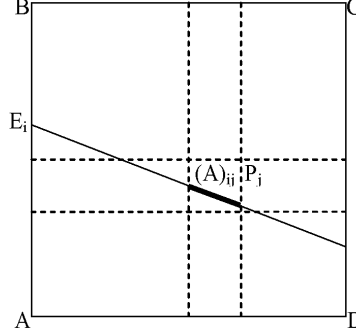


Fig. 4. Matrix coefficients.

corresponding  $(A)_{ij}$  coefficient is set to 0. Following such a construction, the matrix  $A$  becomes sparse, rank-deficient and ill-conditioned (see [3] for details). Moreover, because of measurement errors, the right hand side  $b$  fails to belong to the range of  $A$ , thus the problem (1) becomes inconsistent. In the present paper we shall consider only the consistent case of (1) and let the inconsistent one for a future work.

**Remark 1.** *In any of the above mentioned cases, for the discrete reconstruction problem (1), in practical applications we are looking for its unique minimal norm solution, denoted in what follows by  $x_{LS}$ .*

By “data” associated to problem (1) we understand the matrix  $A$  and the vector  $b$ . Moreover, because the number  $n$  of pixels in the discretization from figure 3 is imposed by technical reasons, we may consider our “data” as essentially determined by  $m$ , the number of X-rays/electromagnetic waves that are used for scanning  $ABCD$ . From a practical point of view, this number is limited at least by the following two reasons:

- for MCT – a too big number  $m$  of X-rays used for scanning a body can become dangerous for its health;
- for EGT – in practice, the underground analysed region  $ABCD$  is very big, thus the length of the holes  $AB$  and  $CD$  is so; then, if we would like a “complete” scanning of the area we would need a very big number of sources and receptors, which is not possible from technical reasons.

From a theoretical view point the above described “data limitation” can be interpreted according to some properties of the fundamental vector subspaces associated to the problem matrix  $A$ . In this sense we shall denote by  $A^t$ ,  $N(A)$ ,  $R(A)$ ,  $A^+$  the transpose, null space, range and Moore-Penrose pseudoinverse of  $A$  and by  $S(A; b)$ , the set of all solutions of (1) in the consistent case, in which (1) can be written in the classical formulation

$$Ax = b. \quad (2)$$

It is well known (see [2]) that

$$S(A; b) = x_{LS} + N(A), \quad x_{LS} \perp N(A), \quad (3)$$

where  $\perp$  denotes the orthogonality w.r.t. the euclidean scalar product  $\langle \cdot, \cdot \rangle$ . According to (3), “data limitation” in problem (2) would mean that the null space  $N(A)$  has an “enough big” dimension (e.g. a factor  $c$  times the total dimension  $n$  of the support space  $\mathbb{R}^n$ ). The above mentioned practical view points about data limitation are fitting into this considerations because, if the number of rays  $m$  is strictly less than the number of pixels  $n$ , then the dimension of  $N(A)$  is positive (but we may have a positive dimension for  $N(A)$  also for  $m > n$ ; see the example from below and the results in Table 1. As a consequence, because any solution  $x^* \in S(A; b)$  is given by

$$x^* = x_{LS} + P_{N(A)}(x^*), \quad (4)$$

where  $P_S(x)$  denotes the (euclidean) orthogonal projection onto the subspace  $S$ , if the null space  $N(A)$  is “big enough” w.r.t the support space  $\mathbb{R}^n$  (see the above considerations and the remark 1) and  $x^*$  has a corresponding “big” component, then the difference

$$x^* - x_{LS} = P_{N(A)}(x^*) \quad (5)$$

becomes important enough to destroy the accuracy of the reconstructed image. All these considerations are illustrated by the following example in which a real image reconstruction EGT problem is simulated (the same procedure will be used in the experiments from section 3 of the paper).

**Example 1.** *Our simulation procedure is the following: we consider an image artificially created (see Figure 5) as a vector  $x^{ex} \in \mathbb{R}^n$ , for a given number  $n \geq 2$  of pixels (as in Figure 3). Each component  $x_i^{ex}$  is a real number in the interval  $[0, 1]$  and corresponds to the grey mapping scale from Figure 6. This gives us the grey original image in Figure 5 (in this case we have  $n = 144$ ). Then, the original image area was scanned as in Figure 2, by using  $p \geq 1$  sources and  $q \geq 1$  receptors, equally distributed on AB and CD. In this way we obtained the  $m \times n$  system matrix  $A$  (see Figure 4) with  $m = pq$ .*

Table 1  
Limited data tests characteristics

Scanning ( $p = q$ )	$m$	$n$	$\text{rank}(A)$	$\dim(N(A))$
$6 \times 6$	36	144	35	109
$8 \times 8$	64	144	61	83
$10 \times 10$	100	144	96	48
$12 \times 12$	144	144	120	24
$16 \times 16$	256	144	131	13
$24 \times 24$	576	144	133	11
$36 \times 36$	1296	144	133	11
$48 \times 48$	2304	144	133	11

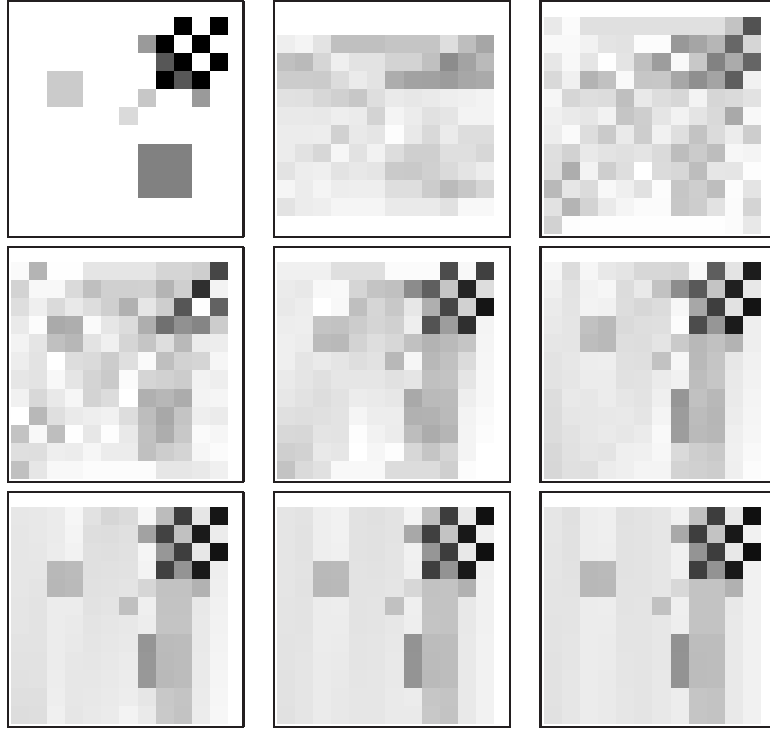


Fig. 5. Original image (upper left) and reconstruction results for  $p$  sources and  $p$  receivers (where  $p \in \{6, 8, 10, 12, 16, 24, 36, 48\}$ ).



Fig. 6. Grey mapping scale

The right hand side  $b \in \mathbb{R}^m$  was defined by

$$b = Ax^{ex}, \quad (6)$$

such that the problem (1) becomes consistent. For the tests from figure 5 we used  $p = q$  (i.e.  $m = p^2$ ) with  $p \in \{6, 8, 10, 12, 16, 24, 36, 48\}$  (for  $p \in \{6, 8, 10, 12\}$  we have  $m \leq n$ , i.e. the “practical” limited data case). The values of the  $\text{rank}(A)$  and  $\text{dim}(A)$  were computed in each case and are presented in Table 1. The original image  $x^{ex}$  from Figure 5 has components in  $N(A)$ , thus (see (5))

$$x^{ex} \neq x_{LS}. \quad (7)$$

We then applied to (1) the classical Kaczmarz algorithm (KA, for short) with the initial approximation  $x^0 = 0$  (which gives us  $x_{LS}$  as the limit of the sequence of approximations; see e.g. [2], [3], [5]). After 100 iterations for the cases presented in Table 1 we got the results from Figure 5. We can observe there that in the “practical limited data” situations ( $p = q = 6, 8, 10, 12$ ) the reconstructed images (which corresponds to  $x_{LS}$ ) are far from the original one (the first 5 images following the original one), whereas for “enough much data” (e.g.  $p = q = 24, 36, 48$ ) the reconstructed images are closer to the original, but still not satisfactory (this because also in these cases – see Table 1 – the null space of  $A$  has dimension 11, which is still big w.r.t. the support space dimension  $n = 144$ ).

Thus, in order to get a good enough approximation of  $x^*$ , with Kaczmarz’s algorithm used in Example 1, we need to start it with an initial approximation  $x^0 \in \mathbb{R}^n$  such that

$$P_{N(A)}(x^0) \approx P_{N(A)}(x^*). \quad (8)$$

According to (7) we will then have

$$\lim_{h \rightarrow \infty} x^h = P_{N(A)}(x^0) + x_{LS} \approx P_{N(A)}(x^*) + x_{LS} = x^*, \quad (9)$$

thus, a much better approximation for  $x^*$ . For generating such a “good” initial approximation  $x^0$  as in (8) we decided to use two evolutionary algorithms. They will be described in the next section, whereas in the last one we shall combine one of them with the previous Kaczmarz algorithm in our numerical experiments.

## 2. Evolutionary Algorithms

Evolutionary algorithms, like **genetic algorithms** (GA) and **particle swarm optimization** algorithms (PSO), can be used to solve problems like ((1)) and ((2)). Evolutionary algorithms are stochastic algorithms that use a set of candidate solutions (called individuals) which evolve in time towards better solutions. Each individual is rated by a fitness function. The algorithms presented in this paper use a fitness function that minimizes the errors defined as

$$f_f(x) = \frac{1}{1 + \|Ax - b\|}, \quad (10)$$

where  $x \in [0, 1]^n$  is the current image (see Example 1).

## 2.1. Genetic Algorithm

The GA, which we tested, is an extension of the unary function optimization GA described in [4]. In the GA view, a possible solution is an organism that is adapting to the its environment in order to fit better within it. A GA's set of possible solutions is called **population**. Each solution from the population is called a **chromosome**. Data structures that define a chromosome are called **genes**. Most genetic algorithms work by evolving their population with the help of three genetic operators: **crossover**, **mutation**, and **selection**.

Our GA implementation uses a fixed size population of *pop\_size* chromosomes, each chromosome being a vector with elements from  $[0, 1]$  (the chromosome's genes). The values of the  $i^{th}$  gene represents the absorption value of the  $i^{th}$  pixel in the image.

Crossover is a binary operator which combines genes from a two chromosomes (called **parents**) to obtain two of new chromosomes (called **offsprings**), which replace their parents in the population. Chromosomes are selected for mating with *cross\_rate* probability. For each pair of chromosomes, a random crossover point is selected and genes to the left of that point are swapped between chromosomes.

Mutation is a unary operator that affects the genes of a single chromosome. Each gene suffers mutation with *mut\_rate* probability. It modifies each gene with a given probability by replacing the old value with a randomly generated new value.

Selection is a population-wide operator that creates a new population based on the old one. There are many possible selection operators available, but we implemented a classical Monte Carlo selection scheme (called **roulette wheel selection**).

For our genetic algorithm the values range of the above described parameters are the following: for *pop\_size* from 10 to 200, *mut\_rate*—1% to 25%, *cross\_rate*—30% to 70%.

## 2.2. Particle Swarm Optimization Algorithm

We tested an  $n$ -dimensional extension of the PSO algorithm described in [4]. In the PSO view, a possible solution is an organism that is moving in the search space in order to find better places than their current location. The set of possible solutions is called **swarm**). Each solution from the swarm is called a **particle**. A particle state is defined by **current position**, **velocity**, **memory**, and **neighbours**.

In our implementation, a particle is a vector of  $n$  tuples of reals from  $[0, 1]$ . The values in the  $i^{th}$  tuple represent the current position (i.e. pixel absorption value), velocity and best so far position of the particle in  $i^{th}$  search dimension.

The number of particles is denoted by *part\_count* parameter. Each particle evolves based on its own memory and its neighbours memory (only one neighbour in our implementation). On each iteration, velocity and position components of particles are updated using the formulas:

$$v'_{t+1} = v_t \cdot inertia + \text{rand}() \cdot cognitive \cdot (b_t - p_t) + \text{rand}() \cdot social \cdot (n_t - p_t), \quad (11)$$

$$v_{t+1} = \min(v_{\max}, \max(-v_{\max}, v'_{t+1})), \quad (12)$$

$$p'_{t+1} = p_t + v_{t+1}, \quad (13)$$

$$p_{t+1} = \min(1, \max(0, p'_{t+1})). \quad (14)$$

The values used for the above described parameters are the following: for *part\_count* from 10 to 100, *inertia*—1% to 50%, *cognitive* and *social*—75% to 200%,  $v_{\max}$ —0.1 to 1.0.

### 2.3. Hybrid Algorithms

As we presented in the experiments section of the paper, the genetic algorithm performed is quite bad compared to PSO and Kaczmarz algorithms. Hybrid algorithms from Kaczmarz and PSO were created in order to combine the benefits of this two algorithms.

First hybrid algorithm (FHA) has two stages: during the first stage it runs the PSO algorithm described earlier; in the second stage, it uses the best solution from the first stage as the starting approximation for Kaczmarz algorithm.

Second hybrid algorithm (SHA) runs Kaczmarz and PSO algorithms in an alternate manner, by applying for each particle, after each iteration of PSO one step of the Kaczmarz algorithm.

## 3. Experiments

We present our results for four image reconstruction experiments. The simulation procedure is the same as in Example 1. The four original images are showed in Figure 6 and their characteristics in Table 2 (left to right, according to Figure 7). For each image, we ran all five algorithms with various settings, and we present the results after 100 iterations with the following parameters:

- KA: no settings required;
- GA: *pop\_size* = 50, *mut\_rate* = 5%, and *cross\_rate* = 70%;
- PSO: *part\_count* = 50, *vmax* = 1.0, *inertia* = 0.3, *cognitive* = 1.2, and *social* = 1.2;
- FHA and SHA: same settings as for PSO.

Table 2  
Test images properties

Image	Size (pixels)	Source	Unique colors
Test 1	$8 \times 8$	Drawing	9
Test 2	$12 \times 12$	Drawing	8
Test 3	$20 \times 20$	Scanned photo	71
Test 4	$40 \times 40$	Scanned photo	177

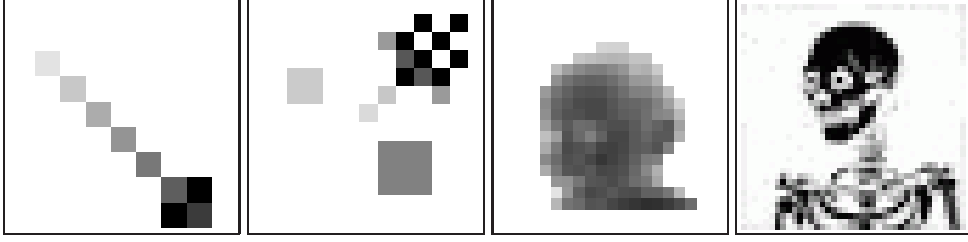


Fig. 7. Test images.

After 100 iterations, GA doesn't produce a good image (Figure 8). We suspect that a different encoding scheme or operators should be tried (there are many genetic operators and they can be combined to form different algorithms).

PSO has some not very satisfactory results for Test 1 and Test 2 (see Figure 9) and has no valid solution for Test 3 and Test 4 images.

For Test 1 and 2, KA found some images which are affected by noise (see Figure 10), which is common in (classical) Kaczmarz image reconstructions (see [3], [2]). This noise is making further improvements difficult. For Test 3 and Test 4 it has no satisfactory results.

FHA performed much better than PSO and a little better than KA (see Figure 11). Reconstructed images are affected by noise of the KA in the second stage of the algorithm.

SHA found an almost perfect Test 1 and Test 2 solution and it found very good solutions for Test 3 and Test 4 (see figure 12). It performed much better than Kaczmarz algorithm alone or the other evolutionary algorithms presented in this article. It seems that KA drives the reconstruction process towards the "good" image, while PSO helps filtering the image in order to eliminate the noise.

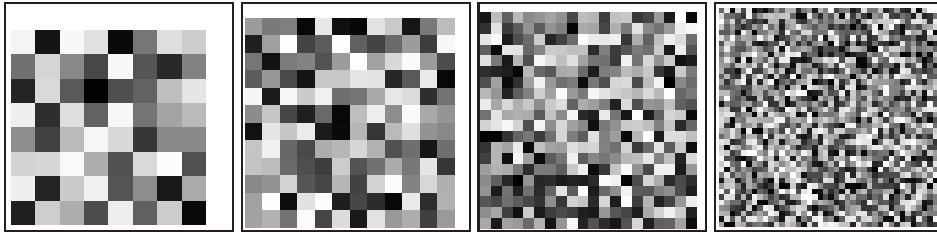


Fig. 8. GA results after 100 iterations.

#### 4. Conclusions and Future work

The consideration and results from this paper are at a very beginning. The first next step will be to apply them to inconsistent least-squares formulations of the type

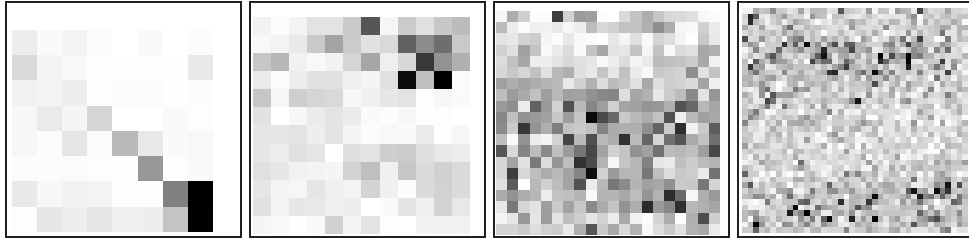


Fig. 9. PSO results after 100 iterations.

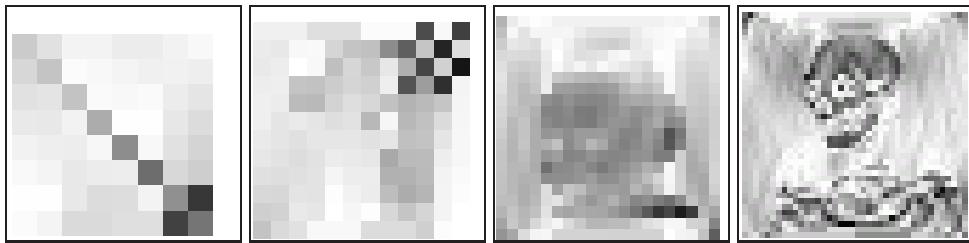


Fig. 10. KA results after 100 iterations.

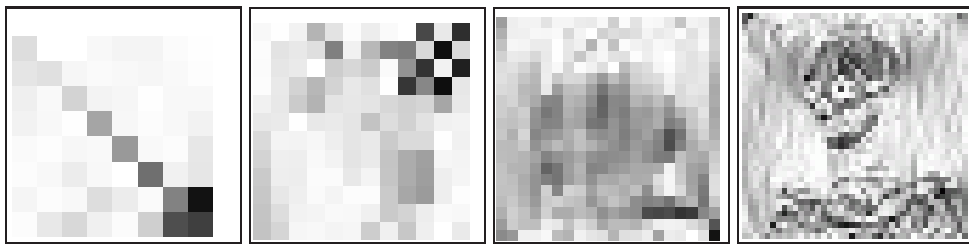


Fig. 11. FHA results after 100 iterations.

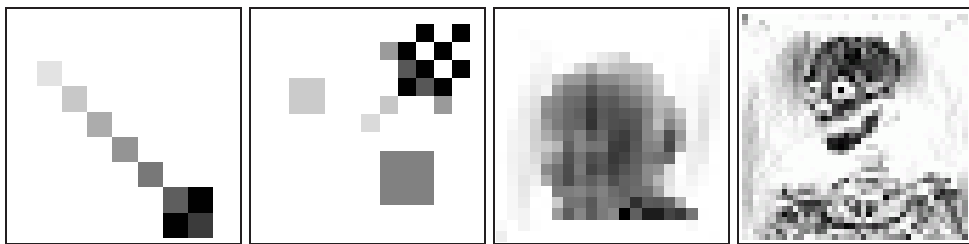


Fig. 12. SHA results after 100 iterations.



(1), whereas a further research will be concerned with a theoretical analysis of the new obtained image reconstruction methods, together with comparisons with other new and efficient techniques in the field (see e.g. [3]).

## References

- [1] Björck, A., *Numerical methods for least squares problems*, SIAM Philadelphia, 1996.
- [2] Censor Y., Stavros A. Z., *Parallel optimization: theory, algorithms and applications*, “Numer. Math. and Sci. Comp.” Series, Oxford Univ. Press, New York, 1997.
- [3] Kennedy, J., Eberhart, R., *Particle Swarm Optimisation*, Proceedings of IEEE International Conference on Neural Networks, Perth, Western Australia, 1995, Vol. 3, 1942–1948.
- [4] Michalewicz, Z., *Genetics Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, New York, 1996.
- [5] Natterer F., *The Mathematics of Computerized Tomography*, John Wiley and Sons, New York, 1986.
- [6] Popa C., Zdunek R., *Kaczmarz extended algorithm for tomographic image reconstruction from limited-data*, Math. and Computers in Simulation, **65** (2004), 579–598.



## The Dynamics of Systems Modeling Acute Inflammation

Cristina Bercia\*

In this paper, we consider three-dimensional non-linear dynamical systems of predator-prey type with nine parameters, which model the acute inflammation of the human body due to an infection, defined clinically as sepsis.

We establish the domains in the parameters space where the equilibrium points exist and the set of conditions for them to be local attractors. We also perform bifurcation analysis and establish the types of dynamics of the systems. We obtain the global phase portrait for each type of dynamics by numerical integration. Varying one or two of the parameters, bifurcation diagrams are presented.

### 1. Introduction

In this paper we have considered two systems of ordinary differential equations with 9 parameters which model the acute inflammation and septic shock of the human body due to an infection or a trauma. The first model is presented in the paper of Brause [1], the second one in Kumar et al. [2].

The differential systems, having 3 variables  $u = (u_1, u_2, u_3)$ , are of predator-prey type

$$u' = F(u), \quad F(u) = \begin{pmatrix} \alpha_1 u_1(1 - u_1) - \alpha_2 u_1 u_2 \\ -\beta_1 u_2 + u_2(1 - u_2)(\beta_2 u_1 + \beta_3 u_3) \\ -\gamma_1 u_3 + \gamma_2 h((u_2 - \theta)/\gamma_3) \end{pmatrix}. \quad (1)$$

Here  $u_1$  and  $u_2$  are concentrations, so they have values in  $[0, 1]$  and the domain of system's variables is  $\mathcal{D} := [0, 1]^2 \times (0, \infty)$ . The parameters  $\alpha_i$ ,  $i = \overline{1, 2}$ ,  $\beta_i$ ,  $\gamma_i$ ,  $i = \overline{1, 3}$ , and  $\theta$  are strictly positive.

---

\* Polytechnica University of Bucharest, Department of Mathematics, Romania, e-mail: [c.bercia@math.pub.ro](mailto:c.bercia@math.pub.ro)

Next, we shall study all the possible dynamics of these systems and for these we need the stability and bifurcation analysis.

## 2. Brause's system

In this case  $u \equiv (P, M, D)$ , where  $P$  represents the influence of the pathogen agent on the organism,  $M$  the immunological response which implies the macrophage cells and  $D$  is the cell damage caused by infection. The number of macrophages grows in the presence of  $P$ , while  $D$  will also cause  $M$  to grow. The amount of additional damage is indicated by a sigmoid function,  $h(x) = \frac{1}{1+\exp(x)}$ , depending on  $M$ .

The system (1) has maximum four equilibria:

- $E_1 (0, 0, D_1)$ , where  $D_1 = \frac{\gamma_2}{\gamma_1 \left(1 + \exp\left(-\frac{\theta}{\gamma_3}\right)\right)}$ ;
- $E_2 (0, M_2, D_2)$ , where  $D_2 = \frac{\beta_1}{\beta_3 (1 - M_2)}$  and  $M_2$  verifies the equation

$$f_1(M) \equiv \gamma_1 \beta_1 \left(1 + \exp\left(\frac{M - \theta}{\gamma_3}\right)\right) - \gamma_2 \beta_3 (1 - M) = 0 \quad (2)$$

- $E_3 (1, 0, D_1)$
- $E_4 (P^*, M^*, D^*)$  – the single interior equilibrium, where  $M^* = \frac{\alpha_1}{\alpha_2} (1 - P^*)$ ,  $D^* = \frac{\gamma_2}{\gamma_1 \left(1 + \exp\left(\frac{\alpha_1(1-P^*) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)\right)}$  and  $P^*$  verifies the equation

$$g_1(P) \equiv \beta_2 P + \frac{\gamma_2 \beta_3}{\gamma_1 \left(1 + \exp\left(\frac{\alpha_1(1-P) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)\right)} - \frac{\alpha_2 \beta_1}{\alpha_2 - \alpha_1 (1 - P)} = 0 \quad (3)$$

For the existence of the four equilibria we shall formulate the following two lemmata.

**Lemma 1.** *The equilibrium points  $E_1$  and  $E_3$  exist in the domain  $D$  for every parameter combinations.*

**Lemma 2.** *The equilibrium  $E_2$  exists in  $\mathcal{D}$  iff*

$$\beta_1 \leq \frac{\gamma_2 \beta_3}{\gamma_1 \left(1 + \exp\left(-\frac{\theta}{\gamma_3}\right)\right)} := \beta_1^{(2)}.$$

*Proof.* The function  $f_1$  in the left hand side of equation (2) is increasing, so the equation has a single solution  $M_2 \in [0, 1]$  if and only if  $f_1(0) \leq 0$  and  $f_1(1) \geq 0$  which are equivalent to  $\beta_1 \leq \beta_1^{(2)}$ . ■

**Lemma 3.** a) *Assume that  $\alpha_1 < \alpha_2$ . An unique interior equilibrium  $E_4$  exists if and only if  $\beta_1^{(1)} < \beta_1 < \beta_1^{(3)}$ , where*

$$\beta_1^{(1)} = \frac{(\alpha_2 - \alpha_1) \gamma_2 \beta_3}{\alpha_2 \gamma_1 \left(1 + \exp\left(\frac{\alpha_1 - \theta \alpha_2}{\alpha_2 \gamma_3}\right)\right)}, \quad \beta_1^{(3)} = \beta_2 + \frac{\gamma_2 \beta_3}{\gamma_1 \left(1 + \exp\left(-\frac{\theta}{\gamma_3}\right)\right)}. \quad (4)$$

b) If  $\alpha_1 \geq \alpha_2$ , then  $E_4$  exists and is unique if and only if  $\beta_1 < \beta_1^{(3)}$ .

*Proof.* The first component of  $E_4$  verifies the equation (3),  $g_1(P^*) = 0$ . But  $g_1'(P) > 0$ ,  $\forall P \in (0, 1)$ , so the equation can have only one solution if and only if  $g_1(0) < 0 < g_1(1)$ . These inequalities take the form  $\frac{\alpha_2 \beta_1}{\alpha_2 - \alpha_1} > \frac{\gamma_2 \beta_3}{\gamma_1 (1 + \exp(\frac{\alpha_1 - \theta \alpha_2}{\alpha_2 \gamma_3}))} \Leftrightarrow \beta_1 > \beta_1^{(1)}$  and  $\beta_2 + \frac{\gamma_2 \beta_3}{\gamma_1 (1 + \exp(-\frac{\theta}{\gamma_3}))} > \beta_1$ . Note that  $M^* \in (0, 1) \Leftrightarrow P^* > 1 - \frac{\alpha_2}{\alpha_1}$  which is satisfied in case a). For b), the function  $g_1$  has the limit  $-\infty$  when  $P$  decrease to  $1 - \frac{\alpha_2}{\alpha_1}$ , so we have an unique solution for  $g_1(P) = 0 \Leftrightarrow g_1(1) > 0$ . ■

**Remark 1.** Always  $\beta_1^{(1)} < \beta_1^{(2)} < \beta_1^{(3)}$ . Also,  $\beta_1^{(1)} \leq 0$  for  $\alpha_1 \geq \alpha_2$ .

So we proved the following

**Proposition 1.** i) For  $\alpha_1 < \alpha_2$ ,  $0 < \beta_1 \leq \beta_1^{(1)}$ , there are only  $E_1$ ,  $E_2$  and  $E_3$ ;  
 ii) For  $\alpha_1 < \alpha_2$ ,  $\beta_1^{(1)} < \beta_1 \leq \beta_1^{(2)}$ , or  $\alpha_1 \geq \alpha_2$ ,  $0 < \beta_1 \leq \beta_1^{(2)}$ , there exist all the four equilibria;  
 iii) For  $\beta_1^{(2)} < \beta_1 < \beta_1^{(3)}$ , there exist only  $E_1$ ,  $E_3$  and  $E_4$ ;  
 iv) For  $\beta_1 \geq \beta_1^{(3)}$ , there are  $E_1$  and  $E_3$ , only.

## 2.1. The stability of the equilibrium points

The Jacobian matrix of the system (1), evaluated at  $E_1$  has the eigenvalues  $\lambda_1 = \alpha_1 > 0$ ,  $\lambda_2 = -\gamma_1 < 0$  and  $\lambda_3 = \frac{\beta_3 \gamma_2}{\gamma_1 (1 + \exp(-\frac{\theta}{\gamma_3}))} - \beta_1$ . So we proved

**Lemma 4.** The equilibrium  $E_1$  is of saddle type, repulsive in the direction  $OP$  and attractive in the direction  $OD$ .

**Lemma 5.** The equilibrium point  $E_2$  is asymptotically stable if and only if  $\beta_1 < \beta_1^{(1)}$ , while for  $\beta_1^{(1)} < \beta_1 < \beta_1^{(3)}$ ,  $E_2$  is of saddle type.

*Proof.* The eigenvalues corresponding to  $E_2$  are  $\lambda_1 = \alpha_1 - \alpha_2 M_2$  and

$$\lambda_2 + \lambda_3 = -\gamma_1 - \frac{\beta_1 M_2}{1 - M_2} < 0, \quad \lambda_2 \lambda_3 = \frac{\gamma_1 \beta_1 M_2}{1 - M_2} \frac{\gamma_2 \beta_3 (\gamma_3 + 1 - M_2) - \gamma_1 \beta_1}{\gamma_2 \gamma_3 \beta_3}.$$

$M_2$  verifies the equation (2), so the product  $\lambda_2 \lambda_3 \geq 0$ . But  $\lambda_2 + \lambda_3 < 0$ , hence  $\text{Re}(\lambda_{2,3}) \leq 0$ . Note that  $\lambda_2$  or  $\lambda_3$  are zero if  $M_2 = 0 \Leftrightarrow \beta_1 = \beta_1^{(2)}$ .

Hence  $E_2$  is asymptotically stable if  $\frac{\alpha_1}{\alpha_2} < M_2$ . Because  $f_1$  defined in (2) is increasing, this inequality is equivalent to  $f_1\left(\frac{\alpha_1}{\alpha_2}\right) < f_1(M_2) = 0 \Leftrightarrow \beta_1 < \beta_1^{(1)}$ . Note that  $\lambda_1 = 0$  if  $\beta_1 = \beta_1^{(1)}$ . ■

The Jacobian of the system (1) at  $E_3$  has the eigenvalues  $\lambda_1 = -\alpha_1$ ,  $\lambda_2 = -\gamma_1$ ,  $\lambda_3 = \beta_2 - \beta_1 + \frac{\beta_3 \gamma_2}{\gamma_1 (1 + \exp(-\frac{\theta}{\gamma_3}))} = \beta_1^{(3)} - \beta_1$ . So, we can formulate

**Lemma 6.** *The equilibrium point  $E_3$  is asymptotically stable if  $\beta_1 > \beta_1^{(3)}$ . For  $\beta_1 < \beta_1^{(3)}$ ,  $E_3$  is of saddle type, attractive in the directions  $OP$  and  $OD$ .*

**Lemma 7.** *If the interior equilibrium  $E_4$  exists, it is asymptotically stable.*

*Proof.* The eigenvalues corresponding to  $E_4$  verify the characteristic equation  $\lambda^3 + A_1\lambda^2 + A_2\lambda + A_3 = 0$ , where  $A_1 = \gamma_1 + \alpha_1 P^* + \frac{\beta_1 \alpha_1 (1-P^*)}{\alpha_2 - \alpha_1 (1-P^*)}$ ,

$$\begin{aligned} A_2 &= \frac{\beta_1 \alpha_1 (\gamma_2 + \alpha_1 P^*) (1-P^*)}{\alpha_2 - \alpha_1 (1-P^*)} + \frac{\alpha_1 \beta_2}{\alpha_2} P^* (1-P^*) (\alpha_2 - \alpha_1 (1-P^*)) \\ &\quad + \frac{\alpha_1 \beta_3 \gamma_2}{\alpha_2^2 \gamma_3} (1-P^*) (\alpha_2 - \alpha_1 (1-P^*)) \frac{\exp\left(\frac{\alpha_1 (1-P^*) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)}{\left(1 + \exp\left(\frac{\alpha_1 (1-P^*) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)\right)^2}, \\ A_3 &= \alpha_1 P^* (1-P^*) \left( \frac{\alpha_1 \beta_1 \gamma_1}{\alpha_2 - \alpha_1 (1-P^*)} + \frac{\beta_2 \gamma_1}{\gamma_2} (\alpha_2 - \alpha_1 (1-P^*)) \right) \\ &\quad + P^* (1-P^*) \frac{\alpha_1^2 \beta_3 \gamma_2}{\alpha_2^2 \gamma_3} \frac{(\alpha_2 - \alpha_1 (1-P^*)) \exp\left(\frac{\alpha_1 (1-P^*) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)}{\left(1 + \exp\left(\frac{\alpha_1 (1-P^*) - \theta \alpha_2}{\alpha_2 \gamma_3}\right)\right)^2}. \end{aligned}$$

The necessary and sufficient condition for  $E_4$  to be asymptotically stable is given by the Ruth-Hurwitz criterion  $A_1 A_2 > A_3$  and  $A_1, A_3 > 0$ . If  $E_4$  exists then the last two inequalities are verified since  $\alpha_2 - \alpha_1 (1-P^*) > 0$  (see Lemma 3). Straightforward computation shows that  $A_1 A_2 - A_3 > 0$ . ■

In consequence, we find the local behavior of the system (1) around the four equilibrium points, depending on  $\beta_1$ .

**Proposition 2.** i) *For  $\alpha_1 < \alpha_2$  and  $0 < \beta_1 < \beta_1^{(1)}$ , the equilibria are:  $E_2$  asymptotically stable and  $E_1, E_3$  saddle points.*

ii) *For  $\max\{0, \beta_1^{(1)}\} < \beta_1 < \beta_1^{(2)}$ , the existing equilibrium points are:  $E_4$  asymptotically stable,  $E_1, E_2$  and  $E_3$  saddle points.*

iii) *For  $\beta_1^{(2)} < \beta_1 < \beta_1^{(3)}$ , there exist the equilibrium  $E_4$  asymptotically stable,  $E_1$  and  $E_3$  saddle points.*

iv) *For  $\beta_1 > \beta_1^{(3)}$ , only  $E_3$  is asymptotically stable and  $E_1$  is saddle point.*

## 2.2. Bifurcation analysis

We consider  $\beta_1$  as a control parameter, while the other parameters are fixed. The differential system (1) takes the form  $u' = G(u, \beta_1)$ ,  $u = (P, M, D)$ . We plot in Fig. 1 (left), the static bifurcation diagram which is the projection of the equilibrium curves  $G(u, \beta_1) = 0$  in the space  $(P, M, \beta_1) \in [0, 1]^2 \times (0, \infty)$ . Solid and broken lines correspond to stable, respectively unstable, equilibrium points. Note that the equilibria exhibit only static bifurcation, since their eigenvalues can't be on the imaginary axis. In the diagram, the points of static bifurcation are  $(P_2, M_2, \beta_1^{(1)})$ ,  $(P_1, M_1, \beta_1^{(2)})$  and  $(P_3, M_3, \beta_1^{(3)})$ . We took  $\alpha_1 = 0.08$ ,  $\alpha_2 = 0.5$ ,  $\beta_2 = 0.2$ ,  $\beta_3 = 0.9$ ,

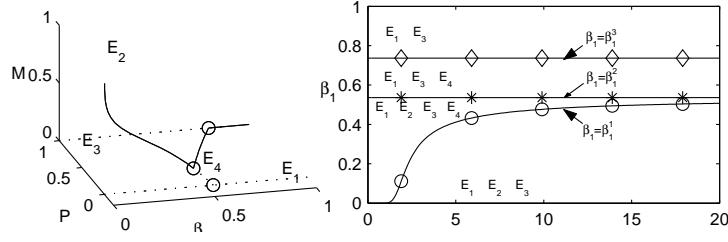


Fig. 1. Left – The bifurcation diagram with  $\beta_1$  as control parameter. Right – The static bifurcation curves  $\beta_1 = \beta_1^{(i)}$ .

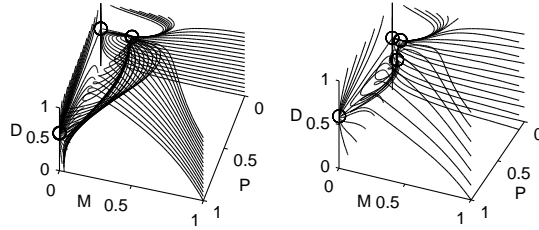


Fig. 2. Phase portrait for  $\beta_1 < \beta_1^{(1)}$  (left) and  $\beta_1 \in (\beta_1^{(1)}, \beta_1^{(2)})$  (right).

$\gamma_1 = 0.05$ ,  $\gamma_2 = 0.03$ ,  $\gamma_3 = 0.1$ ,  $\theta = 0.5$  and we found the bifurcation parameter values  $\beta_1^{(1)} = 0.439$ ,  $\beta_1^{(2)} = 0.5364$ ,  $\beta_1^{(3)} = 0.7364$ .

The points are of transcritical bifurcation type. Indeed, the condition  $\beta_1 = \beta_1^{(1)} \Leftrightarrow g_1(0) = 0$ . Note that  $g_1$  is increasing, so  $P^* = 0 \Rightarrow M^* = \frac{\alpha_1}{\alpha_2}$ ,  $D^* = \frac{\gamma_2}{\gamma_1(1+\exp(\frac{\alpha_1-\theta\alpha_2}{\alpha_2\gamma_3}))}$ . Also  $\beta_1 = \beta_1^{(1)} \Leftrightarrow f_1(\frac{\alpha_1}{\alpha_2}) = 0$ ,  $f_1$  is also increasing so  $M_2 = \frac{\alpha_1}{\alpha_2}$  is the single root for  $f_1(M) = 0$ .

Hence  $D_2 = D^*$  and we proved that the branches of stationary solutions  $E_2$  and  $E_4$  intersect at  $\beta_1 = \beta_1^{(1)}$ . Also we proved in Lemma 2 that one eigenvalue at  $E_2$  is zero. For  $\beta_1 < \beta_1^{(1)}$  the equilibrium  $E_4$  is unphysical.

The condition  $\beta_1 = \beta_1^{(2)} \Leftrightarrow f_1(0) = 0 \Leftrightarrow M_2 = 0$ . So,  $D_2 = D_1$  and the branches of equilibria  $E_1$  and  $E_2$  intersect at  $\beta_1 = \beta_1^{(2)}$ . From Lemma 5 one eigenvalue corresponding to  $E_2$  is zero. For  $\beta_1 > \beta_1^{(2)}$  the equilibrium  $E_2$  is unphysical.

Finally,  $\beta_1 = \beta_1^{(3)} \Leftrightarrow g_1(1) = 0 \Leftrightarrow P^* = 1$ . So  $M^* = 1$  and  $D^* = D_3 = \frac{\gamma_2}{\gamma_1(1+\exp(-\frac{\theta}{\gamma_3}))}$ , in consequence  $E_3$  and  $E_4$  meet at  $\beta_1 = \beta_1^{(3)}$ . The equilibrium  $E_4$  is unphysical for  $\beta_1 > \beta_1^{(3)}$ . From Lemma 6, we noticed that one eigenvalue corresponding to  $E_3$  is zero at  $\beta_1 = \beta_1^{(3)}$ .

Next, we performed numerical integration of the system (1) for  $\beta_1$  corresponding to the four cases presented in proposition 2, so that we obtained phase portraits

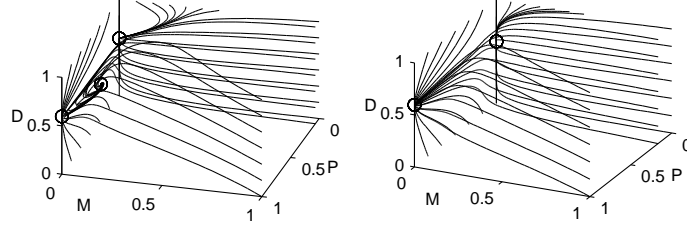


Fig. 3. Phase portrait for  $\beta_1^{(2)} < \beta_1 < \beta_1^{(3)}$  (left) and  $\beta_1 > \beta_1^{(3)}$  (right).

topologically unequivalent (see fig. 2, fig. 3). Our analytical results are verified by the numerical simulations. The phase portraits indicate that the asymptotically stable equilibrium is globally attractive in  $Int(\mathcal{D})$ .

Then, taking as control parameters  $\frac{\alpha_2}{\alpha_1}$  and  $\beta_1$ , we delimited four regions corresponding to different possible dynamics of the system, where the stability or instability of the equilibria are preserved (see Fig. 1 – right).

The biological significance for these four types of dynamics of the system is:

- i) if the mortality rate of the macrophage  $M$  is small enough, i.e.  $\beta_1 < \beta_1^{(1)}$ , then the infection will disappear, remaining an inflammation  $D_2$ ;
- ii) if  $\beta_1 \in (\beta_1^{(1)}, \beta_1^{(2)})$  or  $\beta_1 \in (\beta_1^{(2)}, \beta_1^{(3)})$ , then the infection is defeated by  $M$ , but it remains chronically causing an inflammation of the cells  $D^*$ ;
- iii) if  $\beta_1 > \beta_1^{(3)}$  (which is greater than  $\beta_2$ ), the infection is generalized and the immunity goes to zero, practically the organism dies.

### 3. Kumar's system

In the second system  $u \equiv (P, M, L)$ , where  $P$  is also the infectious pathogen,  $M$  – the early proinflammatory mediators representing a combined effect of immune cells, which attempt to destroy the pathogen.  $M$  activate later inflammatory mediators  $L$  which have effects of tissue damage and dysfunction and can also excite the early mediators. The phenomenon of recruitment of the late mediators  $L$  by  $M$  is modeled through  $h(x) = \frac{1}{1+\exp(-x)}$ .

We consider two natural parameters  $\alpha = \frac{\alpha_2}{\alpha_1}$  and  $\gamma = \frac{\gamma_2}{\gamma_1}$ , both of them will appear in the components and the conditions for the existence of the equilibria. For the rest of the paper we consider  $\theta = 1$ .

The system has the following equilibria:

- 1)  $E_1(P_1, M_1, L_1)$ , where  $P_1 = M_1 = 0$ ,  $L_1 = \gamma \left(1 + \exp\left(\frac{\theta}{\gamma_3}\right)\right)^{-1}$ ;
- 2) Two possible equilibria  $E_2^i(P_2^i, M_2^i, L_2^i)$ ,  $i = \overline{1, 2}$ , where  $P_2^i = 0$ ,



$L_2^i = \frac{\beta_1}{\beta_3(1-M_2^i)}$  and  $M_2^1 \leq M_2^2$  verify the equation

$$f_2(M) \equiv \beta_1 \left( 1 + \exp \left( \frac{\theta - M}{\gamma_3} \right) \right) - \gamma \beta_3 (1 - M) = 0; \quad (5)$$

3)  $E_3(P_3, M_3, L_3)$ , with  $P_3 = 1, M_3 = 0, L_3 = L_1$ ;

4) An interior equilibrium  $E_4(P^*, M^*, L^*)$ , where  $P^* = 1 - \alpha M^*$ ,

$L^* = \gamma \left( 1 + \exp \left( \frac{\theta - M^*}{\gamma_3} \right) \right)^{-1}$  and  $M^*$  verifies the equation

$$g_2(M) \equiv \gamma \beta_3 \left( 1 + \exp \left( \frac{\theta - M}{\gamma_3} \right) \right)^{-1} + \beta_2 (1 - \alpha M) - \frac{\beta_1}{1 - M} = 0. \quad (6)$$

Later, we shall formulate conditions for the existence and uniqueness of the interior equilibrium  $E_4$  which is the case with physiological relevance.

**Lemma 8.** *The equilibrium points  $E_1$  and  $E_3$  exist in the domain  $\mathcal{D}$  for every parameter combinations.*

**Lemma 9.** *Let  $\overline{M} = \theta - \gamma_3 \ln \frac{\gamma \beta_3 \gamma_3}{\beta_1}$ . a) If  $\overline{M} \in [0, 1]$ , the equation (5) has the following number of solutions in the interval  $[0, 1]$ : i) one solution if and only if  $f_2(0) \leq 0$  or  $f_2(\overline{M}) = 0$ ; ii) two solutions if and only if  $f_2(0) \geq 0$  and  $f_2(\overline{M}) < 0$ ; iii) no solutions if  $f_2(\overline{M}) > 0$ . b) If  $\overline{M} \notin [0, 1]$ , the equation  $f_2(M) = 0$  has: i) one solution if  $f_2(0) \leq 0$ ; ii) no solutions if  $f_2(0) > 0$ .*

**Lemma 10.** a)  $f_2(0) > 0 \Leftrightarrow \gamma < \frac{\beta_1}{\beta_3} \left( 1 + \exp \left( \frac{1}{\gamma_3} \right) \right) := \gamma^{(3)}$ . b)  $f_2(\overline{M}) > 0 \Leftrightarrow \gamma < \frac{\beta_1}{\beta_3 \gamma_3 a_0} := \gamma^{(1)}$ , where  $a_0 \approx 0.2784$  is the solution for  $a + \ln a + 1 = 0$ .

*Proof.* a) is obvious. b)  $f_2(\overline{M}) > 0 \Leftrightarrow \beta_1 \left( 1 + \frac{\gamma \beta_3 \gamma_3}{\beta_1} \right) - \gamma \beta_3 \gamma_3 \ln \frac{\gamma \beta_3 \gamma_3}{\beta_1} > 0 \Leftrightarrow 1 + \frac{\beta_1}{\gamma \beta_3 \gamma_3} + \ln \frac{\beta_1}{\gamma \beta_3 \gamma_3} > 0$ . The last inequality holds iff  $\frac{\beta_1}{\gamma \beta_3 \gamma_3} > a_0$ . ■

From the last two lemmas, we deduce the following two propositions.

**Proposition 3.** *For  $\gamma_3 < \frac{1}{1+a_0}$ , there are two equilibrium points  $E_2^1$  and  $E_2^2$  in the plane  $P = 0$  if and only if  $\gamma^{(1)} < \gamma < \gamma^{(3)}$ .*

*Proof.* We observe that  $\overline{M} \in [0, 1] \Leftrightarrow \frac{\beta_1}{\beta_3 \gamma_3} \leq \gamma \leq \frac{\beta_1}{\beta_3 \gamma_3} \exp \left( \frac{1}{\gamma_3} \right)$ , next  $f_2(0) \geq 0 \Leftrightarrow \gamma \leq \frac{\beta_1}{\beta_3} \left( 1 + \exp \left( \frac{1}{\gamma_3} \right) \right)$  and  $f_2(\overline{M}) < 0 \Leftrightarrow \gamma > \frac{\beta_1}{\beta_3 \gamma_3 a_0}$ . We note that  $\gamma_3 < \frac{1}{1+a_0} \Leftrightarrow 1 + \exp \left( \frac{1}{\gamma_3} \right) < \frac{1}{\gamma_3} \exp \left( \frac{1}{\gamma_3} \right)$  and  $\frac{1}{\gamma_3 a_0} \leq 1 + \exp \left( \frac{1}{\gamma_3} \right), \forall \gamma_3 > 0$ . So,  $\frac{\beta_1}{\beta_3 \gamma_3 a_0} < \gamma < \frac{\beta_1}{\beta_3} \left( 1 + \exp \left( \frac{1}{\gamma_3} \right) \right)$  and the proposition is proved. ■

**Proposition 4.** *Assume  $\gamma_3 < \frac{1}{1+a_0}$ . The equation (5) has: a) no solutions in  $[0, 1]$  for  $\gamma < \gamma^{(1)}$ ; b) only one solution  $M_2^2 \in [0, 1]$  if  $\gamma > \gamma^{(3)}$ .*

**Lemma 11.** For  $\alpha > 1$ , the equation (6) has a unique solution  $M^* \in (0, \frac{1}{\alpha})$  if  $\gamma^{(-1)} < \gamma < \gamma^{(2)}$  and  $\gamma_3 > \frac{\alpha\beta_1}{(\alpha-1)(\beta_1+\alpha\beta_2)}$ , where  $\gamma^{(-1)} = \frac{\beta_1-\beta_2}{\beta_3} \left(1 + \exp\left(\frac{1}{\gamma_3}\right)\right)$  and  $\gamma^{(2)} = \frac{\alpha\beta_1}{(\alpha-1)\beta_3} \left(1 + \exp\left(\frac{\alpha-1}{\alpha\gamma_3}\right)\right)$ , in the case when  $\gamma^{(-1)} < \gamma^{(2)}$ .

*Proof.* Due to the condition  $P^* \in (0, 1)$ , we need  $M^* \in (0, \frac{1}{\alpha})$ . Notice that  $g_2(\frac{1}{\alpha}) < 0 \Leftrightarrow \gamma < \gamma^{(2)}$  and  $g_2(0) > 0 \Leftrightarrow \gamma > \gamma^{(-1)}$ . If  $g_2$  is decreasing on  $(0, \frac{1}{\alpha})$ , then the equation  $g_2(M) = 0$  has an unique solution on  $(0, \frac{1}{\alpha})$ . We remark that the derivative of  $g_2$  is a difference between two positive functions,  $\psi_1(M) = \frac{\gamma\beta_3}{\gamma_3} \frac{\exp\left(\frac{1-M}{\gamma_3}\right)}{\left(1 + \exp\left(\frac{1-M}{\gamma_3}\right)\right)^2}$  and  $\psi_2(M) = \beta_2\alpha + \frac{\beta_1}{(1-M)^2}$  with coefficients totally independent on each other. The condition  $\psi_1(\frac{1}{\alpha}) < \psi_2(0) \Leftrightarrow \gamma < \frac{\beta_1+\alpha\beta_2}{\beta_3} \frac{\gamma_3 \left(1 + \exp\left(\frac{\alpha-1}{\alpha\gamma_3}\right)\right)^2}{\exp\left(\frac{\alpha-1}{\alpha\gamma_3}\right)}$  is sufficient for the uniqueness of  $M^*$ . The condition is  $\gamma < \gamma^{(2)}$  for  $\gamma_3 > \frac{\alpha\beta_1}{(\alpha-1)(\beta_1+\alpha\beta_2)}$ . ■

We have discovered so far four values for the parameter  $\gamma$ , where the number of fixed points of the system changes. It can be proved the following

**Proposition 5.** For  $\alpha > 1$  we have: i)  $\gamma^{(1)} \leq \gamma^{(2)}, \gamma^{(-1)} < \gamma^{(3)}$ , for any combination of parameters; ii)  $\gamma^{(2)} < \gamma^{(3)}$  if  $\gamma_3 < \frac{\alpha-1}{\alpha(1+a_0)}$ ; iii)  $\gamma^{(-1)} \leq 0$  for  $\beta_1 \leq \beta_2$ .

*Proof.* a)  $\gamma^{(1)} \leq \gamma^{(2)} \Leftrightarrow \frac{\alpha-1}{\alpha\gamma_3 a_0} \leq 1 + \exp\left(\frac{\alpha-1}{\alpha\gamma_3}\right)$ , which holds, with equality for  $\frac{\alpha-1}{\alpha\gamma_3} = 1 + a_0$ . Then  $\gamma^{(2)} < \gamma^{(3)} \Leftrightarrow \frac{\alpha}{\alpha-1} \left(1 + \exp\left(\frac{\alpha-1}{\alpha\gamma_3}\right)\right) < 1 + \exp\left(\frac{1}{\gamma_3}\right)$ . The function  $\varphi(u) = \frac{1}{u} \left(1 + \exp\left(\frac{u}{\gamma_3}\right)\right) - 1 - \exp\left(\frac{1}{\gamma_3}\right)$  has a minimum for  $u = (1 + a_0)\gamma_3$  and  $\varphi(1) = 0$ . Hence  $\varphi(u) < 0$  for  $u > (1 + a_0)\gamma_3$  and the inequality is proved. The rest of the statements are obvious. ■

**Theorem 1.** Assume that  $\alpha > 1$ ,  $\beta_1 \leq \beta_2$  and  $\frac{\alpha\beta_1}{(\alpha-1)(\beta_1+\alpha\beta_2)} < \gamma_3 < \frac{\alpha-1}{\alpha(1+a_0)}$ . Then a) for  $\gamma < \gamma^{(1)}$ , there are only three equilibrium points  $E_1, E_3$  and  $E_4$ ; b) for  $\gamma^{(1)} < \gamma < \gamma^{(2)}$ , all the five equilibria  $E_1, E_2^1, E_2^2, E_3$  and  $E_4$  exist; c) for  $\gamma^{(2)} < \gamma < \gamma^{(3)}$ ,  $E_1, E_2^1, E_2^2$  and  $E_3$  exist; for  $\gamma > \gamma^{(2)}$  the equilibrium  $E_4$  becomes unphysical; d) for  $\gamma > \gamma^{(3)}$ ,  $E_1, E_2^2$  and  $E_3$  still exist.

### 3.1. The stability of the equilibria and the bifurcation analysis

**Proposition 6.** The equilibrium point  $E_1$  is a saddle-point, always stable in  $L$ -direction and unstable in  $P$ -direction.

*Proof.* The Jacobian matrix of the system (1) evaluated at  $E_1$  has the eigenvalues  $\lambda_1 = \alpha_1$ ,  $\lambda_2 = -\gamma_1$ ,  $\lambda_3 = \beta_3\gamma \left(1 + \exp\left(\frac{\theta}{\gamma_3}\right)\right)^{-1} - \beta_1$ . The eigenvectors for  $\lambda_1$  and  $\lambda_2$  are  $v_1 = (1, 0, 0)^T$  and, respectively,  $v_2 = (0, 0, 1)^T$ . ■

**Proposition 7.** The equilibrium point  $E_3$  is saddle for  $\gamma > \gamma^{(-1)}$  and positive attractor for  $\gamma < \gamma^{(-1)}$ . The plane  $M = 0$  is always its stable invariant manifold.

*Proof.* For  $E_3$  the eigenvalues are

$$\lambda_1 = -\alpha_1, \lambda_2 = -\gamma_1, \lambda_3 = \beta_2 - \beta_1 + \beta_3\gamma \left(1 + \exp\left(\frac{\theta}{\gamma_3}\right)\right)^{-1}.$$

$E_3$  is positive attractor only if  $\lambda_3 < 0 \Leftrightarrow \gamma < \gamma^{(-1)}$ . ■

**Lemma 12.** *For the equilibrium points  $E_2^i, i = \overline{1, 2}$ , the eigenvalues  $\lambda_1^i, \lambda_2^i, \lambda_3^i$  of the Jacobian of the system, have the following properties: i)  $\lambda_1^i = \alpha_1 - \alpha_2 M_2^i$  are positive for  $\alpha \leq 1$  and have variable sign for  $\alpha > 1$ ; ii)  $\lambda_2^1 \lambda_3^1 < 0$  and  $\text{Re} \lambda_2^2, \text{Re} \lambda_3^2 < 0$ ; iii)  $\exists i \in \{1, 2\}$  such that  $\lambda_1^i = 0 \Leftrightarrow f_2\left(\frac{1}{\alpha}\right) = 0 \Leftrightarrow \gamma = \gamma^{(2)}$ ; iv)  $\exists i \in \{1, 2\}$  such that  $\lambda_2^i = 0$  or  $\lambda_3^i = 0 \Leftrightarrow f_2(\overline{M}) = 0 \Leftrightarrow \gamma = \gamma^{(1)}$ .*

*Proof.* Suppose  $E_2^1, E_2^2$  exist.  $\lambda_1^i = \alpha_1 - \alpha_2 M_2^i < 0 \Leftrightarrow M_2^i < \frac{1}{\alpha}$  and we get i).  $\lambda_2^i + \lambda_3^i = -\gamma_1 - \frac{\beta_1 M_2^i}{1 - M_2^i} < 0$ ,  $\lambda_2^i \lambda_3^i = \frac{\gamma_1 \beta_1 M_2^i}{\gamma \gamma_3 \beta_3} \frac{\gamma \beta_3 \gamma_3 - \beta_1 \exp\left(\frac{1 - M_2^i}{\gamma_3}\right)}{1 - M_2^i} > 0 \Leftrightarrow M_2^i > \overline{M}$  given by lemma 9. Hence we deduce ii). For iii) we observe that  $\lambda_1 = 0 \Leftrightarrow \exists i = 1, 2$  such that  $M_2^i = \frac{1}{\alpha} \Leftrightarrow f_2\left(\frac{1}{\alpha}\right) = 0$ . ■

**Proposition 8.** *Assume  $\alpha > 1$  and  $\gamma_3 < \frac{\alpha - 1}{\alpha(a_0 + 1)}$ . If the equilibria  $E_2^1$  and/or  $E_2^2$  exist, then  $E_2^1$  is saddle and  $E_2^2$  is positive attractor.*

*Proof.* From the Lemma 12, we deduce that  $E_2^1$  is a saddle point for  $\alpha > 1$  and both equilibria are saddles for  $\alpha \leq 1$ . We study now only the stability of  $E_2^2$  for  $\alpha > 1$ , so it follows  $\lambda_1^2 < 0 \Leftrightarrow M_2^2 > \frac{1}{\alpha}$ . We notice that the condition  $M_2^1 < \frac{1}{\alpha} < M_2^2 \Leftrightarrow f_2\left(\frac{1}{\alpha}\right) < 0 \Leftrightarrow \gamma > \gamma^{(2)}$ . Then  $M_2^2 > \frac{1}{\alpha} \Leftrightarrow \left\{\frac{1}{\alpha} \in (M_2^1, M_2^2) \text{ or } \frac{1}{\alpha} < \overline{M}\right\} \Leftrightarrow \gamma > \gamma^{(2)}$  or  $\gamma < \frac{\beta_1}{\beta_3 \gamma_3} \exp\left(\frac{\alpha - 1}{\alpha \gamma_3}\right)$ . Observe that  $\gamma^{(2)} < \frac{\beta_1}{\beta_3 \gamma_3} \exp\left(\frac{\alpha - 1}{\alpha \gamma_3}\right) \Leftrightarrow \gamma_3 < \frac{\alpha - 1}{\alpha(a_0 + 1)}$ . In conclusion, if  $E_2^2$  exists (see Proposition 3 and 4) and  $\gamma_3 < \frac{\alpha - 1}{\alpha(a_0 + 1)}$ , then it is a positive attractor. ■

For the interior equilibrium  $E_4$ , the eigenvalues of the Jacobian matrix of the system verify the characteristic equation  $\lambda^3 + A_1 \lambda^2 + A_2 \lambda + A_3 = 0$ , where

$$A_1 = \alpha_1 - \alpha_2 M^* + \frac{\beta_1 M^*}{1 - M^*} + \gamma_1 > 0,$$

$$A_2 = -\gamma_1 M^* (1 - M^*) (g_2'(M^*) + \alpha \beta_2) + (\alpha_1 - \alpha_2 M^*) \cdot \left( \gamma_1 + \frac{\beta_1 M^*}{1 - M^*} + \alpha \beta_2 M^* (1 - M^*) \right),$$

$$A_3 = -\gamma_1 M^* (\alpha_1 - \alpha_2 M^*) \cdot (1 - M^*) g_2'(M^*) > 0$$

in the conditions of Lemma 11 for the existence of  $E_4$ . Using Routh-Hurwitz criterion,  $E_4$  is asymptotic stable if and only if  $A_1 A_2 > A_3$ .

Taking  $\gamma$  as a control parameter of the system, we found three points of static bifurcation, as follows:

**Proposition 9.** *If a)  $\gamma_3 < \frac{\alpha-1}{\alpha(a_0+1)}$ ,  $\alpha > 1$ , or b)  $\gamma_3 < \frac{1}{a_0+1}$ ,  $\alpha \leq 1$ , then  $E_2^1$  and  $E_2^2$  appear at  $\gamma = \gamma^{(1)}$  through a saddle-node bifurcation.*

*Proof.* For  $\gamma = \gamma^{(1)}$  we have  $f_2(\overline{M}) = 0 \Leftrightarrow M_2^1 = M_2^2 = \overline{M} \Leftrightarrow M_2^i = 1 + \gamma_3 \ln a_0 = 1 - \gamma_3(a_0 + 1)$ , which belongs to  $(0, 1)$  in conditions a) and b). Then  $P_2^1 = P_2^2 = 0$ ,  $L_2^1 = L_2^2 = \frac{\beta_1}{\beta_3 \gamma_3(a_0+1)}$ , hence  $E_2^1 = E_2^2$  for  $\gamma = \gamma^{(1)}$ .

The eigenvalues for the linearized system  $u' = D_u F(E_2^1, \gamma^{(1)})u$  are

$$\lambda_1 = \alpha_1(1 - \alpha + \alpha\gamma_3(a_0 + 1)) = \begin{cases} < 0, & \text{in case a)} \\ > 0, & \text{in case b)} \end{cases}, \lambda_2 = 0,$$

$$\lambda_3 = -\gamma_1 + \beta_1 \left(1 - \frac{1}{\gamma_3(a_0 + 1)}\right) < 0.$$

So,

(SN1):  $D_u F(E_2^1, \gamma^{(1)})$  has  $k$  eigenvalues with negative real parts and a single eigenvalue 0, with right eigenvector

$$v = \left(0, \gamma_3^2(a_0 + 1)^2, \frac{\beta_1}{\beta_3}\right)^T$$

and left eigenvector

$$w = \left(\frac{-\beta_2}{\alpha_1 - \alpha_2(1 - \gamma_3(a_0 + 1))}, \frac{1}{\gamma_3(a_0 + 1)(1 - \gamma_3(a_0 + 1))}, \frac{\beta_2}{\beta_1}\right);$$

$$(SN2): \quad w \cdot \frac{\partial F}{\partial \gamma}(E_2^1, \gamma^{(1)}) = \frac{\beta_3 a_0}{a_0 + 1} > 0;$$

$$(SN3): \quad w \cdot (D_{xx} F(E_2^1, \gamma^{(1)})(v, v)) = -w_3 \frac{\gamma_1 \beta_1 (a_0 - 1)}{\gamma_3^2 \beta_3 (a_0 + 1)^3} v_2^2 < 0.$$

Accordingly to a theorem due to Sotomayor [4], the conditions (SN1–SN3) are sufficient for a static bifurcation point to be of saddle node type. The stability of the bifurcate branches was analysed in proposition 9. ■

**Proposition 10.** i) *For  $\alpha > 1$  under the hypothesis of Theorem 1, at  $\gamma = \gamma^{(2)}$ , the equilibria  $E_2^1$  and  $E_4$  meet through a transcritical bifurcation.*

ii) *At  $\gamma = \gamma^{(3)}$ , the equilibrium points  $E_2^1$  and  $E_1$  coincide through a transcritical, assuming the hypotheses of Proposition 9.*

*Proof.* i). For  $\gamma = \gamma^{(2)}$ ,  $f_2(\frac{1}{\alpha}) = 0 \Rightarrow M_2^1 = \frac{1}{\alpha}$ . Note that  $M_2^2 > \frac{1}{\alpha}$  if  $\gamma_3 < \frac{\alpha-1}{\alpha(a_0+1)}$  (see the proof of Proposition 9). Also  $\gamma = \gamma^{(2)} \Rightarrow g_2(\frac{1}{\alpha}) = 0 \Rightarrow M^* = \frac{1}{\alpha}$ , since  $g_2(M) = 0$  has a single solution. We have  $L_2^1 = L^* = \frac{\beta_1 \alpha}{\beta_3(\alpha-1)}$  and  $P_2^1 = P^*$ . So,  $E_2^1$  coincides with  $E_4$  at  $\gamma = \gamma^{(2)}$ , when one eigenvalue corresponding to these equilibria is zero (see Lemma 12.) ii). For  $\gamma = \gamma^{(3)}$ ,  $f_2(0) = 0 \Rightarrow M_2^1 = 0 \Rightarrow L_2^1 = L_1$ , hence  $E_2^1 \equiv E_1$ . From proposition 6, we deduce that only one eigenvalue corresponding to  $E_1$  is zero for  $\gamma = \gamma^{(3)}$ .

The system has to be reduced to the central manifold in the neighborhood of each bifurcation point and we obtain transcritical bifurcations. ■

The equilibrium point  $E_4$  is the only one that can experience Hopf bifurcation. We shall investigate this numerically.

### 3.2. Bifurcation diagram and numerical results

We use  $\gamma$  as a control parameter and the fixed parameters respect the conditions of Theorem 1. Then we plot the bifurcation diagram for the system (1) and make its projection into the plane  $(\gamma, M)$ , where  $\gamma \in (0, \infty)$  and  $M \in [0, 1]$ .

The bifurcation diagram contains the branches of fixed points and also the detection of the Hopf bifurcation point for  $E_4$  (see Fig. 4–left).

The necessary condition for equilibrium point  $E_4$  to have a Hopf bifurcation is:  $\exists \lambda_{1,2} = \pm i\omega, \omega > 0$ , solution of the characteristic equation, which is equivalent to  $A_2 > 0$  and  $A_1 A_2 = A_3$ . So we find a Hopf bifurcation point for  $\gamma = \gamma_H$  at the intersection of the curve  $A_1 A_2 = A_3$  (where  $A_i$  are functions of  $\gamma$  and  $M$ ) with the branch  $E_4$ , when it exists, i.e. for  $0 < \gamma < \gamma^{(2)}$ . Moreover the condition  $A_1 A_2 > A_3$  is fulfilled for  $\gamma < \gamma_H$ , so that  $E_4$  is a stable equilibrium for  $\gamma < \gamma_H$ .

In the diagram, solid and broken lines correspond to the stable and unstable (or unphysical) equilibria, respectively. For the numerical simulations we took as fixed parameters  $\alpha_1 = 3$ ,  $\alpha = 10$ ,  $\beta_1 = 0.8$ ,  $\beta_2 = 5$ ,  $\beta_3 = \gamma_1 = 1$ ,  $\gamma_3 = 0.25$ . We obtained  $\gamma_H = 29.8707$ . The Hopf bifurcation is subcritical, since we found an unstable limit-cycle emerging from  $E_4$  for  $\gamma < \gamma_H$  (see Fig. 4–right and Fig. 5–left). Trajectories within the limit-cycle spiral into  $E_4$  which is a focus. In consequence, there exists an interval  $(\gamma^{(1)}, \gamma_H)$  where two stable branches of solutions coexist ( $E_2^2$  and  $E_4$ ). This is an interval of bistability (see Fig. 4–left). For  $\gamma < \gamma^{(1)}$ , we found that  $E_4$  is a global attractor (Fig. 5–right).

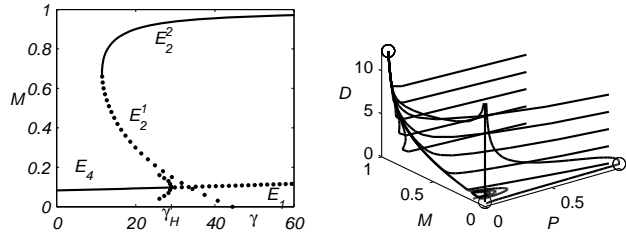


Fig. 4. The bifurcation diagram for  $\alpha > 1$  and  $\gamma$  as a control parameter (left). The phase portrait for  $\gamma = 29 \in (\gamma^{(1)}, \gamma_H)$ . Trajectories which tend to  $E_2^2$  where  $P \rightarrow 0$ , but  $M$  and  $L$  remain elevated, are interpreted as persistent non-infectious inflammation (right).

In conclusion, we have examined the behavior of the system (1) by varying one parameter  $\gamma$  and we obtained the picture of the global dynamics of the system which captures the important clinically scenarios presented in [2], for  $\beta_1 \leq \beta_2$  and  $\alpha > 1$ .

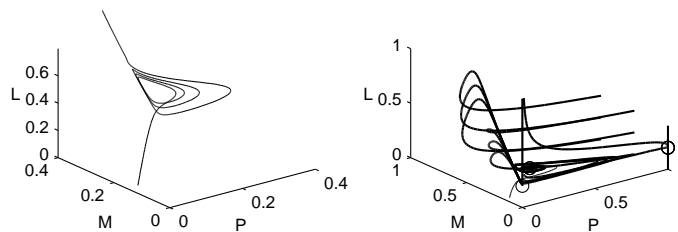


Fig. 5. One trajectory which spiral outside the unstable limit-cycle, zoom on Fig. 4–right. During the oscillations,  $P$  falls below a threshold. The trajectory is interpreted as healthy response (left). The phase portrait for  $\gamma = 11 < \gamma^{(1)}$ , showing that  $E_4$  is a focus. Trajectories which tend to  $E_4$ , are interpreted as recurrent infection (right).

## References

- [1] R. Brause, *Adaptive modeling of biochemical pathways*, Intl. J. Artificial Intelligence Tools, **13** (2003), 4, 851–862.
- [2] R. Kumar, G. Clermont, Y. Vodovotz, C.C. Chow, *The dynamics of acute inflammation*, J. Theor. Biology, **230** (2004), 3, 145–155.
- [3] Kuznetsov, Y., *Elements of Applied Bifurcation Theory*. Springer-Verlag, 1995.
- [4] Sotomayor, J., *Generic bifurcations of dynamical systems*, in M.M. Peixoto (ed.), *Dynamical Systems*, Acad. Press, 1973.

## Self-Propulsion of an Oscillatory Wing Including Ground Effects

Adrian Carabineanu<sup>\*†</sup>

In the framework of the small perturbations theory, we study the motion of an uniform stream past an oscillating thin wing including ground effects. Using the theory of distributions we deduce the integral equation for the jump of the pressure past the wing. We solve the integral equation numerically and we calculate the average drag coefficient. We find that for some kind of wings there appears a propulsive force and this force increases when the wing is close to the ground.

### 1. Introduction

The problem of the oscillatory wings in subsonic flow (and implicitly in incompressible flow) was studied, among others by Watkins, Runyan and Woolston [14], Laschka [12], and Landahl [11]. In their theory the integral equation, originally performed by Küssner [10], is obtained by determining the acceleration potential due to a continuous distribution of doublets on the wing.

Latter, L. Dragoș [4] studied the problem of oscillating thick wings by means of the fundamental solutions method.

D. Homentcovschi [8] utilized the Fourier transform for obtaining the fundamental solutions of the linear Euler system and then the general integral equation relating the jump of the pressure and the downwash distributions for the unsteady flow past a lifting surface, moving over an abstract fixed cylindrical surface. From this equation it was deduced the integral equation for the oscillating wing.

In our paper we employ the theory of distributions in order to deduce the integral

---

<sup>\*</sup> University of Bucharest, Faculty of Mathematics and Informatics and Institute of Statistics and Applied Mathematics of Romanian Academy, Bucharest, Romania, e-mail: [acara@fmi.unibuc.ro](mailto:acara@fmi.unibuc.ro)

<sup>†</sup> Supported by CNCSIS Grant 33379/217, 2005.

equation of the oscillating thin wing. For taking into account the ground effect we utilize the image method, i.e. we consider that another oscillatory wing, symmetrical with the original one with respect to the plane  $z = -d$  is immersed into the fluid.

Many numerical methods were developed for solving the integral equation of steady or oscillatory lifting surface equation. Among them we mention the methods considered by A. Ichikawa [9], Ueda and Dowell [13], Eversman and Pitt [6], etc.

The great number of papers devoted to the numerical methods used for integrating the lifting surface integral equation is justified by the difficulties caused by the singularities of the kernel.

In our paper, in order to discretize the integral equation, we split the kernel of the equation into several kernels for which we provide appropriate quadrature formulas depending on the type of singularity of the kernel. By solving the discretized integral equation we calculate the jump of the pressure over the delta wing.

After obtaining the pressure field we calculate, by performing a numerical integration the average drag. We study an example of oscillatory motion of the delta wing and we notice that if the frequency surpasses a critical value, the drag becomes negative, i.e. it appears a propulsive force. We also notice that the propulsive force increases when the wing is situated closer to the ground.

## 2. The integral equation of the problem

We consider the continuity and Euler equations for incompressible flow in a fixed Cartesian frame of reference  $Oxyz$ ,

$$\begin{aligned} \operatorname{div} \mathbf{v} &= 0, \mathbf{v} = (u, v, w), \\ \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \operatorname{grad}) \mathbf{v} + \frac{1}{\rho_0} \operatorname{grad} p &= 0, \end{aligned} \quad (1)$$

The wing and its symmetric with respect to  $z = -d$  plane have the equations:

$$\begin{aligned} S : F(x, y, z, t) &= z - h(x, y, t), (x, y) \in D(t), \\ S' : F'(x, y, z, t) &= z + h(x, y, t) - 2d, (x, y) \in D(t). \end{aligned}$$

We assume that the wing is thin, i.e. there is a small parameter  $0 < \varepsilon \ll 1$  such that

$$|h| < \varepsilon, \left| \frac{\partial h}{\partial x} \right| < \varepsilon, \left| \frac{\partial h}{\partial y} \right| < \varepsilon.$$

The coordinates of the normal at the wing surface  $S$  are:

$$\mathbf{n} = (n_x, n_y, n_z, n_t) = \left( -\frac{\partial h}{\partial x}, -\frac{\partial h}{\partial y}, 1, -\frac{\partial h}{\partial t} \right). \quad (2)$$

We linearize the equations around the rest state (neglecting the products of the perturbation quantities) and we write them into distributions:

$$\frac{\partial u}{\partial t} + \frac{1}{\rho_0} \frac{\partial p}{\partial x} = \left( [u]_S n_t + \frac{1}{\rho_0} [p]_S n_x \right) \delta_{S \cup S'} = 0, \quad (3)$$



$$\frac{\partial v}{\partial t} + \frac{1}{\rho_0} \frac{\partial p}{\partial y} = \left( [v]_S n_t + \frac{1}{\rho_0} [p]_S n_y \right) \delta_{S \cup S'} = 0, \quad (4)$$

$$\frac{\partial w}{\partial t} + \frac{1}{\rho_0} \frac{\partial p}{\partial z} = \left( [w]_S n_t + \frac{1}{\rho_0} [p]_S n_z \right) \delta_{S \cup S'} = f \delta_D - f \delta_{D'}, \quad f = \frac{[p]_S}{\rho_0}, \quad (5)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = ([u]_S n_x + [v]_S n_y + [w]_S n_z) \delta_{S \cup S'} = 0, \quad (6)$$

where  $\mu \delta_S$  represents the simple layer distribution with density  $\mu$  and  $[\cdot]_S$  is the jump over the surface  $S$ .

From (3)–(6) we get

$$\left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) p = \frac{\partial}{\partial z} (\rho_0 f \delta_D - \rho_0 f \delta_{D'}). \quad (7)$$

Since

$$\left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \left( -\frac{\delta(t)}{4\pi |\mathbf{x}|} \right) = \delta(x, y, z) \delta(t), \quad |\mathbf{x}| = \sqrt{x^2 + y^2 + z^2}, \quad (8)$$

we get

$$p = -\rho_0 \frac{\delta(t)}{4\pi} \frac{\partial}{\partial z} \frac{1}{|\mathbf{x}|} * (f \delta_D - f \delta_{D'}), \quad (9)$$

whence, taking into account (3), we deduce

$$\frac{\partial w}{\partial t} = \frac{\delta(t)}{4\pi} \frac{\partial}{\partial z^2} \frac{1}{|\mathbf{x}|} * (f \delta_D - f \delta_{D'}) + f \delta_D. \quad (10)$$

From (9)–(10) we have, denoting by  $H(t)$  Heaviside's function,

$$\begin{aligned} w(x, y, z, t) &= -\frac{H(t)}{4\pi} \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \frac{1}{|\mathbf{x}|} * (f \delta_D - f \delta_{D'}) = \\ &= -\frac{1}{4\pi} \int_{-\infty}^{\infty} H(t - t') dt' \int \int_{D(t') \cup D'(t')} \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \frac{f(\mathbf{x}', t)}{|\mathbf{x} - \mathbf{x}'|} dx' dy' = \\ &\stackrel{z \neq 0}{=} \frac{1}{4\pi} \int_{-\infty}^t \int \int_{D(t')} \frac{\partial^2}{\partial z^2} \left[ \frac{f(\mathbf{x}', t)}{|\mathbf{x} - \mathbf{x}'|} - \frac{f(\mathbf{x}', t)}{|\mathbf{x} - \mathbf{x}' + 2d\mathbf{k}|} \right] dx' dy' dt'. \end{aligned} \quad (11)$$

In the sequel we shall introduce a new system of coordinates  $O^{(1)}x^{(1)}y^{(1)}z^{(1)}$ , related to the lifting surface. We have the relations

$$\begin{aligned} x^{(1)} &= x + V_0 t, \quad y^{(1)} = 1, \quad z^{(1)} = z, \\ x'^{(1)} &= x' + V_0 t', \quad u^{(1)} = x^{(1)} - x'^{(1)} - V_0(t - t'). \end{aligned} \quad (12)$$

In the new coordinates, the integral representation (11) becomes, for  $z^{(1)} \neq 0$ ,

$$w(\mathbf{x}^{(1)}, t) = \frac{1}{4\pi V_0} \int_{-\infty}^{x^{(1)}-x'^{(1)}} du^{(1)} \int_{D^{(1)}} \frac{\partial^2}{\partial z^{(1)2}} \left( \frac{f(\mathbf{x}'^{(1)}, t)}{\sqrt{u^{(1)2} + (y^{(1)} - y'^{(1)})^2 + z^{(1)2}}} - \frac{f(\mathbf{x}'^{(1)}, t)}{\sqrt{u^{(1)2} + (y^{(1)} - y'^{(1)})^2 + (z^{(1)} + 2d)^2}} \right) dx'^{(1)} dy'^{(1)}, \quad (13)$$

where  $D^{(1)}$ , the projection of the lifting surface onto the  $O^{(1)}x^{(1)}y^{(1)}$ -plane, is a fixed surface.

Considering the lifting surface subjected to harmonic oscillations, we set

$$w(\mathbf{x}^{(1)}, t) = d^{(1)}(\mathbf{x}^{(1)}) \exp(i\omega t), \quad f(x^{(1)}, y^{(1)}, t) = f(x^{(1)}, y^{(1)}) \exp(i\omega t), \quad (14)$$

whence it follows

$$\int \int_{D^{(1)}} dx'^{(1)} dy'^{(1)} \int_{-\infty}^{x^{(1)}-x'^{(1)}} f(x'^{(1)}, y'^{(1)}) \exp\left(-i\frac{\omega}{V_0}(x^{(1)} - x'^{(1)} - u^{(1)})\right) \cdot \frac{\partial^2}{\partial z^{(1)2}} \left[ \frac{1}{\sqrt{u^{(1)2} + (y^{(1)} - y'^{(1)})^2 + z^{(1)2}}} - \frac{1}{\sqrt{u^{(1)2} + (y^{(1)} - y'^{(1)})^2 + (z^{(1)} + 2d)^2}} \right] du^{(1)} = 4\pi V_0 d^{(1)}(\mathbf{x}^{(1)}). \quad (15)$$

Denoting by  $a$  the length of the wing and by  $2b$  the chord, we introduce the dimensionless coordinates

$$(x, y, z, u, \xi, \eta) = \left( \frac{x^{(1)}}{a}, \frac{y^{(1)}}{b}, \frac{z^{(1)}}{a}, \frac{u^{(1)}}{a}, \frac{x'^{(1)}}{a}, \frac{y'^{(1)}}{b} \right). \quad (16)$$

For the sake of simplicity we use again the notation  $(x, y, z)$  which must not be confounded with the notations for the variables of the fixed system  $Oxyz$ . Denoting

$$D = \left\{ (x, y) ; (ax, by) \in D^{(1)} \right\}$$

and passing to limit for  $z \rightarrow 0$  we get from (15):

$$\frac{ab}{4\pi V_0} \int \int_D^* f(a\xi, b\eta) \exp\left(-i\frac{\omega}{V_0}a(x - \xi)\right) \left[ \int_{-\infty}^{a(x-\xi)} \exp(i\frac{\omega}{V_0}u) \dots \left( \frac{1}{(u^2 + b^2(y - \eta)^2)^{3/2}} + \frac{8d^2 - u^2 - b^2(y - \eta)^2}{(u^2 + b^2(y - \eta)^2 + 4d^2)^{5/2}} \right) \right] du d\xi d\eta = -d(x, y). \quad (17)$$

In the framework of the linearized theory,

$$d(x, y) \exp(i\omega t) = w(x^{(1)}, y^{(1)}, t), \quad (18)$$

where  $w$  is the projection of the velocity on the  $Oz^{(1)}$ -axis.

The velocity field is

$$\mathbf{V} = V_0 \mathbf{i} + \mathbf{v}, \quad \mathbf{v} = u \mathbf{i} + v \mathbf{j} + w \mathbf{k},$$

where  $\mathbf{v}$  is the perturbation velocity of the fluid.

For calculating the downwash distribution, we employ the slipping condition

$$\mathbf{V} \cdot \mathbf{n} |_{D^{(1)}} = - \frac{\frac{\partial F}{\partial t}}{|\text{grad} F|} \quad (19)$$

with

$$\mathbf{n} = \frac{\text{grad} F}{|\text{grad} F|} = - \frac{\partial h^{(1)}}{\partial x^{(1)}} \exp(i\omega t) \mathbf{i} - \frac{\partial h^{(1)}}{\partial y^{(1)}} \exp(i\omega t) \mathbf{j} + \mathbf{k}. \quad (20)$$

Since

$$\frac{\partial F}{\partial t} = -i\omega h^{(1)}(x^{(1)}, y^{(1)}) \exp(i\omega t), \quad (21)$$

from (19)–(20) we obtain the linearized condition

$$w = \left( V_0 \frac{\partial h^{(1)}}{\partial x^{(1)}} + i\omega h^{(1)} \right) \exp(i\omega t). \quad (22)$$

Denoting

$$h(x, y) = \frac{h^{(1)}(x^{(1)}, y^{(1)})}{a}, \quad \tilde{\omega} = \frac{\omega a}{V_0},$$

from (18) and (22) it follows

$$d(x, y) = V_0 \left( \frac{\partial h(x, y)}{\partial x} + i\tilde{\omega} h(x, y) \right). \quad (23)$$

Introducing the dimensionless functions and variables

$$\tilde{d} = \frac{d}{V_0}, \quad \tilde{f}(x, y) = \frac{f(ax, by)}{V_0^2}, \quad x_0 = x - \xi, \quad y_0 = y - \eta,$$

eq. (17) becomes

$$\begin{aligned} & \frac{\varpi}{4\pi} \int \int_D^* \tilde{f}(\xi, \eta) \exp(-i\tilde{\omega} x_0) \cdot \\ & \left[ \int_{-\infty}^{x_0} \exp(i\tilde{\omega} u) \left( \frac{1}{(u^2 + \varpi^2 y_0^2)^{3/2}} + \frac{8d^2 - u^2 - \varpi^2 y_0^2}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} \right) du \right] d\xi d\eta = \\ & = \frac{\partial h(x, y)}{\partial x} + i\tilde{\omega} h(x, y). \end{aligned} \quad (24)$$

where  $(x, y) \in D$  if and only if  $(x^{(1)}, y^{(1)}) \in D^{(1)}$ .

The star indicates the finite part in the Hadamard sense of the integral,  $\varpi$  is the *aspect ratio* and  $\tilde{\omega}$  is the *reduced frequency*.

### 3. The discretization of the integral equation for the symmetrical oscillating delta flat plate

We consider the oscillating delta wing. The equations of the leading edge of  $D^{(1)}$  are

$$y_{\pm}^{(1)}(x^{(1)}) = \pm \frac{b}{a}(x^{(1)}); \quad x^{(1)} \in [0, a] \quad (25)$$

and the equations of the leading edge of  $D$  are

$$y_{\pm}(x) = \pm x; \quad x \in [0, 1]. \quad (26)$$

For solving numerically the integral equation, we have to discretize the left hand member in order to obtain an algebraic system of equations. To this aim we split, like in [2], the kernel

$$K(x, y; \xi, \eta) = \int_{-\infty}^{x_0} \exp(i\tilde{\omega}u) \left( \frac{1}{(u^2 + \varpi^2 y_0^2)^{3/2}} + \frac{8d^2 - u^2 - \varpi^2 y_0^2}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} \right) du$$

into several kernels in order to put into evidence the kind of singularities we are dealing with and to find afterwards the most convenient quadrature formulas.

We have step by step:

$$\begin{aligned} \int_{-\infty}^{x_0} \frac{\exp(i\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2)^{3/2}} du &= \int_{-\infty}^{x_0} \frac{\exp(i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du + \\ &+ \frac{1}{\varpi^2 y_0^2} \left( 1 + \frac{x_0}{|x_0|} \right) + \frac{1}{\varpi^2 y_0^2} \left( \frac{x_0}{\sqrt{x_0^2 + \varpi^2 y_0^2}} - \frac{x_0}{|x_0|} \right), \end{aligned} \quad (27)$$

$$\int_{-\infty}^{x_0} \frac{\exp(i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du = \int_0^{x_0} \frac{\exp(i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du + \int_0^{\infty} \frac{\exp(-i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du, \quad (28)$$

$$\begin{aligned} &\int_0^{\infty} \frac{\exp(-i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du = \\ &= -\frac{1}{\varpi^2 y_0^2} + \int_0^{\infty} \frac{\cos \tilde{\omega}u}{(u^2 + \varpi^2 y_0^2)^{3/2}} du - i \int_0^{\infty} \frac{\sin \tilde{\omega}u}{(u^2 + \varpi^2 y_0^2)^{3/2}} du. \end{aligned} \quad (29)$$

The integrals from the right hand part of (29) represent the *sine* and *cosine Fourier transforms* of  $(u^2 + \varpi^2 y_0^2)^{-3/2}$  and in [3] one shows that

$$\int_0^{\infty} \frac{\cos \tilde{\omega}u}{(u^2 + \varpi^2 y_0^2)^{3/2}} du = \frac{\tilde{\omega}}{\varpi |y_0|} K_1(\tilde{\omega} \varpi |y_0|), \quad (30)$$

$$\int_0^{\infty} \frac{\sin \tilde{\omega}u}{(u^2 + \varpi^2 y_0^2)^{3/2}} du = \frac{\pi}{2} \frac{\tilde{\omega}}{\varpi |y_0|} (L_{-1}(\tilde{\omega} \varpi |y_0|) - I_1(\tilde{\omega} \varpi |y_0|)), \quad (31)$$

where  $L_{-1}$  is a Strouve function and  $I_1, K_1$  are Bessel functions and their series expansions are

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}, \quad (32)$$

$$K_1(x) = I_1(x) \ln \frac{x}{2} + \frac{1}{x} - \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!} (\psi(k+1) + \psi(k+2)), \quad (33)$$

$$L_{-1}(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k}}{\Gamma(k + \frac{3}{2})\Gamma(k + \frac{1}{2})}. \quad (34)$$

where  $\psi$  represents the logarithmic derivative of Euler's  $\Gamma$  function.

We also have

$$\begin{aligned} \int_0^{x_0} \frac{\exp(i\tilde{\omega}u) - 1}{(u^2 + \varpi^2 y_0^2)^{3/2}} du &= \int_0^{x_0} \frac{\exp(i\tilde{\omega}u) - 1 - i\tilde{\omega}u + \tilde{\omega}^2 u^2/2}{(u^2 + \varpi^2 y_0^2)^{3/2}} du - \\ &\quad - \frac{i\tilde{\omega}}{(x_0^2 + \varpi^2 y_0^2)^{1/2}} + \frac{i\tilde{\omega}}{|y_0|} + \frac{\tilde{\omega}^2 x_0}{2(x_0^2 + \varpi^2 y_0^2)^{1/2}} - \\ &\quad - \frac{\tilde{\omega}^2}{2} \ln \left( x_0 + \sqrt{(x_0^2 + \varpi^2 y_0^2)} \right), \quad (35) \\ \int_{-\infty}^{x_0} \exp(i\tilde{\omega}u) \frac{8d^2 - u^2 - \varpi^2 y_0^2}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du &= \\ = \int_0^{x_0} \exp(i\tilde{\omega}u) \frac{8d^2 - u^2 - \varpi^2 y_0^2}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du + \\ + 12d^2 \int_0^{\infty} \frac{\cos(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du - 12id^2 \int_0^{\infty} \frac{\sin(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du - \\ - \int_0^{\infty} \frac{\cos(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{3/2}} du + i \int_0^{\infty} \frac{\sin(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{3/2}} du, \\ \int_0^{\infty} \frac{\cos(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{3/2}} du &= \frac{\tilde{\omega}}{\sqrt{\varpi^2 y_0^2 + 4d^2}} K_1 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right), \\ \int_0^{\infty} \frac{\cos(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du &= \frac{2\tilde{\omega}}{3\sqrt{\varpi^2 y_0^2 + 4d^2}} K_2 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right), \\ \int_0^{\infty} \frac{\sin(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{3/2}} du &= \\ = \frac{\tilde{\omega}\pi}{2\sqrt{\varpi^2 y_0^2 + 4d^2}} \left[ L_{-1} \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) - I_1 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) \right], \\ \int_0^{\infty} \frac{\sin(\tilde{\omega}u)}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du &= \\ = \frac{\tilde{\omega}^2 \pi}{6(\varpi^2 y_0^2 + 4d^2)} \left[ I_2 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) - L_{-2} \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) \right], \end{aligned}$$

$$I_2(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+2}}{k!(k+2)!}, \quad (36)$$

$$K_2(x) = -I_2(x) \ln \frac{x}{2} + \frac{2}{x^2} - \frac{1}{2} + \frac{(-1)^n}{2} \sum_{k=0}^{\infty} \frac{(x/2)^{2k+2}}{k!(k+2)!} (\psi(k+1) + \psi(k+3)), \quad (37)$$

$$L_{-2}(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k-1}}{\Gamma(k + \frac{3}{2})\Gamma(k - \frac{1}{2})}, \quad (38)$$

$$\Gamma\left(k + \frac{3}{2}\right) = \frac{\sqrt{\pi}(2k+1)!}{2^{2k+1} \cdot k!}, \quad \Gamma\left(k - \frac{1}{2}\right) = \frac{\sqrt{\pi}(2k-2)!}{2^{2k-2} \cdot (k-1)!}.$$

Hence

$$K(x, y; \xi, \eta) = K_1(x, y; \xi, \eta) + \dots + K_{14}(x, y; \xi, \eta)$$

and the integral equation becomes

$$\begin{aligned} \frac{\varpi}{4\pi} \sum_{i=1}^{14} \int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_i(x, y; \xi, \eta) d\xi d\eta = \\ = - \left( \frac{\partial h(x, y)}{\partial x} + i\tilde{\omega}h(x, y) \right) \exp(i\tilde{\omega}x). \end{aligned} \quad (39)$$

In the sequel we shall provide adequate quadrature formulas for the integrals from the left hand part of the equation (39) in order to discretize it. Let

$$K_1(x, y; \xi, \eta) = \frac{1}{\varpi^2 y_0^2} \left( \frac{x_0}{\sqrt{x_0^2 + \varpi^2 y_0^2}} - \frac{x_0}{|x_0|} \right),$$

$$K_2(x, y; \xi, \eta) = \frac{1}{\varpi^2 y_0^2} \left( 1 + \frac{x_0}{|x_0|} \right),$$

$$K_3(x, y; \xi, \eta) = \frac{-i\tilde{\omega}}{\sqrt{x_0^2 + \varpi^2 y_0^2}},$$

$$K_4(x, y; \xi, \eta) = -\frac{\tilde{\omega}^2}{2} \frac{x_0}{|x_0|} \ln \left( |x_0| + \sqrt{x_0^2 + \varpi^2 y_0^2} \right),$$

$$K_5(x, y; \xi, \eta) = \frac{\tilde{\omega}^2}{2} \ln(\varpi |y_0|) \left( 1 + \frac{x_0}{|x_0|} \right),$$

$$K_6(x, y; \xi, \eta) = \frac{\tilde{\omega}^2 x_0}{\sqrt{x_0^2 + \varpi^2 y_0^2}},$$

$$K_7(x, y; \xi, \eta) = \frac{\tilde{\omega}}{\varpi |y_0|} K_1(\tilde{\omega}\varpi |y_0|) - \frac{1}{\varpi^2 y_0^2} - \frac{\tilde{\omega}^2}{2} \ln \frac{\tilde{\omega}\varpi |y_0|}{2} + \frac{\tilde{\omega}^2}{2} \ln \frac{\varpi}{2},$$

$$\begin{aligned}
K_8(x, y; \xi, \eta) &= \int_0^{x_0} \frac{\exp(i\tilde{\omega}u) - 1 - i\tilde{\omega}u + \tilde{\omega}^2 u^2/2}{(u^2 + \varpi^2 y_0^2)^{3/2}} du, \\
K_9(x, y; \xi, \eta) &= \int_0^{x_0} \exp(i\tilde{\omega}u) \frac{8d^2 - u^2 - \varpi^2 y_0^2}{(u^2 + \varpi^2 y_0^2 + 4d^2)^{5/2}} du, \\
K_{10}(x, y; \xi, \eta) &= -\frac{\tilde{\omega}}{\sqrt{\varpi^2 y_0^2 + 4d^2}} K_1 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right), \\
K_{11}(x, y; \xi, \eta) &= \frac{6\tilde{\omega}d^2}{\sqrt{\varpi^2 y_0^2 + 4d^2}} K_2 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right), \\
K_{12} &= \frac{\tilde{\omega}\pi i}{2\sqrt{\varpi^2 y_0^2 + 4d^2}} \left[ L_{-1} \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) - I_1 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) \right], \\
K_{13} &= \frac{3\tilde{\omega}^2 \pi i d^2}{2(\varpi^2 y_0^2 + 4d^2)} \left[ L_{-2} \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) - I_2 \left( \tilde{\omega} \sqrt{\varpi^2 y_0^2 + 4d^2} \right) \right], \\
K_{14}(x, y; \xi, \eta) &= \frac{i\pi\tilde{\omega}^2}{2\varpi|y_0|} \left( I_1(\tilde{\omega}\varpi|y_0|) - L_{-1}(\tilde{\omega}\varpi|y_0|) + \frac{2}{\pi} \right).
\end{aligned}$$

The analytical results from [1] suggest us to presume the following behaviour of the unknown function

$$\tilde{f}(\xi, \eta) = \frac{g(\xi, \eta)}{\sqrt{\xi^2 - \eta^2}},$$

whence we have

$$\begin{aligned}
&\int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_1(x, y; \xi, \eta) d\xi d\eta = \\
&= \frac{1}{\varpi^2} FP \int_{-1}^1 \frac{1}{y_0^2} \left( \int_{|\eta|}^1 \frac{g(\xi, \eta)}{\sqrt{\xi^2 - \eta^2}} \exp(i\tilde{\omega}\xi) \left( \frac{x_0}{\sqrt{x_0^2 + \varpi^2 y_0^2}} - \frac{x_0}{|x_0|} \right) d\xi \right) d\eta, \quad (40)
\end{aligned}$$

where  $FP$  stands for the finite part of the hypersingular integral as it is introduced by Ch. Fox in [7]. Taking into account that  $x(1) = x(-1) = 1$  we assume the following behaviour

$$\begin{aligned}
&\int_{|\eta|}^1 \frac{g(\xi, \eta)}{\sqrt{\xi^2 - \eta^2}} \exp(i\tilde{\omega}\xi) \left( \frac{x_0}{\sqrt{x_0^2 + \varpi^2 y_0^2}} - \frac{x_0}{|x_0|} \right) d\xi = \\
&= \sqrt{1 - \eta^2} G(x, y; \eta), \quad (41)
\end{aligned}$$

where  $G(x, y; \eta)$  is finite in  $\eta = \pm 1$ . We consider on  $D$  a net consisting of the nodes (grid points, controll points)  $(x_i, \bar{y}_j) = \left( \frac{i}{n}, \frac{2j+1}{2n} \right)$ ,  $i = 1, \dots, n$ ,  $j = -i, -i+1, \dots, i-1$ . For the hypersingular integral occuring in (40) we may use the quadrature formula for equidistant controll points given by Dumitrescu [5]

$$FP \int_{-1}^1 \frac{\sqrt{1 - \eta^2} G(x_k, \bar{y}_l; \eta)}{(\bar{y}_l - \eta)^2} d\eta = \sum_{j=-n}^{n-1} G(x_k, \bar{y}_l; \bar{y}_j) A_{lj}, \quad (42)$$

$$A_{lj} = -\arccos(y_j) + \arccos(y_{j+1}) + \frac{\sqrt{1-y_j^2}}{y_j - \bar{y}_l} - \frac{\sqrt{1-y_{j+1}^2}}{y_{j+1} - \bar{y}_l} - \frac{\bar{y}_l}{\sqrt{1-\bar{y}_l^2}} \ln \left| \frac{C_{l(j+1)}}{C_{lj}} \right| \quad (43)$$

with

$$C_{lj} = \frac{\sqrt{1-y_j} \cdot \sqrt{1+\bar{y}_l} - \sqrt{1+y_j} \cdot \sqrt{1-\bar{y}_l}}{\sqrt{1-y_j} \cdot \sqrt{1+\bar{y}_l} + \sqrt{1+y_j} \cdot \sqrt{1-\bar{y}_l}}. \quad (44)$$

We have the quadrature formula

$$G(x_k, \bar{y}_l; \bar{y}_j) = \sum_{i=j}^n g_{ij} B_{ijkl} \quad (45)$$

with

$$g_{ij} = g(\bar{x}_{ij}, \bar{y}_j),$$

$$\bar{x}_{ij} = \begin{cases} x_i - \frac{1}{2n}, & -i < j < i-1, \\ x_i - \frac{1}{4n}, & j \in \{-i, i-1\}, \end{cases} \quad \tilde{x}_{ij} = \begin{cases} x_i - \frac{1}{n}, & -i < j < i-1, \\ x_i - \frac{1}{2n}, & j \in \{-i, i-1\}, \end{cases} \quad (46)$$

$$B_{ijk} = \frac{E_{ij} D_{ijkl}}{\sqrt{1-\bar{y}_j^2}}, \quad (47)$$

$$E_{ij} = \exp(i\tilde{\omega}\bar{x}_{ij}) \left[ \ln \left( x_i + \sqrt{x_i^2 - \bar{y}_j^2} \right) - \ln \left( \bar{x}_{ij} + \sqrt{\tilde{x}_{ij}^2 - \bar{y}_j^2} \right) \right], \quad (48)$$

$$D_{ijkl} = \left( \frac{x_k - \bar{x}_{ij}}{\sqrt{(x_k - \bar{x}_{ij})^2 + \varpi^2 (\bar{y}_l - \bar{y}_j)^2}} - \frac{x_k - \bar{x}_{ij}}{|x_k - \bar{x}_{ij}|} \right), \quad -i < j < i-1. \quad (49)$$

Finally we deduce

$$\int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_1(x_k, \bar{y}_l; \xi, \eta) d\xi d\eta = \sum_{i=1}^n \sum_{j=-i}^{i-1} g_{ij} K_{ijkl}^{(1)}, \quad (50)$$

where

$$K_{ijkl}^{(1)} = \frac{A_{lj} B_{ijkl}}{\varpi^2}. \quad (51)$$

Let

$$K_2(x, y; \xi, \eta) = \frac{1}{\varpi^2 y_0^2} \left( 1 + \frac{x_0}{|x_0|} \right). \quad (52)$$



We have

$$\begin{aligned} & \int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_2(x, y; \xi, \eta) d\xi d\eta = \\ & = \frac{2}{\varpi^2} F P \int_{-x}^x \frac{1}{y_0^2} \left( \int_{|\eta|}^x \frac{g(\xi, \eta) \exp(i\tilde{\omega}\xi)}{\sqrt{\xi^2 - \eta^2}} d\xi \right) d\eta. \end{aligned} \quad (53)$$

Assuming the behaviour

$$\int_{|\eta|}^{x_k} \frac{g(\xi, \eta) \exp(i\tilde{\omega}\xi)}{\sqrt{\xi^2 - \eta^2}} d\xi = \sqrt{x_k^2 - \eta^2} G^{(k)}(x_k; \eta), \quad (54)$$

we have

$$\begin{aligned} & \int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_2(x_k, \bar{y}_l; \xi, \eta) d\xi d\eta = \\ & = \frac{2}{\varpi^2} F P \int_{-x_k}^{x_k} \frac{\sqrt{x_k^2 - \eta^2} G^{(k)}(x_k; \eta)}{(\bar{y}_l - \eta)^2} d\eta = \frac{2}{\varpi^2} \sum_{j=-k}^{k-1} G^{(k)}(x_k; \bar{y}_j) A_{lj}^{(k)}, \end{aligned} \quad (55)$$

where

$$\begin{aligned} A_{lj}^{(k)} &= -\arccos\left(\frac{y_j}{x_k}\right) + \arccos\left(\frac{y_{j+1}}{x_k}\right) + \\ &+ \frac{\sqrt{x_k^2 - y_j^2}}{y_j - \bar{y}_l} - \frac{\sqrt{x_k^2 - y_{j+1}^2}}{y_{j+1} - \bar{y}_l} - \frac{\bar{y}_l}{\sqrt{x_k^2 - \bar{y}_l^2}} \ln \left| \frac{C_{l(j+1)}^{(k)}}{C_{lj}^{(k)}} \right| \end{aligned} \quad (56)$$

with

$$C_{lj}^{(k)} = \frac{\sqrt{x_k - y_j} \cdot \sqrt{x_k + \bar{y}_l} - \sqrt{x_k + y_j} \cdot \sqrt{x_k - \bar{y}_l}}{\sqrt{x_k - y_j} \cdot \sqrt{x_k + \bar{y}_l} + \sqrt{x_k + y_j} \cdot \sqrt{x_k - \bar{y}_l}}. \quad (57)$$

For calculating  $G^{(k)}(x_k; \bar{y}_j)$  we employ the quadrature formula

$$G^{(k)}(x_k; \bar{y}_j) = \frac{1}{\sqrt{x_k^2 - \bar{y}_j^2}} \int_{|\bar{y}_j|}^{x_k} \frac{g(\xi, \bar{y}_j) \exp(i\tilde{\omega}\xi)}{\sqrt{\xi^2 - \bar{y}_j^2}} d\xi = \sum_{i=|j|}^k g_{ij} \frac{E_{ij}}{\sqrt{x_k^2 - \bar{y}_l^2}}. \quad (58)$$

At last we find

$$\int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_2(x_k, \bar{y}_l; \xi, \eta) d\xi d\eta = \sum_{i=1}^n \sum_{j=-i}^{i-1} g_{ij} K_{ijkl}^{(2)} \quad (59)$$

with

$$K_{ijkl}^{(2)} = \begin{cases} \frac{2}{\varpi^2} A_{lj}^{(k)} \frac{E_{ij}}{\sqrt{x_k^2 - \bar{y}_l^2}}; & i \leq k, \\ 0; & i > k. \end{cases} \quad (60)$$

The kernels  $K_3$  and  $K_4$  are singular. We divide and multiply them by  $y_0^2$  for obtaining quadrature formulas similar to the formulas for  $K_1$ . We get

$$\begin{aligned} \int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) [K_3(x_k, \bar{y}_l; \xi, \eta) + K_4(x_k, \bar{y}_l; \xi, \eta)] d\xi d\eta = \\ = \sum_{i=1}^n \sum_{j=-i}^{i-1} g_{ij} K_{ijkl}^{(3,4)} \end{aligned} \quad (61)$$

with

$$K_{ijkl}^{(3,4)} = \begin{cases} \frac{A_{lj} (\bar{y}_l - \bar{y}_j)^2}{\sqrt{1 - \bar{y}_j^2}} [K_3(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j) + K_4(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j)] E_{ij}; & i \neq j \\ 0; & i = j. \end{cases} \quad (62)$$

Although it may happen that  $y_0 = 0$ , the kernels  $K_5, K_6, K_7$  and  $K_8$  have integrable singularities and the kernels  $K_9, K_{10}, K_{11}, K_{12}, K_{13}$  and  $K_{14}$  are continuous and we utilize the quadrature formulas

$$\int \int_D^* \tilde{f}(\xi, \eta) \exp(i\tilde{\omega}\xi) K_p(x_k, \bar{y}_l; \xi, \eta) d\xi d\eta = \sum_{i=1}^n \sum_{j=-i}^{i-1} g_{ij} K_{ijkl}^{(p)}, \quad p = 5, \dots, 14$$

where

$$K_{ijkl}^{(5)} = \begin{cases} \tilde{\omega}^2 B_{jl}^{(k)} E_{ij}, & i \leq k \\ 0, & i > k, \end{cases} \quad (63)$$

$$B_{jl}^{(k)} = (y_{j+1} - \bar{y}_l) \ln |y_{j+1} - \bar{y}_l| - (y_j - \bar{y}_l) \ln |y_j - \bar{y}_l|, \quad (64)$$

$$K_{ijkl}^{(p)} = E_{ij} K_p(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j) / n, \quad p = 6, \dots, 14. \quad (65)$$

For calculating  $K_7(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_l)$  we use the series expansions of the Bessel and Struve functions and we take into account that

$$\begin{aligned} K_7(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_l) = -\frac{\tilde{\omega}^2 (\psi(1) + \psi(2))}{4} + \frac{\pi i \tilde{\omega}}{4}, \\ \psi(1) = -0.5772, \psi(2) = 0.4228. \end{aligned} \quad (66)$$

The kernels  $K_8(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j)$  and  $K_9(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j)$  are integrals which are evaluated numerically with the trapezoidal rule.

For calculating the Bessel (MacDonald) functions  $K_1$  and  $K_2$  we may utilize the series expansions (33) and (37). We may also utilize the libraries offered by MATLAB.

For calculating the kernels  $K_{12}(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j)$  and  $K_{14}(x_k, \bar{y}_l; \bar{x}_{ij}, \bar{y}_j)$  we use the integral representations

$$I_\nu(x) = \frac{(x/2)^\nu}{\sqrt{\pi} \Gamma(\nu + 1/2)} \int_{-1}^1 \cosh xs (1 - s^2)^{\nu-1/2} ds =$$

$$= \frac{(x/2)^\nu}{\sqrt{\pi}\Gamma(\nu+1/2)} \int_0^{\pi/2} (\exp(x \cos t) + \exp(-x \cos t)) \sin^\nu t dt, \quad \nu > -1/2, \quad (67)$$

$$L_\nu(x) = \frac{(x/2)^\nu}{\sqrt{\pi}\Gamma(\nu+1/2)} \int_0^{\pi/2} (\exp(x \cos t) - \exp(-x \cos t)) \sin^\nu t dt. \quad (68)$$

From (34), (38) and (68) we deduce

$$L_{-1}(x) = \frac{2}{\pi} - \frac{x}{\pi} \int_0^{\pi/2} (\exp(-x \cos t) - \exp(x \cos t)) \sin^2 t dt,$$

$$L_{-2}(x) = -\frac{2}{\pi x} + \frac{2x}{3\pi} - \frac{x^2}{3\pi} \int_0^{\pi/2} (\exp(-x \cos t) - \exp(x \cos t)) \sin^4 t dt,$$

whence, taking into account (67) it follows

$$L_{-1} - I_1 = \frac{2 \exp(-x)}{\pi} + \frac{2x}{\pi} \int_0^{\pi/2} \exp(-x \cos t) (\sin t - \sin^2 t) dt, \quad (69)$$

$$L_{-2} - I_2 = -\frac{2}{\pi x} + \frac{2x \exp(-x)}{3\pi} + \frac{2x^2}{3\pi} \int_0^{\pi/2} \exp(-x \cos t) (\sin t - \sin^4 t) dt, \quad (70)$$

and the integrals are evaluated numerically with the trapezoidal rule.

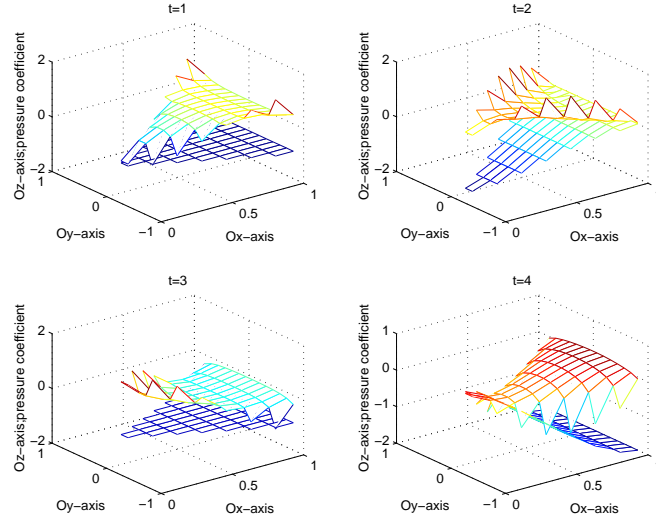


Fig. 1. Pressure coefficient field for  $d = 10$ .

Denoting

$$K_{ijkl} = K_{ijkl}^{(1)} + K_{ijkl}^{(2)} + K_{ijkl}^{(3,4)} + K_{ijkl}^{(5)} + \dots + K_{ijkl}^{(14)},$$

we obtain, discretizing the two-dimensional integral equation (24):

$$\frac{\omega}{4\pi} \sum_{i=1}^n \sum_{j=-i}^{i-1} g_{ij} K_{ijkl} = - \left( \frac{\partial h(x_k, \bar{y}_l)}{\partial x} + i\tilde{\omega} h(x_k, \bar{y}_l) \right) \exp(i\tilde{\omega} x_k). \quad (71)$$

After solving this equation we may obtain

$$\tilde{f}(x_k, \bar{y}_l) = \frac{g_{ij}}{\sqrt{x_k^2 - \bar{y}_l^2}}.$$

#### 4. The average drag coefficient and the propulsive force. Numerical results

In the sequel we shall deal with the pressure coefficient

$$C_p(x^{(1)}, y^{(1)}, t) = \text{Re}[\tilde{f}(x, y) \exp(i\omega t)]. \quad (72)$$

Among the aerodynamic characteristics of the wing, in this paper we are interested in the drag coefficient

$$C_D(t) = -2 \iint_D n_x C_p(ax, by, t) dx dy. \quad (73)$$

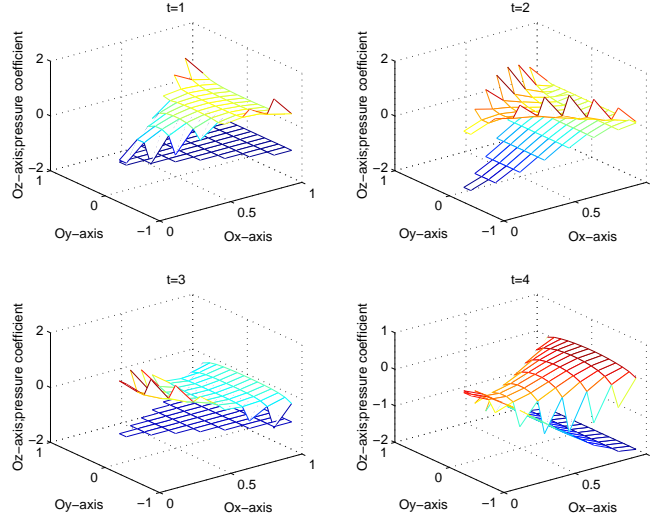


Fig. 2. Pressure coefficient field for  $d = 1$ .

We consider the oscillating delta wing whose equation is

$$0 = z^{(1)} - \alpha \exp(i\omega_1 x^{(1)} + i\omega t); \quad (x^{(1)}, y^{(1)}) \in D^{(1)} \quad (74)$$

whence

$$h(x, y) = \alpha \exp(i\tilde{\omega}_1 x), \quad \tilde{\omega}_1 = a\omega_1; \quad (x, y) \in D. \quad (75)$$

We have therefore

$$n_x = -\alpha\tilde{\omega}_1 \operatorname{Re} \left[ \exp(i\omega_1 x^{(1)} + i\omega t) \right],$$

whence

$$C_D(t) = 2\alpha\tilde{\omega}_1 \int \int_D \operatorname{Re} \left[ \exp(i\omega_1 x^{(1)} + i\omega t) \right] \operatorname{Re}[\tilde{f}(x, y) \exp(i\omega t)] dx dy.$$

Denoting

$$T = \frac{2\pi}{\omega},$$

the average drag coefficient is

$$\tilde{C}_D = \frac{1}{T} \int_0^T C_D(t) dt = \alpha\tilde{\omega}_1 \int \int_D \operatorname{Im}[\tilde{f}(x, y) \exp(-i\tilde{\omega}_1 x)] dx dy.$$

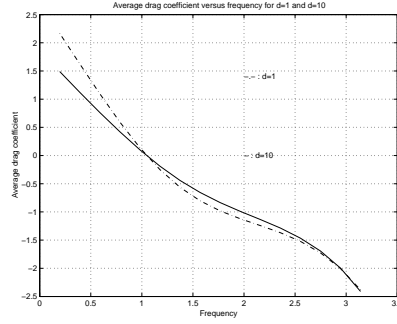


Fig. 3. Average drag coefficient versus frequency.

In Fig. 1 and Fig. 2 we present the pressure coefficient fields (divided by  $\alpha/2$ ) for  $t \in \{1, 2, 3, 4\}$ . In fig. 1 we considered  $d = 10$  and in fig. 2 we took  $d = 1$ . We also present the wings translated along the  $Oz$ -axis with  $\Delta z = -1$ . We may notice that the two pressure coefficients fields are very close but not identical. In fig. 3 we present the average drag coefficient (divided by  $\alpha^2$ ) against the frequency. For  $\omega > 1.1$  the average drag coefficient is negative, i.e. it appears a propulsive force and we notice that the propulsive force is bigger for  $d = 1$  (dash-dot line) than for the case  $d = 10$  (continuous line) i.e. it is bigger when the oscillatory wing is closer to the ground.

## References

- [1] Carabineanu, A., *Incompressible flow past oscillatory wings of low aspect ratio by the integral equations method*, Int. J. Numer. Engng., **45** (1999), 1187–1201.

- [2] Carabineanu, A., *Self-Propulsion of Aquatic Animals by Undulatory or Oscillatory Motion*, Workshop on Math. Modelling Environ. Life Sci. Problems, Constanța, (2004), 127–140.
- [3] Ditzkin, V. and Proudnikov, A., *Transformations intégrales et calcul opérationnel*, Editions Mir, Moscou, 1978 .
- [4] Dragos, L., *The theory of oscillating thick wings in subsonic flow. Lifting line theory*, Acta Mechanica, **54** (1985), 221–238.
- [5] Dumitrescu, D.-F., *Three methods for solving Prandtl's equation*, ZAMM, **76**, 6 (1996), 1–4.
- [6] Eversman, W., and Pitt, D., *Hybrid doublet lattice/doublet point method for lifting surfaces in subsonic flow*, J. Aircraft, **28**, 9 (1991), 572–578.
- [7] Fox, Ch., *A generalisation of the Cauchy principal value*, Canadian J. Math., **9** (1957), 110–115.
- [8] Homencovschi, D., *Theory of the lifting surface in unsteady motion in an inviscid fluid*, Acta Mechanica, **27** (1977), 205–216.
- [9] Ichikawa, A., *Doublet strip method for oscillating swept tapered wings in incompressible flow*, J. Aircraft, **22**, 11 (1985), 1008–1012.
- [10] Küssner, H.G., *Allgemeine Tragflächentheorie*, Luftfahrtforschung **17**, (1940), 370–378.
- [11] Landahl, M.T., *Kernel function for nonplane oscillating surfaces in a subsonic flow*, AIAA J, **5** (1967), 1054–1054.
- [12] Laschka, B., *Das Potential und das Geschwindigkeitsfeld der harmonisch schwingenden tragenden Fläch bei Unterschallströmung*, ZAMM, **43** (1963), 325–335.
- [13] Ueda, T. and Dowel, E.H., *A new solution method for lifting surfaces in subsonic flow*, AIAA J., **20**, 3 (1982), 348–355.
- [14] Watkins, C. E., Runyan, H. L. and Woolston, D. S. *On the kernel function of the integral equation relating the lift and downwash distributions of oscillating finite wings in subsonic flow*, NACA T.R. 1234 (1955).

## **Thermal Coupling Numerical Models for Boundary Layer Flows over a Finite Thickness Plate Exposed to a Time-Dependent Temperature**

**Emilia Mladin Cerna<sup>\*†</sup> and Dorin Stanciu<sup>\*</sup>**

The present study treats the case of a finite thickness planar plate of known material and dimensions, exposed to a ramp change in the temperature imposed at the plate back surface. The flow was considered laminar and of constant velocity. The temperature temporal variation was modeled as a ramp-up, a smooth function that realistically replaces the theoretical step change. The initial system state is of thermal equilibrium. Two numerical approaches have been used to model the heat transfer performance between the fluid and the plate: the Karman-Pohlhausen integral method and the finite-volume modeling built in the commercial FLUENT code. The surface heat flux was found sensitive to the plate thickness and material, as well as to the imposed temperature ramp duration and amplitude. Although a multitude of fluid-solid combinations may be considered in the analysis, a water flow over a steel plate was analysed here. Results are expressed in terms of a correction factor defined as the ratio between the surface heat flux associated with the finite thickness plate of specified material and the surface heat flux associated with the zero thickness plate. The model was validated against differential method and integral method results reported in the literature for the zero thickness plate and stationary regimes, the maximum error being about 6.5%.

### **1. Introduction**

Most of the previous works on heat convection in parallel flows over bodies use various boundary conditions at the contact surface. In all such cases, the plate thermal resistance is not encountered in calculus, although heat transfer may be

---

<sup>\*</sup> “Politehnica” University, Mechanical Engineering Department, Bucharest, Romania.

<sup>†</sup> e-mail: [mladin@yahoo.com](mailto:mladin@yahoo.com)

highly influenced by the impact body geometry and material. In practical applications however, it is most probably that the boundary conditions are known at the accessible surfaces, i.e. the ones that are not in contact with the flow. Use of common measuring instruments at the contact surface between the fluid and the body would clearly disturb the boundary layers and thus the measurements will be erroneous.

In the present paper, the authors aim to study the dynamics of the heat transfer in a parallel steady laminar flow over a finite thickness plate. The transient regime results from a ramp change in the temperature imposed at the bottom plate surface (the one that is not in contact with the fluid). Two mathematical approaches have been used for this purpose. One relies on a previously developed model [1, 2], based on the Karman-Pohlhausen integral methodology to formulate ordinary differential governing equations. The model was however modified to include a forcing function for the time-dependent boundary condition. The second approach uses the built-in finite volume conservation equations of the FLUENT code that is based on the Patankar algorithms for incompressible flows [3]. Due to paper length restrictions, numerical solutions are reported for a water flow over a steel plate of specified thickness. However, the two models may be equally used for other combinations of fluids and solids, as long as the fluid Prandtl numbers are greater than 0.7. Figure 1 schematically presents the physical system. The incompressible fluid flow is stationary and laminar and has a constant temperature  $T_\infty$ . Its velocity  $U_\infty$  is constant as no pressure gradients are assumed. The flow is parallel with a plate of thickness  $E$ , which is much smaller than its length. The plate bottom surface temperature has an imposed temporal variation. The initial state is of thermal equilibrium in the entire system. Therefore, when the plate bottom surface temperature changes, the generated heat flux penetrates the plate and gives rise to a thermal boundary layer developing in the fluid. While the hydrodynamic boundary layer thickness  $\delta(x)$  and velocity profile  $u(x, y)$  are constant in time, the thermal boundary layer thickness  $\delta_t(x, t)$  and temperature profile  $T(x, y, t)$  are time-dependent as long as the transients last. The instantaneous temperature distribution within the plate  $T_p(x, y, t)$  is distinctively illustrated in Fig. 1 for both the penetration phase and after penetration phase.

## 2. Description of the models and solution methodologies

Exact analytical solutions are often impossible to find when dealing with transient phenomena. On the other hand, where transient effects can be incorporated with similarity methods for example, the resulting solutions may impose particular relationships between variables. For these reasons, the authors chose to use two different approaches: (i) a semi-analytical methodology, based on the integral method of Karman-Pohlhausen; (ii) a numerical method based on finite volumes, built-in the FLUENT code. The first one, although approximate to some extent, has the advantage of providing ordinary differential equations that governs the system behavior. Such ODE's can then be easily integrated to study nonlinear dynamics effects associated with transients and embedded nonlinearities in the governing equations.



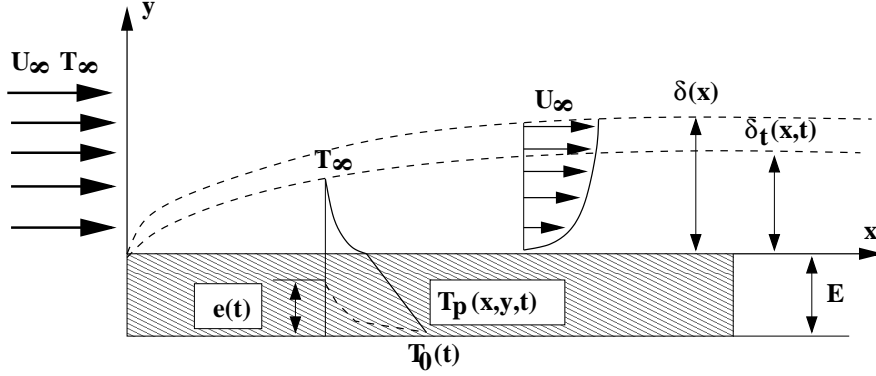


Fig. 1. Description of the physical system.

The second approach is more accurate but the use of FLUENT code does not allow enough flexibility and hides the governing equations, making thus the embedded nonlinearities invisible. For example, energy equation is linear in temperature but highly nonlinear in the thermal boundary layer thickness [1, 4], which is not a variable in the finite-volume approach.

Momentum and energy conservation equations within the fluid and the plate, as well as the energy conservation equation at the interface are used in both approaches. The basic assumptions considered here for the heat transfer modeling are: (i) incompressible fluid with constant thermo-physical properties; (ii) negligible viscous dissipation; (iii)  $\delta_t \leq \delta$ , ( $Pr \geq 0.7$ ); (iv) constant plate thermal conductivity; (v) one-dimensional conduction and no heat sources within the plate.

### 2.1. Integral method

Equations for momentum and energy conservation in their integral and differential forms were used with temporally adaptive profiles for fluid and temperature to obtain governing equations for the thermal boundary layer response [1, 2, and 5]. The integral equations were solved together with proposed velocity and temperature profiles for the fluid and plate material. The profiles were modeled as high order polynomials, according to the Karman-Pohlhausen methodology. The time dependent polynomial coefficients allow the instantaneous adaptation of the profiles to the transient boundary conditions. In dimensionless format and in connection to the boundary conditions, the fourth-order polynomials for the fluid profiles are [1, 5].

$$u_* = 2\eta - 2\eta^3 + \eta^4, \quad \eta = \frac{y}{\delta}, \quad (1)$$

$$\theta = \theta_s - \left(2\theta_s + \frac{1}{3}\omega\right) \beta + \omega \beta^2 + (2\theta_s - \omega) \beta^3 + \left(-\theta_s + \frac{1}{3}\omega\right) \beta^4, \quad (2)$$

where

$$\omega \equiv \frac{x_* \text{Pr} \Delta^2}{2} \frac{\partial \theta_s}{\partial \tau}, \quad \beta = \frac{y}{\delta_t}.$$

In the impact plate case, two distinct temporal phases were considered separately: (i) the initial phase of plate penetration, treated as conduction through a semi-infinite body with imposed thermal condition at  $y = -E$ , and (ii) after-penetration phase, associated with the thermal boundary layer development within the fluid (see Fig. 1). In the first phase, the instantaneous penetration depth is  $e(t) \leq E$  and the boundary conditions allow the plate temperature profile modeling as a third polynomial [6]. After the heat flux reaches the front plate surface (Fig. 1), the boundary conditions at  $y = 0$ ,  $\theta_p = \theta_s$ , lead to a different temperature profile inside the plate [2]. The resulting polynomials are presented below:

$$\theta_p^i = \theta_0 - \left( \frac{3}{2} \theta_0 + \frac{1}{2} \omega_e \right) \cdot \frac{y + E}{e} + \omega_e \cdot \left( \frac{y + E}{e} \right)^2 + \frac{1}{2} (\theta_0 - \omega_e) \cdot \left( \frac{y + E}{e} \right)^3 \quad (3)$$

with  $\omega_e \equiv \frac{e_*^2 A \text{Pr}}{2} \frac{\partial \theta_0}{\partial \tau}$  and  $-E \leq y \leq -E + e(t)$ .

$$\theta_p = \theta_s + \left( \theta_s + \frac{2}{3} \omega_p + \frac{\text{Pr} A E_*^2}{6} \theta'_0 - \theta_0 \right) \frac{y}{E} + \omega_p \left( \frac{y}{E} \right)^2 + \frac{2 \omega_p - \text{Pr} A E_*^2 \cdot \theta'_0}{6} \left( \frac{y}{E} \right)^3 \quad (4)$$

with  $\omega_p \equiv \frac{E_*^2 A \text{Pr}}{2} \frac{\partial \theta_s}{\partial \tau}$  and  $-E \leq y \leq 0$ .

Temperature polynomial profiles are illustrated in Figure 2, next to those obtained with the aid of FLUENT code.

The penetration time is calculated with the differential equation governing the instantaneous penetration depth  $e$  and corresponds to the condition  $e = E$ . This equation has been derived by using the temperature profile (4) in the plate energy conservation integral equation [1],

$$\frac{\partial}{\partial \tau} \left( \frac{e_*}{E_*} \right) \left( 6 \theta_0 - \frac{e_*^2}{E_*^2} \text{Pr} A E_*^2 \frac{\partial \theta_0}{\partial \tau} \right) = \frac{24 \theta_0}{\text{Pr} A E_* e_*} - \frac{2 e_*}{E_*} \frac{\partial \theta_0}{\partial \tau} + \frac{\text{Pr} A E_*^2}{3} \left( \frac{e_*}{E_*} \right)^3 \frac{\partial^2 \theta_0}{\partial \tau^2}. \quad (5)$$

The use of polynomial (7) in the fluid momentum integral equations lead to the hydrodynamic boundary layer thickness:

$$\delta = 5,83 \sqrt{\nu \cdot x_* / U_\infty} = C \sqrt{x_*}. \quad (6)$$

For the energy conservation equations within the fluid and at the interface, two more assumptions have been made in addition to the temperature profile (2) [5]:

(i) The thermal boundary layer thickness varies with the spatial coordinate  $x$  in a similar way as the hydrodynamic boundary layer thickness; it results that their ratio is  $x$ -independent,  $\Delta(\tau) \equiv \delta_t(x, \tau) / \delta(x, \tau)$ ;

(ii) The temperature distribution within the fluid may be expressed as a product between the steady-state solution and a transient correction factor, by using the

variable separation,  $\theta_s(x, t) = \theta_s^{ss}(x) \cdot \theta_s^t(t)$ . In this way, the spatial coordinate  $x$  is treated as a parameter and the resulting governing equations (7) and (8) for the fluid are ordinary differential equations with respect to time only.

$$\begin{aligned} \frac{\partial \Delta}{\partial \tau} \left( \frac{3}{10} \theta_s - \frac{1}{40} x_* \text{Pr} \frac{\partial \theta_s}{\partial \tau} \Delta^2 \right) &= \frac{\theta_s}{x_*} \left[ \frac{2}{\text{Pr}} \frac{1}{\Delta} - \frac{5.83^2}{2} (2 - \theta_s^{ss}) \Delta \cdot \varphi_1 \right] \\ &+ \left[ \frac{5.83^2}{4} \text{Pr} (4 - \theta_s^{ss}) \Delta^3 \varphi_2 - \frac{2}{15} \Delta \right] \frac{\partial \theta_s}{\partial \tau} + \frac{x_* \text{Pr}}{120} \Delta^3 \cdot \frac{\partial^2 \theta_s}{\partial \tau^2}, \end{aligned} \quad (7)$$

where  $\varphi_1 = 2/15 \cdot \Delta - 3/140 \cdot \Delta^3 + \Delta^4/180$  and  $\varphi_2 = \Delta/90 - \Delta^3/420 + \Delta^4/1512$ .

$$\frac{d\theta_s}{d\tau} \left( 2 + \frac{\Delta \sqrt{x_*} \Lambda}{AE_*} \right) = \frac{6}{\text{Pr} AE_*^2} \left[ \left( \theta_0 - \frac{1}{6} \text{Pr} AE_*^2 \cdot \frac{d\theta_0}{d\tau} \right) - \theta_s \left( \frac{2\Lambda \cdot E_*}{\Delta \cdot x_*} + 1 \right) \right]. \quad (8)$$

It is remarkable that the governing equations (7) and (8) are coupled and highly nonlinear. However, they can be integrated by numerical techniques that are commonly used for ordinary differential equations. The steady-state forms and solutions are readily obtained by cancelling the time-derivatives.

The laminar flow condition imposes a Reynolds number less than the critical value  $Re_{x,cr} = 5 \cdot 10^5$ , which also limits the spatial coordinate at a maximum value of  $x^* = 14700$ .

The transient solutions were obtained by integrating the ordinary differential equations by Runge-Kutta algorithms of fourth and fifth order. The integration was performed with different time steps, depending on the variable time responses. Most commonly, very small time steps were used at the beginning, due to the rapid increase of the thermal boundary layer thickness. The subsequent system dynamics allowed for larger time steps and thus for reasonable computational durations. The singularities present at  $\tau = 0$  were avoided by considering limiting values for  $\Delta$  and  $\theta_s$ .

## 2.2. Finite-volume approach

For this second method, numerical solutions were obtained with the commercial code FLUENT 6.0. It is based on the finite control volume formulation of Navier-Stokes equations on an unstructured grid. For incompressible flows, the code uses the segregated technique that consists in sequentially and iteratively solving the momentum, pressure correction and energy equations. This solver code provides many spatial and temporal discretization procedures, which can be selected according to the particular case under consideration. Thus, there were employed here the second order upwind scheme for the spatial discretization of momentum and energy equations, and the PISO algorithm with neighbour correction for the pressure correction equation. The temporal discretization was performed with Euler second order implicit scheme.

The full fluid Navier-Stokes system was solved within a rectangle computation domain, whose horizontal boundary (containing the upper plate surface) was extended with half plate length at both the leading edge and trailing edge. The vertical boundary was about one hundred times greater than the hydrodynamic boundary layer

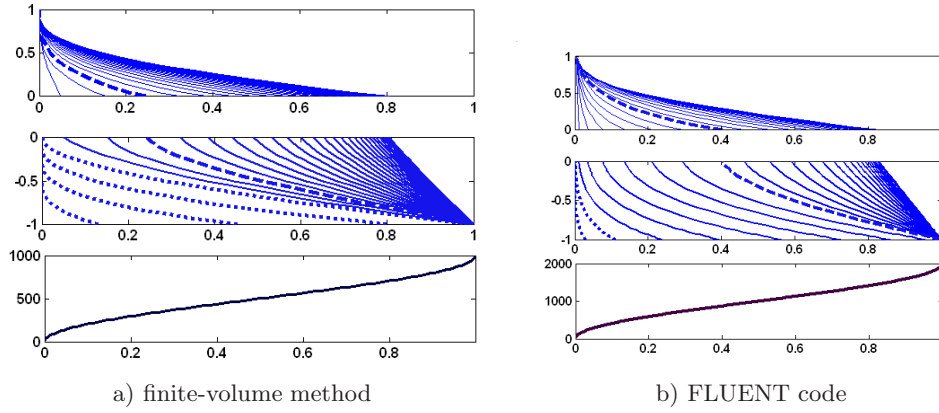


Fig. 2. Temperature profiles.

thickness. The plate computational domain coincided with its real geometry and was used to solve the solid energy conservation equation. The two domains were meshed with 2D quadrilateral cells having a variable density in the  $x$  and  $y$  directions around the fluid-solid interface.

The initial solution was obtained by a steady state calculation for an isothermal boundary layer flow developed on the finite thickness flat plate. Once the time-dependent thermal boundary condition  $T_0(t)$  was set at the plate bottom surface ( $y = -E$  in Fig. 1), the unsteady calculation was performed and transient solutions resulted. For the particular case of a step change in the surface temperature, the grid independency on the numerical solution was obtained through the grid adaptation technique of the initial mesh at the solid-fluid boundary interface. Then, the same mesh was used for all other calculations.

The Figure 2 illustrates the temperature profiles within the fluid as well as within the plate, as obtained with the finite-volume approach and FLUENT code.

### 3. Heat transfer performance

This study analyzes the transient heat transfer performances in parallel flows over a flat plate. Most of the correlations reported in the literature do not consider the conduction through the plate, or, otherwise said, consider the plate of zero thickness and impose a condition (e.g. temperature or heat flux) at the contact surface. However, real situations deal with finite thickness plates:  $E \neq 0$ . Then, the surface temperature varies along the plate and the local Nusselt number ( $\equiv h x / \lambda$ ) used to compute the local heat transfer coefficient  $h$ , becomes insufficient for using the Newton's law of cooling:  $q_s(x) = h(x) [T_s(x) - T_\infty]$ . In order to point out the impact plate influence on the heat transfer, a correction factor is defined as the ratio between the instantaneous heat flux associated with a finite thickness plate and the instan-

taneous heat flux associated with a zero thickness plate, both at the same spatial location. The surface heat flux is easily obtained from the Fourier's law and the temperature profile within the thermal boundary layer. The integral method provides the following expression based on Eq. (2):

$$CF \equiv \frac{q_s(E \neq 0)}{q_s(E = 0)} = \left( \theta_s + \frac{\omega}{6} \right) \frac{\Delta(E = 0)}{\Delta(E \neq 0)}. \quad (9)$$

For a finite thickness plate, the interface temperature  $\theta_s$  is always inferior to unity. As time goes to infinity, the correction factor reaches its steady-state value ( $\omega = 0$ ), which is also inferior to unity:

$$CF^{ss} = \theta_s^{ss} \frac{\Delta^{ss}(E = 0)}{\Delta^{ss}(E \neq 0)} < 1. \quad (10)$$

Knowledge of the steady-state correction factor would allow the use of the present correlations derived for a zero thickness and isothermal impact plate.

#### 4. Selected forcing function and system parameters

In the present study, the temperature imposed at the plate bottom surface will follow a temporal ramp variation of finite duration. The step change may be viewed as a limiting case, i.e., a ramp of zero duration. This type of forcing function has been chosen as it models the real variations which are never totally abrupt. However, the developed models can be used with any other temporal variation of the imposed boundary condition so long as the variations are piece-wise smooth.

The ramp function is characterized by its duration  $D$ . In dimensionless variables, it starts from an initial value of zero, corresponding to thermal equilibrium, and a final value of unity associated with the steady-state conditions [1].

$$\theta_0 = \begin{cases} \frac{1}{2} \left[ 1 + \sin \left( \frac{\pi \cdot \tau}{D} - \frac{\pi}{2} \right) \right], & 0 \leq \tau < D \\ 1, & \tau \geq D \end{cases} \quad (11)$$

The ramp profile is presented in Figure 1, in the bottom panels.

Solutions were obtained for various system parameters but will be reported here only for an illustrative case: *water flow over a steel plate*. The considered mean thermophysical properties are presented in Table 1. An extended paper may cover multiple other combinations. Table 2 presents the conversion of a few non-dimensional values into the physical values considered in this study, fact that will enable the interpretation of results in section 6. The flow velocity was chosen  $U_\infty = 1$  m/s.

#### 5. Model validation

Under steady-state conditions, the integral method provided the following expression for the Nusselt, as derived from its definition:

Table 1

	water	Steel
$k$ [W / mK]	0.59	14.7
$\alpha$ [m <sup>2</sup> / s]	$0.142 \cdot 10^{-6}$	$4 \cdot 10^{-6}$
Pr	7.0	—

Table 2

$C = 5,83\sqrt{\nu/U_\infty}$	$= 0,00583$
$\tau \equiv t \cdot \nu / C^4 = 10^4$	$t = 11.5$ sec
$E_* \equiv E / C^2 = 100$	$E = 3.4$ mm
$x^* \equiv x / C^2 = 7000$	$x = 238$ mm

$$Nu_x \equiv \frac{hx}{k} = 0,343Re_x^{1/2} \frac{1}{\Delta_{ss}}, \quad (12)$$

where  $h \equiv -k(\partial T / \partial y)_{y=0} / (T_s - T_\infty)$ .

On the other hand, the FLUENT code provided values for the same Nusselt numbers. All values were validated against other solutions previously reported for steady-state conditions and a zero thickness plate. Particular values and associated errors are shown in Table 3 for water flows, location  $x^* = 7000$ ,  $Re_x = 237922$  and the system parameters specified in Tables 1 and 2.

Table 3

Correlation	Value and relative error
Exact solution (Succes 1985): $Nu_x \equiv \frac{hx}{k} = 0,332Re_x^{1/2} Pr^{1/3}$	309.774 ( <b>0%</b> )
Integral solution (Padet 1998): $Nu_x \equiv \frac{hx}{k} = 0,343Re_x^{1/2} Pr^{1/3}$	320.038 ( <b>3.3%</b> )
Equation (12), $E_* = 0, \Delta^{ss} = 0.5076$ for $Pr = 7$ .	329.600 ( <b>6.4%</b> )
Fluent code	311.013 ( <b>0.4%</b> )

The assumption that the boundary layer thickness ratio  $\Delta$  is not a function of  $x^*$  in Eq. (7) was checked for  $x^* = 1000 - 13000$  and a plate thickness  $E_* = 100$ . The steady-state values  $\Delta^{ss}$  ranged from 0.4576 to 0.4875, which leads to an interval error of 6.2%.

The error levels of less than 6.5% rend the methodologies used here appropriate for engineering applications. The steady-state error analysis confers credibility to the transient solutions used further to characterize the instantaneous heat transfer performance and which cannot be compared to other previously reported results.

## 6. Results

First, the two methodologies were compared under steady-state conditions. The figure 3 presents the local temperature at the contact surface ( $y = 0$  in Fig. 1). It is obvious that the differences are so very small that they hardly can be noticed. This result gives credit to the approximate integral method, which in turn can provide analytical expressions for heat transfer system performances. The thermal boundary

layer thicknesses were not compared because the full Navier-Stokes solutions, obtained with FLUENT, do not imply the use of this variable. Under transient conditions,

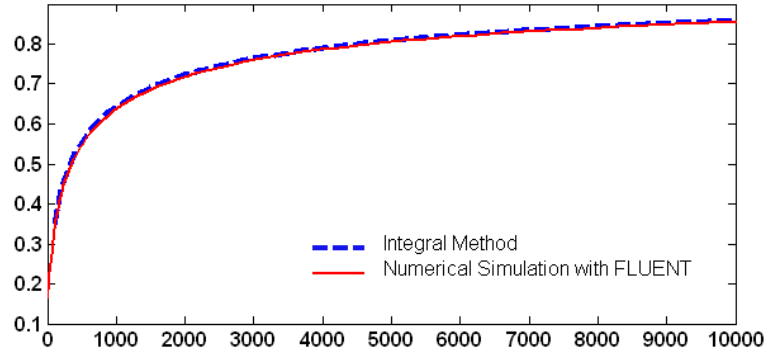


Fig. 3. Steady-state local surface temperature.

Figure 4 illustrates the surface temperature dynamics for different ramp durations and for both methods. It appears that the major differences occur in the penetration times, they being much smaller when calculated with FLUENT solver. This result may be attributed to the  $3^{rd}$  polynomial profile imposed for the plate temperature during the penetration time, as well as to singularities encountered in the governing differential equations (7) and (8) when the thermal boundary layer starts developing.

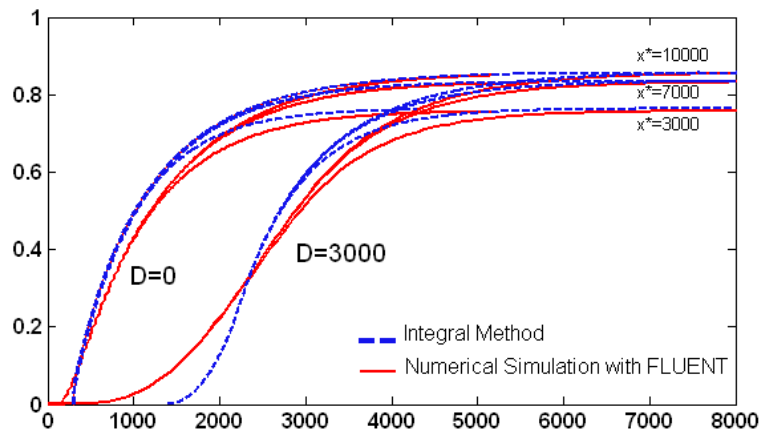


Fig. 4. Surface temperature dynamics for two ramp durations and three x-locations.

The influence of the impact plate on the heat transfer performance represents

the most significant aspect of the problem. As stated earlier, results were expressed as a correction factor CF defined in Eq. (9). In this way, the influence of thermal coupling is indicated by the difference between the CF value and unity. The plate material and thickness, as well as the type of fluid, have a major influence on the heat transfer rates, facts that were reported in a previous publication [1,2, and 5]. The purpose of this paper is to compare the approximate integral method with the finite-volume numerical method which is built-in the widely used commercial code FLUENT. The comparison of the two types of solutions is shown in Figure 5 for different ramp durations in the forcing function of  $T_0$  and for the location  $x^* = 7000$ .

It is obvious that all the CF-values are inferior to unity even under steady-state conditions. However, the FLUENT code provided a value ( $CF = 0.9228$ ) that is 4.7% higher than that obtained with the integral method ( $CF = 0.8791$ ).

The maximum zones (bumps) indicated at the end of transients derive from the fact that during the penetration times, there is a heat flux to the fluid for the zero-thickness plate but no heat transfer for the finite-thickness plate. Equation (9) indicates that CF depends on the ratio of the boundary layer thicknesses associated with  $E = 0$  and  $E \neq 0$ , respectively. This ratio is obviously higher than unity during the penetration times and induces also higher values afterwards. The “bumps” are shown to decrease as the ramp duration increases and have lower values at more remote locations from the plate leading edge.

Under transient conditions, the instantaneous differences are significant, especially due to the high penetration times related to the integral approach. However, even shifted in time, the CF-growths deduced from this method are more abrupt, rapidly approaching the steady-state values. Table 4 presents the time-averages of the correction factor for the curves of Figure 5. Transient time was defined as the time needed to reach 95% of the steady-state value.

Table 4

	$D = 0$	$D = 1\,000$	$D = 2\,000$	$D = 3\,000$
Integral method	0.8226	0.7064	0.6617	0.6011
FLUENT	0.8425	0.7057	0.6741	0.5855

Results indicate that the time-average differences between the two approaches range from  $-2.4\%$  for the step change in  $T_0$  ( $D = 0$ ) to  $2.6\%$  for a ramp change of duration  $D = 3\,000$ . If the transient time is redefined, i.e. the time needed to reach 90% or 99% of the steady-state value, the time-averages diminish or increase accordingly.

## 7. Conclusion

The study aimed to characterize the heat transfer performance in the case of a parallel laminar and stationary flow over a finite thickness plate. Two methodologies were applied: a semi-analytical one based on the integral approach of Karman-



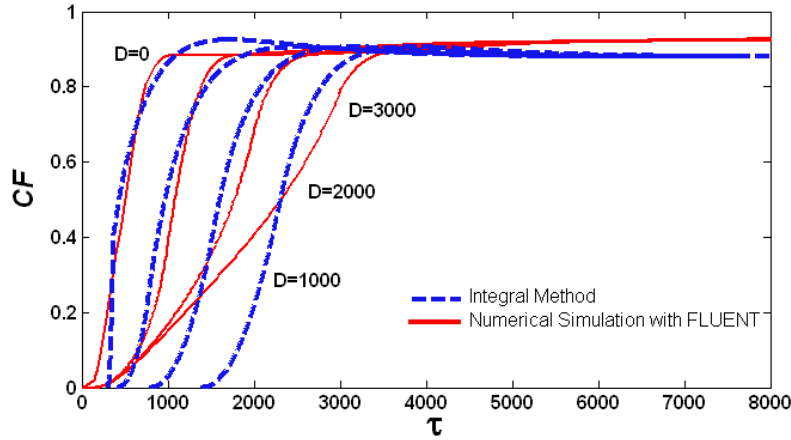


Fig. 5. Correction factor CF for the different ramp durations and  $x^* = 7000$ .

Pohlhausen, and a numerical simulation using the finite-volume discretization of the system domain with the aid of the FLUENT commercial solver. Both methods were validated against results previously reported for a zero-thickness plate and stationary conditions. The paper presents an illustrative case of a water flow of 1 m/s incident velocity over a steel plate, 0.4 m long and 3.4 mm thick. The forcing function was chosen to be a ramp change in the temperature imposed at the plate surface, that is not in contact with the fluid.

The interface temperature was shown to significantly vary with the spatial coordinate parallel to the plate, fact that does not allow anymore the simple use of a Nusselt number correlation for the heat transfer rate calculus. For this reason, the heat transfer results were reported as a correction factor defined as the ratio of the heat flux associated with the finite thickness plate and the heat flux associated with the zero thickness plate. In this way, two aspects are addressed: first, the departure of the correction factor from unity indicates the plate influence on the heat transfer performances; second, the current correlations derived for a zero thickness plate can still be used and then corrected with the correction factor as indicated here. Results related to the correction factor dynamics indicated that the two methods agree within 5% under steady-state conditions and within 2.6% for the transients, with higher differences for ramps of higher durations.

The numerical solutions indicated the magnitude of the impact plate influence on the heat transfer performances as compared to the zero thickness plate. For example, for a ramp change of  $D = 3000$  (3 s), the correction factor had a time-average of about 0.59, meaning 41% less heat transferred to the fluid during the transient regime. This result suggests that for frequent changes (on-off regimes) in the applied temperature, or for longer ramps, the neglecting of the plate influence could lead to great errors in the engineering application design or operation.

Despite the fact that the two solution methods used here lead to similar results,

it is obvious that the integral method is more suitable for dynamical analysis, while the FLUENT code should be preferred for more precise engineering calculations.

## LIST OF SYMBOLS

- $A = a/a_p$  – ratio of fluid and plate diffusivities  
 $C$  – proportionality factor [eq. (6)] [ $\text{m}^{1/2}$ ]  
 $E$  – plate thickness [m]  
 $E_* = E/C^2$  – dimensionless plate thickness  
 $CF = q_s(E \neq 0)/q_s(E = 0)$  – correction factor  
 $D$  – ramp duration [eq. (11)]  
 $h$  – convective heat transfer coefficient [ $\text{W/mK}$ ]  
 $k$  – thermal conductivity [ $\text{W/mK}$ ]  
 $Nu_x = hx/k$  – local Nusselt number  
 $q_s$  – contact surface heat flux ( $y = 0$ ) [ $\text{W/m}^2$ ]  
 $T$  – fluid temperature in the boundary layer [K]  
 $T_p$  – plate temperature [K]  
 $T_\infty$  – freestream fluid temperature [K]  
 $T_0$  – bottom plate surface temperature [K]  
 $u_* = u/U_\infty$  – dimensionless fluid velocity  
 $U_\infty$  – freestream fluid velocity [m/s]  
 $x^* = x/C^2$ , dimensionless coordinate  
 $y^* = y/E$ , dimensionless coordinate  
*Greek symbols:*  
 $\beta = y/\delta_t$  – dimensionless coordinate [eq. (2)]  
 $\delta$  – velocity boundary layer thickness [m]  
 $\delta_t$  – thermal boundary layer thickness [m]  
 $\eta = y/\delta$  – dimensionless coordinate [eq. (7)]  
 $\Delta = \delta_t/\delta$  – ratio of boundary layers thicknesses  
 $\theta = (T - T_\infty)/(T_0^{ss} - T_\infty)$  – non-dimensional temperature in the thermal boundary layer  
 $\theta_0 = (T - T_\infty)/(T_0^{ss} - T_\infty)$  – instantaneous plate temperature at its bottom surface ( $y = -E$ )  
 $\theta_s = (T_s - T_\infty)/(T_0^{ss} - T_\infty)$  – dimensionless temperature at the interface ( $y = 0$ )  
 $\theta_p = (T_p - T_\infty)/(T_0^{ss} - T_\infty)$  – dimensionless plate temperature  
 $\Lambda = k/k_p$  – ratio of fluid and plate thermal conductivities  
 $\tau = t \cdot \nu/C^4$  – dimensionless time  
 $\tau_{f1}$  – plate penetration dimensionless time  
*Indices:*  
 $ss$  – steady state  
 $p$  – relative to plate  
 $s$  – relative to the contact surface ( $y = 0$ )  
 $i$  – relative to the plate penetration time

## References

- [1] Mladin, E.C., Radulescu M., *Thermal Convection in Parallel Flows over a Finite Thickness Plate and Ramp Temperature Change*, Proceedings of BIRAC'02, 2002, 115–124.
- [2] Mladin, E.C., Lachi, M., Padet, J., *Transfert de chaleur couplé conduction-convection en régime instationnaire, induit par une température imposée sur une plaque d'épaisseur finie*, Congrès Français de Thermique, SFT Nantes, 2001, 87–92.
- [3] Patankar, V.S., *Numerical Heat Transfer and Fluid Flow*, Hemisphere Publishing Corp. New York (1980).
- [4] Mladin, E.C., Padet, J., *Unsteady planar stagnation flow on a heated plate*, Int. J. of Thermal Sciences (Révue Générale de Thermique), Vol. **40**, No. 7 (2001) 638–64.
- [5] Mladin, E.C., Radulescu, M., Rebay, M., Padet, J., *Transient thermal convection in a parallel flow over a finite thickness plate and an isothermal surface*, Travaux du Colloque Franco-Roumain COFRET'02, Bucarest, April 25–27, 2002, pp. 326–334.
- [6] Özisik, M.N., *Heat Conduction*, 2<sup>nd</sup> Edition, John Wiley & Sons, Inc., New York, 1993.
- [7] Leca, A., Mladin, E.C., Stan, M., 1998, *Transfer de căldură și masă*, Ed. Tehnică, București.
- [8] Padet, J., *Principes des transferts convectifs*, Ed. Polytechnica, Paris, 1998.
- [9] Sucec, J., *Heat Transfer*, W.M.C. Brown Publishers, Iowa, 1985.



## Mathematical Modeling of the Dynamic Crack Propagation in a Double Cantilever Beam

Eduard-Marius Craciun<sup>\*†</sup>, Tudor Udrescu<sup>\*</sup>, George Cîrlig<sup>\*</sup>

We consider a rapid crack propagation along its line in an elastic body subject to a plane strain loading with constant velocity. We present the path independence of the  $J$ -integral in the dynamic case. The velocity and acceleration of a crack in a double cantilevers beam (DCB) are studied using analytical and numerical methods.

### 1. Introduction

Basic solutions of the elastodynamic crack fields are presented in the second section. For a rapid crack propagation in an elastic body subjected to a plane strain loading we obtain the representation of the stress and displacement fields in terms of two displacement potentials. Using the polar coordinates we obtain for the Mode I of classical fracture crack the tip stress fields.

In the third Section we present the energetical aspects of dynamic crack propagation.

In the last part of the paper we study the behavior of the velocity and acceleration of a right crack in a DCB made by a non-linear material and we show the representation of the velocities and of the accelerations, versus the ratio of initial and current lengths crack and versus the coefficient  $\beta$  of exponent of the strain from the stress-strain relation of two nonlinear material.

---

<sup>\*</sup> “Ovidius” University of Constanța, Romania.

<sup>†</sup> e-mail: [mcraciun@univ-ovidius.ro](mailto:mcraciun@univ-ovidius.ro)

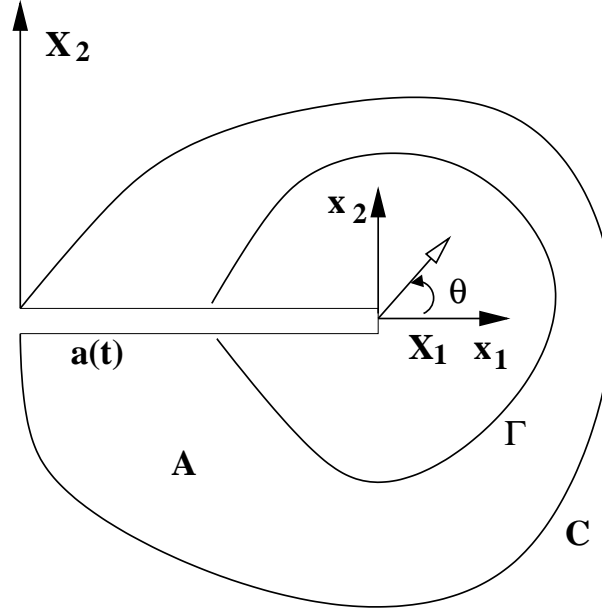


Fig. 1. Coordinate axes for a rapidly crack propagation. The  $X_1X_2$ -axes are fixed in space and  $x_1x_2$  axes are attached to the crack tip.

## 2. Basic solutions of the elastodynamic crack fields

In the dynamic problems, the equation of equilibrium are replaced by the equation of motion. In the absence of body forces we have

$$\sigma_{ij,j} - \rho \ddot{u}_i = 0, \quad i, j = \overline{1, 3}, \quad (1)$$

where  $\sigma_{ij}$  and  $u_i$  are the components of stress and displacement tensor,  $x_j$  represents the moving rectangular coordinates and dot indicates a time derivative.

For a linear elastic material we have the following equation in terms of displacements and elastic constants:

$$\mu \frac{\partial^2 u_i}{\partial x_j^2} + (\lambda + \mu) \frac{\partial^2 u_j}{\partial x_i \partial x_j} = \rho \ddot{u}_i, \quad (2)$$

where  $\mu$  and  $\lambda$  are *Lamé* constants.

We consider a rapid crack propagation in a body subjected to a plane strain loading, with constant velocity  $V$  in the  $X_1$  direction as in Figure 1.

We have

$$x_1 = X_1 - a(t), \quad x_2 = X_2, \quad a(t) = Vt. \quad (3)$$

Introducing two displacement potentials  $\Psi_1$  and  $\Psi_2$  defined by

$$u_1 = \frac{\partial \Psi_1}{\partial X_1} + \frac{\partial \Psi_2}{\partial X_2}, \quad u_2 = \frac{\partial \Psi_1}{\partial X_2} - \frac{\partial \Psi_2}{\partial X_1} \quad (4)$$

and introducing in Eq. (2) we get

$$\frac{\partial^2 \Psi_1}{\partial X_1^2} + \frac{\partial^2 \Psi_1}{\partial X_2^2} = \frac{1}{c_1^2} \ddot{\Psi}_1, \quad \frac{\partial^2 \Psi_2}{\partial X_1^2} + \frac{\partial^2 \Psi_2}{\partial X_2^2} = \frac{1}{c_2^2} \ddot{\Psi}_2. \quad (5)$$

The wave speeds, for plane strain, are given by

$$c_1^2 = \frac{\lambda + \mu}{\rho}, \quad c_2 = \frac{\mu}{\rho}. \quad (6)$$

We have the following representation of the stress field in terms of displacement potentials (see [1]):

$$\begin{aligned} \sigma_{X_1 X_1} + \sigma_{X_2 X_2} &= 2(\lambda + \mu) \left( \frac{\partial^2 \Psi_1}{\partial X_1^2} + \frac{\partial^2 \Psi_1}{\partial X_2^2} \right), \\ \sigma_{X_1 X_1} - \sigma_{X_2 X_2} &= 2\mu \left( \frac{\partial^2 \Psi_1}{\partial X_1^2} - \frac{\partial^2 \Psi_1}{\partial X_2^2} + 2 \frac{\partial^2 \Psi_2}{\partial X_1 \partial X_2} \right), \\ \sigma_{X_1 X_2} &= \mu \left( \frac{\partial^2 \Psi_2}{\partial X_2^2} - \frac{\partial^2 \Psi_2}{\partial X_1^2} + 2 \frac{\partial^2 \Psi_1}{\partial X_1 \partial X_2} \right). \end{aligned} \quad (7)$$

Taking into account (3) the rate of change of each wave potential can be written as:

$$\frac{d\Psi_i}{dt} = \frac{\partial \Psi_i}{\partial t} - V \frac{\partial \Psi_i}{\partial x_1}, \quad i = 1, 2. \quad (8)$$

Differentiating Eq. (8) and introducing the second order derivatives of  $\ddot{\Psi}_i$ ,  $i = 1, 2$  in Eqs. (5) we get the following governing equations:

$$\begin{aligned} \beta_1^2 \frac{\partial^2 \Psi_1}{\partial x_1^2} + \frac{\partial^2 \Psi_1}{\partial x_2^2} &= 0, \\ \beta_2 \frac{\partial^2 \Psi_2}{\partial x_1^2} + \frac{\partial^2 \Psi_2}{\partial x_2^2} &= 0, \end{aligned} \quad (9)$$

where

$$\beta_1^2 = 1 - \left( \frac{V}{c_1} \right)^2, \quad \beta_2^2 = 1 - \left( \frac{V}{c_2} \right)^2. \quad (10)$$

We try to determine the wave potentials as real and respectively imaginary parts of two unknown complex functions  $F(z_1)$  and, respectively,  $G(z_2)$ ,

$$z_1 = x_1 + ix_2^1, \quad z_2 = x_1 + ix_2^2. \quad (11)$$

Expressing the boundary conditions for a stationary crack in mode I of classical fracture  $\sigma_{x_2 x_2} = \sigma_{x_1 x_2} = 0$ , in terms of second derivatives of the unknown functions  $F$  and  $G$  for  $x_2 = 0$  and  $x_1 < 0$  we have

$$\begin{aligned}
(1 + \beta_2^2) [(F''(x)^+ + (F''(x)^-)] + 2\beta_2 [(G''(x)^+ + (G''(x)^-)] &= 0, \\
2\beta_1 [(F''(x)^+ - (F''(x)^-)] + (1 + \beta_1^2) [(G''(x)^+ - (G''(x)^-)] &= 0,
\end{aligned} \tag{12}$$

where  $^+$  and  $^-$  correspond to upper and lower faces of the crack, respectively.

Adding and subtracting Eqs. (12) we obtain two Riemann-Hilbert problems (see [1], [2]), with the solutions:

$$F''(z_1) = \frac{C}{\sqrt{z_1}} G''(z_2) = \frac{-2\beta_2 C}{(1 + \beta_2^2)\sqrt{z_2}}, \tag{13}$$

where  $C$  is a constant.

Making the substitution  $z_1 = r_1 e^{i\theta_1}$ ,  $z_2 = r_2 e^{i\theta_2}$  we obtain the following expression for the Mode I crack tip stress fields:

$$\begin{aligned}
\sigma_{x_1 x_1} &= \frac{K_1(t)}{\sqrt{2\pi r}} \frac{1 + \beta_2^2}{D(t)} \left[ (1 + 2\beta_1^2 - \beta_2^2) \sqrt{\frac{r}{r_1}} \cos \frac{\theta_1}{2} - \frac{4\beta_1 \beta_2}{1 + \beta_2^2} \sqrt{\frac{r}{r_2}} \cos \frac{\theta_2}{2} \right], \\
\sigma_{x_2 x_2} &= \frac{K_1(t)}{\sqrt{2\pi r}} \frac{1 + \beta_2^2}{D(t)} \left[ -(1 + \beta_2^2) \sqrt{\frac{r}{r_1}} \cos \frac{\theta_1}{2} + \frac{4\beta_1 \beta_2}{1 + \beta_2^2} \sqrt{\frac{r}{r_2}} \cos \frac{\theta_2}{2} \right], \\
\sigma_{x_1 x_2} &= \frac{K_1(t)}{\sqrt{2\pi r}} \frac{2\beta_1 (1 + \beta_1^2)}{D(t)} \left[ \sqrt{\frac{r}{r_1}} \sin \frac{\theta_1}{2} - \sqrt{\frac{r}{r_2}} \sin \frac{\theta_2}{2} \right],
\end{aligned} \tag{14}$$

where

$$D(t) = 4\beta_1 \beta_2 - (1 + \beta_2^2)^2.$$

### 3. Crack tip energy release rate

We consider the case of a crack in a two-dimensional body where the crack is propagating along the  $x_1$ -axis and the origin is attached to the crack tip (see Figure 1). Let us consider a contour  $\mathcal{C}$  that contains the propagating crack and bounds an area  $\mathcal{A}$ . The crack tip is surrounded by a small contour,  $\Gamma$ , that is fixed in size and moves with the crack.

From the equation of motion (1) making the inner product of both sides with displacement rate,  $\dot{u}_i$ , we obtain (see [3])

$$\frac{\partial(\sigma_{ij} \dot{u}_i)}{\partial x_j} = \rho \ddot{u}_i + \sigma_{ji} \frac{\partial \dot{u}_i}{\partial x_j} = \dot{T} + \dot{\omega}, \tag{15}$$

where  $T$  is kinetic energy and  $\omega$  represents stress work density and are given by

$$T = \frac{1}{2} \rho \frac{\partial u_i}{\partial t} \frac{\partial u_i}{\partial t}, \quad w = \int_0^{\varepsilon_{ij}} \sigma_{ij} d\varepsilon_{ij}. \tag{16}$$



Integrating the general balance law (15) over an arbitrary area  $\mathcal{A}$  and applying the convergence Green theorem we get the following balance law:

$$\int_C \sigma_{ij} n_j \dot{u}_i ds = \frac{d}{dt} \int \int_A \rho E dA + \frac{1}{2} \frac{d}{dt} \int \int_\Gamma \rho \dot{u}_i \dot{u}_i dA + VG \quad (17)$$

with  $\mathbf{n}$  the outward unit normal vector of curve  $\mathcal{C}$  and  $G$  having the following form (see [13])

$$G = \int_C \left( \omega + \frac{1}{2} \rho V^2 \frac{\partial u_i}{\partial x_i} \frac{\partial u_i}{\partial x_i} \right) dx_2 - \int_C T_i \frac{\partial u_i}{\partial x_i} ds. \quad (18)$$

The first term in Eq. (17) represents the work traction across  $\mathcal{C}$ ; the first and second terms from the right side are the rate of increase of internal and kinetic energies stored inside the region  $\mathcal{A}$  and the third term represent the energy dissipated by the moving crack.

Taking into account that the displacement rate can be written as

$$\dot{u}_i = -V \frac{\partial u_i}{\partial x_1} + \frac{\partial u_i}{\partial t}$$

and under steady state condition the second term in above equation vanishes. Close to the crack tip, we have in a small contour  $\Gamma$ , for the energy release rate defined as

$$J = \frac{G}{V} \quad (20)$$

the following expression

$$J = \lim_{\Gamma \rightarrow 0} \int_\Gamma \left[ (\omega + T) dx_2 - \sigma_{ij} n_j \frac{\partial u_i}{\partial x_i} d\Gamma \right] \quad (21)$$

known in the literature as  $J$ -integral for the dynamic case. For the quasi-static case Rice [4] showed that the corresponding  $J$ -integral is path independent. In dynamic case when the crack propagation is steady-state, *i.e.*  $\frac{\partial u_i}{\partial t} = 0$ , Eq. (21) is also path independent.

#### 4. Crack velocity and acceleration in a DCB

We consider a double cantilever beam (DCB) of height  $2h$  with a crack of length  $a$ , as in Figure 2, made by a nonlinear material characterized by the equation

$$\sigma = \alpha \varepsilon^{\frac{\beta+1}{2}},$$

where  $\alpha$  measures the stiffness of material (see [3], [5]). The DCB is subjected to an end load  $P$ , that remains constant during rapid crack propagation. Let  $a_0$  denote the initial crack length and  $P_c$  the load at crack propagation.

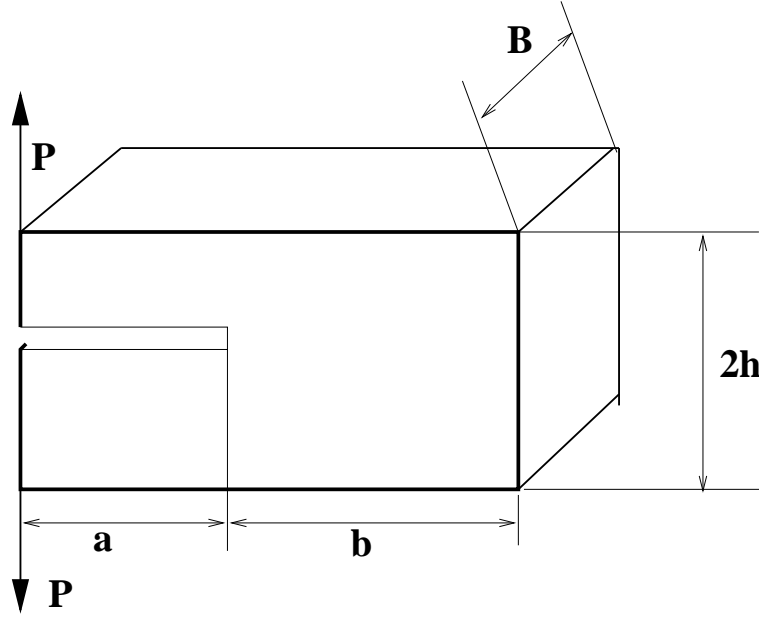


Fig. 2. A double cantilever beam subjected to a load  $P$ .

The energy balance equation during crack propagation takes the form

$$P_c(u - u_c) = U(a) - U(a_0) + K + \gamma(a - a_0). \quad (22)$$

The left side of Eq. (22) represents the work supplied to the system during the growth of the crack from its initial length  $a_0$  to the length  $a$ . The right side is composed by the change of strain energy  $U(a) - U(a_0)$ , kinetic energy  $K$  and the change of surface energy  $\gamma(a - a_0)$ .

From the beam theory we have for the stress  $\sigma$ , strain  $\varepsilon$  and deflection  $x_2 = x_2(x_1)$  at position  $x_1$  of DCB the following expressions:

$$\sigma = \frac{Px_1x_2}{I} (x_2^2)^{\frac{\beta-1}{2}}, \quad \varepsilon = x_2 \left( \frac{\alpha I}{Px_1} \right)^{-\frac{1}{\beta}}, \quad (23)$$

$$x_2(x_1) = \frac{\beta+1}{\beta} \left( \frac{P_c}{\alpha I} \right)^{\frac{1}{\beta}} \left[ \frac{2\beta}{\beta+1} x_1^{\frac{2\beta+1}{\beta}} - a^{\frac{\beta+1}{\beta}} x_1 + \frac{\beta+1}{2\beta+1} a^{\frac{2\beta+1}{\beta}} \right], \quad (24)$$

where

$$I = 2 \int_0^{\frac{h}{2}} x_2^{\beta+1} dx_2. \quad (25)$$

The strain energy of the beam is given by

$$U(a) = \int_V \left( \int_0^\varepsilon \sigma d\varepsilon \right) dV. \quad (26)$$

Using the values of  $\sigma, \varepsilon$  and  $U = x_2(0)$  from Eqs. (23)–(24) we get

$$U(a) = \frac{P_c u}{\beta + 1}. \quad (27)$$

Taking into account Eq. (24) in the expression of the kinetic energy

$$K = \frac{1}{2} \int_0^a \rho h \left( \frac{dx_2}{dt} \right) dx_1, \quad (28)$$

we obtain

$$K = \frac{1}{6} \rho h \left( \frac{P_c}{\alpha I} \right)^{2\beta} a^{\frac{2+3\beta}{\beta}} V^2, \quad (29)$$

with  $V = \frac{da}{dt}$  representing crack velocity. Taking into account Eqs. (24), (27) and (29) from the energy balance equation we obtain for crack speed during crack propagation

$$V^2 = \frac{6\beta^2}{(2\beta+1)(\beta+1)} \frac{(\alpha I)^{\frac{1}{\beta}}}{\rho h} P_c^{\frac{\beta-1}{\beta}} a^{-\frac{\beta+1}{\beta}} \left[ 1 - \left( \frac{a_0}{a} \right)^{\frac{2\beta+1}{\beta}} - n \left( 1 - \frac{a_0}{a} \right) \left( \frac{a_0}{a} \right)^{\frac{\beta+1}{\beta}} \right] \quad (30)$$

and by differentiation of Eq. (30) with respect to time we obtain the crack acceleration  $a_c = \frac{dV}{dt}$ ,

$$\begin{aligned} a_c = & \frac{3\beta^2}{(2\beta+1)(\beta+1)} \frac{(\alpha I)^{\frac{1}{\beta}}}{\rho h} P_c^{\frac{\beta-1}{\beta}} a^{-\frac{2\beta+1}{\beta}} \left[ (1-n) \frac{3\beta+2}{\beta} \left( \frac{a_0}{a} \right)^{\frac{2\beta+1}{\beta}} + \right. \\ & \left. + \frac{\beta+1}{\beta} \left[ 2n \left( \frac{a_0}{a} \right)^{\frac{\beta+1}{\beta}} - 1 \right] \right], \end{aligned} \quad (31)$$

where

$$na_0^{\frac{\beta+1}{\beta}} = \frac{(\beta+1)(2\beta+1)}{\beta^2} \gamma (\alpha I)^{\frac{1}{\beta}} (P_c)^{-\frac{1+\beta}{\beta}}. \quad (32)$$

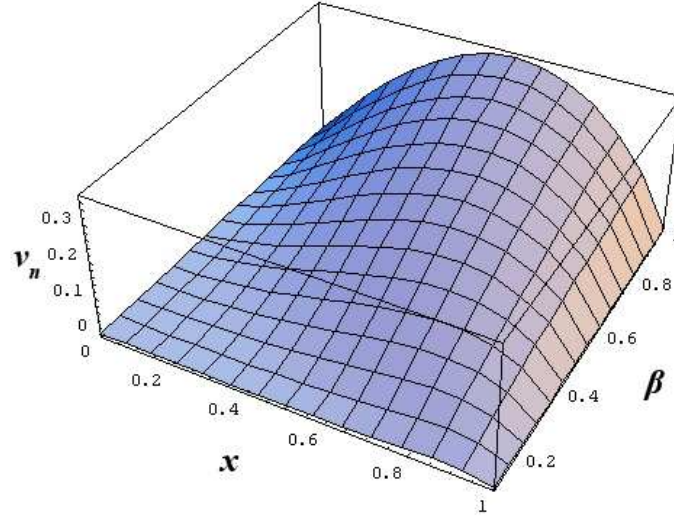
The initial acceleration of the crack  $a_c = a_c(a_0)$  is positive and we obtain the following restriction between  $n$  and  $\beta$ :

$$n < \frac{2\beta+1}{\beta}. \quad (33)$$

Numerical results using Mathematica for  $\frac{a_0}{a} \in [0, 1]$ ,  $\beta \in [0, 1]$  and  $n = n_1 \frac{2\beta+1}{\beta}$  with  $n_1 = 0.9$  are presented in Figure 3, for the variation of normalized crack speed  $v_n$ ,

$$\begin{aligned} v_n = & \frac{\sqrt{\rho h}}{(\alpha I)^{\frac{1}{2\beta}}} a_0^{\frac{\beta+1}{2\beta}} P_c^{\frac{1-\beta}{2\beta}} v = \\ = & \frac{6\beta^2}{(2\beta+1)(\beta+1)} \left( \frac{a_0}{a} \right)^{\frac{\beta+1}{2\beta}} \left[ 1 - \left( \frac{a_0}{a} \right)^{\frac{2\beta+1}{\beta}} - n \left( 1 - \frac{a_0}{a} \right) \left( \frac{a_0}{a} \right)^{\frac{\beta+1}{\beta}} \right]. \end{aligned} \quad (34)$$

We conclude that crack speed increases during the crack propagation from zero at

Fig. 3. 3D Plot for  $n_1 = 0.9$ .

$\frac{a_0}{a} = 1$ , and reaches a maximum for  $\frac{a_0}{a} > 0.5$ .

In Figure 4, we present the variation of normalized crack acceleration  $a_n$ ,

$$a_n = \frac{\sqrt{\rho h}}{(\alpha I)^{\frac{1}{2\beta}}} a_0^{\frac{2\beta+1}{\beta}} P_c^{\frac{1-\beta}{\beta}} a_c = \frac{3\beta^2}{(2\beta+1)(\beta+1)} \left(\frac{a_0}{a}\right)^{\frac{2\beta+1}{\beta}} \cdot \left\{ (1-n) \left(\frac{3\beta+2}{\beta}\right) \left(\frac{a_0}{a}\right)^{\frac{2\beta+1}{\beta}} + \frac{\beta+1}{\beta} \left[ 2n \left(\frac{a_0}{a}\right)^{\frac{\beta+1}{\beta}} - 1 \right] \right\}. \quad (35)$$

The crack first accelerates from zero at  $\frac{a_0}{a} = 1$ , reaches a maximum for  $\frac{a_0}{a} \in (0.8, 0.9)$ , then decelerates before coming to a complete stop at  $\frac{a_0}{a} \rightarrow 0$ .

We conclude that the crack propagation is slowing as  $\beta$  decreases, *i.e.* the material becomes stiffer with decreasing  $\beta$ .

## 5. Final remarks

In this paper we studied the dynamic crack propagation of an elastic body subjected to a constant velocity. Basic solution of elastodynamic crack fields are presented. Mode I stress fields in a neighborhood of the crack tip and path independence of  $J$ -integral in the dynamic case were obtained.

Analytical results were obtained for velocity and acceleration of the crack tip with particular case of a DCB made by nonlinear material.

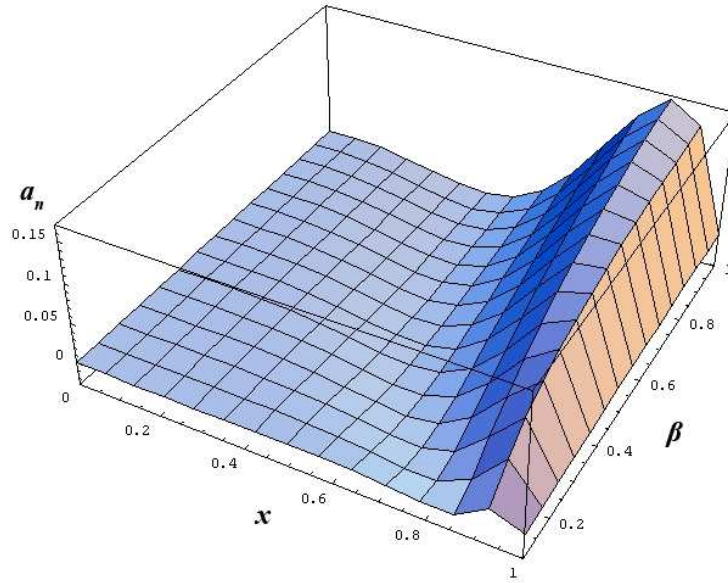


Fig. 4. 3D Plot of acceleration for  $n_1 = 0.9$ .

Numerical representation using Mathematica for different ratios of the initial and current velocities and for different values of the coefficient  $\beta$  from the exponent of the stress from stress-strain relation allowed us to make an image of the crack tip. We obtained that the crack speed and acceleration increase during crack growth to a maximum value and then decrease before reaching a complete stop.

**Acknowledgements.** The authors gratefully acknowledge the support provided for this research by the Romanian Academy under GAR 96, Contract 13/2005.

## References

- [1] T.L. Anderson, *Fracture Mechanics*, Boca Raton, FL: CRC Press, 1975.
- [2] N.I. Muskhelishvili, *Some Basic Problems of Mathematical Theory of Elasticity*, Nordhoff, Germany, 1957.
- [3] E. Gdoutos, *Fracture Mechanics, An Introduction*, Kluwer Academic Publishers, 1993.
- [4] J. Rice, *A path independent integral and the approximate analysis of strain concentration by notches and cracks*, J. of Appl. Mech., **35**, pp. 379–386, 1968.
- [5] G.C. Sih, Dynamical aspects of crack propagation, in *Inelastic Behavior of Solids*, Mc Graw-Hill, pp. 607–639, 1970.



## Noise prediction model for wind turbines

Alexandru Dumitrache<sup>\*†‡</sup> and Horia Dumitrescu<sup>\*‡</sup>

Aerodynamic noise is generated when the rotor encounters smooth flow. It contains airfoil self-noise and turbulence inflow noise. The present semiempirical model is coupled with CFD and aerodynamic calculation so as to improve the accuracy of the prediction model. By doing CFD computations, boundary layer parameters for some relevant airfoil profiles are stored as a database which is used directly for the noise prediction model. The total noise spectrum for a given wind turbine is compared with experiment and encouraging result is obtained.

### 1. Introduction

The future of using wind energy is so exciting that it has many advantages such as: no air pollution, no need of any fossil or nuclear resources during operation. But with the fast increase of wind energy development, it also gives rise to the problems concerning public acceptance of wind energy. The noise and visual impact are the main drawbacks. The noise from wind turbines may annoy people who live around. Therefore it's necessary to estimate the noise level in the field where the wind turbines are installed. Generally, the noise from wind turbine is composed with mechanical noise and aerodynamic noise. The mechanical noise is caused by the different operating machine elements which can be reduced efficiently by many engineering methods and will certainly not reduce the power output. However, the way to reduce aerodynamic noise from wind turbines should be studied together with power efficiency. A fully established method to predict the wind turbine noise is still

---

<sup>\*</sup> “Gheorghe Mihoc–Caius Iacob” Institute of Statistical Mathematics and Applied Mathematics, Bucharest, Romania.

<sup>†</sup> e-mail: [dalex@gheorghita.ima.ro](mailto:dalex@gheorghita.ima.ro)

<sup>‡</sup> Partially supported by Research Contract 149/2004.

limited. For engineering purpose of use, several semi-experimental noise prediction models are available. Some of the models are originally developed for application on helicopter and aircraft wings. One of the first model was carried out by Grosveld [1] in 1985. In 1981, Viterna [2] applied a method to the low-frequency noise estimation from a wind turbine. Brooks, Pope and Marcolini [3] performed a set of experiments for NACA0012 airfoil sections. However, the aerodynamic and acoustic measurements were only based on NACA0012 airfoil which may not be suitable for other airfoil profiles. Therefore, the boundary layer parameters at trailing edge should be calculated instead of using experimental data from NACA0012 airfoil. Unweighted, or linear-weighting, was used in the presentation of data, while the standard regulations, typically used in measuring the acoustic emitting from wind turbines, specify A-weighting, which de-emphasizes frequencies below 1 000 Hz and correlates extremely well with human subjective response.

## 2. Noise mechanisms of wind turbines

Nowadays, the size of wind turbine and the capacity of wind farms are becoming larger. The noise generated from wind turbines is considerably higher than before. A large amount of effort has to be gone into reducing noise emission from wind turbines to make wind energy the really green energy.

The noise is generated from the wind turbine blades, gearbox and generator. There are two potential types of noise from wind turbines *mechanical* and *aerodynamic noise*. Mechanical noise comes from the metal components moving or knocking against each other. Aerodynamic noise is caused by the blade passing through air.

The noises caused by wind turbine mechanical components are of tonal property, e.g. noise from the meshing gears. Due to the better engineering practices, mechanical noise of modern wind turbine has been dropped to a very low level and is not the main problem any longer.

### 2.1. Aerodynamic noise

The causes of aerodynamic noise are mainly divided into three types:

- Low-frequency noise,
- Turbulent inflow noise,
- Airfoil self-noise [10].

**Low-frequency noise** from wind turbine is originated by the changes of the wind speed experienced by the blades due to the presence of the tower and the wind shear. It may excite vibrations for buildings around the wind turbines.

**Turbulent inflow noise.** The broadband noise from wind turbines also depends on the turbulent inflow characteristics, especially in the case of large wind turbines. The inflow turbulence creates broadband noise which are perceived by the observers as swishing noise.

**Turbulent boundary layer trailing edge (TBL-TE) noise.** Trailing edge noise has been long recognized as another major source of airfoil self-noise. It is



generated as the blade-attached turbulent boundary layer converts into the wake at the airfoil trailing edge. The noises on suction side and on pressure side are major noise source at low angle of attacks. However, at high angle of attacks the boundary layer separation occurs which generates separated-flow-noise.

The trailing edge noise is also broadband.

**Laminar boundary layer vortex shedding (LBL-VS) noise.** A wind turbine can be operated under a Reynolds range from  $10^5$  to  $10^6$  because of the changes of relative wind speed and the chord length at different blade radius [14]. Therefore, the flow conditions on each blade sections are different from each other. If the laminar boundary layer exist on one or both sides of the airfoil and cover the most of the airfoil surface, a resonant interaction between the unsteady laminar-turbulent transition with the trailing-edge noise will occur. This is termed as laminar boundary layer vortex shedding noise.

It is a tonal noise that can be significant at certain operating conditions for modern wind turbines which have low angular speed and large blade radius.

**Tip vortex formation noise.** Evidence exists that the flow around the tip is three dimensional, thus the pressure difference between the suction and the pressure sides result in a rotational flow region over the airfoil. This flow is described as a vortex with a thick viscous turbulent core. The interaction between the tip vortex and the trailing edge has the same manner as the interaction between turbulent boundary layer and the trailing edge. Tip noise level strongly depends on the geometry details of the blade tip.

**Trailing edge bluntness vortex shedding (TEB-VS) noise.** The TEB-VS noise results by the vortex shedding from the blunted trailing edges. The noise level varies with the bluntness thicknesses at each blade sections especially near the tip of the blade.

### 3. Noise prediction model for wind turbines

Numerical simulation of far-field sound on a large computational domain is expensive and very difficult even for simple flow conditions. Some features of wind turbine noise are of considerable importance in subjective response for wind turbine noise. Therefore available theories require reinterpretation for application to predict wind turbine noise. The well-know Lighthill's acoustic theory is the theoretical basis for most of the prediction models. The self-noise prediction model introduced in this section is based on the experimental work of Brooks, Pope, and Marcolini [3]. The turbulent-inflow noise model is based on Amiet [7].

**Turbulence characteristic.** Prediction of the turbulence characteristic is crucial to predict the noise. The turbulence intensity and the length scale are depended on the evaluate height above the ground and also the meteorological conditions at the given site. The height of the wind turbine above the ground is fixed, thus the turbulence might be considered isotropic which indicates that the fluctuations are

approximately the same in all directions. The mathematical description is given as:

$$w = \bar{w}e^{i\omega_z(t-z/V_0)}, \quad (1)$$

where  $w$  is the turbulence velocity,  $z$  is the down-stream direction,  $\omega_z$  is the longitudinal frequency and  $V_0$  is the mean free stream wind velocity. This simplification is valid for horizontal axis wind turbines. The longitudinal turbulence is the most important component and it is assumed to be a horizontal sinusoidal gust of the form as equation 1. The mean square turbulence fluctuation at the height  $h$  is given by [8].

A method of computing the intensity is introduced here which is used in the present noise prediction model. The mean wind speed varies with height and it's often described with the power law relationship

$$V_z = V_{ref}(Z/Z_{ref})^\gamma, \quad (2)$$

where  $\gamma$  is the power law factor which gives the amount of the shear:

$$\gamma = 0.24 + 0.096 \log_{10} z_0 + 0.016(\log_{10} z_0)^2, \quad (3)$$

Turbulence intensity can be found using the relationship:

$$\bar{w}/\bar{V} = \gamma[\ln(30/z_0)]/[\ln(z/z_0)], \quad (4)$$

To characterize the turbulence of wind, the turbulence length scales also play an important role. The turbulence length scale is the measurement of averaged size of a gust in a certain directions which is used to determine how rapidly the gust properties vary in space. The turbulence length scale, given by ESDU [25], is formulated as following:

$$L_{ESDU} = 25z^{0.35}z_0^{-0.063}. \quad (5)$$

**Inflow noise prediction.** The adopted prediction model for turbulence inflow-noise in this paper is based on the model on Amiet [7]. A semi-empirical model was given which was valid against wind tunnel measurements with a single airfoil section under turbulent inflow. For case of rotating wind turbines, a corrected model was given by Lowson [10]. The model from Amiet can be used for each blade segments along the blade span. For both high and low frequency regions, Lowson shown a model with smooth transition between the two regions:

$$L_{p,INF} = L_{p,INF}^H + 10 \log_{10}(K_c/(1 + K_c)), \quad (6)$$

where  $L_{p,INF}^H$  is the sound pressure level for high frequency region:

$$L_{p,INF}^H = 10 \log_{10}[\rho_0^2 c_0^2 l (\Delta L / r^2 M^3 I^2 \hat{k}^3) (1 + \hat{k}^2)^{-7/3}] + 58.4, \quad (7)$$

where  $l$  denotes the turbulence length scale and  $I$  denotes the turbulence intensity.  $\Delta L$  is the blade segment semi-span. The low frequency correction  $K_c$  in equation (6) is given as  $K_c = 10S^2 M \hat{k}^2 / \beta^2$ , where  $S$  is a function which denote the compressibility of the flow. The formula is suggested by Amiet:

$$S^2 = (2\pi \hat{k} / \beta^2 + (1 + 2.4 \hat{k} / \beta^2)^{-1})^{-1}, \text{ where } \beta^2 = 1 - M^2. \quad (8)$$

The wave number is given by Lowson which is corrected from Amiet:  $\hat{k} = \pi f c / V_{rel}$ .

**Prediction model for TBL-TE noise and separation-stall noise.** The scaling laws for self-noise mechanisms are based on the results of Ffowcs-Williams and Hall [5] which has been mentioned in section 2. The scaling law applied for the TBL-TE noise is basically described by:

$$p^2 \propto \rho_0^2 \bar{w}^2 (U_c^3 / c_0) (\bar{D} \Delta L \ell / r^2), \quad (9)$$

where  $p^2$  is the mean square sound pressure at distance  $r$  from the trailing edge,  $\bar{D}$  is the directivity parameter with the value equals 1 if the observer is normal to the trailing edge surface.  $\ell$  is the correlation factor for the turbulence. It is approximated by [3] that  $\ell \propto \delta^*$  or  $\delta$ .

From this scaling law the total TBL-TE noise together with separation/stall noise spectrum in 1/3-octave band is predicted by [3].

**Prediction model for LBL-VS noise.** The scaling method for LBL-VS noise is similar with that used for TBL-TE noise. The scaling parameters are boundary layer parameters, Mach number, angle of attack and Reynolds number. The boundary layer thickness at trailing edge is used for LBL-VS noise prediction instead of using boundary layer displacement thickness.

**Prediction model for tip vortex formation noise.** The flow around the tip is highly turbulent and the tip noise is generated due to the passage of this turbulence over the TE edge into the wake. The TIP noise model is developed by Brooks, Pope and Marcolini [3].

$$SPL_{TIP} = 10 \log(M^2 M_{max}^3 l'^2 \bar{D}_h / r^2) - 30.5 (\log(St'' + 0.3))^2 + 126, \quad (10)$$

where  $l'$  is the spanwise extent of tip vortex at trailing edge.

**Prediction model for TEB-VS noise.** The TEB-VS noise is scaled using the method as TBL-TE noise and LBLVS noise. The sound pressure level frequency and the spectral shapes are modelled as function of angle of attack and the airfoil trailing edge parameters. The model is based on the measurement of airfoil NACA0012.

The noise spectrum are predicted as:

$$SPL_{Blunt} = 10 \log(M^{5.5} \Delta L \bar{D}_h / r^2) + G_4(h / \delta_{arg}^*, \Psi) + G_5(h / \delta_{arg}^*, \Psi, St''' / St_{peak}''') \quad (11)$$

where  $h$  is the bluntness thickness, used for calculating the Strouhal number instead of using boundary layer thickness parameter,  $St''' = fh/U$ ,  $G_4$  is the peak level spectrum and  $G_5$  is modified as the shape of the spectrum.

**Boundary layer thickness calculations.** For each airfoil, the boundary layer displacement thicknesses are calculated at both pressure side and suction side for Reynolds number range  $10^6 \sim 2 \times 10^6$  and for the attack angle range  $-5^\circ \sim 25^\circ$ , by using the 2D airfoil code XFOIL [11]. To model the transition conditions, the  $e^N$  method is used in XFOIL.  $n_{crit} = 9$  corresponds to the standard situation which is assumed to be the untripped case. Calculation of  $n_{crit} = 4$  is also performed which represent the tripped case. Therefore there is a trigger in the noise prediction code to specify whether the flow is tripped or untripped.

**Sound directivity.** The sound directivity at both high and low frequencies are based on the research of Amiet [7]. It can be normalized by the trailing edge noise emitted at the position of  $\Theta = 90^\circ$  and  $\Phi = 90^\circ$ . Therefore, at this position the sound directivity reaches the maximum value of 1 (valid at low frequencies). The low frequency directivity is applied in case of separation stall noise. The reason is that the turbulence eddies are comparable in size with the airfoil chord length and the eddy convection speeds are low.

The low frequency directivity is applied in case of separation stall noise. The reason is that the turbulence eddies are comparable in size with the airfoil chord length and the eddy convection speeds are low.

#### 4. Analysis of the noise prediction model

In our present work, the aerodynamic code BEM [12] is coupled with the aerodynamic noise prediction model. The induced velocities are computed by BEM code using a new tip correction method [12]. Also, for wind turbines operating at certain yaw angle or/and tilt angle, the velocity field is accurately computed using the coordinate transformation matrix.

In this section some results are illustrated using different input parameters (unfortunately, not all figures shown here due to lack of space). Therefore, the model is validated against measurements and also some principle trends are studied. All the analysis are based on the Bonus Combi 300 kW wind turbine since some measurements have already been performed [13].

**The effect by different airfoil contours.** The first calculation is based on the original experimental data from NACA0012 airfoil by Brooks, Pope and Marcolini [3]. The result in general fits the measured curve. In another case, when NACA634xx series airfoils are used instead of the original NACA0012 airfoil, the predicted result agree with measurement data well at low and high frequency range.

The boundary layer parameters at trailing edge play an important role of the total noise radiation level. In general, the prediction for untripped curve has better result at high frequency range and the tripped curve fits well with experiment curve at low frequency range. The measured overall sound power level at wind speed 8 m/s is 99.1 dB(A). The predicted overall sound power level for NACA634xx are: 97.02 dB(A) (untripped boundary layer) and 96.81 dB(A) (tripped boundary layer).

**Effect of the tip pitch angle.** It should be mentioned here that the pitch angle of Bonus Combi 300 kW wind turbine is fixed. The motivation of changing the pitch setting is only from an analysis point of view. The sound power level decreases by increasing the tip pitch angle. The differences between these curves are mainly observed in the intermediate frequency range. The reason is that the effect is only due to the changes of trailing edge noise and the separation-stall noise.

After doing the aerodynamic and noise calculation together, it can be seen in this case that for the present wind turbine the optimized pitch setting is around 0 to  $-1$  degree which produce low noise and normal energy output.

**Effect of sound directivity.** It is known that sound pressure level changes

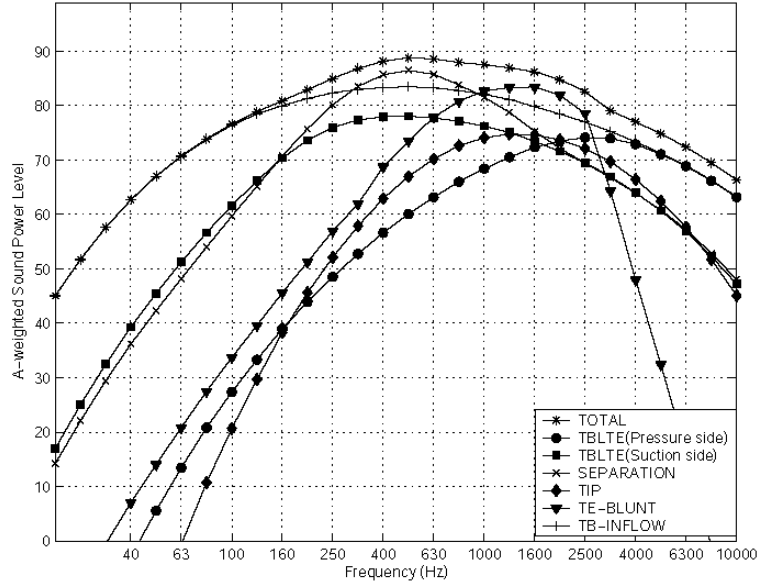


Fig. 1. Noise spectra from all source mechanisms.

with the observer positions by effect of the sound directivity. For a single wind turbine case, we can conclude that the noise radiation property behaves like a dipole sound source. Since the blades are twisted, there are also nonsymmetrical properties. Therefore, there is no doubt that the best location for the residents is in tangential direction of the rotor plane.

Noise from all sources are plotted together in figure 2. The result curves are obtained using noise prediction model together with BEM code. In general the larger the Mach number, the louder the noise radiates from wind turbine. Therefore, the noise level is very sensitive with the rotating speed and the size of the wind turbine.

## 5. Conclusion

In the present prediction model, the mechanical noise from the wind turbine is neglected. The unsteadiness includes turbulence inflow which is time depended, wind shear, pitching and yawing blades and wake dynamics. Boundary layer on every airfoil section is controlled by different Mach number and Reynolds number. The surface roughness of the airfoil could also change the flow condition on the airfoil surface. In general, the laminar boundary layer vortex shedding noise is not included for wind turbines with high tip speed ratio.

In general, the larger the Mach number, the louder the noise radiates from wind turbine. The noise level is very sensitive with the rotating speed and the size of

the wind turbine. Boundary layer thickness is required as the basic input for all the self-noise prediction laws. For the wind turbine designer, the noise prediction model should be always coupled together with a comprehensive aerodynamic computation to make sure that the electric power efficiency is not change too much. For wind turbines to be sited, the prediction model can be helpful to avoid non-necessary annoyance for the inhabitants.

## References

- [1] Grosveld, F.W., *Prediction of Broadband Noise from Horizontal Axis Wind Turbines*, Journal of Propulsion and Power, **1**, 4, 292–299, 1985.
- [2] Viterna, L.A., *The NASA LeRC Wind Turbine Sound Prediction Code*, Journal of Propulsion and Power, NASA CP-2185, pp. 410–418, 1981.
- [3] Brooks, T. F., Pope, D. S., Marcolini, M. A., *Airfoil Self-Noise and Prediction*, NASA Reference Publication 1218, 1989.
- [4] Lighthill, M.J., *On Sound Generated Aerodynamically*, Proceedings of the Royal Society of London, Series A, **211**, pp. 564–587, 1952.
- [5] Ffowcs-Williams, J.E, Hawkings, D.L, *Sound Generation by Turbulence and Surfaces in Arbitrary Motion*, Philosophical Transactions of the Royal Society of London, **264**, A 1151, pp. 321–342, 1969.
- [6] Ciskowsky, R.D., Brebbia, C. A., *Boundary Element Methods in Acoustics*, Computational Mechanics Publications, Elsevier Applied Science, 1991.
- [7] Amiet, R.K., *Acoustic Radiation from an Airfoil in a Turbulent Stream*, Journal of Sound and Vibration, **41**, 4, pp. 407–420, 1975.
- [8] Ferdinand, W. G., *Prediction of Broadband Noise from Horizontal Axis Wind Turbines*, Journal of Propulsion and Power, **1**, 4, pp. 292–299, 1984.
- [9] ESDU-Report, *Characteristics of Atmospheric Turbulence Near the Ground, Part II: single point data for strong winds (neutral atmosphere)*, No. 85020, 1993.
- [10] Lowson, M.V., *Assessment and Prediction Model for Wind Turbine Noise: 1. Basic Aerodynamic and Acoustic Models*, Flow Solution Report 93/06, pp. 1–46, 1993.
- [11] <http://raphael.mit.edu/XFOIL>.
- [12] Mikkelsen, R., Soerensen, J. N., Shen W. Z., *Modelling and analysis of the flow field around a coned rotor*, Wind Energy, **4**, 3, pp. 121–135, 2001.
- [13] Jakobsen, J., Andersen, B., *Aerodynamical Noise from Wind Turbine Generators*, EFP j.nr.1364/89-5, JOUR-CT 90-0107, 1993.
- [14] Dumitrescu, H., Cardos, V., Dumitrache, Al., *Aerodynamics of Wind Turbine*, (in Romanian), Editura Academiei Române, București, 2001.

## Quasimonotone ODE Approximation of Nonlinear Diffusion Process

Stelian Ion<sup>\*‡</sup>

The paper deal with the numerical approximation of a class of nonlinear diffusion process that includes the water flow through porous media. The ordinary differential equations used as approximation of the diffusion equation is obtained applying the finite volume scheme for space derivatives discretization and keeping the continuum time derivative. We prove that in certain conditions concerning the diffusion flux, its numerical approximation and the mesh size one obtains a quasimonotone ODE system. Using the quasimonotone property we investigate some theoretical properties of the approximative solution in the case of Richards' equation.

### 1. Introduction

In this paper we develop a numerical approximation scheme for a class of parabolic nonlinear diffusion equation and we prove a comparison theorem for the approximation models. The mathematical model is given by

$$\begin{cases} \frac{\partial b(u)}{\partial t} - \operatorname{div} \mathbf{a}(b(u), \nabla u) = 0, & t > 0, x \in \Omega, \\ u = u_D, & t > 0, x \in \partial\Omega, \\ u(0, x) = u_0(x), & x \in \Omega, \end{cases} \quad (1)$$

where  $\Omega$  is a domain in  $\mathbb{R}^d$ ;  $\operatorname{div}$  and  $\nabla$  are with  $x \in \mathbb{R}^d$ ; and  $u(t, x)$  is the scalar unknown function. Such models are widely used in soil science, heat transfer theory, Stephen problems, etc. In the heat transfer theory one has  $b(s) = s$ ,  $\mathbf{a}(u, \nabla u) =$

---

<sup>\*</sup> “Gheorghe Mihoc–Caius Iacob” Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: [istelian@ima.ro](mailto:istelian@ima.ro)

<sup>‡</sup> Supported by CERES Grant no. 4-187/2004.

$k(u)\nabla u$  so that evolution of the temperature is governed by the equation

$$\frac{\partial u}{\partial t} - \operatorname{div}(k(u)\nabla u) = 0. \quad (2)$$

The unsaturated water flow through porous media is described by the well known Richards' equations [3]

$$\frac{\partial \theta(h)}{\partial t} - \operatorname{div}(K(\theta)\nabla h + \mathbf{e}_3 K(\theta)) = 0, \quad (3)$$

where  $\theta$  represents the relative volumetric water content,  $h$  represents the pressure,  $K$  is the hydraulic conductivity and  $\mathbf{e}_3$  is the upward vertical versor. In this paper we consider that the equation of the mathematical model (1) takes only one of the form (2) or (3). For more different examples of diffusion-convections operator, containing many other references see Diaz [5].

Let us present the sufficient conditions for the existence of the weak solutions and the comparison principle. We essentially adopt the frame of Alt and Luckhaus in [1].

*Assumptions on constitutive functions, boundary data, initial data and geometry of domain  $\Omega$ :*

**A1**  $b : \mathbb{R} \rightarrow \mathbb{R}$  is a continuous and nondecreasing function.

The empirical models of soil sciences use the functions  $\theta(h)$  that are monotone and bounded,

$$0 \leq \theta(h) \leq 1.$$

The diffusion flux  $\mathbf{a}$  satisfies:

$$\mathbf{A2} \quad \left\{ \begin{array}{l} (1) \mathbf{a} : \mathbb{R} \times \mathbb{R}^d \text{ is a continuous function in } (u, \xi) \\ (2) |\mathbf{a}(b(u), \xi)| \leq C(1 + B(u)^{(p-1)/p} + |\xi|^{p-1}), \forall u, \forall \xi, \\ (3) (\mathbf{a}(u, \xi) - \mathbf{a}(u, \xi^*)) \cdot (\xi - \xi^*) > c|\xi - \xi^*|^p, \forall \xi \neq \xi^* \in \mathbb{R}^d \end{array} \right.$$

with  $1 < p < \infty$  and

$$B(z) = b(z) \cdot z - \int_0^z b(s) ds.$$

In the Richards' equation model the growth condition (2) is ready satisfied, the hydraulic function used in the soil sciences are upper bounded, but the uniform elliptic conditions (3) is not satisfied by all hydraulic models since the hydraulic conductivity becomes zero as water content approach the residual values, see [3]. Also the continuity condition (1) is satisfied for homogeneous soil but it is not satisfied for layered soils.

**A3**  $u_D \in L^p((0, T) : W^{1,p}(\Omega)) \cap L^\infty((0, T) \times \Omega)$ ,  $\frac{\partial u_D}{\partial t} \in L^1((0, T) : L^\infty(\Omega))$ .

**A4**  $B(u_0) \in L^1(\Omega)$ .

**A5**  $\Omega \in \mathbb{R}^d$  is open, bounded, and connected with Lipschitz boundary.

*The existence of weak solutions [1].*

If the data satisfy **A1**–**A5**, then there exists a weak solution of the Cauchy problem (1). A function  $u(t, x)$  is a weak solution if the following properties are fulfilled:



- 1)  $u - u_D \in L^p((0, T) : W^{1,p}(\Omega))$ ,  
 2)  $b(u) \in L^\infty(0, T : L^1(\Omega))$  and  $\frac{\partial b(u)}{\partial t} \in L^q(0, T : W^{-1,q}(\Omega))$  with initial values  $b(u_0)$ , that is,

$$\int_0^T \left\langle \frac{\partial b(u)}{\partial t}, v \right\rangle dt + \int_0^T \int_\Omega (b(u) - b(u_0)) \frac{\partial v}{\partial t} dx dt = 0 \quad (4)$$

for every  $v \in L^p(0, T : W_0^{1,p}(\Omega)) \cap W^{1,1}(0, T : L^1(\Omega))$ ,  $v(T, \cdot) \equiv 0$ .

- 3)  $\mathbf{a}(u, \nabla u) \in L^q((0, T) \times \Omega)$  and  $u$  satisfies the differential equations, that is,

$$\int_0^T \left\langle \frac{\partial b(u)}{\partial t}, v \right\rangle dt + \int_0^T \int_\Omega \mathbf{a}(u, \nabla u) \cdot \nabla v dx dt = 0 \quad (5)$$

for every test function  $v \in L^p(0, T : W_0^{1,p}(\Omega))$ .

*Pointwise comparison principle*

We present here a comparison result concerning the infiltration problem, the result is a variant of the comparison results obtained in [1] and [5]. The infiltration problem read as

$$\begin{cases} \frac{\partial \theta(h)}{\partial t} - \operatorname{div}(\kappa(h) \nabla h + \mathbf{e}_3 \kappa(h)) = 0, & t > 0, x \in \Omega, \\ h = h_D, & t > 0, x \in \partial\Omega, \\ h(0, x) = h_0(x), & x \in \Omega. \end{cases} \quad (6)$$

Assume that: the boundary data, initial data and the domain  $\Omega$  satisfy the conditions **A3**, **A4** and **A5**, respectively, the function  $\theta(h)$  satisfies the conditions **A1** and the hydraulic conductivity function satisfies:

$$\mathbf{A2'} \quad \left\{ \begin{array}{l} (1) \kappa : \mathbb{R} \rightarrow \mathbb{R}_+, \kappa(h) \geq \eta, \\ (2) |\kappa(h_1) - \kappa(h_2)| < C|h_1 - h_2|^\gamma, \gamma \geq \frac{1}{2}, \forall h_1, h_2 \in \mathbb{R}. \end{array} \right.$$

**Theorem 1** (COMPARISON THEOREM). *Let  $(h_D, h_0)$ ,  $(\widehat{h}_D, \widehat{h}_0)$  be such that  $h_D \leq \widehat{h}_D, h_0 \leq \widehat{h}_0$ . Let  $(h, \widehat{h})$  be two bounded weak solutions of infiltration problem (6) associated to  $(h_D, h_0)$  and  $(\widehat{h}_D, \widehat{h}_0)$ , respectively. Assume, in addition, that*

$$\theta(h)_t, \theta(\widehat{h})_t \in L^1((0, T) \times \Omega).$$

*Then*

$$\text{a) } \theta(h) \leq \theta(\widehat{h}).$$

*If in addition  $\theta(h)$  and  $\kappa(h)$  satisfy*

$$\mathbf{A2''} \quad \text{for any } z_1, z_2 \text{ such that } \theta(z_1) = \theta(z_2) \text{ results } \kappa(z_1) = \kappa(z_2)$$

*then we have*

$$\text{b) } h \leq \widetilde{h} \text{ on } (0, T) \times \Omega.$$

*Proof.* In view of the comparison theorem in the paper [1] the result prove true if one observes that the solution  $h$  corresponding to the datum  $(h_D, h_0)$  is a subsolution of the infiltration problem associate to the datum  $(\widehat{h}_D, \widehat{h}_0)$ . A direct proof is as follow.

For any  $\delta > 0$  let  $\Psi_\delta(\alpha) = \min(1, \max(0, \alpha/\delta))$ . The function  $w = \Psi_\delta(h - \widehat{h})$  belongs to  $L^2(0, T : W_0^{1,2}(\Omega))$  and its gradient is given by

$$\nabla w = \begin{cases} \frac{1}{\delta} (\nabla h - \nabla \widehat{h}), & \text{if } 0 < h - \widehat{h} < \delta, \\ 0, & \text{otherwise.} \end{cases}$$

Set  $w$  as test function in (5). Then

$$\begin{aligned} & \int_0^t \int_{\Omega} (\theta(h)_t - \theta(\widehat{h})_t) w dx dt + \\ & + \underbrace{\frac{1}{\delta} \int_0^t \int_{\Omega_\delta} [\kappa(h) \nabla h - \kappa(\widehat{h}) \nabla \widehat{h} + (\kappa(h) - \kappa(\widehat{h})) \mathbf{e}_3] \cdot \nabla (h - \widehat{h}) dx dt}_I = 0, \end{aligned} \quad (7)$$

where  $\Omega_\delta := \{x | 0 < h - \widehat{h} < \delta\}$ . The integral  $I$  can be rewritten as

$$I = \int_0^t \int_{\Omega_\delta} \kappa(h) \|\nabla(h - \widehat{h})\|^2 dx dt + \int_0^t \int_{\Omega_\delta} (\kappa(h) - \kappa(\widehat{h})) (\nabla \widehat{h} + \mathbf{e}_3) \cdot \nabla (h - \widehat{h}) dx dt.$$

Using Young inequality,  $ab \leq C(\epsilon)p^{-1}a^p + \epsilon q^{-1}b^q$  and **A2'**-(1) we obtain

$$I \geq \left(\eta - \frac{\epsilon}{2}\right) \int_0^t \int_{\Omega_\delta} \|\nabla(h - \widehat{h})\|^2 dx dt - \frac{C(\epsilon)}{2} \int_0^t \int_{\Omega_\delta} (\kappa(h) - \kappa(\widehat{h}))^2 \|\nabla \widehat{h} + \mathbf{e}_3\|^2 dx dt.$$

Taking  $\epsilon < 2\eta$  and using **A2'**-(2) we have

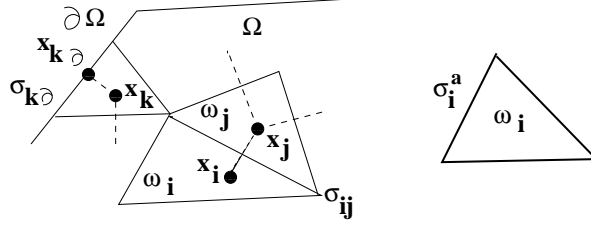
$$\int_0^t \int_{\Omega} (\theta(h)_t - \theta(\widehat{h})_t) w dx dt + \frac{c}{\delta} \int_0^t \int_{\Omega_\delta} \|\nabla(h - \widehat{h})\|^2 dx dt \leq C\delta^{2\gamma-1} \int_0^t \int_{\Omega_\delta} \|\nabla \widehat{h} + \mathbf{e}_3\|^2 dx dt.$$

The term on the right tends to zero as  $\delta \rightarrow 0$ , and the first integral on the left converges to

$$\int_0^t \int_{h > \widehat{h}} (\theta(h) - \theta(\widehat{h}))_t dx dt = \int_0^t \int_{\Omega} \partial_t \max(0, \theta(h) - \theta(\widehat{h})) dx dt$$

so, we have

$$\int_{\Omega} \max(0, \theta(h) - \theta(\widehat{h})) (t, x) dx \leq 0,$$

Fig. 1. Triangulation of polygonal domain in  $\mathbb{R}^2$ .

that led to  $\theta(h) \leq \theta(\tilde{h})$ .

From the **A1** and **A2''** we see that  $\theta(h) = \theta(\tilde{h})$  and  $\kappa(h) = \kappa(\tilde{h})$  in the set  $\{0 < h - \tilde{h}\}$  so that the equality (7) is reducing to

$$\int_0^t \int_{\Omega_\delta} \|\nabla(h - \tilde{h})\|^2 dx dt = 0,$$

which lead to  $\nabla(h - \tilde{h}) = 0$  in  $\{0 < h - \tilde{h}\}$ . Boundary conditions, the continuity of the solutions with respect with space variable and  $\nabla(h - \tilde{h}) = 0$  in  $\{0 < h - \tilde{h}\}$  imply  $h \leq \tilde{h}$ . This end the proof of the comparison theorem.

By considering that for many interesting PDE we known only approximative solutions it is desirable to set up a numerical algorithm which preserves some relevant properties of exact models. The comparison principle is not only useful to demonstrate the uniqueness of the solution but it is also physical relevant. In the next we build up a numerical scheme that preserves this principle.

## 2. Quasimonotone ODE Approximation. Method of Line

By the method of lines, one can associate an ordinary differential system of equations (ODE) to a parabolic partial differential equation. The MOL consists in the discretization of the space variable using one of the standard methods as finite element, finite differences or finite-volume method (FVM).

Using the FVM, one introduce a net of the inner knots  $x_i$  and a set of the control volumes  $\omega_i$ .

**Definition 1** (ADMISSIBLE MESHES). *The triangulation  $\mathcal{T} = \{(\omega_i, x_i)\}_{i \in I}$  is calling an admissible meshes if it satisfies:*

- $$\left| \begin{array}{l} \omega_i \text{ is open polygonal set } \subseteq \Omega, x_i \in \overline{\omega}_i \\ (1) \bigcup_{i \in I} \omega_i = \overline{\Omega} \\ (2) \forall i \neq j \in I \text{ and } \overline{\omega}_i \cap \overline{\omega}_j \neq \emptyset, \text{ either } \mathcal{H}_{N-1}(\overline{\omega}_i \cap \overline{\omega}_j) = 0 \text{ or} \\ \quad \sigma_{ij} := \overline{\omega}_i \cap \overline{\omega}_j \text{ is } (N-1) - \text{side of } \omega_i \text{ and } \omega_j \\ (3) \forall \sigma_{ij}, [x_i, x_j] \perp \sigma_{ij} \\ (4) [x_{i\partial}, x_i] \perp \sigma_{\partial i} := \overline{\omega}_i \cap \partial\Omega. \end{array} \right.$$

The space discretized equations are derived from the integral form of (1) for each control volume  $\omega_i$  (see Fig. 1),

$$\int_{\omega_i} \frac{\partial b(u)}{\partial t} dx - \int_{\partial \omega_i} \mathbf{a}(u, \nabla u) \cdot \mathbf{n} da = 0, \quad \forall i \in I.$$

Consider the infiltration problem (6). By a proper approximations of the volume integral and line integral one obtains

$$\begin{cases} m(\omega_i) \frac{d\theta(h_i)}{dt} - \sum_{j \in \mathcal{N}(i)} \left[ K_{ij} m(\sigma_{ij}) \left( \frac{h_j - h_i}{d_{ij}} + \mathbf{e}_3 \cdot \mathbf{n}_{ij} \right) \right] = 0, \\ h_i|_{t=0} = h_{0i}, \end{cases} \quad (8)$$

for  $t > 0$  and for any  $i \in I$ .  $\mathcal{N}(i)$  denotes all neighbours of  $\omega_i$ . For those  $\omega_i$  whose boundary  $\sigma_i^a$  lay on boundary  $\partial\Omega$  the corresponding term in the sum is given by

$$K(h_i, h_{Di}) m(\sigma_i^a) \left( \frac{h_{Di} - h_i}{d_{\partial i}} + \mathbf{e}_3 \cdot \mathbf{n}_{ij} \right). \quad (9)$$

Let

$$f_i(\mathbf{h}; \mathbf{h}_D) = \sum_{j \in \mathcal{N}(i)} \left[ K(h_i, h_j) m(\sigma_{ij}) \left( \frac{h_j - h_i}{d_{ij}} + \mathbf{e}_3 \cdot \mathbf{n}_{ij} \right) \right],$$

then the discrete infiltration problem read as

$$m(\omega_i) \frac{d\theta(h_i)}{dt} = f_i(\mathbf{h}; \mathbf{h}_D), \quad h_i|_{t=0} = h_{0i}. \quad (10)$$

We will show that, for a suitable definition of the numerical hydraulic conductivity,  $K(\cdot, \cdot)$ , and by imposing an upper bound on the mesh size, the function  $\mathbf{f}(\cdot, \mathbf{h}_D)$  verifies Kamke conditions. Then using this conditions we will prove that there exists a discrete comparison result.

For that, define

$$K(u, v) = \max(\kappa(u), \kappa(v), \eta_0), \quad (11)$$

and assume that the step size of the mesh,  $\Delta := \max(d_{ij})$ , and hydraulic conductivity,  $\kappa$ , satisfy

$$\mathbf{A2}''' \left\{ \begin{array}{l} (1) \kappa : \mathbb{R} \rightarrow \mathbb{R}_+, \text{ bounded and nondecreasing} \\ (2) \eta_0 |h_1 - h_2| > \Delta |\kappa(h_1) - \kappa(h_2)|, \forall h_1, h_2 \in \mathbb{R}. \end{array} \right.$$

We note that the second requirement implies that  $\kappa$  is a Lipschitz function which is a stronger condition than  $\mathbf{A2}'$ -(2). The threshold  $\eta_0$  in (11) is necessary if the function  $\kappa$  do not satisfy the second condition  $\mathbf{A2}'$ -(1).

Also we suppose that in addition to  $\mathbf{A1}$  the water content function is a differentiable function on  $(-\infty, 0)$  and  $\theta(h) = 1$  on  $(0, \infty)$ ,

$$\mathbf{A1}' \quad \frac{d\theta}{dh} > 0 \text{ on } (-\infty, 0).$$

**Proposition 1 (QUASSIMONOTONY).** *Let  $\mathcal{T}$  be an admissible mesh and let  $K(\cdot, \cdot)$  be given by (11). Assume **A2'''**. Then the following statements hold:*

- (a) *If  $v < w$  then  $K(u, w)(w - u) - K(u, v)(v - u) > \eta(w - v)$ .*
- (b) *If  $\hat{\mathbf{h}}_D \geq \mathbf{h}_D$  then*

$$f_i(\mathbf{h}, \hat{\mathbf{h}}_D) \geq f_i(\mathbf{h}, \mathbf{h}_D).$$

- (c) *Kamke condition. If  $\hat{\mathbf{h}} \geq \mathbf{h}$  and  $\hat{h}_i = h_i$  then*

$$f_i(\hat{\mathbf{h}}, \mathbf{h}_D) \geq f_i(\mathbf{h}, \mathbf{h}_D).$$

*Proof.* The statement (a) is a simply consequence of **A2'''**-(1) and the definition (11). The statement (b) results from (9).

(c) We have

$$f_i(\hat{\mathbf{h}}, \mathbf{h}_D) - f_i(\mathbf{h}, \mathbf{h}_D) = \sum_{j \in \mathcal{N}(i)} m(\sigma_{ij}) \Gamma_{i,j}(\hat{\mathbf{h}}, \mathbf{h})$$

with

$$\Gamma_{i,j} = K(h_i, \hat{h}_j) \frac{\hat{h}_j - h_i}{d_{ij}} - K(h_i, h_j) \frac{h_j - h_i}{d_{ij}} + (K(h_i, \hat{h}_j) - K(h_i, h_j)) \mathbf{e}_3 \cdot \mathbf{n}_{ij}.$$

Using (a) one obtains

$$\Gamma_{i,j} \geq \eta \frac{\hat{h}_j - h_j}{\Delta} - (K(h_i, \hat{h}_j) - K(h_i, h_j)) \geq \eta \frac{\hat{h}_j - h_j}{\Delta} - (\kappa(\hat{h}_j) - \kappa(h_j)).$$

From the condition **A2'''**-(2) we see that  $\Gamma_{i,j} \geq 0$ . Hence, (c) prove true.

The results presented in the proposition improve our previous results [7] in sense that  $\mathbf{f}$  verifies the Kamke condition in less restrictive conditions on hydraulic conductivity function than those imposed there.

**Theorem 2 (COMPARISON THEOREM. DISCRETE CASE).** *Assume the hypothesis of the proposition and **A1'**. Let  $\mathbf{h}(t)$  and  $\hat{\mathbf{h}}(t)$ ,  $t \in (0, T)$ , be the solutions of the problem (10) associated to  $(\mathbf{h}_D, \mathbf{h}_0)$  and  $(\hat{\mathbf{h}}_D, \hat{\mathbf{h}}_0)$ , respectively. Suppose that*

$$\mathbf{h}_D \leq \hat{\mathbf{h}}_D < 0, \quad \mathbf{h}_0 \leq \hat{\mathbf{h}}_0 < 0.$$

*Then*

$$\mathbf{h} \leq \hat{\mathbf{h}} \text{ on } (0, T).$$

*Proof.* Let  $\alpha < 0$  be an upper bound for boundary data and initial data, that is

$$\hat{h}_{Di} \leq \alpha, \quad \hat{h}_{0i} \leq \alpha$$

and let  $T_1$  be such that the solutions  $\mathbf{h}(t)$  and  $\hat{\mathbf{h}}(t)$  stay less than 0 on the interval  $(0, T_1)$  [2], [4]. As  $\theta'(h) > 0$  on  $(0, T_1)$  one can apply Nickel's theorem ([8], [9]) and one obtains that

$$\mathbf{h}(t) \leq \hat{\mathbf{h}}(t) \text{ on } (0, T_1).$$

Moreover, observe that the solution  $\mathbf{h}_\alpha(t)$  associate to the boundary data  $h_{Di} = \alpha$  and initial data  $h_{0i} = \alpha$  is a constant solution  $\mathbf{h}(t)_i = \alpha$  ( $f_i(\alpha; \alpha) = 0!$ ). Using the inequality (b) of the proposition we have

$$\mathbf{h}(t) \leq \widehat{\mathbf{h}}(t) \leq \alpha \text{ on } (0, T_1).$$

The inequality show that  $T = T_1$ .

## References

- [1] H. W. Alt, S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., **183** (1983), 311–341.
- [2] V. Barbu, *Ecuatii diferențiale*, Editura Junimea, Iași, 1985.
- [3] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, 1988.
- [4] K.E. Brenan, S.L. Campbell and L.R. Petzold, *Numerical Solution of Initial Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics, SIAM, 1996.
- [5] J.I. Diaz, *Qualitative Study of Nonlinear Parabolic Equations: an Introduction*, Extracta Mathematicae, **16** (3) (2001), 303–341.
- [6] R. Eymard, Th. Gallouet and R. Herbin, *Finit Volume Method*, in Handbook of Numerical Analysis, G. Ciarlet and J.L Lions eds., North Holland, 2000.
- [7] S. Ion, D. Homentcovschi and D. Marinescu, *Method of Lines for Solving Richards' Equation*, Proc. of 5<sup>th</sup> International Seminare on “Geometry, Continua and Microstructures”, eds. Sanda Cleja-Tigoiu and V. Tigoiu, Editura Academiei Române, 2002, pp. 125–132.
- [8] V. Lakshmikantham and S. Leela, *Differential and Integral Inequalities: Theory and Applications*. Academic Press, New York, 1969.
- [9] K. Nickel, *Bounds for the Set of Solutions of Functional-Differential Equations*, MRC Tech. Summary Report No. 1782, Univ. of Wisconsin, Madison, 1977.
- [10] F. Otto,  *$L^1$ -contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, **131** (1) (1986), 20–38.

## Minimum Free Energy Configuration of the Planar Lipidic Bilayer. Analytical Solutions

Stelian Ion<sup>\*†‡</sup> and Dumitru Popescu<sup>\*‡</sup>

The paper deal with the calculus of the shape of the monolayers, which constitute the two leaflets of the planar lipidic bilayer (BLM), that undergo external constrains. The deformed configuration minimises the free energy of the BLM and it is determined as the solution of a forth order elliptic equation subject to certain boundary conditions. For planar radial symmetrical problems the general solutions of elliptic equation are obtained as a power series expansion and it is shown they belong to the family of the Bessel functions.

### 1. Introduction

The main point of the paper is to obtain the solutions of the linear elliptic equation of the forth order

$$\Delta^2 u - 2\xi \Delta u + u = 0, \quad (1)$$

where  $\xi$  is a real positive constant and  $\Delta$  represents the differential Laplace operator.

The equations like (1) are frequently used to study the space configuration of the planar lipid bilayer and its free energy response to the deformation ([6], [5], [7]). Briefly, the bilayer lipidic membranes (BLM) is a biological sheet like structure that consists in two lipidic monolayer combined tail to tail. The monolayers can be slightly stretched or compressed and undergo curvature deformations and there exists a free energy response to these deformations. The mechanical properties of BLM are specific to smectic liquid crystal of A type ([4], [6] and [2], [3] for more general theory of liquid crystal).

---

\* “Gheorghe Mihoc–Caius Iacob” Institute of Mathematical Statistics and Applied Mathematics, Buchares, Romania.

† e-mail: [istelian@ima.ro](mailto:istelian@ima.ro)

‡ Supported by PNCDI–VIASSAN Grant NO. 344/2004.

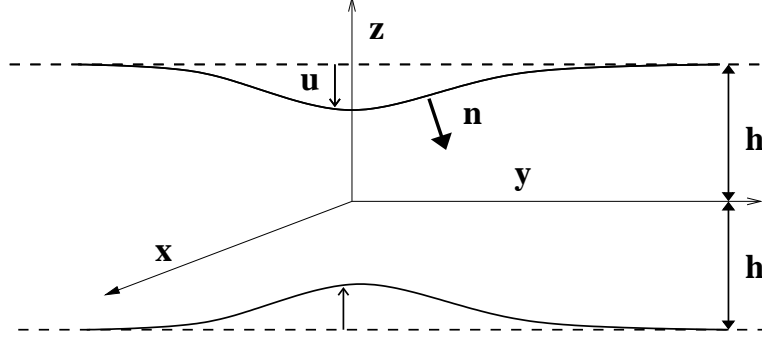


Fig. 1. The unperturbed (dashed line) and perturbed (solid line) position of the two monolayer. The bilayer is  $2h$  width in the initial state and it is infinite spread in the  $x, y$  directions in the both states. The vector  $\mathbf{n}$  represents the normal to the deformed surface.

Let us consider the planar structure given by the plane surface  $z = \pm h$  as unperturbed position of the BLM (see fig. 1). We suppose that the two surfaces  $z = \pm h$  undergo a displacement with same distance  $u$  but in opposite directions. Then deformation free energy density of unit unperturbed area can be written as [6]

$$\rho(x, y) = hB \left( \frac{u}{h} \right)^2 + hK_1 (\Delta u)^2 + \gamma |\nabla u|^2, \quad (2)$$

where:  $h$  is the monolayer equilibrium thickness,  $B$  is the compression-extension modulus constant,  $K_1$  is splay modulus constant and  $\gamma$  represents the interfacial tension constant.

The deformation free energy associated to a domain  $D$ ,  $\mathcal{F}(D)$ , is defined as the surface integral

$$\mathcal{F}(D) = \int_D \rho(x, y) d\sigma. \quad (3)$$

Passing to the dimensionless variable  $u = hu_*$ ,  $x = Lx_*$ ,  $y = Ly_*$  and choosing the characteristic length  $L$  such that  $L^2 = h\sqrt{K_1/B}$  the relation (2) becomes

$$\rho = hB \left( (\Delta_* u_*)^2 + 2\xi (\nabla_* u_*)^2 + u_*^2 \right), \quad (4)$$

where  $2\xi = \gamma/\sqrt{K_1 B}$ . The  $*$  subscript denote the dimensionless quantity and it will be suppressed in the sequel.

Let us define the bilinear form  $a(u, v)$ , the linear operator  $L[u]$  by

$$\begin{aligned} a(u, v) &= \Delta u \Delta v + 2\xi \nabla u \cdot \nabla v + uv, \\ L[u] &= \Delta^2 u - 2\xi \Delta u + u, \end{aligned}$$

and the dimensionless energy  $J(D, u)$  by

$$J(D, u) = \frac{1}{2} \int_D a(u, u) d\sigma.$$



The deformation free energy  $\mathcal{F}(D)$  can be expressed as

$$\mathcal{F}(D, u) = 2h^2 \sqrt{K_1 B} J(D, u).$$

Performing some calculus we obtain

$$\begin{aligned} J(D, v) - J(D, u) &= \int_D L[u](v - u) d\sigma + \\ &+ \int_{\partial D} \left( \Delta u \frac{\partial(v - u)}{\partial n} - (v - u) \frac{\partial \Delta u}{\partial n} + 2\xi(v - u) \frac{\partial u}{\partial n} \right) ds \\ &+ J(D, v - u). \end{aligned} \quad (5)$$

This identity shows that the solution of the equation  $L[u] = 0$  that satisfies the boundary conditions

$$\begin{cases} u|_{\partial D} = u_0, \\ \frac{\partial u}{\partial n} \Big|_{\partial D} = u_1 \end{cases} \quad (6)$$

minimises the free energy  $J$  in the class of the functions that satisfy same boundary conditions.

## 2. Analytical solutions

We shall consider the radial symmetric function defined in the whole plane. As in [1] we search for a real power series solution of the equation (1),

$$\Phi(r) = \sum_{m=0}^{\infty} c_m r^{\alpha+m},$$

where the exponent  $\alpha$  is a real number and  $c_0 \neq 0$ . Taking into account that in polar coordinate the Laplace operator  $\Delta$  is given by

$$\Delta = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right),$$

one obtains

$$\begin{aligned} r^4 L[\Phi] &= \alpha^2 (\alpha - 2)^2 c_0 r^\alpha + (\alpha + 1)^2 (\alpha - 1)^2 c_1 r^{\alpha+1} + \\ &+ \alpha^2 [(\alpha + 2)^2 c_2 - 2\xi c_0] r^{\alpha+2} + (\alpha + 1)^2 [(\alpha + 3)^2 c_3 - 2\xi c_1] r^{\alpha+3} + \\ &+ \sum_{m=0}^{\infty} \{ (\alpha + m + 2)^2 [(\alpha + m + 4)^2 c_{m+4} - 2\xi c_{m+2}] + c_m \} r^{\alpha+m+4}. \end{aligned}$$

The exponent  $\alpha$  and coefficients  $c_m$  will be determined equating the coefficients of  $r^{\alpha+m}$ ,  $m = 0, 1, \dots$  to zero.

One obtains a recurrence relation for the coefficients  $c_m$ ,

$$(\alpha + m + 4)^2(\alpha + m + 2)^2 c_{m+4} - 2\xi(\alpha + m + 2)^2 c_{m+2} + c_m = 0 \quad (m = 0, 1, \dots) \quad (7)$$

and four equation for  $\alpha, c_0, c_1, c_2, c_3$ .

Let us firstly study the solutions of the recurrence relation.

We choose  $c_1 = c_3 = 0$  and it follows that  $c_{2m+1} = 0, m = 2, 3, \dots$  successively.

To find the even rank terms  $c_{2m}$  we introduce another sequence  $\{b_m\}_{m=0,\infty}$  and write  $c_{2m}$  as

$$c_0 = b_0, \\ c_{2m} = \frac{b_m}{(\alpha + 2)^2 \dots (\alpha + 2m)^2} \quad (m = 1, 2, \dots).$$

Substituting this expression into recurrence relation (7) we find that the terms  $b_m$  verify a linear recurrence relation

$$b_{m+2} - 2\xi b_{m+1} + b_m = 0 \quad (m = 0, 1, \dots). \quad (8)$$

We note that there always exist two real solutions of this recurrence relation and we postpone to write their explicit solutions.

It follows that the power series  $\Phi$  has the form

$$\Phi(\alpha, r) = r^\alpha \left( b_0 + \sum_{m=1} r^{2m} \frac{b_m}{(\alpha + 2)^2 \dots (\alpha + 2m)^2} \right)$$

and satisfies

$$r^4 L[\Phi(\alpha, r)] = r^\alpha \alpha^2 [(\alpha - 2)^2 c_0 + ((\alpha + 2)^2 c_2 - 2\xi c_0)] r^2.$$

One observes that  $\alpha = 0$  is a double root of the polynomial on right hand side.

The solutions that we are looking for can be obtained as in the following proposition

**Proposition 1.** *If the sequence  $\{b_m\}_{m=0,\infty}$  satisfies the recurrence relation (8), then the functions  $\mathcal{I}(r; \mathbf{b})$ ,  $\mathcal{K}(r; \mathbf{b})$  given by*

$$\mathcal{I}(r; \mathbf{b}) \equiv \Phi(0, r) = \sum_{m=0} \left( \frac{r}{2} \right)^{2m} \frac{b_m}{(m!)^2}, \quad (9)$$

$$\mathcal{K}(r; \mathbf{b}) \equiv \frac{\partial \Phi}{\partial \alpha} \Big|_{\alpha=0} = \ln(r) \mathcal{I}(r; \mathbf{b}) - \sum_{m=1} \left( \frac{r}{2} \right)^{2m} \frac{b_m H(m)}{(m!)^2}, \quad (10)$$

represent the solutions of the equations (1).

The function  $H(m)$  is given by  $H(m) = 1 + \frac{1}{2} + \dots + \frac{1}{m}$ .

To obtain the general solution  $b_m$  of the recurrence relation (8) we consider the characteristic equation

$$t^2 - 2\xi t + 1 = 0.$$

Depending on the  $\xi$  this equation can have real or complex solutions.

a) **Case**  $\xi > 1$

There exists two real positive solutions

$$\lambda = \sqrt{\xi - \sqrt{\xi^2 - 1}}, \quad \beta = \sqrt{\xi + \sqrt{\xi^2 - 1}}$$

and the general solution of recurrence equation is given by

$$b_m = C_1 \lambda^{2m} + C_2 \beta^{2m}.$$

Consequently, we obtain that the functions  $\mathcal{I}(r, \mathbf{b})$  are nothing but the Bessel functions with imaginary argument of order zero  $I_0$  [9],

$$\begin{aligned} I_0(\lambda r) &= \mathcal{I}(r, \lambda) = \sum_{m=0}^{\infty} \left( \frac{r\lambda}{2} \right)^{2m} \frac{1}{(m!)^2}, \\ I_0(\beta r) &= \mathcal{I}(r, \beta). \end{aligned} \quad (11)$$

One can obtain another two independent solutions by linear combination of the function  $\mathcal{I}$  and  $\mathcal{K}$ . Among all linear combination there exist one that led to the Bessel function  $K_0$  which is given by

$$\begin{aligned} K_0(\lambda r) &= -\mathcal{K}(r; \lambda) + \ln(\lambda/2) \mathcal{I}(r; \lambda) - \gamma \mathcal{I}(r; \lambda) \\ &= -\ln(\lambda r/2) I_0(\lambda r) + \sum_{m=0}^{\infty} \left( \frac{\lambda r}{2} \right)^{2m} \frac{\psi(m+1)}{(m!)^2}, \\ K_0(\beta r) &= \dots, \end{aligned} \quad (12)$$

where  $\psi(m+1) = H(m) - \gamma$ ,  $\psi(1) = -\gamma$ , and  $\gamma$  is Euler's constant.

b) **Case**  $0 < \xi < 1$

There exists two complex solutions and the general solution of recurrence relation is given by

$$b_m = A \cos(m\theta) + B \sin(m\theta),$$

where the angle  $\theta$  is given by

$$\cos(\theta) = \xi.$$

In this case the functions  $\mathcal{I}(r, \mathbf{b})$  become

$$\begin{aligned} \text{Ber}(r; \theta) &= \sum_{m=0}^{\infty} \left( \frac{r}{2} \right)^{2m} \frac{\cos(m\theta)}{(m!)^2}, \\ \text{Bei}(r; \theta) &= \sum_{m=1}^{\infty} \left( \frac{r}{2} \right)^{2m} \frac{\sin(m\theta)}{(m!)^2}. \end{aligned} \quad (13)$$

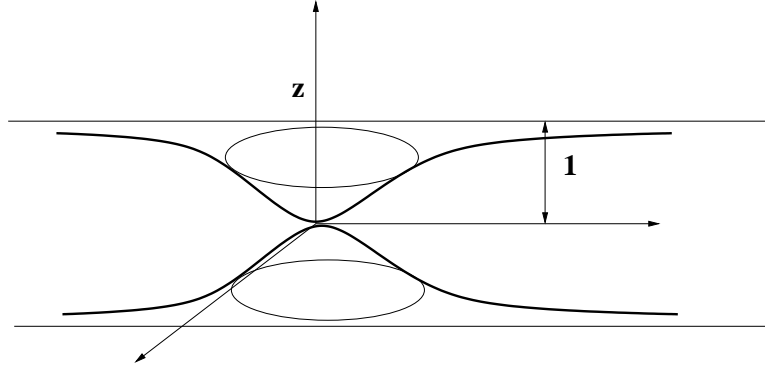


Fig. 2. Perforation problem. The two monolayers of the bilayer have a contact point and at great distance the monolayers rest unperturbed.

As in the first case, by linear combination, we can also obtain the following solutions

$$\begin{aligned} \text{Ker}(r; \theta) &= -\ln \frac{r}{2} \text{Ber}(\cdot; \cdot) + \theta/2 \text{Bei}(\cdot; \cdot) + \text{Rc}(\cdot; \cdot), \\ \text{Kei}(r; \theta) &= -\ln \frac{r}{2} \text{Bei}(\cdot; \cdot) - \theta/2 \text{Ber}(\cdot; \cdot) + \text{Rs}(\cdot; \cdot). \end{aligned} \quad (14)$$

The series  $\text{Rs}(\cdot; \cdot)$  and  $\text{Rc}(\cdot; \cdot)$  are given by

$$\text{R}_s^c(r; \theta) = \sum_{m=0}^{\infty} \left(\frac{r}{2}\right)^{2m} \frac{\frac{\cos(m\theta)}{\sin(m\theta)}}{(m!)^2} \psi(m+1)$$

The real functions (13) verify the relation

$$\text{Ber}(r; \theta) + i \text{Bei}(r; \theta) = \text{I}_0 \left( e^{i\theta/2} r \right) \quad (15)$$

and the real functions (14) verify the relation

$$\text{Ker}(r; \theta) + i \text{Kei}(r; \theta) = \text{K}_0 \left( e^{i\theta/2} r \right). \quad (16)$$

For  $\theta = \pi/4$  they are known as Thomson's function and Rusell's function, respectively [9].

### 3. Perforation of the planar bilayer problem

As an application of the previous results formulae we present here the solution of the perforation bilayer problem (see fig. 2). The perforation of the BLM can be generated by the thermal fluctuations of its compounds and in certain circumstances the perforation produce a pore in the BLM [8].

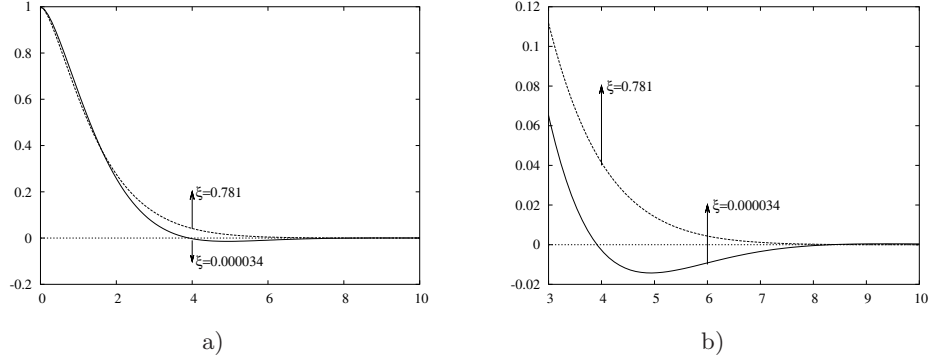


Fig. 3. The solutions of the perforation problem (17) for two values of the parameter  $\xi$  a) and detail b).

Mathematically, the perforation problem can be formulate as follows. Find  $u : \mathbf{R}^2 - \{0,0\} \rightarrow \mathbf{R}$  such that:

$$\begin{cases} L[u] = 0, & \text{in } \mathbf{R}^2 - (0,0), \\ u|_{r=0} = 1, & u|_{r=\infty} = 0, \\ \frac{\partial u}{\partial r}\bigg|_{r=0} = 0, & \frac{\partial u}{\partial r}\bigg|_{r=\infty} = 0, \end{cases} \quad (17)$$

Depending of the parameters  $\xi$  we find out that the solution of the problem is

$$\text{Case } \xi > 1 : \quad u(r) = (K_0(\lambda r) - K_0(\beta r)) / (\ln \beta - \ln \lambda),$$

$$\text{Case } \xi < 1 : \quad u(r) = -\frac{2}{\theta} \text{Kei}(r; \theta).$$

Free energy of the perforated BLM is given by

$$J(u) = \pi r \left( \Delta u \frac{\partial u}{\partial r} - u \frac{\partial \Delta u}{\partial r} + 2\xi u \frac{\partial u}{\partial r} \right) \bigg|_0^\infty = \pi r \frac{\partial \Delta u}{\partial r} \bigg|_{r=0}.$$

Then we have

$$J(u) = \begin{cases} \pi \frac{\sqrt{\xi^2 - 1}}{\ln(\xi + \sqrt{\xi^2 - 1})}, & \xi > 1, \\ 2\pi \frac{\sqrt{1 - \xi^2}}{\arccos \xi}, & \xi < 1. \end{cases}$$

For the following parameters of the BLM [8]:  $\gamma \approx 15 \times 10^{-4} \text{ Nm}^{-1}$ ,  $K_1 \approx 0.93 \times 10^{-11} \text{ N}$ ,  $B \approx 5.36 \times 10^7 \text{ Nm}^{-2}$ ,  $h \approx 11.3 \times 10^{-10} \text{ m}$  and  $\xi = \frac{\gamma}{2\sqrt{K_1 B}} \approx 3.4 \times 10^{-2}$

one obtain  $J(u) \approx 4$  and then

$$\mathcal{F}/kT = 2h^2 \sqrt{K_1 B} J(u)/kT \approx 50.$$

## References

- [1] W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, John Wiley & Sons, 1976.
- [2] J. L. Ericksen, *Conservation laws for liquid crystals*, Tran. Soc. Rheol. **5**, 1961, pp. 23–34.
- [3] J. L. Ericksen, *Liquid crystals with variable degrees of orientation*, Arch. Rat. Mech. Anal., **113**, 1991, pp. 97–120.
- [4] P. G. deGennes, *The Physics of Liquid Crystals*, Clarendon Press, Oxford, 1974.
- [5] P. Helfrich and E. Jakobsson, *Calculation of Deformation Energies and Conformations in Lipid Membranes Containing Gramicidin Channels*, Biophys. J., Vol. **57**, May 1990, pp. 1075–1084.
- [6] H. W. Huang, *Deformation Free Energy of Bilayer Membrane and its Effect on Gramicidin Channel Lifetime*, Biophys. J., Vol. **50**, December 1986, pp. 1061–1070.
- [7] C. Nielsen, M. Goulian and O. Andersen, *Energetics of Inclusion-Induced Bilayer Deformation*, Biophys. J., vol.**74** (1998), pp. 1966–1983.
- [8] D. Popescu, S. Ion, A.I. Popescu, and L. Movileanu, *Elastic properties of bilayer lipid membranes and pore formation*, in Planar Lipid Bilayer(BLMS) and their Applications, H. TiTien and A. Ottowa, eds., Elsevier Sciences Publishers, 2003, pp. 173–204.
- [9] G. N. Watson, *Theory of Bessel Functions*, Cambridge, 1966.

## A Numerical Study of Axisymmetric Slow Viscous Flow Past Two Spheres

Gheorghe Juncu\*

This paper presents a computational study of the axisymmetric, slow, viscous flow (Stokes flow) around two spheres placed parallel to their line of centers. The fourth-order stream function equation in bispherical coordinates, was split into its coupled form, by defining the fluid vorticity. The central, second order accurate, finite difference scheme was used to discretize the model equations. Two numerical algorithms were employed to solve the discrete equations: multigrid (MG) and preconditioned conjugate gradient squared (PCGS). The preconditioners tested are approximations (ILU and multigrid iterations) of the symmetric part of the discrete Stokes operator in bispherical coordinates. The mesh behaviour of the convergence rate of MG and PCGS was analysed for different values of the model parameters.

### 1. Introduction

The hydrodynamic interaction of two spherical particles moving slowly in a viscous fluid is of fundamental importance to meteorology, colloid chemistry, flow of suspensions and other fields from nuclear and chemical engineering, environmental sciences, etc. The solution to the problem of two spheres rotating with constant angular velocities around their line of centres was obtained first by Jeffery [1], assuming negligible inertial effects (Stokes flow). The first solution of the associated axisymmetrical problem when the spheres translate with the same velocity along their line of centres in Stokes flow was obtained by Stimson and Jeffery [2]. Stimson and Jeffery [2] considered only the case of equal-sized spheres in non-contact. Starting from these two classical articles, other creeping flow solutions were developed by Goldman et al. [3], Wakia [4], Cooley and O'Neill [5], Davis [6], O'Neill and Majumdar [7],

---

\* “Politehnica” University of Bucharest, Polizu 1, 78126 Bucharest, Romania, e-mail: [juncugh@netscape.net](mailto:juncugh@netscape.net), [juncu@easynet.ro](mailto:juncu@easynet.ro)

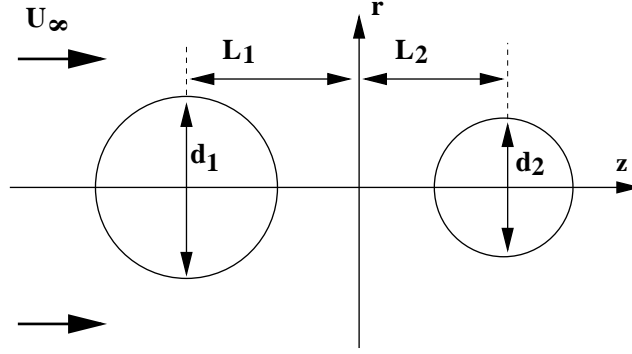


Fig. 1. Schematic of the problem.

Rushton and Davies [8], Haber et al. [9], Hetsroni and Haber [10] and Zinchenko [11]. A detailed account of the work done on the problem of two spheres external to each other is given by Happel and Brenner [12] and Kim and Karrila [13]. We restricted our citation only to problems similar to that solved in this work. For this reason, related problems as electrophoretic or thermocapillary migration of two spheres are not mentioned here.

In [1–9, 11], the momentum balance equations (the Stokes and continuity equations) were solved analytically in general orthogonal curvilinear coordinate systems (bispherical coordinate system, [1–4], [6–9], [11] and tangent-sphere coordinate system [5]). The suitable analytical solutions in terms of these coordinates have been shown to be infinite series. The method of reflection was used in [10]. The application of the boundary integral methods to Stokes flow was discussed by Roumeliotis [14]. For a pair of equal spheres in tandem, the Navier-Stokes equations were solved numerically in bispherical coordinate system at  $Re = 40$  by Tal et al. [15]. The aim of this work is to solve numerically the Stokes flow equations for two spheres in-line. The vorticity–stream function formulation of the flow equations in bispherical coordinates was chosen. The central, second order accurate, finite difference scheme was used to discretize the model equations. Two numerical algorithms were tested: multigrid (MG) and preconditioned conjugate gradient squared (PCGS). The convergence rate of MG and PGCG was analysed for different spheres spacing and sizes.

## 2. Statement of the problem

We suppose that a homogeneous viscous fluid of constant density and viscosity flows past two rigid spheres of diameters  $d_i$ ,  $i = 1, 2$ , as illustrated in figure 1. The diameters of the spheres are assumed considerably higher than the molecular mean free path of the surrounding fluid. Oscillations and rotation of the spheres do not occur during the flow. The flow is considered steady, laminar and axisymmetric. Also, the inertial effects are negligible in comparison with the viscous forces (creeping or



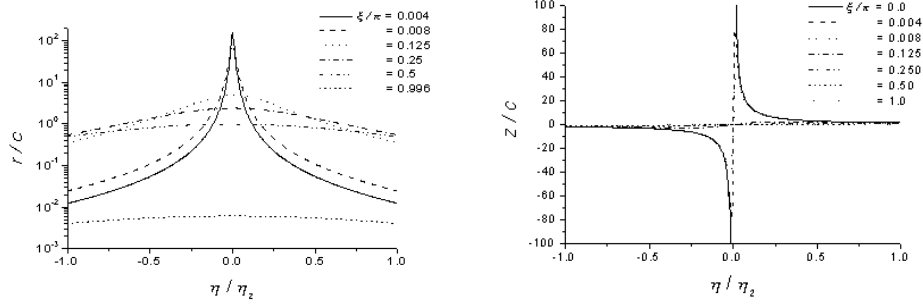


Fig. 2. Cylindrical coordinates  $(r, z)$  versus bispherical coordinates  $(\eta, \xi)$ ;  $d_1 = d_2 = d$ .

Stokes flow).

Let a system of cylindrical coordinates  $(r, z)$  be chosen so that the centres of the spheres lie along the  $z$ -axis (see figure 1). In order to facilitate the numerical solution of this problem, we introduce the bispherical coordinates  $(\eta, \xi)$  defined by (see also figure 2)

$$r = \frac{c \sin \xi}{\cosh \eta - \cos \xi}; \quad z = \frac{c \sinh \eta}{\cosh \eta - \cos \xi},$$

where  $c > 0$  is a characteristic length. This transformation maps the right half of the  $rz$ -plane (from which the surface occupied by the spheres is excluded) into the rectangle  $\eta_1 \leq \eta \leq \eta_2$ ,  $0 \leq \xi \leq \pi$ ,  $\eta_1 < 0$ ,  $\eta_2 > 0$ . The surfaces of the spheres are located at  $\eta = \eta_1$  and  $\eta = \eta_2$ . The relations between  $\eta_i$ , the diameters of the spheres  $d_i$  and the distances  $L_i$  of their centers from the origin of the coordinates system are:

$$\frac{d_i}{2} = \frac{c}{\sinh |\eta_i|}; \quad L_i = c \coth |\eta_i|, \quad i = 1, 2.$$

The following scaling is chosen: the radius of the up-stream sphere  $d_1/2$  for length and  $4/U_\infty d_1^2$  for stream function ( $U_\infty$  being the free stream velocity).

In terms of dimensionless stream function,  $\psi$ , the partial differential equation of incompressible viscous hydrodynamics in the time-independent Stokes approximation is

$$E^4 \psi = 0. \quad (1)$$

The boundary conditions considered are:

– spheres surfaces ( $\eta = \eta_i$ ,  $i = 1, 2$ )

$$\psi = 0; \quad (2)$$

– free stream ( $\eta \rightarrow 0$ ,  $\xi \rightarrow 0$ )

$$\psi \rightarrow \frac{1}{2} \frac{\sin^2 \xi}{A^2}; \quad (3)$$

– symmetry axis ( $\xi = 0$  and  $\eta \neq 0$ ,  $\xi = \pi$ )

$$\psi = 0, \quad (4)$$

where

$$A = \frac{\cosh \eta - \cos \xi}{\bar{c}}, \quad \bar{c} = \frac{2c}{d_1}.$$

The operator  $E^2$  assumes the following form in axisymmetric bispherical coordinates:

$$E^2 = A^2 \frac{\partial^2}{\partial \eta^2} + \frac{A \sinh \eta}{\bar{c}} \frac{\partial}{\partial \eta} + A^2 \frac{\partial^2}{\partial \xi^2} + \frac{A (1 - \cosh \eta \cos \xi)}{\bar{c} \sin \xi} \frac{\partial}{\partial \xi}. \quad (5)$$

### 3. Numerical methods

The fourth-order stream function equation (1) was split into its coupled form

$$E^2 \psi = \omega, \quad (6)$$

$$E^2 \omega = 0 \quad (7)$$

by defining the fluid vorticity  $\omega$ . Note that the vorticity defined previously is not identical to the exact fluid vorticity (i.e. is the curl of the velocity). The relation between these two quantities is  $\omega = h_3$ , where  $h_3$  is the metric coefficient,

$$h_3 = c \frac{\sin \xi}{\cosh \eta - \cos \xi}.$$

It is convenient numerically to work with the deviation from the uniform flow  $\psi^*$ ,

$$\psi^* = \psi - \frac{1}{2} \frac{\sin^2 \xi}{A^2}.$$

After  $\psi^*$  is substituted for  $\psi$  in (4a), the final form of the mathematical model is:

$$E^2 \psi^* = \omega, \quad (8)$$

$$E^2 \omega = 0 \quad (9)$$

with the boundary conditions:

– spheres surfaces ( $\eta = \eta_i$ ,  $i = 1, 2$ )

$$\psi^* = -\frac{1}{2} \frac{\sin^2 \xi}{A^2}; \quad (10)$$

– free stream ( $\eta \rightarrow 0$ ,  $\xi \rightarrow 0$ )

$$\psi \rightarrow 0, \quad \omega \rightarrow 0; \quad (11)$$

– symmetry axis ( $\xi = 0$  and  $\eta \neq 0$ ,  $\xi = \pi$ )

$$\psi^* = \omega = 0. \quad (12)$$

The two-dimensional region  $(\eta_1, \eta_2) \times (0, \pi)$  was transformed into the unit square by well known elementary relations. Equations (5) were discretized with the central second order accurate finite difference scheme on uniform meshes with  $N \times N$  points and  $h = (N - 1)^{-1}$ ,  $N = 5, 9, 17, 33, 65$  and  $129$ . Two algorithms were employed to solve the discrete equations: (1<sup>o</sup>) the preconditioned conjugate gradient squared (PCGS) [16] and (2<sup>o</sup>) multigrid (MG).

The preconditioners used are the incomplete LU factorisation (with two extra diagonals; algorithm IC (1, 3) from [17]) and the multigrid approximation (two multigrid cycles) of the symmetric part of the discrete operator  $E^2$ . These preconditioners are similar with those analysed theoretically in [18–20] and tested experimentally in [21–25]. The structure of the MG cycle used in preconditioning is: (1.) cycle of type V; (2.) smoothing by point Gauss-Seidel; one smoothing step is performed before the coarse grid correction and one after in the opposite direction; (3.) prolongation by bilinear interpolation; (4.) restriction by full weighting; (5.) the coarse grid has  $5 \times 5$  points. The stopping criterion used for PCGS is

$$\frac{\|r_i\|}{\|r_0\|} \leq 10^{-6},$$

where  $r_i$  is the residual after  $i$  iterations and  $\|\cdot\|$  the discrete Euclidean norm. The maximum number of iterations allowed is 5000.

The MG algorithm used is the nested FAS technique, [26], [27]. The structure of the MG cycle is: 1) cycle of type V; 2) smoothing by point Gauss-Seidel; two smoothing steps are performed before the coarse grid correction and one after; 3) prolongation by bilinear interpolation for corrections and cubic interpolation for solution; 4) restriction of residuals by full weighting. Three levels were used in the numerical experiments.

The convergence rate of PCGS is monitored by the cumulative number of multiplications per grid point and iteration step. The following convention was adopted in the calculation of this quantity: the number of multiplications of non-preconditioned CGS was considered as the measure unit. The numerical performances of MG are expressed by two quantities: (1) the average reduction factor,  $\rho$ , and (2) the efficiency,  $\tau$ . The average reduction factor and the efficiency are given by

$$\rho = (\|r_{i+j}\| / \|r_i\|)^{1/j}, \tau = -W / \ln \rho,$$

where  $r_i$  is the residual after  $i$  iterations,  $\|\cdot\|$  the discrete Euclidean norm and  $W$  the number of multiplications per grid point and iteration step. The efficiency is also used for the comparison between PCGS and MG.

The computations were made using double-precision Fortran on a PC computer (COMPAQ – Presario 5358 Series, Pentium-Celeron processor).

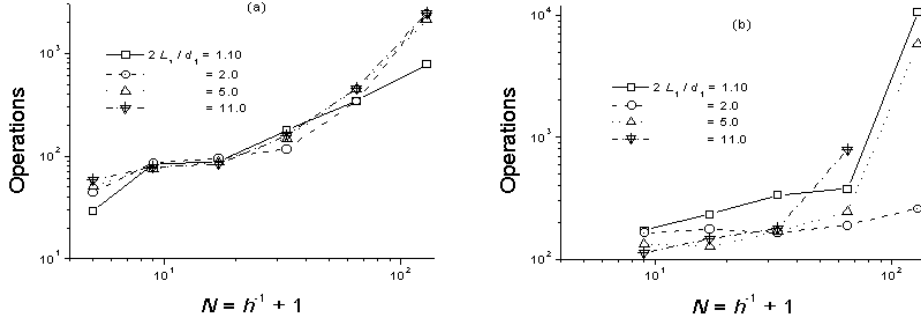


Fig. 3. Mesh behaviour of the convergence rate of PCGS for  $d_1 = d_2 = d$ ;  
(a) ILU preconditioning; (b) MG preconditioning.

#### 4. Numerical experiments

Two geometrical quantities were considered significant for the present problem: (i) the distance between the spheres (expressed by mean of  $2L_i/d_i$ ) and (ii) the diameters of the spheres. First, we will consider the case of equal-sized spheres, e.g.  $d_1 = d_2 = d$ . For this case, the influence of the spheres spacing on the convergence rate is analysed. Secondly, for a given value of  $2L_1/d_1$ , the influence of  $d_1/d_2$  on the convergence rate of the numerical algorithms will be investigated. For the ratio  $d_1/d_2$  two values different from 1 were considered: 0.5 and 2.

The first task in any numerical work is to validate the code's ability to reproduce published results accurately. For the problem studied in this work, one of the most important characteristic quantities is the force  $F_i$  acting on either sphere [2],

$$F_i = \mu \pi U_\infty \frac{d_1}{2} \int_{\gamma} r^3 \frac{\partial}{\partial n} \left( \frac{E^2 \psi}{r^2} \right) d\gamma,$$

where  $\mu$  is the dynamic viscosity of the fluid,  $\partial/\partial n$  is the normal derivative and the integral is taken around any meridian  $\gamma$  of the sphere in a sense which is right-handedly related to the outward drawn normal  $n$  to the sphere. A comparison of the present results with published solutions [6] shows a good agreement.

Figures 3 show the convergence behaviour of PCGS (figure 3a for ILU preconditioning and figure 3b for MG preconditioning) for  $d_1 = d_2 = d$ . We must mention that CGS (i.e. the non-preconditioned algorithm) converges only on  $N = 5$  and  $N = 9$  meshes. Also, PCGS preconditioned with MG does not converge for  $2L/d = 11$  on the  $N = 129$  mesh. Based on extensive numerical experiments and on the results plotted in figures 3, we can note the following facts:

- The influence of the spheres spacing on the convergence rate of PCGS is significant;
- Both preconditionings improve the robustness of the algorithm;

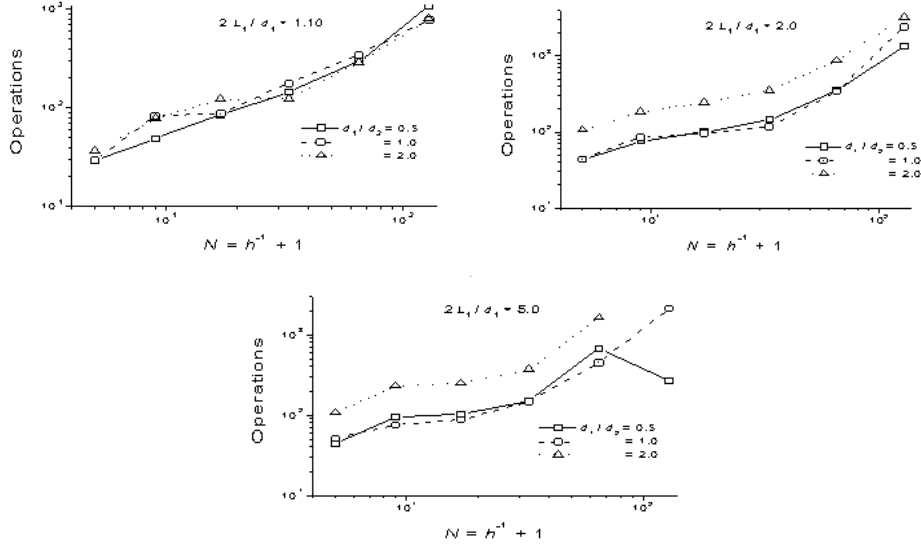


Fig. 4. Mesh behaviour of the convergence rate of PCGS for different spheres spacing and sizes; ILU preconditioning.

- Preconditioning by incomplete factorisation performs better for small values of  $2L/d$ ;

- PCGS preconditioned by MG(2) works very well for  $2L/d = 2$ .

The influence of the spheres sizes on the PCGS convergence rate is plotted in figures 4 and 5. From the data depicted in these figures it is difficult to draw a general conclusion concerning the influence of the diameters ratio on the convergence rate of PCGS. Figures 4 and 5 show that:

- for ILU preconditioning, the ratio  $d_1/d_2$  does not influence significantly the convergence rate at small spheres spacing (i.e.  $2L_1/d_1 = 1.10$ );
- for values of  $d_1/d_2$  greater than one and  $2L_1/d_1 \geq 2.0$ , the convergence rate of ILU preconditioning decreases;
- for MG preconditioning, we can note the decrease in the convergence rate for  $2L_1/d_1 = 2.0$  and  $d_1/d_2 = 0.5$  and the increase in the convergence rate for  $2L_1/d_1 = 5.0$  and  $d_1/d_2 = 0.5$ .

Numerical experiments with MG were made considering for the finest mesh  $N = 65, 129$  and  $257$ . The influence of the spheres spacing and sizes on the convergence rate of the MG algorithm is not significant. The values obtained for the average reduction factor fall inside the interval  $[0.8, 0.87]$ . Note that the average reduction factor for Gauss-Seidel iteration is greater than  $0.990$ . The values computed for the efficiency vary between  $255.0$  and  $410.0$ . It must be mentioned that for  $2L_1/d_1 = 5.0$ , MG does not converge on the finest mesh with  $N = 257$ .

For the comparison between PCGS and MG we consider that the solution is

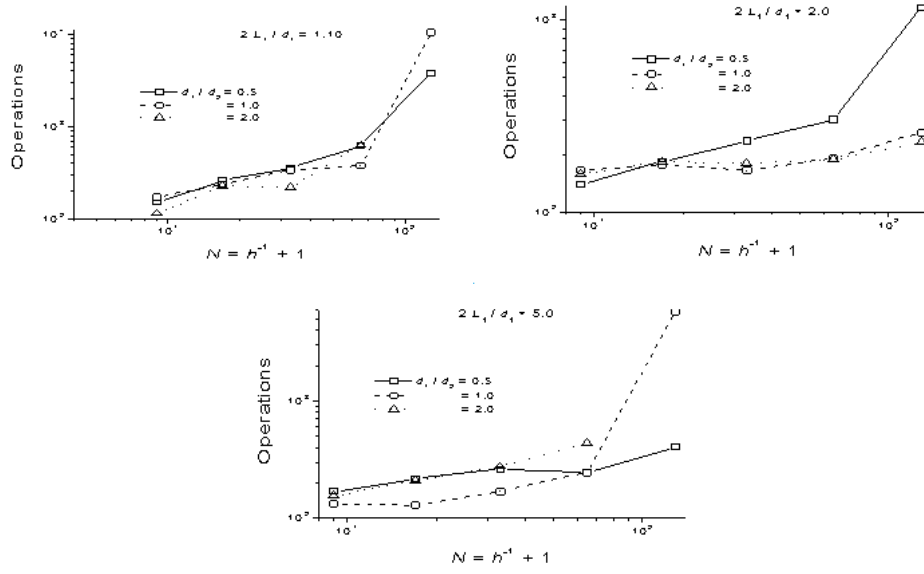


Fig. 5. Mesh behaviour of the convergence rate of PCGS for different spheres spacing and sizes; MG preconditioning.

desired on the  $N = 129$  mesh. Function of the values of  $2L_1/d_1$ ,  $d_1/d_2$  and the type of preconditioner, the efficiency of PCGS for  $N = 129$  is comprised between 180.2 and 14 000. The MG efficiency is that presented previously. Thus, for  $N = 129$ , PCGS is a competitor for MG only for  $2L_1/d_1 = 2.0$ ,  $d_1/d_2 \geq 1$  and MG (2) preconditioning.

## References

- [1] G.B. Jeffery, *On the steady rotation of a solid of revolution in a viscous fluid*, Proc. London Math. Soc., **14** (1915), 327.
- [2] M. Stimson, G.B. Jeffery, *The motion of two spheres in a viscous fluid*, Proc. R. Soc., **A11** (1926), 110.
- [3] A.J. Goldman, R.G. Cox, H. Brenner, *The slow motion of two identical arbitrarily oriented spheres through a viscous fluid*, Chem. Eng. Sci., **21** (1966), 1151.
- [4] S. Wakiya, *Slow motions of a viscous fluid around two spheres*, J. Phys. Soc. Japan, **22** (1967) 1101.
- [5] M.D.A. Cooley, M.E. O'Neill, *On the slow motion of two spheres in contact along their line of centres through a viscous fluid* Proc. Camb. Phil. Soc., **66** (1969), 407.

- [6] M.H. Davis, *The slow translation and rotation of two unequal spheres in a viscous fluid*, Chem. Eng. Sci., **24** (1969), 1769.
- [7] M.E. O'Neill, S.R. Majumdar, *Asymmetrical slow viscous fluid motion caused by the translation or rotation of two spheres, I + II.*, ZAMP, **21** (1970), 164.
- [8] E. Rushton, G.A. Davies, *The slow unsteady settling of two fluid spheres along their line of centres*, Appl. Sci. Res., **28** (1973), 37.
- [9] S. Haber, G. Hetsroni, A. Solan, *On the low Reynolds number motion of two droplets*, Int. J. Multiphase Flow, **1** (1973), 57.
- [10] G. Hetsroni, S. Haber, *Low Reynolds number motion of two drops submerged in an unbounded arbitrary velocity field*, Int. J. Multiphase Flow, **4** (1978), 1.
- [11] A.Z. Zinchenko, *The slow asymmetric motion of two drops in a viscous medium*, Prikl. Matem. Mekhan., **44** (1980), 49.
- [12] J. Happel, H. Brenner, *Low Reynolds Number Hydrodynamics*, Martinus, Nijhoff, 1983.
- [13] S. Kim, S. Karrila, *Microhydrodynamics. Principles and Selected Applications*, Butterworth-Heinemann, 1991.
- [14] J. Roumeliotis, *A boundary integral method applied to Stokes flow*, Ph. D. Thesis, University of New South Wales, Australia, 2000.
- [15] R. Tal (Thau), D.N. Lee, W.A. Sirignano, *Heat and momentum transfer around a pair of spheres in viscous flow*, Int. J. Heat Mass Transfer, **27** (1984), 1953.
- [16] P. Sonneveld, *CGS, A fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., **10** (1989), 36.
- [17] J.A. Meijerink, H.A. van der Vorst, *Guideliness for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems*, J. Comput. Phys., **44** (1981), 134.
- [18] D. Adam, *Mesh independence of Galerkin approach by preconditioning*, Int. J. Comput. Math., **58** (1989), 117.
- [19] C. Popa, *Mesh independence of the condition number of discrete Galerkin systems by preconditioning*, Int. J. Comput. Math., **51** (1994), 127.
- [20] C. Popa, *Preconditioning conjugate gradient method for non-symmetric systems*, Int. J. Comput. Math., **58** (1995), 117.
- [21] Gh. Juncu, I. Iliuta, *Preconditioned cg-like methods for solving non-linear convection-diffusion equations*, Int. J. Num. Meth. Heat Fluid Flow, **5** (1995), 239.
- [22] Gh. Juncu, *Preconditioning by approximations of the discrete Laplacian for 2-D non – linear free convection elliptic equations*, Int. J. Num. Meth. Heat Fluid Flow, **9** (1999), 586.

- [23] Gh. Juncu, C. Popa, *Preconditioning by approximations of the Gram matrix for convection-diffusion equations*, Mathematics and Computers in Simulations, **48** (1998), 225.
- [24] Gh. Juncu, C. Popa, *Numerical experiments with preconditioning by Gram matrix approximation for non-linear elliptic equations*, Mathematics and Computers in Simulations, **52** (2000), 253.
- [25] Gh. Juncu, C. Popa, *Preconditioning by Gram matrix approximation for diffusion-convection-reaction equations with discontinuous coefficients*, Mathematics and Computers in Simulation, **60** (2002), 487.
- [26] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comput., **33** (1977), 333.
- [27] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer, Berlin, 1985.



## Approach to nonstationary (transient) Birth-Death Processes

Alexei Leahu<sup>\*‡</sup>

In this paper we show how, in certain cases, nonstationary (transient) Birth-Death Processes (BDP) may be best approximated by their finite projection, which are stationary (recurrent positive) BDP. To illustrate, we consider nonstationary M/M/c Queueing Systems.

### 1. Introduction

Many mathematical models in Life Sciences, Physics, Computer Sciences, etc. are described by means of nonstationary BDP. Such processes are more difficult to approach than stationary BDP. Our aim is to find stationary BDP nearest (in certain sense) to the initial BDP.

### 2. Solution of the Problem

Let's consider BDP  $X = (X(t))_{t \in [0, +\infty)}$  defined on the probability space  $(\Omega, \mathcal{F}, \mathbf{P})$  with Birth intensities  $(\lambda_n)_{n \geq 0}$  and Death intensities  $(\mu_n)_{n \geq 0}$ . The set of states being  $E = \{0, 1, 2, \dots\}$ , we suppose that  $\mathbf{P}(X(0) = 0) = 1$ .

In order to achieve our aim, we observe that, according to the terminology of the book [1], BDP  $X$  may be interpreted as *regenerative process* with *renewal cycles*  $(Y_n)_{n \geq 0}$  (or *embedded renewal process*  $(S_n)_{n \geq 0}$ ) as the sequence of consecutive intervals  $Y_n$  between the moments  $S_0, S_1, S_2, \dots$  when BDP  $X$  falls into the state 0, i.e.,  $S_0 = Y_0 = 0$ ,  $S_n = \sum_{k=0}^n Y_k$  and  $(Y_n)_{n \geq 1}$  are nonnegative, independents, identically distributed random variables with distribution function  $F(y) = \mathbf{P}(Y_n <$

---

<sup>\*</sup> “Ovidius” University of Constanța, Romania, e-mail: [alexeleahu@univ-ovidius.ro](mailto:alexeleahu@univ-ovidius.ro)

<sup>‡</sup> Partially supported by Romanian Academy Grant 410/216, 2005.

$y) = \mathbf{P}(Y_1 < y)$ ,  $n \geq 1$ . According to the same terminology, in the case then  $\mathbf{P}(Y_n < +\infty) = 1$ ,  $n \geq 1$ , the processes  $X$  and  $S$  are recurrent regenerative and renewal processes respectively. From the point of view of real applications it is normally to consider that  $\mathbf{P}(Y_n = 0) \neq 1$ , i.e.  $X$  is a *regular* BDP. Because we deals with nonstationary BDP, in fact what means that  $S$  is a *stopped (transient) renewal process*, i.e.  $\mathbf{P}(Y_n < +\infty) = \lim_{y \nearrow +\infty} F(y) = F(+\infty) < 1$ ,  $n \geq 1$ . Then the corresponding process  $X$  will be called *stopped (transient or nonstationary) regenerative process* (BDP) and according to the [2]  $X(t)$  may be transcribed in this way:  $X(t) = X(t - S_{N(t)})$ , where  $N(t) = \max\{n : S_n \leq t\}$  is a *counting renewal process* corresponding to the process  $X$ .

Now we may construct finite projection  $X^*$  of the BDP  $X$  according to the definition of finite projection of regenerative processes [2], i.e., the process  $X^* = (X^*(t))_{t \in [0, +\infty)}$ , such that  $X^*(t) = X(t - S_{N^*}^*(t))$ , where  $S_0^* = Y_0^* = 0$ ,  $S_n^* = \sum_{k=0}^n Y_k^*$  and  $(Y_n^*)_{n \geq 1}$  are nonnegative, independents, identically distributed random variables with distribution function

$$F^*(y) = \mathbf{P}(Y_n^* < y) = \mathbf{P}(Y_n < y) / \mathbf{P}(Y_n < +\infty) = \mathbf{P}(Y_1 < y) / \mathbf{P}(Y_1 < +\infty), n \geq 1.$$

So,  $X^*$  is regenerative process too and, more than as, it is recurrent process because

$$F^*(+\infty) = \lim_{y \nearrow +\infty} F^*(y) = \mathbf{P}(Y_n^* < +\infty) = \mathbf{P}(Y_n < +\infty) / \mathbf{P}(Y_n < +\infty) = 1.$$

**Remark 1.** *The above mentioned definition of finite projection was constructive variant of definition, but we may use too the descriptive variant of the same definition [3] based on the*

**Proposition 1.** *Does exists unique probabilistic measure  $\mathbf{P}^*$  defined on the measurable space  $(\Omega, \mathcal{F})$  such that transient regenerative process  $X$  considered as a process defined on the probability space  $(\Omega, \mathcal{F}, \mathbf{P}^*)$  become a recurrent regenerative process.*

*Proof.* Indeed, as a consequence of the Ionescu-Tulcea's Theorem [4], it is sufficient to consider the family of finite dimensional distributions

$$\{\mathbf{P}^*(X(t_1) \in B_1, \dots, X(t_k) \in B_k, S_n \leq t_1 < \dots < t_k \leq S_{n+1})\}$$

for any borelian sets  $B_i \in \mathcal{B}(\mathbb{R})$ ,  $t_i \in [0, +\infty)$ ,  $i = \overline{1, k}$ ,  $t_1 < \dots < t_k$ ,  $k \geq 1$ ,  $n \geq 0$ , where

$$\begin{aligned} & \mathbf{P}^*(X(t_1) \in B_1, \dots, X(t_k) \in B_k, S_n \leq t_1 < \dots < t_k \leq S_{n+1}) = \\ & = \mathbf{P}(X(t_1) \in B_1, \dots, X(t_k) \in B_k, S_n \leq t_1 < \dots < t_k \leq S_{n+1} / S_{n+1} < +\infty). \quad \square \end{aligned}$$

**Descriptive definition of finite projection.** *Probability  $\mathbf{P}^*$  determined by Proposition 1 will be called finite projection of probability  $\mathbf{P}$  with respect to the regenerative process  $X$  and stochastic process  $X^*$  defined by probability distribution  $\mathbf{P}^*$  will be called finite projection of regenerative process  $X$ .*

So, trajectories of the finite projection  $X^*$  are the trajectories of the regenerative process  $X$  observed under condition that corresponding renewal moments  $S_1, S_2, \dots$  (or renewal intervals  $Y_1, Y_2, \dots$ ) are finite. This justify the name “finite projection” as a generic name for a special class of conditioned processes.

The following Theorem show us that in fact finite projection  $X^*$  of BDP is BDP too.

**Theorem 1** [3]. *Finite projection  $X^*$  of transient BDP  $X$  with Birth intensities  $(\lambda_n)_{n \geq 0}$  and Death intensities  $(\mu_n)_{n \geq 0}$  is BDP with Birth intensities  $(\lambda_n^*)_{n \geq 0}$  and Death intensities  $(\mu_n^*)_{n \geq 0}$ , where  $\lambda_0^* = \lambda_0$ ,  $\lambda_n^* = \lambda_n \rho_{n+1}$  and  $\mu_n^* = \mu_n / \rho_n$ ,  $n \geq 1$  and*

$$\rho_n = \mathbf{P}\{\text{Sojourn time of the process } X \text{ in the set of states } \{n, n+1, \dots\} \text{ is finite}\}$$

*may be calculated as continued fraction*

$$\rho_n = \frac{\mu_n}{\mu_n + \lambda_n - \frac{\lambda_n \mu_{n+1}}{\mu_{n+1} + \lambda_{n+1} - \frac{\lambda_{n+1} \mu_{n+2}}{\mu_{n+2} + \lambda_{n+2} - \dots}}}, \quad n \geq 1. \quad (1)$$

**Remark 2.** *Finite projection  $X^*$  of transient BDP  $X$ , being BDP too, is recurrent BDP. Counterexample given in the paper[3] confirms that above formulated Theorem does not guarantee that finite projection  $X^*$  is always recurrent positive (ergodic or stationary) BDP. So, we need to know the conditions of stationarity for BDP  $X^*$ . Such conditions are given in the above mentioned paper by the following*

**Proposition 2.** *For finite projection  $X^*$  of transient (nonstationary) BDP  $X$  with Birth intensities  $(\lambda_n)_{n \geq 0}$  and Death intensities  $(\mu_n)_{n \geq 0}$  to be recurrent positive (ergodic or stationary) BDP it is sufficient to be satisfied one of the following conditions:*

- a)  $\sup_n \rho_n \neq 1$  and  $\inf_n \lambda_n \neq 0$ ;
- b)  $\sup_n \rho_n \neq 1$  and  $\inf_n \mu_n \neq 0$ ;
- c)  $\sum_{n=1}^{\infty} (\lambda_n - \lambda_n \rho_n)^{-1} < +\infty$ ;
- d)  $\sum_{n=2}^{\infty} (\mu_n - \mu_n \rho_{n-1})^{-1} < +\infty$ ,

where  $\rho_n$ ,  $n \geq 1$  are determined by formula (1).

Now, let's introduce and comment the main mathematical objects to make understandable the meaning of the words “Transient (nonstationary) BDP  $X$  may best approximated by its finite projection  $X^*$ ”. First of all, let's denote by  $\mathcal{F}(n)$  the  $\sigma$ -algebra generated by transient BDP  $X$  until the moment  $S_n$ ,  $n \geq 1$ , and  $\mathcal{P} = \{ \mathbf{Q} \mid \mathbf{Q} \text{ is a probability defined on the measurable space } (\Omega, \mathcal{F}) \}$ . Evidently, function

$$d : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}_+, \quad d(\mathbf{P}_1, \mathbf{P}_2) = \sum_{n=1}^{\infty} \sup_{A \in \mathcal{F}(n)} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| / n!$$

posses all properties of distance between probability measures  $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{P}$ . That means  $(\mathcal{P}, d)$  is a metric space.

In add, let's introduce subfamily

$$\mathcal{PR} = \{ \mathbf{Q} \in \mathcal{P} \mid \mathbf{Q}(S_n < +\infty) = 1, n \geq 1 \}.$$

So,  $\mathcal{PR}$  consists of all probability measures  $Q$ , including finite projection  $\mathbf{P}^*$  of probability  $\mathbf{P}$ , such that BDP  $X$  became recurrent BDP with respect to them.

**Theorem 2.** *Finite projection  $\mathbf{P}^*$  of probability  $\mathbf{P}$  is one of the best approximation for probability distribution  $\mathbf{P}$  of transient BDP  $X$  by means of elements from  $\mathcal{PR}$  in sense that*

$$\min_{\mathbf{Q} \in \mathcal{PR}} d(\mathbf{P}, \mathbf{Q}) = d(\mathbf{P}, \mathbf{P}^*) = e - \rho_1 e^{\rho_1},$$

where

$$\rho_1 = \mathbf{P}(S_1 < +\infty) = \mathbf{P}(Y_1 < +\infty) =$$

$$\mathbf{P}\{\text{Sojourn time of the process } X \text{ in the set of states } \{1, 2, \dots\} \text{ is finite}\}$$

and may be calculate by formula (1).

*Proof.* Let's consider  $\mathbf{Q} \in \mathcal{PR}$  and  $A_n = \{S_n = \infty\} \in \mathcal{F}(n)$ ,  $n \geq 1$ , then

$$\begin{aligned} d(\mathbf{P}, \mathbf{Q}) &= \sum_{n=1}^{\infty} \sup_{A \in \mathcal{F}(n)} |\mathbf{P}(A) - \mathbf{Q}(A)| / n! \geq \\ &\geq \sum_{n=1}^{\infty} |\mathbf{P}(A_n) - \mathbf{Q}(A_n)| / n! = \sum_{n=1}^{\infty} \mathbf{P}(A_n) / n! = \\ &= \sum_{n=1}^{\infty} (1 - \mathbf{P}(S_n < \infty)) / n! = \sum_{n=1}^{\infty} (1 - \rho_1^n) / n! = e - \rho_1 e^{\rho_1}, \end{aligned}$$

because  $S_n = \sum_{k=0}^n Y_k$  and the fact that  $(Y_n)_{n \geq 1}$  are nonnegative, independents, identically distributed random variables with distribution function

$$F(y) = \mathbf{P}(Y_n < y) = \mathbf{P}(Y_1 < y)$$

imply

$$\{S_n < \infty\} \iff \{Y_1 < \infty, \dots, Y_n < \infty\},$$

i.e.,

$$\mathbf{P}(S_n < \infty) = \mathbf{P}(Y_1 < \infty, \dots, Y_n < \infty) = \rho_1^n,$$

where

$$\rho_1 = \mathbf{P}(S_1 < +\infty) = \mathbf{P}(Y_1 < +\infty) =$$

$$= \mathbf{P}\{\text{Sojourn time of the process } X \text{ in the set of states } \{1, 2, \dots\} \text{ is finite}\}$$

and may be calculated by formula  $\rho_1 = \mathbf{P}(S_1 < +\infty) = \mathbf{P}(Y_1 < +\infty) = (1)$ .

On the other hand,  $\mathbf{P}^* \in \mathcal{PR}$  and for  $A \in \mathcal{F}(n)$

$$\begin{aligned} |\mathbf{P}(A) - \mathbf{P}^*(A)| &= \\ &= |\mathbf{P}(S_n < \infty) \mathbf{P}\{A/S_n < \infty\} + \mathbf{P}(S_n = \infty) \mathbf{P}\{A/S_n = \infty\} - \mathbf{P}^*(A)| = \\ &= \mathbf{P}(S_n = \infty) |\mathbf{P}\{A/S_n = \infty\} - \mathbf{P}^*(A)| \leq \mathbf{P}(S_n = \infty) = 1 - \rho_1^n. \end{aligned}$$

So,  $d(\mathbf{P}, \mathbf{P}^*) \leq e - \rho_1 e^{\rho_1}$  and finally

$$\min_{\mathbf{Q} \in \mathcal{PR}} d(\mathbf{P}, \mathbf{Q}) = d(\mathbf{P}, \mathbf{P}^*) = e - \rho_1 e^{\rho_1}. \quad \square$$

**Corollary.**  $d(\mathbf{P}, \mathbf{P}^*) \rightarrow 0$ , when  $\rho_1 \rightarrow 1$ .

**Example.** To illustrate our theoretical considerations let's take, as example, nonstationary  $M/M/c$  Queueing Systems widely applied in Computer Science. That means flow of the arrivals is Poissonian with parameter  $\lambda > 0$  ("parameter of Birth"), service time of each customer on the each of  $c$  servers is exponentially distributed random variable with parameter  $\mu > 0$  ("parameter of Death"). So, interpreting  $X(t)$  as the total number of customers in the system at the moment  $t$ , our system may be described by nonstationary BDP  $X = (X(t))_{t \in [0, +\infty)}$  with Birth intensities  $(\lambda_n)_{n \geq 0}$  and Death intensities  $(\mu_n)_{n \geq 0}$ , where  $\lambda_n = \lambda, \forall n \geq 0$ , and

$$\mu_n = \begin{cases} n\mu, & \text{if } 1 \leq n \leq c, \\ c\mu, & \text{if } n > c. \end{cases}$$

According to the Theorem 2, the best approximation for  $M/M/c$  Queueing System described by nonstationary BDP  $X$  is the  $(M/M/c)^*$  Queueing Systems described by finite projection  $X^*$  of process  $X$ . By definition  $(M/M/c)^*$  will be called finite projection of the  $M/M/c$  Queueing System.

On the base of Theorem 1 and Proposition 2 we deduce the following

**Proposition 3.** *Finite projection of nonstationary  $M/M/c$  Queueing System is a stationary  $(M/M/c)^*$  Queueing System described by BDP  $X^*$  defined by parameters*

$$\lambda_n^* = \begin{cases} \lambda, & \text{if } n = 0, \\ \lambda \rho_{n+1}, & \text{if } 1 \leq n < c-1, \\ c\mu, & \text{if } n \geq c-1, \end{cases} \quad \mu_n = \begin{cases} n\mu/\rho_n, & \text{if } 1 \leq n < c, \\ \lambda, & \text{if } n \geq c, \end{cases}$$

where

$$\rho_n = \begin{cases} \eta_n(1 + \eta_n)^{-1}, & \text{if } 1 \leq n < c, \\ c\mu/\lambda, & \text{if } n \geq c, \end{cases}$$

and

$$\eta_n = \sum_{i=n}^{c-1} \frac{i!}{(n-1)!} \left(\frac{\mu}{\lambda}\right)^{i-n+1} + \frac{c!}{(n-1)!} \left(\frac{\mu}{\lambda}\right)^{c-n+1} \frac{1}{1 - n\mu/\lambda}.$$

In conclusion, research of the nonstationary  $M/M/c$  Queueing Systems being difficult, may be reduced to the research of stationary  $(M/M/c)^*$  Queueing Systems which are more simple to investigate due to their stationarity. Quality of this approximation depends of parameter  $\rho_1$  and it is increasing with increasing  $\rho_1 \nearrow 1$ .

## References

- [1] Asmussen, S., *Applied Probability and Queues*, Wiley, New York, 1987.

- [2] Leahu, A., Matveev, V., *On finite projection of regenerative processes*, Bulet. Acad. Sci. Rep. Moldova, Ser. Math., **3** 1991, 74–77.
- [3] Leahu, A., Hlibiciuc, A., *Finite projection and dual of Birth-Death processes*, Bulet. Acad. Sci. Rep. Moldova, Ser. Math., **3** (1994), 5–14.
- [4] Loève, M., *Probability Theory*, Van Nostrand, Princeton, 1960.

## Statistical simulation and analysis of some software reliability models

Alexei Leahu<sup>\*†‡</sup> and Elena Carmen Lungu<sup>\*‡</sup>

In our work we develop statistical simulation and analysis of some software reliability models in the case when initial number of errors is constant or random number. Some theoretical and practical aspects of the problems will be discussed.

### 1. Introduction

The computer revolution is fueled by an ever more rapid technological advancement. Today, computer hardware and software permeates our modern society. Computers are embedded in telephones, home appliances, buildings, automobiles and aircraft. Science and technology demand high-performance hardware and high-quality software for making improvements and breakthroughs. Software reliability is generally accepted as the key factor in software quality since it quantifies software failure—which can make a powerful system inoperative or, even, deadly.

Concerning software reliability, basically, the approach is to apply mathematics and statistics to model past failure data to predict future behavior of a component or system. Our aim is to illustrate power of the statistical simulation method in the comparative analysis of some software reliability mathematical models and verification of maximum likelihood procedure of statistical estimation. Of course, other estimation procedures are also applicable (e.g., method-of-moments, method-of-least-squares), but in this work we'll concentrate on maximum likelihood estimates because their many desirable properties (e.g., asymptotic normality, asymptotic efficiency and invariance).

---

<sup>\*</sup> “Ovidius” University of Constanța, Romania.

<sup>†</sup> e-mail: [alexeleahu@univ-ovidius.ro](mailto:alexeleahu@univ-ovidius.ro)

<sup>‡</sup> Partially supported by Romanian Academy Grant 12/25.07.2005.

## 2. Solution of the Problem

In our work we take, as example, some variants of Jelinski-Moranda (**JM**) models. More exactly, we propose to analyze the software reliability models based on the following hypotheses:

1. The total number  $N$  of errors existing initially in the software is a constant (unknown) number or variant

1'. The total number  $N$  of errors existing initially in the software is a Poisson distributed random variable (r.v.) with parameter  $a > 0$ , i.e.,  $P(N = n) = \frac{a^n}{n!}e^{-a}$ ,  $\forall n = 0, 1, 2, \dots$ ;

2. Each error is eliminated with probability  $p = 1$ , independently of the past trials, repair of the error being snapshot and without introduction of the new errors or variant

2'. Each error is eliminated with probability  $p$ ,  $0 < p < 1$ , independently of the past trials, repair of the error being snapshot and without introduction of the new errors;

3. The time intervals between two successive failures of the software are independent exponentially distributed random variables with parameter  $\mu > 0$  or variant

3'. The time intervals between two successive failures of the software are independent Erlang distributed random variables with parameters  $r$  and  $\mu > 0$ ;

4. Distribution's parameter of the interval between two successive failures of the software, i.e. rate or intensity of the failures is directly proportional with the number of non eliminated errors in the software at the begin of this interval.

So, we have (**JM**)<sub>1</sub> model if the hypotheses 1, 2, 3, 4 are valid [1], (**JM**)<sub>2</sub> model if the hypotheses 1, 2', 3, 4 are valid [1], (**JM**)<sub>3</sub> model if the hypotheses 1, 2', 3', 4 are valid and (**JM**)<sub>4</sub> model if the hypotheses 1', 2', 3', 4 are valid.

For beginning we present some results of Monte-Carlo simulations in order to estimate probability distribution of the number  $S_n(t)$  of non eliminated errors at the moment  $t$ ,  $t \geq 0$ , by means of relative frequencies  $fm(S_N(t) = k)$ ,  $k = 0, 1, 2, \dots$

### CASE I. Model (**JM**)<sub>4</sub>

Table 1

$a = 1, \mu = 2, p = 0.5, t = 1$

$r$	$k$	0	1	2	3	4
1	$fm(S_N(t) = k)$	0,692	0,255	0,047	0,006	$5,58 \cdot 10^{-4}$
2	$fm(S_N(t) = k)$	0,82	0,173	0,007	$1,22 \cdot 10^{-4}$	$4,00 \cdot 10^{-6}$
3	$fm(S_N(t) = k)$	0,93	0,067	$3,38 \cdot 10^{-4}$	0	0
4	$fm(S_N(t) = k)$	0,982	0,018	$8,00 \cdot 10^{-6}$	0	0

Table 2

$a = 1, \mu = 2 \cdot r, p = 0.5, t = 1$

$r$	$k$	0	1	2	3	4
1	$fm(S_N(t) = k)$	0,692	0,255	0,047	0,0058	$5,58 \cdot 10^{-4}$
2	$fm(S_N(t) = k)$	0,897	0,101	0,0024	$1,80 \cdot 10^{-5}$	0
3	$fm(S_N(t) = k)$	0,97	0,03	$5,40 \cdot 10^{-5}$	0	0
4	$fm(S_N(t) = k)$	0,993	0,007	$2,00 \cdot 10^{-6}$	0	0



**Remark 1.** Table 1 show us how look the probabilities  $\mathbf{P}(S_N(t) = k)$  (or relative frequencies  $fm(S_N(t) = k), k = 0, 1, 2, \dots$ ) depending of  $r = \overline{1, 4}$ , i.e., when mean value of interval between two successive excluded errors, being equal with  $r/\mu = r/2$ , increases. In the same way, table 2 show us the behavior of probabilities  $\mathbf{P}(S_N(t) = k)$  (or relative frequencies  $fm(S_N(t) = k), k = 0, 1, 2, \dots$ ) depending of  $r = \overline{1, 4}$ , i.e., when mean value of interval between two successive excluded errors, being equal with  $r/\mu$ , remain the same:  $r/\mu = r/2r = 1/2$ . Remark 1 include also the Case II and in the both cases we have number of trials  $m = 500\,000$  (Error  $\pm 10^{-2}$  with level of confidence 0.99).

### CASE II. Model (JM)<sub>3</sub>

Table 1

$\mu = 2, p = 0.5, t = 1, N = 5$

$r$	$k$	0	1	2	3	4	5
1	$fm(S_N(t) = k)$	0, 101	0, 294	0, 342	0, 199	0, 058	0, 007
2	$fm(S_N(t) = k)$	$8 \cdot 10^{-4}$	0, 0422	0, 217	0, 398	0, 277	0, 065
3	$fm(S_N(t) = k)$	$6 \cdot 10^{-6}$	0, 00235	0, 065	0, 328	0, 444	0, 164
4	$fm(S_N(t) = k)$	0	$8, 2 \cdot 10^{-5}$	0, 013	0, 204	0, 514	0, 27

Table 2

$\mu = 2 \cdot r, p = 0.5, t = 1, N = 5$

$r$	$k$	0	1	2	3	4	5
1	$fm(S_N(t) = k)$	0, 101	0, 294	0, 342	0, 199	0, 058	0, 007
2	$fm(S_N(t) = k)$	0, 056	0, 306	0, 393	0, 199	0, 044	0, 003
3	$fm(S_N(t) = k)$	0, 038	0, 31	0, 415	0, 196	0, 039	0, 003
4	$fm(S_N(t) = k)$	0, 03	0, 311	0, 426	0, 194	0, 036	0, 002

In order to verify experimentally maximum likelihood procedure of statistical estimation let us remember that in our algorithm of Monte-Carlo simulation we use fundamentally the following proposition proved in the paper [2].

**Proposition 1.** If  $(X_k)_{k \geq 1}$  are i.i.d.r.v. such that  $X_k \sim \exp(\mu)$ ,  $\mu > 0$ ,  $k = 1, 2, \dots$  and  $K$  is a r.v geometrically distributed with parameter  $p$ ,  $0 < p \leq 1$ , independently of  $(X_k)_{k \geq 1}$ , then  $X_1 + X_2 + \dots + X_K \sim \exp(\mu \cdot p)$ , i.e., is exponentially distributed r.v. with parameter  $\mu \cdot p$ .

Let's us consider that during the time interval  $T$ ,  $T > 0$ , of error detections and their eliminations we observes intervals of length  $t_1, t_2, \dots, t_n$ , where  $n$  is the total number of eliminated errors until the moment  $T$ . In this case, as a consequence of above formulated Proposition 1 we have the following

**Proposition 2.** The likelihood function  $L(t_1, t_2, \dots, t_n; \mu_0, N)$  for (JM)<sub>2</sub> model is the same as the likelihood function for (JM)<sub>1</sub> model with the parameter  $\mu_0$ , where  $\mu_0 = \mu \cdot p$ .

That means likelihood equations to be solve in the model (JM)<sub>2</sub> for prediction of initial number of errors  $N$  (remainder number of errors  $N - S_N(T)$ ) are the same as for model (JM)<sub>1</sub>, i.e. (see [1]),

$$\begin{cases} \frac{\partial \ln L}{\partial N} = \sum_{i=1}^n \frac{1}{N-i+1} - \mu_0 \sum_{i=1}^n t_i = 0, \\ \frac{\partial \ln L}{\partial \mu_0} = \frac{n}{\mu_0} - \sum_{i=1}^n t_i (N-i+1) = 0. \end{cases} \quad (1)$$

Below, in the Case III, we present some results of simulations which reflects possibility of prediction for parameters  $\mu_0, N$ .

**CASE III. Model (JM)<sub>2</sub>**

$$\mu_0 = \mu \cdot p = 2, T = 3$$

*Table 1*  
 $p = 1$

N	$\hat{\mu}_0$	$\hat{N}$
2	4.51	2.79
3	1.23	3.30
4	2.03	3.85
5	2.77	4.68
6	1.45	5.56
7	1.97	7.60
8	1.42	8.02
9	1.91	8.50
10	2.65	9.32

*Table 2*  
 $p = 1/2$

N	$\hat{\mu}_0$	$\hat{N}$
2	1.06	2.11
3	.07	2.59
4	1.2	4.34
5	0.85	5.28
6	1.42	6.12
7	0.92	6.94
8	1.27	8.01
9	0.82	8.69
10	0.99	10.27

*Table 3*  
 $p = 1/4$

N	$\hat{\mu}_0$	$\hat{N}$
2	$\emptyset$	$\emptyset$
3	0.50	3.48
4	0.56	3.61
5	0.65	5.36
6	0.69	6.57
7	0.53	7.35
8	0.67	8.27
9	0.72	9.17
10	0.8	10.54

### 3. Conclusions

Likelihood equations (1) give as sufficiently good estimator for prediction of initial number of errors  $N$  (or remaining number of errors) but estimations of parameter  $\mu_0 = \mu \cdot p$  is not acceptable, i.e., in order to estimate  $\mu_0$  we need to change design of the experiment.

### References

- [1] Vaduva, I., *Fiabilitatea programelor*, Editura Univiversității București, 2003.
- [2] Leahu, A., Lupu, E.C., *Software's Reliability with the Random Number of Errors*, Proceedings of Workshops Mathematical Modeling of Environmental and Life Sciences Problems, Editura Acadademeiei Române, 2004, pp. 167–172.

## **A Mathematical Model Describing the Vulnerability to Pollution of Groundwater in the Proximity of Slatina Town**

**Anca Marina Marinov\*** and **Victor Moldoveanu\*\***

This work deals with groundwater quality forecasting in the proximity of **Olt** River (pumping system for the town “Slatina” water supply). Using geomorphologic data and measured levels in the wells from the region, we predict the vulnerability of groundwater to pollution. We calculate the discharge lost by drainage between the two layers existing in this area, the influence of pumped discharges on the direction of the flow and on the increase vulnerability to pollution. Our work is based on a mathematical model describing the water advance in a saturated porous soil. For a two-dimensional system a steady flow in a homogeneous and isotropic porous layer with constant thickness is considered for two cases: a confined aquifer and an unconfined one. The flow system includes a lateral flow from the natural limit of the aquifers, the Olt River and the pumped wells existing in the proximity of the river. The path lines for individual fluid particles through the flow system provide the points of emergence at outflow boundaries are determined by the path function. The stream function gives the flux rate at the outflow boundary as well as throughout the entire flow system. Our results give important information regarding the vulnerability of groundwater to pollution in the vicinity of Olt River.

### **1. Introduction**

Groundwater constitutes an important component of many water resource systems. Due to good purification properties of the soil, groundwater is generally a very good source of drinking water.

---

\* “Politehnica” University of Bucharest; “Gheorghe Mihoc–Caius Iacob” Institute of Mathematical Statistics and Applied Mathematics.

\*\* S.C. CONFORD PROD S.R.L.

The quantity and quality problems cannot be separated. The drinking water for Slatina town is pumped from the groundwater. A big number of drinking wells pump the water from the phreatic aquifer and from the confined aquifer. The wells are disposed in lines on the two borders of the river Olt.

The natural river Olt has been modified and as a result, near Slatina town, the river became the lakes Arcești and Slatina-Slătioara. The two lakes are connected to the groundwater.

Water is pumped from well lines displayed in Table 1 which contains the total number of wells in use and the pumped discharges in each well line. The wells are drilled in the two aquifers existing around the river Olt.

*Table 1*  
Pumped discharges for Slatina town

Pumping line	Execution year	Total number of wells (in use)	Discharge [l/s]
Salcia-Slătioara	1977–1978, 1987, 1990	39(25)	113
Noua	1981–1982	13(12)	50
„B”	1987–1988	32(22)	75
Zăvoi	1986	4(3)	15
„D”	1987	4(3)	16
Curtișoara - phreatic	1974–1975	44(34)	118
Curtișoara - confined aquifer	1974–1975	24(18)	74
<b>Total</b>		160(117)	460

## 2. Hydrogeological Conditions

A cross section through the two aquifers is presented in Fig. 1 [3].

The first layer-the phreatic aquifer – with a 5–12 m depth, with alluvium materials (coarse sand and gravel) – is in connection to the lakes. The transmissivity has values between 50 and 1 000 m<sup>2</sup>/day. The wells are drilled in the phreatic aquifer in Curtișoara-Teslui area.

The confined aquifer is made up of several layers whose heights are of 5–11m, having average hydraulic transmissivity  $T = 350$  m<sup>2</sup>/day and the hydraulic gradient  $I = 1.2$ –10 [3].

The confined aquifer storages water under pressure (Fig. 1) in the Căndești layers (15–160m). Between the phreatic and the confined aquifer there is an aquitard with low hydraulic conductivity values and 2–25 m thickness values. The hydraulic vertical conductivity of the aquitard is approximately  $K' = 2.5$ – $3.0 \times 10^{-4}$  m/day with some local values of  $5 \times 10^{-2}$  m/day. The aquitard drainage parameter is  $K'/M' = 1.5$ – $3.0 \times 10^{-5}$  day<sup>-1</sup>.

The well lines are made up of small depth wells, of 10–15 m (Curtișoara line),

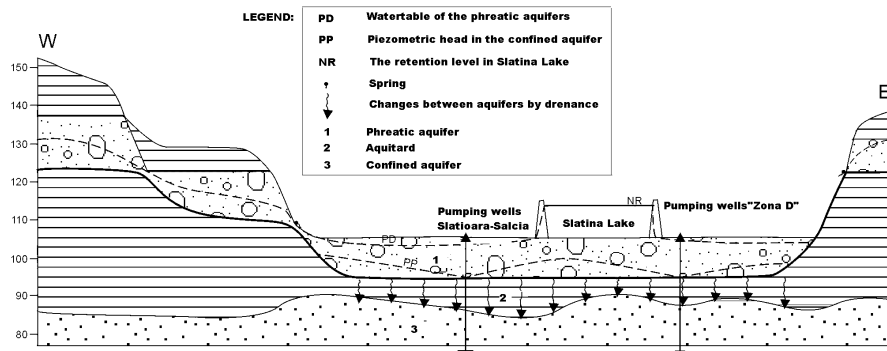


Fig. 1. Hydrogeological section showing local groundwater flow system in Slătioara-Salcia area.

average depth wells of 50–120 m (Curtișoara-Teslui, Salcia-Slătioara), as well as great depth ones.

Geo-morphologically, the pumping area is characterised by the presence of river Olt meadow and terraces.

The hydropower planning of the downstream area of river Olt has imposed the formation of Străjești, Arcești, and Slatina lakes. Their presence has altered the natural groundwater flow. The phreatic aquifer is hydraulically connected to the hydrographic net elements in the area. The water supply of the phreatic aquifer is mainly achieved by means of rainfall and the discharge from the hydrographic net components whose breath surfaces are situated at higher quotas than the water table of the aquifer.

In the dam walls there are some impervious less spaces allowing water to flow from the lake to the groundwater and vice versa.

### 3. Mathematical Modeling of Groundwater Flow in the Slatina Area

The objectives of modeling this area are:

- understanding the effect of pumping on the hydrologic flow regime;
- assessing the feasibility of increased pumping in time, especially in terms of water quality.

The groundwater flow simulation consists of the prediction of quantities of interest based upon an equation or series of equations that describe system behaviour under a set of assumed simplifications. A numerical method that approximates the governing PDE using the finite element code is considered.

The aims of our study are:

- Investigating and explaining the flow patterns at different layers;

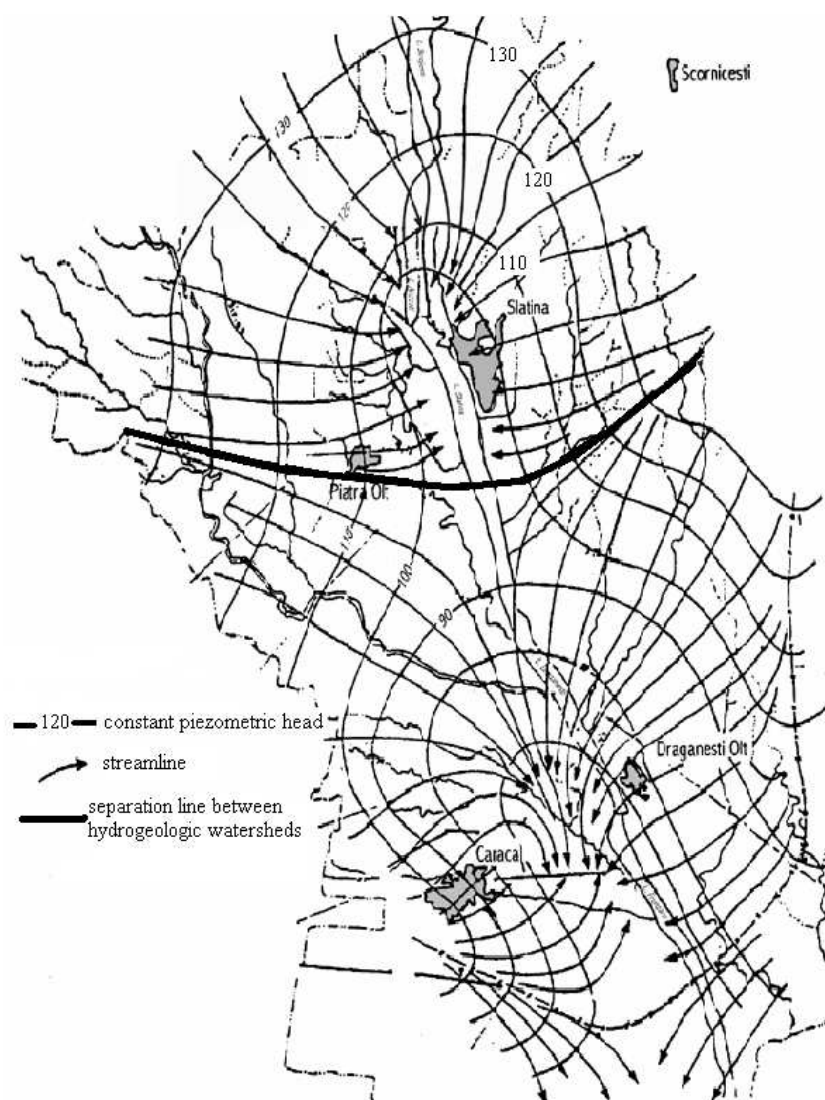


Fig. 2. The measured equipotential lines in hydrologic watersheds: Slatina and Caracal.

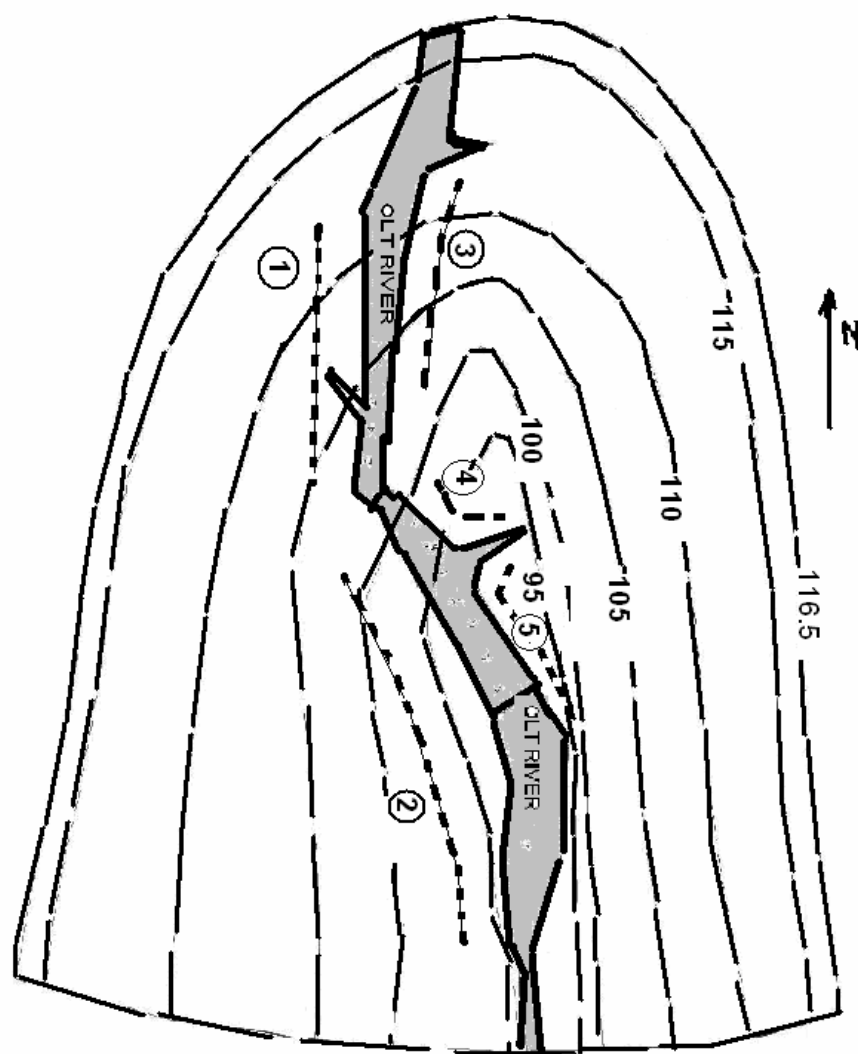


Fig. 3. The measured equipotential lines (constant head) for the confined aquifer Slatina, and wells' location: 1= Area „B”; 2= Salcia-Slătioara; 3= Curtișoara – average depth wells; 4= Area „D”; 5= Area „Noua”, Area „Zvoi”.

- Analyzing how the flow pattern differs with discharges in wells;
- Analyzing where the groundwater discharges into the river;
- Identifying the flux coming in/out of the river or of the boundaries;
- Identifying the flux coming in/out of the phreatic aquifer to the confined aquifer or vice versa.

The groundwater flow can be described by combining two equations. By introducing into the continuity equation the velocity given by Darcy's law we obtain the equation of diffusivity [2]

$$\operatorname{div}(\overline{T} \operatorname{grad} h) = S \frac{\partial h}{\partial t} + Q. \quad (1)$$

$\overline{T}$  [ $L^2 T^{-1}$ ] is the transmissivity tensor defined by

$$\overline{T} = \int_b^a \overline{K} dz, \quad (2)$$

componentwise:

$$T_{xx} = \int_b^a K_{xx} dz, T_{yy} = \int_b^a K_{yy} dz, T_{zz} = \int_b^a K_{zz} dz, \quad (3)$$

where  $a, b$  [L] are the layer's limits;  $K_{xx}, K_{yy}, K_{zz}$  [ $LT^{-1}$ ] are the hydraulic conductivity in directions  $x, y$  and  $z$ .

$S$  [ $L^3 / L^3$ ] is the storage coefficient or storativity (equal to the effective porosity for the phreatic aquifer);

$$S = \int_b^a S_s dz, \quad (4)$$

$h$  [L] is the hydraulic head (piezometric head),  $h = \frac{p}{\rho \cdot g} + z$ ;

$Q$  [ $LT^{-1}$ ] is the discharge incoming into an elementary volume on the unit surface;

$S_s$  [1/L] is the specific storage of a saturated aquifer (the volume that a unit volume of aquifer released from storage for a unit decline in head).

For a homogeneous and isotropic confined aquifer the diffusivity equation becomes:

$$\operatorname{div}(\operatorname{grad} h) = \frac{S}{T} \cdot \frac{\partial h}{\partial t} + \frac{Q}{T} \quad (5)$$

or

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} + \frac{\partial^2 h}{\partial z^2} = \frac{S}{T} \cdot \frac{\partial h}{\partial t} + \frac{Q}{T}. \quad (6)$$

The ratio  $\frac{T}{S}$  is called the aquifer's diffusivity.

For the phreatic aquifers the storage is primarily done by filling up and draining of pores. The storage coefficient is therefore equivalent to the storage effective porosity  $n_c$  and thus the diffusivity equation becomes:



$$\frac{\partial}{\partial x} \left( K(h - h_s) \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left( K(h - h_s) \frac{\partial h}{\partial y} \right) = n_c \frac{\partial h}{\partial t} + Q. \quad (7)$$

$h(x, y, t)$  [L] is the groundwater surface level,  $h_s(x, y)$  [L] is the level of the aquitard,  $K(x, y)$  [L/T] is the hydraulic conductivity,  $n_c(x, y)$  is the storage-effective porosity and  $Q(x, y, t)$  is the source term.

We consider separately the phreatic aquifer (Fig. 3) and the confined one (Fig. 2). The steady flow is considered.

For the confined aquifer, the diffusivity equation (6) is integrated using the boundary conditions which characterise the flow in Slatina watershed. Fig. 2 contains equipotential lines drawn using measured levels in the observation wells. This figure gives us the boundary conditions for our problem. Similarly, for the phreatic aquifer the diffusivity equation, (7) is integrated using the boundary conditions from Fig. 3.

For the equation integration, an executable code is used. The code is based on the finite element method and is used separately for the phreatic aquifer on the one hand and for the confined one on the other.

In the pre-processing stage the necessary data is introduced into the program: the flow type (steady or unsteady); the aquifer type (phreatic or confined); the domain's dimensions; pumping wells' position; discharge in each well; network step dimension (triangles); finite elements network generation.

Using the hydro-geological description in Fig. 1 we can choose the following attributes of

- **elements**: transmissivity, ratio between secondary and main direction of transmissivity, storage coefficient;
- **nodes**: initial head, aquifer thickness.

The boundary conditions can be: "imposed head" or "imposed discharge".

After the integration, the post-processing stage allows the drawing of equal piezometric head lines in the aquifer, the streamlines, the velocity vectors (module and direction). For the steady flow, the streamlines coincide with the pathlines. On each streamline the values of travel time is plotted.

Using the real boundary conditions we change the pumped discharges in the wells to obtain the piezometric head lines with the same aspect as the real ones. The difference between the real pumped discharge and the calculated one is the lost discharge from an aquifer to the other. The flow between the aquifers is possible by draining through the aquitard when the piezometric head is different in the phreatic and the confined aquifer.

#### 4. Flow Analysis in the Confined Aquifer

Using the measured data in the observation wells the equal-piezometric head lines are plotted for the confined aquifer. From these equal head lines we have chosen the boundary conditions for the aquifer.

The analyzed area has: the length  $L = 21\,000$  m, the width  $l = 21\,000$  m, the thickness  $M = 10$  m, the hydraulic conductivity  $K = 30$  m/day, the transmissivity  $T$

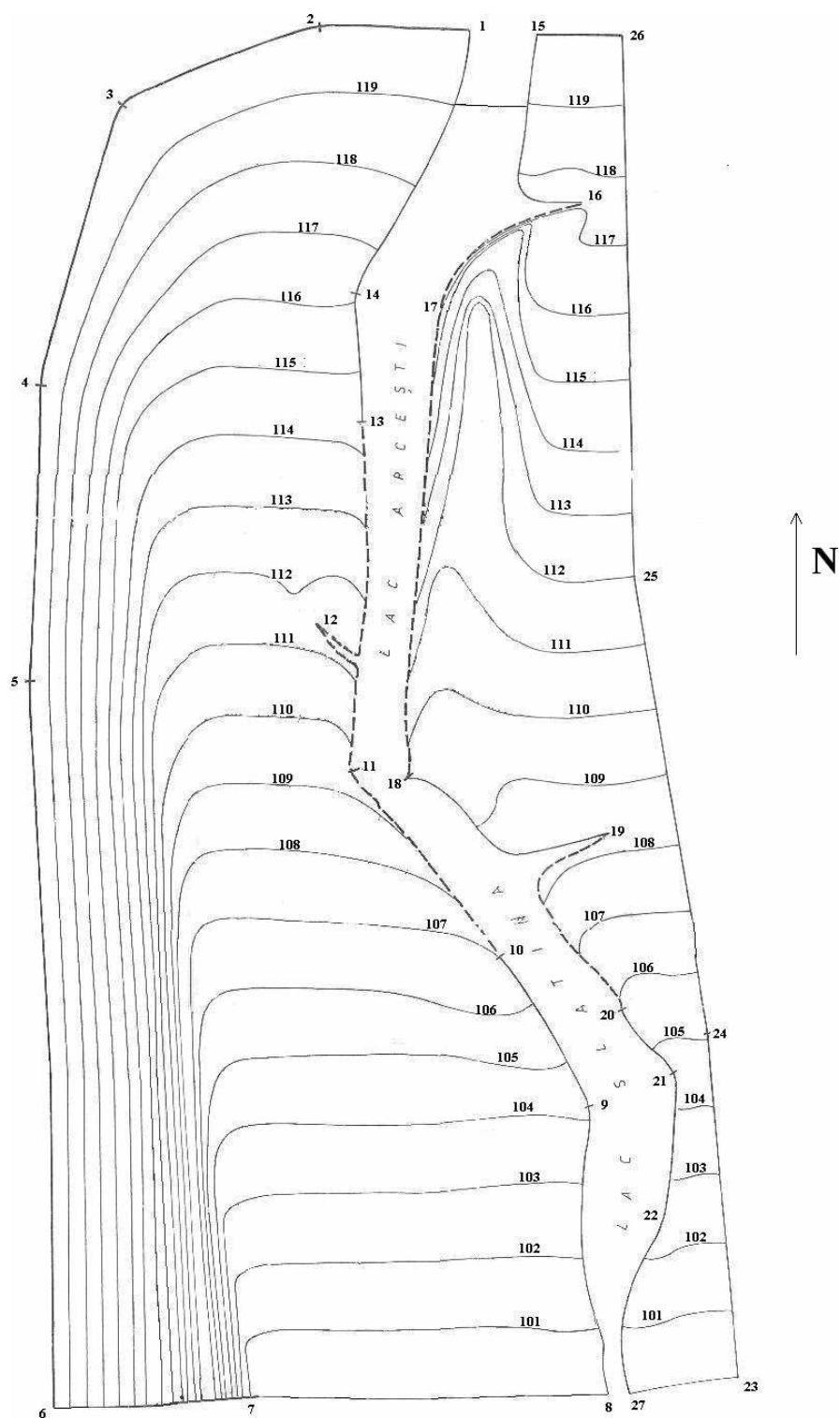


Fig. 4. The measured equipotential lines for the phreatic aquifer, in Slatina area.  
(The river flow direction is N-S).

$= 300 \text{ m}^2/\text{day}$ , the effective porosity  $n = 0.2$ , the storage coefficient  $S = 0.009$ . With these data we have obtained the flow pattern presented in Fig. 4.

The flow pattern obtained by using the real pumped discharges (Fig. 4) looks very similar to the real one (Fig. 2).

This pattern can be used to calculate the velocity at each point in the aquifer, the direction of the flow, as well as the discharge at the boundaries. The piezometric level varies within a minimum value of 97.67 m (in the Noua and Zăvoi areas) and a maximum of 117.5 m (on the chosen boundaries). The vertical distance between the piezometric lines is 1 m.

If we are to change the pumped values of discharges, the flow net will obviously be different. In order to obtain the real one we have to know the real pumped discharges. In the Northern part of the domain the flow has a N-S direction. In the Eastern part, the flow direction is from the Eastern boundary towards the West (the left hand of Olt well lines pump all the water coming from the East boundary). On the right-hand bank of Olt the flow direction is from the West boundary towards the East.

The piezometric surface levels in the confined aquifer have smaller values than those of the free surface of the phreatic aquifer existing above it. A flow between the two layers therefore continually tries to equalise the piezometric levels and the free surface levels. That flow's direction is from the phreatic to the confined aquifer, through the aquitard. The values of that discharge can be calculated analysing the flow pattern of the phreatic aquifer.

The phreatic aquifer is exposed to pollution, on account of its being at a small depth under the soil surface. Due to the above-explained connection between the two layers the confined aquifer runs the risk of pollution. The Olt river (the lakes Arcești and Slatina) has the free surface at higher levels than the phreatic aquifer's. Hence a flow from the Olt river to the phreatic is very likely to occur. The concentration of pollutants in the water of Olt thus determines the groundwater quality.

## 5. Analysis in the Unconfined Aquifer in Slatina Area

The analyzed area has: the length  $L = 21\,000 \text{ m}$ , the width  $l = 21\,000 \text{ m}$ , the thickness  $M = 10 \text{ m}$ , the hydraulic conductivity  $K = 35 \text{ m/day}$ , the transmissivity  $T = 350 \text{ m}^2/\text{day}$ , the effective porosity  $n = 0.2$ . With this data we have obtained the flow pattern presented in Fig. 5 (the right hand of the river Olt). Fig. 4 shows the equal level lines of water-table in the phreatic aquifer, measured in observation wells.

The phreatic aquifer is divided into two areas: one to the right of the Olt river and the other to its left (Fig. 4). Each area is separately considered. The boundary conditions take into account the impervious-less spaces in the dam walls ( $Q \neq 0$ ), the impervious ones ( $Q = 0$ ), and the values of the hydraulic head (free surface levels) from the measured levels map (Fig. 4). For the phreatic aquifer the procedure is identical to the one for the confined aquifer (only the diffusivity equation is different).

The phreatic aquifer is less pumped than the confined one. On the right hand domain there are only domestic wells which don't modify the flow net of groundwater.

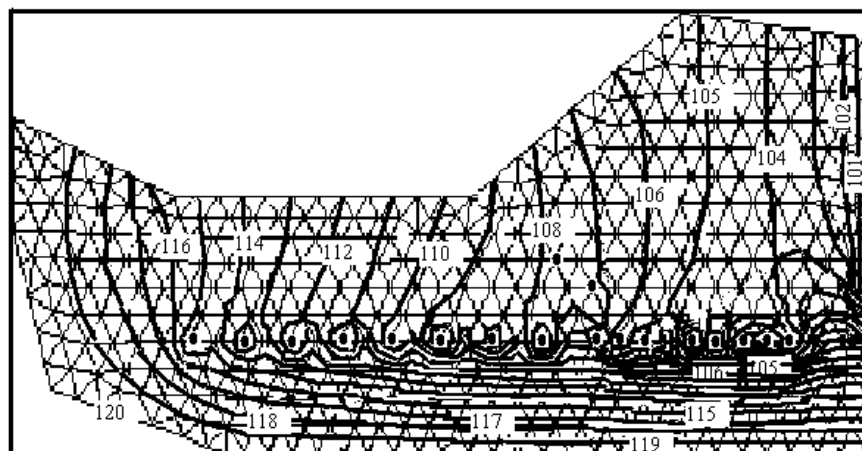


Fig. 5. Equal head lines (the equal level of groundwater table) in the phreatic-(right side of Olt) ( $h_{\max}=120$  m,  $h_{\min}=100$  m,  $\text{dist}=1$  m), in the case without pumping wells.

To the left, in Curtișoara area, the water is pumped from an important well line. The pumped discharges in this area influence the aspect of the flow net and determine the groundwater pollution in case the Olt water is polluted. So, the best method to understand the vulnerability to pollution in Curtișoara area is to calculate the discharge values from the Olt to the groundwater, using the flow net obtained with the above-presented code.

The level of phreatic free surface varies between 120 m (in the N) and 100 m (in the S). The phreatic is in direct connection with the lake so the lake's level is a boundary condition for the aquifer.

For the right hand phreatic of Olt, using the boundary conditions from Fig. 3 (imposed head conditions) in the numerical code, we have obtained the equipotential lines (different from the measured ones). To obtain a flow net similar to the real one we have inserted in the domain a number of fictitious pumping wells. We have changed the position and the discharges in the wells in order to obtain the same equipotential lines as in the Fig. 3 (the right hand of Olt). We have thus obtained approximately, the discharges drained from the phreatic to the confined aquifer. This drainage is produced by the decrease of the piezometric level in the confined aquifer during the pumping process from that aquifer.

In Fig. 6 the equipotential lines (equal level of water table) are plotted in the phreatic aquifer. In the Curtișoara wells we have considered pumping discharges greater than the real ones (the values in  $\text{m}^3/\text{day}$ : 2000, 2000, 5000, 5000, 5000, 3000, 1500), to obtain the measured values of the water table (Fig. 4). The pumped discharge in Curtișoara wells is  $Q = 1181/\text{s} = 10195.2\text{m}^3/\text{day}$  and the discharge used in the model is  $Q_p = 23\,500\text{ m}^3/\text{day}$ . The difference  $Q_p - Q = 13304.8\text{ m}^3/\text{day}$  is lost by drainage in the confined aquifer (depressurized by pumping).

The velocity field in Curtișoara area shows the flow direction from the river Olt into the well lines. So, the river quality influences the pumped water quality. We have proved that a quantity of water is drained toward the confined aquifer. The quality of that one is also influenced by the Olt water.

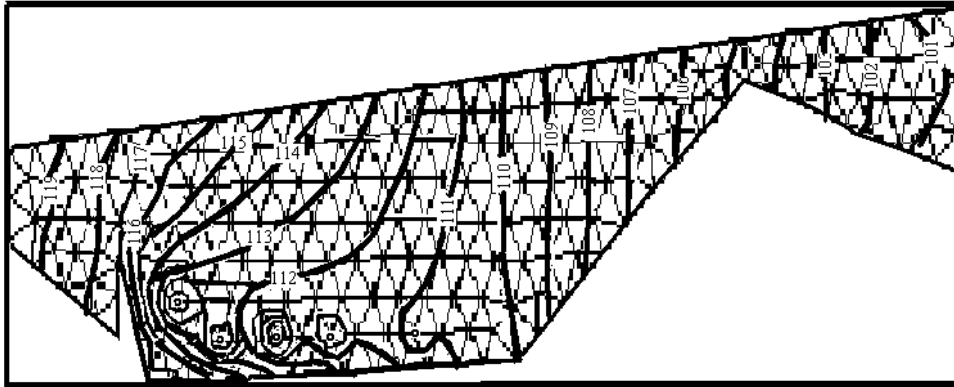


Fig. 6. Equal head lines (the equal level of groundwater table) in the phreatic-(left side of Olt) ( $h_{\max} = 120$  m,  $h_{\min} = 100$  m, distance between head lines =1 m), in the case with pumping wells.

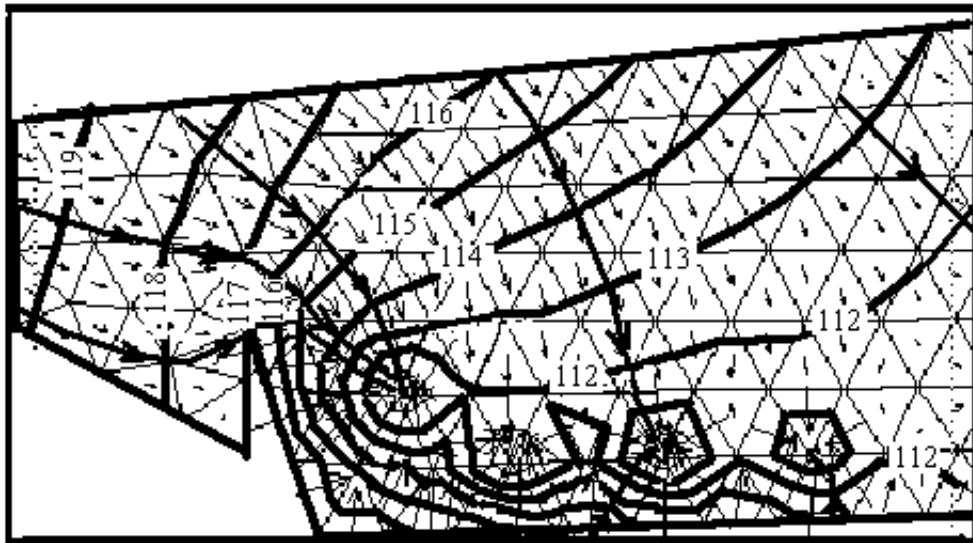


Fig. 7. Equal head lines (the equal level of groundwater table) and the velocity field in the phreatic-(right side of Olt) ( $h_{\max} = 120$  m,  $h_{\min} = 100$  m, distance between head lines =1 m), in the case with pumping wells in Curtișoara area.

## 6. Conclusions

The groundwater quality in Slatina area is influenced by the water quality of river Olt, especially in Curtișoara area. If the phreatic aquifer is affected by certain surface pollution sources, then the confined aquifer will as a rule be affected as well.

A similar study could be carried out if we had a better hydrogeological description of aquifers. The code also allows for the study of the anisotropic case and it is to be expected that the results will be closer to the real ones.

## References

- [1] Fried, J.J., *Groundwater pollution*, Elsevier S PC, New York, 1975,
- [2] G.Marsily, *Quantitative Hydrogeology*, Academic Press Inc., 1986.
- [3] Moldoveanu, V., Niculae, A., Rotar, C., *Optimizarea exploatarei surselor de apa subterana potabila ale municipiului Slatina.*, Conferinta Zilele Hidraulicii, Ingineria Resurselor de Apă, UCB, 28-29 iunie 2001.

## **Analysis of a Preconditioned CG method for an Inverse Bioelectric Field Problem**

**Marcus Mohr<sup>\*</sup>, Constantin Popa<sup>\*\*</sup> and Ulrich Rüde<sup>\*\*\*</sup>**

This paper is a continuation of our previous analysis from [4] related to the electrocardiographic (ECG) inverse problem. In that paper we formulated the inverse ECG problem as a differential inverse problem and derived an appropriate simulation procedure. As numerical solver we employed the Conjugate Gradient algorithm for the normal equations (CGNE) together with a stopping test constructed following the discrepancy principle by Morozov. In the current paper we consider a preconditioned version of the CGNE algorithm. The preconditioner is constructed using the Cholesky factors of the discrete Laplacian which forms a block of the original system matrix. We derive some theoretical results concerning the efficiency and also the limitations of the preconditioner. Numerical experiments and comparisons are presented for the cases analysed in [4].

### **1. Introduction**

This paper is a continuation of our previous analysis from [4] related to the inverse problem of electrocardiography (ECG). We will therefore only briefly replay the formulation of the problem and refer the reader to [4, 5] for further details. In the inverse ECG problem one attempts to determine from voltage measurements on a person's torso the underlying electric behaviour of the heart, see e.g. [6]. One approach is the so called cardiac imaging where one tries to re-construct from the measurements the electric potential on the epicardium, the outer surface of the heart.

---

<sup>\*</sup> Department for Sensor Technology, University of Erlangen–Nuremberg, Germany; for this author the paper was partially supported by the DFG Junior Research Group Grant Ka 1778/1

<sup>\*\*</sup> “Ovidius” University of Constanța, Romania; for this author the paper was supported by the PNCDI INFOSOC Grant 131/2004, e-mail: [cpopa@univ-ovidius.ro](mailto:cpopa@univ-ovidius.ro)

<sup>\*\*\*</sup> IMMD 10- System Simulation Group, University of Erlangen–Nuremberg, Germany.

In [4] we formulated a simplified 2D version of this problem in the following form: find a function  $u : \bar{\Omega} \rightarrow \mathbb{R}$  such that

$$\begin{aligned} \Delta u &= 0, \text{ in } \Omega, \\ \frac{\partial u}{\partial n} &= 0, \text{ on } \Gamma \setminus \Gamma_1, \\ u &= d, \text{ on some } \Gamma_2 \subsetneq \Gamma. \end{aligned} \quad (1)$$

Here  $\Omega := (0, 1) \times (0, 1)$ ,  $\Gamma$  is the boundary of  $\Omega$ ,  $\Gamma_1 \cup \Gamma_2 \subset \Gamma$  with  $\Gamma_2 \cap \Gamma_1 = \emptyset$  and  $d$  is a given function on  $\Gamma_2$ . In this paper we assume the configuration is as given in Figure 1. Applying a standard discretisation of the differential parts of problem (1) by Finite Differences on a regular grid of mesh-width  $h = 1/n$  we arrive at an over-determined linear system

$$Ax = b, \quad (2)$$

where  $A$  is an  $M \times N$  matrix of the form

$$A = \begin{bmatrix} \Delta_h & E \\ S & 0 \end{bmatrix} \quad (3)$$

with  $M = (n-1)^2 + (n-1)$  and  $N = (n-1)^2 + (n-3)$ . The sub-matrices of  $A$  are given by

$$E = \begin{bmatrix} 0 \\ -1 \\ \ddots \\ -1 \\ 0 \end{bmatrix} \in \mathbb{R}^{(n-1)^2 \times (n-3)} \quad (4)$$

and

$$S = \begin{bmatrix} 1 & 0 \cdots 0 \\ 1 & 0 \cdots 0 \\ \ddots & \vdots \\ 1 & 0 \cdots 0 \end{bmatrix} \in \mathbb{R}^{(n-1) \times (n-1)^2}. \quad (5)$$

The matrix  $\Delta_h$  finally is simply the well-known 5-point discretisation of the 2D Laplacian (we assume here that both  $\Delta_h$  and  $E$  include a scaling by the factor  $h^2$ ). The right-hand side of the system is given by

$$b = (0, \dots, 0, d_1, \dots, d_{n-1})^t,$$

where  $d_k$  denotes evaluation of the prescribed function  $d$  at a given node of the discrete grid, see also Figure 2. In this setup the problem (2) is consistent. In the real application, however, the measurements  $d$  will contain noise and we will consider instead, the least-squares formulation

$$\|Ax - b\|_2 = \min! \quad (6)$$



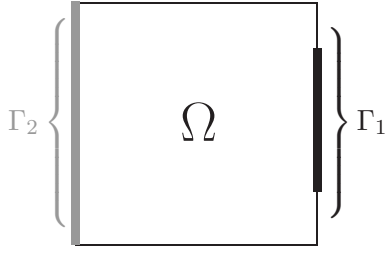


Fig. 1. Domain of the continuous problem.

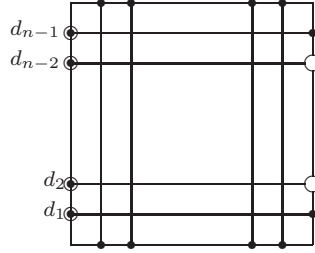


Fig. 2. Pattern of discrete grid:  $\bullet$  represent Neumann,  $\odot$  Neuman + Dirichet boundary conditions and  $\circ$  points on  $\Gamma_1$ .

## 2. The preconditioned CG algorithm

In [4] we investigated the solution of the linear system (2) by means of the CG algorithm applied to the normal equations  $A^t A x = A^t b$ . We employed the CGNE algorithm in the following form (see [1])

$$\begin{aligned}
 d_0 &= b - A x_0, \quad p_0 = A^t d_0 \\
 \left. \begin{aligned}
 \alpha_i &= \frac{\|A^t d_i\|_2^2}{\|A p_i\|_2^2} \\
 x_{i+1} &= x_i + \alpha_i p_i \\
 d_{i+1} &= b - A x_{i+1} = d_i - \alpha_i A p_i \\
 \beta_i &= \frac{\|A^t d_{i+1}\|_2^2}{\|A^t d_i\|_2^2} \\
 p_{i+1} &= A^t d_{i+1} + \beta_i p_i
 \end{aligned} \right\} \text{ for } i = 0, 1, \dots
 \end{aligned} \tag{7}$$

While the results obtained in [4] were satisfying with respect to the reconstruction of the “shape” of the exact solution, the high number of CGNE iterations was not. This aspect is directly related to the ill-conditioning of  $A$  from (3). In order to eliminate the latter aspect we investigate here a special form of (right-)preconditioning of the problem (2). Instead of the least-squares formulation (6) of (2) we consider the minimization problem

$$\|Bz - b\|_2 = \min!, \tag{8}$$

where

$$B = A P^{-1} \quad \text{and} \quad P^{-1} z = x \iff P x = z, \tag{9}$$

with a square and invertible matrix  $P$ . With this the preconditioned version of the CGNE algorithm (7) can be written as (see e.g. [1])

$$\begin{aligned} \bar{d}_0 &= b - Bz_0 = b - AP^{-1}z_0 \\ \bar{p}_0 &= B^t \bar{d}_0 = P^{-t}(A^t \bar{d}_0) \\ \left. \begin{aligned} \bar{\alpha}_i &= \frac{\|B^t \bar{d}_i\|_2^2}{\|B \bar{p}_i\|_2^2} = \frac{\|B^t \bar{d}_i\|_2^2}{\|AP^{-1} \bar{p}_i\|_2^2} \\ z_{i+1} &= z_i + \bar{\alpha}_i \bar{p}_i \\ \bar{d}_{i+1} &= \bar{d}_i - \bar{\alpha}_i B \bar{p}_i \\ \bar{\beta}_i &= \frac{\|B^t \bar{d}_{i+1}\|_2^2}{\|B^t \bar{d}_i\|_2^2} = \frac{\|P^{-t}(A^t \bar{d}_{i+1})\|_2^2}{\|B^t \bar{d}_i\|_2^2} \\ \bar{p}_{i+1} &= B^t \bar{d}_{i+1} + \bar{\beta}_i \bar{p}_i \end{aligned} \right\} \text{ for } i = 0, 1, \dots, i_{\text{final}} \end{aligned} \quad (10)$$

and then put

$$x_{i_{\text{final}}} = P^{-1} z_{i_{\text{final}}} . \quad (11)$$

**Remark 1.** *If we restrict ourselves to symmetric matrices  $P$  then we must solve two systems of the form  $Pv = w$  in each iteration step of algorithm (10). Another such system must be solved once for computing the final approximate  $x_{i_{\text{final}}}$ .*

It is well known, see e.g. [3], that the error reduction formula for the algorithm (7) is of the form

$$\|x_i - u_{LS}\|_2^2 \leq \frac{C(x_0)}{\lambda_{\min}(A^t A)} \left( \frac{\lambda_{\max}(A^t A) - \lambda_{\min}(A^t A)}{\lambda_{\max}(A^t A) + \lambda_{\min}(A^t A)} \right)^{2i}, \quad (12)$$

where  $C(x_0) \geq 0$  is a constant depending on the initial approximation  $x_0$  and  $\lambda_{\min}(A^t A)$ ,  $\lambda_{\max}(A^t A)$  are the minimal and maximal nonzero eigenvalues of  $A^t A$ , respectively. Let us denote in the following by  $\mathcal{I}_k$  the identity matrix from  $\mathbb{R}^{k \times k}$ . Using this notation we can write  $A^t A$  for our matrix  $A$  from (3) as

$$A^t A = \begin{bmatrix} \Delta_h^2 + \tilde{I}_1 & C \\ C^t & \mathcal{I}_{(n-3)}, \end{bmatrix} \quad (13)$$

where  $\mathcal{I}_{(n-3)} = E^t E$ , the square matrix  $\tilde{I}_1$  is given by

$$\tilde{I}_1 = S^t S = \begin{bmatrix} \mathcal{I}_{(n-1)} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n-1)^2 \times (n-1)^2}$$

and  $C$  is the  $(n-1)^2 \times (n-3)$  matrix defined as

$$C = \Delta_h E . \quad (14)$$

A (classical) good preconditioning of the form (9) for  $A$  will be one which improves the condition number of the upper-left block  $\Delta_h^2 + \tilde{I}_1$ . Thus, if we denote by  $\tilde{\Delta}_h^2$  this matrix, i.e.

$$\tilde{\Delta}_h^2 := \Delta_h^2 + \tilde{I}_1 , \quad (15)$$

an “almost the best” choice for  $P$  in (9) will be

$$P = \begin{bmatrix} \tilde{\Delta}_h & 0 \\ 0 & \mathcal{I}_{(n-3)} \end{bmatrix}. \quad (16)$$

In this case we have

$$B = A \begin{bmatrix} \tilde{\Delta}_h^{-1} & 0 \\ 0 & \mathcal{I}_{(n-3)} \end{bmatrix} \quad (17)$$

and

$$B^t B = \begin{bmatrix} \mathcal{I}_{(n-1)^2} \tilde{\Delta}_h^{-1} C \\ C^t \tilde{\Delta}_h^{-1} \mathcal{I}_{(n-3)} \end{bmatrix}. \quad (18)$$

If we denote by  $T$  the matrix

$$T = \tilde{\Delta}_h^{-1} C \quad (19)$$

the following result holds.

**Proposition 1.** *With the above definitions and notations we have the equivalence*

$$\{\lambda \in \sigma(B^t B)\} \iff \{\lambda = 1 \text{ or } \lambda = 1 + \sqrt{\mu} \text{ or } \lambda = 1 - \sqrt{\mu} \geq 0, \mu \in \sigma(T^t T)\}. \quad (20)$$

*Proof.* A simple computation (for the case  $\lambda \neq 1$ ) gives us the following sequence of equivalences

$$\begin{aligned} & \lambda \in \sigma(B^t B) \\ \iff & \lambda \in \sigma \left( \begin{bmatrix} \mathcal{I}_{(n-1)^2} & T \\ T^t & \mathcal{I}_{(n-3)} \end{bmatrix} \right) \\ \iff & 0 = \det \left( \begin{bmatrix} \mathcal{I}_{(n-1)^2} & T \\ T^t & \mathcal{I}_{(n-3)} \end{bmatrix} - \lambda \begin{bmatrix} \mathcal{I}_{(n-1)^2} & 0 \\ 0 & \mathcal{I}_{(n-3)} \end{bmatrix} \right) \\ \iff & 0 = \det \left( \begin{bmatrix} (1-\lambda)\mathcal{I}_{(n-1)^2} & T \\ T^t & (1-\lambda)\mathcal{I}_{(n-3)} \end{bmatrix} \right) \\ \iff & 0 = \det \left( \begin{bmatrix} (1-\lambda)\mathcal{I}_{(n-1)^2} & 0 \\ T^t & (1-\lambda)\mathcal{I}_{(n-3)} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{I}_{(n-1)^2} & \frac{1}{1-\lambda}T \\ 0 & \mathcal{I}_{(n-3)} - \frac{1}{(1-\lambda)^2}T^t T \end{bmatrix} \right) \\ \iff & 0 = \det \left( \mathcal{I}_{(n-3)} - \frac{1}{(1-\lambda)^2} T^t T \right) \\ \iff & 0 = \det (T^t T - (1-\lambda)^2 \mathcal{I}_{(n-3)}) \\ \iff & (1-\lambda)^2 \in \sigma(T^t T) \\ \iff & 1-\lambda = \pm\sqrt{\mu} \text{ with } \mu \in \sigma(T^t T) \end{aligned}$$

this, together with that fact that  $\sigma(B^t B) \subset (0, \infty)$ , proves (20).

**Remark 2.** From the definitions of  $C$ ,  $\tilde{\Delta}_h^2$  and  $T$  in (14), (15) and (19) we obtain

$$\begin{aligned} T^t T &= E^t \Delta_h (\Delta_h^2 + \tilde{I}_1)^{-1} \Delta_h E = \\ &= E^t [\mathcal{I}_{(n-1)^2} + \Delta_h^{-1} \tilde{I}_1 \Delta_h^{-1}]^{-1} E = E^t (\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1} E, \end{aligned} \quad (21)$$

where we denoted by  $\Delta_1$  the symmetric non-negative definite matrix

$$\Delta_1 = \Delta_h^{-1} \tilde{I}_1 \Delta_h^{-1}. \quad (22)$$

Because of the special structure of  $E$ , see (4) we obtain from (18) that the matrix  $T^t T$  is singular. Thus, for  $0 = \mu \in \sigma(T^t T)$  we get one more time  $\lambda = 1$  in (20). Let now  $\mu \neq 0$  be another eigenvalue of  $T^t T$  corresponding to an eigenvector  $v \neq 0$ , i.e.

$$(T^t T)v = \mu v. \quad (23)$$

Because  $\mu \neq 0$  we have

$$w := Ev \neq 0. \quad (24)$$

Thus

$$\frac{\langle (T^t T)v, v \rangle}{\langle w, w \rangle} = \frac{\langle (\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1} w, w \rangle}{\langle w, w \rangle} = \frac{\mu \langle v, v \rangle}{\langle E^t E v, v \rangle} = \frac{\mu \langle v, v \rangle}{\langle \mathcal{I}_{n-3} v, v \rangle} = \mu$$

and

$$\mu = \frac{\langle (\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1} w, w \rangle}{\langle w, w \rangle}.$$

Since  $(\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1}$  is a symmetric and positive definite matrix we obtain that

$$\lambda_{\min}[(\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1}] \leq \mu \leq \lambda_{\max}[(\mathcal{I}_{(n-1)^2} + \Delta_1)^{-1}] \quad (25)$$

or

$$\frac{1}{1 + \lambda_{\max}(\Delta_1)} \leq \mu \leq \frac{1}{1 + \lambda_{\min}(\Delta_1)}. \quad (26)$$

Thus, if  $\mu$  is “close” to the right bound in (26) i.e. (for a small  $\varepsilon > 0$ )

$$\mu = \frac{1 - \varepsilon}{1 + \lambda_{\min}(\Delta_1)}, \quad (27)$$

then from (20) we can have

$$\begin{aligned} \lambda &= 1 - \sqrt{\mu} = 1 - \frac{\sqrt{1 - \varepsilon}}{\sqrt{1 + \lambda_{\min}(\Delta_1)}} = \\ &= \frac{\lambda_{\min}(\Delta_1) + \varepsilon}{\sqrt{1 + \lambda_{\min}(\Delta_1)}(\sqrt{1 + \lambda_{\min}(\Delta_1)} + \sqrt{1 - \varepsilon})}. \end{aligned} \quad (28)$$

If  $\varepsilon$  is of order  $\lambda_{\min}(\Delta_1)$  and if  $\lambda_{\min}(\Delta_1)$  is small, which we expect from (21) for large values of  $n$ , then  $\lambda$  will be as small as  $\lambda_{\min}(\Delta_1)$ , i.e. no essential improvement

Table 1

Spectral properties of the matrix  $A$  and the matrix  $B$  preconditioned as in (17)

$n$	$z_{\max}(A)$	$z_{\max}(B)$	$z_{\min}(A)$	$z_{\min}(B)$	$k(A)$	$k(B)$
8	7.63	1.41	$1.87 \cdot 10^{-05}$	$1.3 \cdot 10^{-05}$	$\approx 10^{05}$	$\approx 10^{05}$
16	7.81	1.41	$5.21 \cdot 10^{-12}$	$3.7 \cdot 10^{-12}$	$\approx 10^{12}$	$\approx 10^{12}$
24	7.96	1.41	$9.00 \cdot 10^{-17}$	$3.9 \cdot 10^{-17}$	$\approx 10^{17}$	$\approx 10^{17}$
32	7.97	1.41	$1.40 \cdot 10^{-19}$	$2.8 \cdot 10^{-18}$	$\approx 10^{20}$	$\approx 10^{18}$

(“compression”) of  $\sigma(A^t A)$  will be obtained by the preconditioning (16)–(17). Some results in this sense are presented in table 1. We denoted by  $z_{\max}(A)$ ,  $z_{\max}(B)$  and  $z_{\min}(A)$ ,  $z_{\min}(B)$  the largest and smallest singular value of  $A$  and  $B$  respectively. The numbers  $k(A)$ ,  $k(B)$  are defined by

$$k(x) = \frac{z_{\max}(x)}{z_{\min}(x)} . \quad (29)$$

**Remark 3.** Concerning the preconditioning (17) we want to point out the following:

- (i) It is not important that in (16) we used the identity matrix  $\mathcal{I}_{n-3}$ . This choice was made only to simplify the theoretical considerations that followed. For different choices we got results similar to those in Table 1.
- (ii) Unfortunately, from a practical point of view, the computation of  $\tilde{\Delta}_h^{-1}$  in (17) can be very expensive. A “less costly” choice is

$$P = \begin{bmatrix} \Delta_h & 0 \\ 0 & \mathcal{I}_{(n-3)} \end{bmatrix} . \quad (30)$$

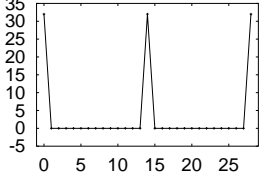
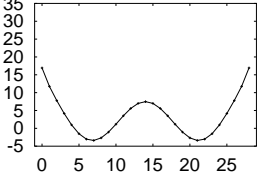
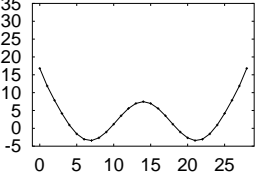
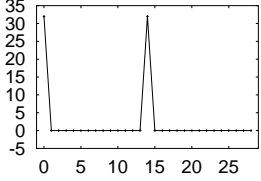
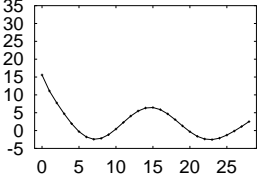
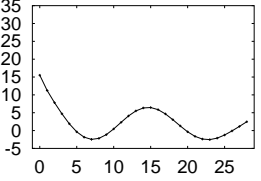
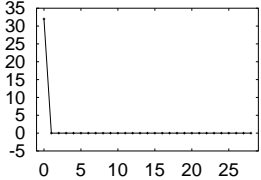
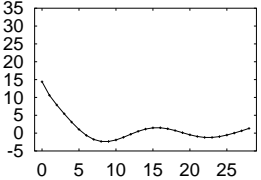
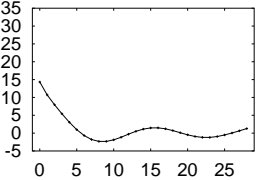
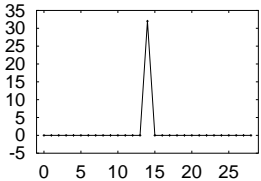
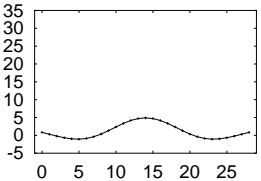
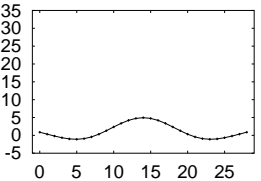
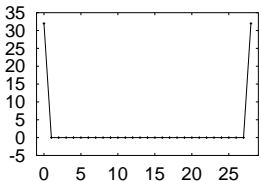
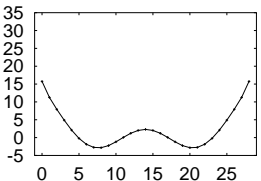
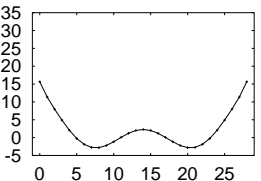
In this case, at least up to some point, a theoretical analysis as in the above Remark 2 can be made, with similar conclusions with respect to  $\sigma(B^t B)$ . We restrict ourselves here to present in Table 2 some numerical results regarding  $\sigma(B^t B)$  for the choice (30).

Table 2

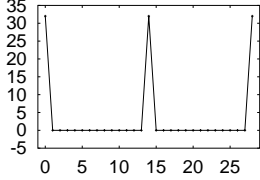
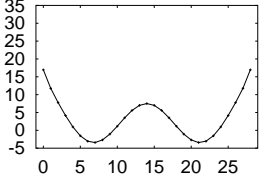
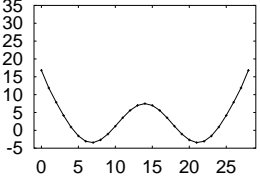
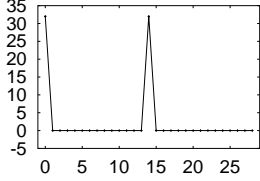
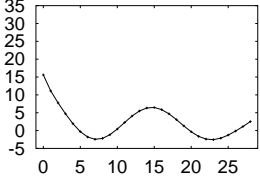
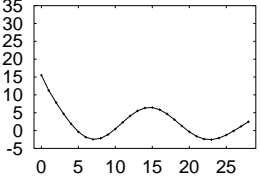
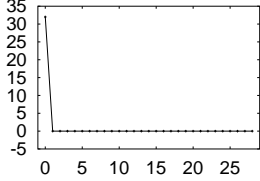
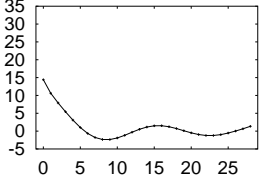
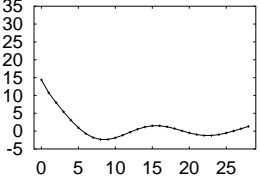
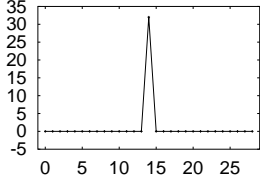
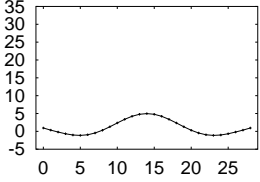
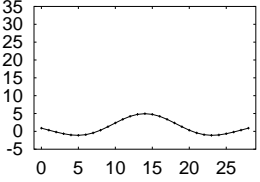
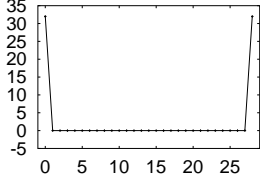
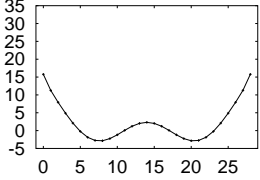
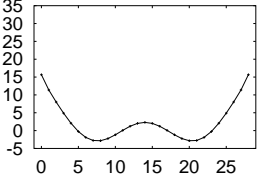
Spectral properties of the matrix  $A$  and the matrix  $B$  preconditioned as in (30)

$n$	$z_{\max}(A)$	$z_{\max}(B)$	$z_{\min}(A)$	$z_{\min}(B)$	$k(A)$	$k(B)$
8	7.63	13.37	$1.87 \cdot 10^{-05}$	$1.30 \cdot 10^{-05}$	$\approx 10^{05}$	$\approx 10^{05}$
16	7.81	36.15	$5.21 \cdot 10^{-12}$	$3.00 \cdot 10^{-12}$	$\approx 10^{12}$	$\approx 10^{13}$
24	7.96	66.50	$9.00 \cdot 10^{-17}$	$2.00 \cdot 10^{-17}$	$\approx 10^{17}$	$\approx 10^{18}$
32	7.97	102.69	$1.40 \cdot 10^{-19}$	$3.46 \cdot 10^{-19}$	$\approx 10^{20}$	$\approx 10^{21}$

*Table 3*  
CG versus preconditioned CG using stopping test (31)

prescribed shape	shape approx. by CG	shape approx. by PCG
	 nits = 1574	 nits = 52
	 nits = 3041	 nits = 119
	 nits = 3035	 nits = 118
	 nits = 1569	 nits = 52
	 nits = 1637	 nits = 53

*Table 4*  
CG versus preconditioned CG using stopping test (32)

prescribed shape	shape approx. by CG	shape approx. by PCG
	 nits = 1637	 nits = 60
	 nits = 3045	 nits = 119
	 nits = 3038	 nits = 118
	 nits = 1566	 nits = 52
	 nits = 1644	 nits = 64

**Remark 4.** *There is also a “good” aspect concerning the above preconditionings with respect to our problem (6). Indeed, after the transformation (17) with  $P$  from (16) or (30), a “large” part of the eigenvalues of  $B^t B$  will be 1. Typically only up to  $2n$  eigenvalues will be different from 1 and of those only up to  $(n - 3)$  will have values in the interval  $(0, 1)$ , see Proposition 1 in this respect. Thus, following the theoretical considerations from [2], the “filter factors” associated with the eigenvalues of  $B^t B$  greater or equal to 1 will become very small in only “few” iterations of the preconditioned CG algorithm. Thus, we will obtain an approximation of  $x_{LS}$ , i.e. the minimal norm solution of (6), in significantly fewer iterations than with the non-preconditioned algorithm and this approximation will have the same quality of reconstruction.*

We can see this improvement in the following tests. We use the verification procedure detailed in [4, 5], i.e. we prescribe Dirichlet values along the boundary  $\Gamma_1$  and solve the forward problem associated with (1) to obtain “measurement” values  $d$ . In order to simulate noise we add a random vector  $x^p \in \mathbb{R}^{(n-1)}$  to  $d$ ,

$$d_i^p = d_i + \varepsilon x_i^p, \quad i = 1, \dots, n - 1.$$

We then solve (6) with  $b$  derived from  $d^p$  using CG and CG preconditioned with  $P$  according to (30). For the experiments we use  $\varepsilon = 0.01$  and test both stopping rules from [4]. The results for the first stopping rule are given in Table 3. Here we terminated the iterations once

$$\| r^{(k)} \|_2 + N \cdot \| r^{(k)} \|_2 \cdot \text{sstop} \leq \text{neps}, \quad (31)$$

where  $r^{(k)}$  is the residual of the current approximation,  $\text{neps} = \varepsilon \| x^p \|_2 \approx 0.03$  and

$$\text{sstop} = \| \text{cu}^{(k)} - \text{cuex} \|_\infty.$$

Here  $\text{cuex}$  denotes the values of the least-squares solution of the inverse problem for the unperturbed data  $d$  along  $\Gamma_1$ , while  $\text{cu}^{(k)}$  denotes the values of the current approximate solution on  $\Gamma_1$ . Since  $\text{cuex}$  is not available in practice we devised a second stopping test that is only based on computable values, namely

$$\| r^{(k)} \|_2 + N \cdot \| r^{(k)} \|_2 \cdot \| \text{cu}^{(k)} \|_\infty \leq \text{neps}. \quad (32)$$

The results for this second rule are given in Table 4. Both experiments demonstrate that the preconditioning yields a significant reduction in the number of iterations (nits), while retaining the quality of the reconstructed shape.

## References

- [1] Golub, G.H. and Van Loan, C.F., *Matrix computations*, The John’s Hopkins Univ. Press, Baltimore, 1983.



- [2] Hanke, M. and Hansen, P.C., *Regularization methods for large-scale problems*, Surv. Math. Ind. **3** (1993), 253–315.
- [3] Kammerer, W.J. and Nashed, M.K., *On the convergence of the conjugate gradient method for singular linear operator equations*, SIAM J. Numer. Anal., **9**(1) (1972), 165–181.
- [4] Mohr, M., Popa, C. and Rüde, U., *A Differential Inverse Problem from Cardiac Imaging*, in *Proceedings of the Third Workshop on Mathematical Modelling of Environmental and Life Sciences Problems*, Constanța, Romania, 27-30 May 2004; Editura Academiei Române, București 2004, 189–204.
- [5] Mohr, M., Popa, C. and Rüde, U., *An experimental analysis of a differential inverse problem*, Technical Report **99-1** (1999), Lehrstuhl für Informatik 10 (Systemsimulation), Friedrich-Alexander-Universität Erlangen-Nürnberg.
- [6] MacLeod, R. and Brooks, D., *Recent Progress in Inverse Problems in Electrocardiology*, IEEE Engineering in Medicine and Biology, **17**(1) (1998), 73–83.



## **Dosimetric Estimates in Biological Tissue Exposed to Microwave Radiation in the Near Field of an Antenna**

**Mihaela Morega<sup>\*</sup>, Alina Machedon<sup>\*</sup> and Marius Neagu<sup>\*</sup>**

Human exposure to electromagnetic field (including the microwave radiation range) is limited by international safety guidelines, based on health considerations. Thermal health effects are commonly considered for the radiofrequency range; in particular, for microwave exposure, the absorbed energy that produces heat is quantified by the specific energy absorption rate, as dosimetric reference. Induced electric and magnetic field strengths are also restricted by the exposure guidelines. We report here a numerical study of the electromagnetic field induced in several biological models by a common microwave applicator. The sensitivity of dosimetric parameters and the compliance with exposure guidelines are evaluated. We have examined the influence of dielectric properties dispersion on dosimetric quantities, useful in the design and validation of experimental settings and numerical models.

### **1. Introduction**

A wide debate developed over the last two decades, both in scientific and social forums, on the possible health effects of human exposure to non-ionizing electromagnetic fields continues to concentrate attention without concluding results. Research activity was therefore developed by the international scientific community aimed at evaluating the risk associated with exposure to this type of radiation. At the same time, various international authorities began to issue recommendations on exposure limits valid for workers and for the population in the frequency range 0 Hz÷300 GHz. The limits specified by the guidelines are settled both at workplaces and in the living environment. The specified accepted limits are intended to be used as a basis for planning work procedures, and designing protective facilities, as much as in the

---

<sup>\*</sup> “Politehnica” University of Bucharest, Romania.

assessment of the efficacy of protective measures and practices, or in the guidance on health surveillance.

The most known and accepted are the guidelines developed by the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [1]. Other internationally recognized documents, such as those developed by the Institute of Electrical and Electronics Engineers (IEEE) and American National Standardization Institute (ANSI) [2] in the USA, by the Australian Radiation Protection and Nuclear Safety Agency (ARPANSA) [3], or by the National Radiological Protection Board (NRPB) in the UK adopt the same basic approach of ICNIRP, although some differences exist in numerical values, and they will be discussed in this paper. In 1999 the Council of the European Union has issued a Recommendation to Member States [4] to adopt a common frame of norms on exposure of the general public to electromagnetic fields that made precisely the indications supplied by ICNIRP for the protection of the population. Presently, about 30 countries have adopted ICNIRP guidelines as national regulations.

The market of wireless devices is presently highly dynamic and competitive. Producers have to find new and commercial solutions, as an optimum of cost, performance, modern design, miniaturization and multifunctionality. The compliance with the safety guidelines is also a restriction, and the manufacturer is required to include the *SAR* limit in the technical specification of each product. This value is then compared by the consumer safety authorities with the limits stated by standards. Also that seems to be a simple and non-controversial process, our study investigates the so called “technical accuracy” of safety and performance parameters assessment, based on valid standards.

## 2. Basic restrictions and reference levels presently stated by standards

The ICNIRP guidelines [1], as well as the other international standards [3–5], are based on a two-level structure. *Basic restrictions* are defined in terms of “dosimetric quantities” that are directly related to biological effects; these quantities are: the current density ( $J$ ) for low-frequency electric and magnetic fields, and the specific energy absorption rate (*SAR*) and the power density ( $PD$ ) for high-frequency electromagnetic fields (including microwaves). The limits are defined for exposure of all, or only a part, of the human body. For practical reasons, *reference levels* are derived from basic restrictions, through appropriate dosimetric models (simplified computational and experimental models). Reference levels are expressed in terms of physical quantities (electric field strength, magnetic field strength, and equivalent plane wave power density) that can be directly measured outside the exposed body and inside experimental phantoms.

Given the conservative hypotheses assumed in dosimetric models, exposures to fields that are below the reference levels necessarily comply with basic restrictions, but the vice-versa is not true. Even when the reference levels are exceeded, the

standard may be complied with, provided it can be proved that basic restrictions are not exceeded under the specific exposure conditions.

We are interested here by the high-frequency electromagnetic field, the microwave range with applications in wireless communications technologies, 0.5–3 GHz. The basic restriction for localized exposure (head and trunk) in this frequency range is the specific energy absorption rate (*SAR*) which is set in terms of maximum mass-normalized quantity, as follows:

- 2 W/kg for “any 10 g of contiguous tissue” in the ICNIRP guidelines [1] and the European recommendation [4], while “any 10 g of contiguous tissue in the shape of a cube” in the Australian standard [3];
- 1.6 W/kg for “1 g of tissue in the shape of a cube” in the ANSI/IEEE standard [2].

Basic references [1], [2] and [4] are issued in the same period of time, 1998-1999, are based on the research and documentation literature available at the time and are still valid. At the first glance the specifications do not seem to be contradictory; however, we found some important differences in their practical use, that we attempt to emphasize on a case study presented further.

### 3. Physical properties of the model and general assumptions

The work presented here examines the electromagnetic field penetration in human tissue considering several conditions and particularities related to wireless communications in the microwave frequency range [7–9]:

(1) An antenna is the electromagnetic radiation source in our study. The electromagnetic field produced by an antenna can be described as having several components; only one of these actually propagates through space, and this component is called the *radiated field* or *the far field*. The strength of the radiated field does decrease with distance, since the energy must spread as it travels. The other components of the electromagnetic field remain near the antenna and do not propagate. There are generally two other components: *the static field* and *the induction field*, and their strength decreases very rapidly with distance. The entire field (all of the components) near the antenna is called the *near field*. In this region, approximately one wavelength in extent, the electric field strength can be relatively high and pose a hazard to the human body. The dipole configuration is the most common and conventional type for near field human exposure related to wireless personal communication systems in the GSM frequency range (0.5 to 3 GHz); the harmonic waveform is considered in our study. In numerical and experimental models, the length of the antenna is usually adjusted at the half wavelength, both because of modeling reasons (like symmetry conditions) and to maximize the efficiency of the emission.

(2) The exposed body is represented by a layered biological structure, with an idealized shape, inspired by the human anatomy (head or trunk); the electromagnetic field penetration depth in dispersive dielectrics, like animal tissue, depends on the ex-

ternal shape of the body, on the electric conductivities and dimensions of tissue layers; the skin and fat peripheral layers present screening effect for the electric component of the incident field.

(3) Biological tissues are nonmagnetic and dispersive dielectric materials; for the purpose of this study, they are considered linear, isotropic and homogeneous materials; dielectric properties are expressed in the form of the *complex permittivity*  $\underline{\varepsilon} = \varepsilon - i\sigma/\omega$ , or the *complex conductivity*  $\underline{\sigma} = \sigma + i\omega\varepsilon$ , where  $\varepsilon$  is the dielectric permittivity,  $\sigma$  is the electric conductivity and  $\omega = 2\pi f$  is the angular frequency of the electromagnetic field. The specific values for  $\sigma$  and  $\varepsilon$  considered in our study correspond to data in literature [6]. For illustration, Fig. 1 shows the frequency dependence of the electric conductivity  $\sigma$  and relative dielectric permittivity  $\varepsilon_r$  of several anatomical tissues involved in the human body exposure to electromagnetic field (skin, fat, bone, dura, cerebro-spinal-fluid, brain and muscle); the microwave frequency range used by the GSM communication system is considered.

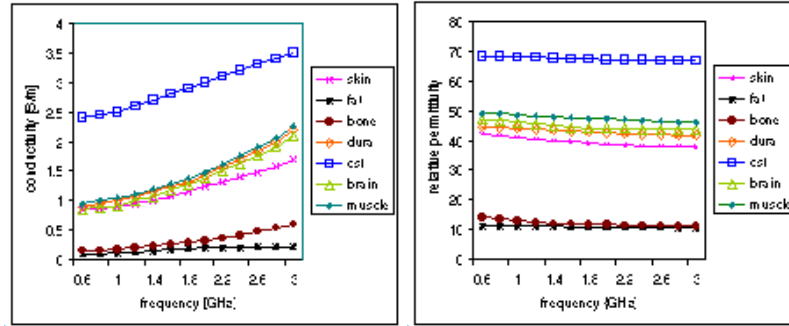


Fig. 1. Frequency dependence of the conductivity and the permittivity of several tissues, in the GSM frequency range.

Our research is focused on the study of two idealized anatomic structures:

- **model A\_6** – a human (adult) head with external ellipsoidal shape, composed by six tissue layers (skin, fat, bone, dura, csf and brain) [8, 9], and
- **model B\_4** – a planar layered structure representing the human trunk, composed by four tissue layers (skin, fat, muscle and bone).

(4) The electromagnetic coupling between the antenna and the exposed body is quantified by the so called *power transfer factor*, defined as the ratio between the emitted and the absorbed time averaged active powers; this factor depends on: the antenna type, the shape and structure of the body and the distance between the antenna and the body. We adopted for this study a constant emitted power of 1 W, regardless the type of the antenna, and for a more realistic evaluation we have expressed the dosimetric quantities in rated (scaled) values, as it is seen in the results section.

(5) Our purpose is to determine some quantitative and qualitative information on dosimetric parameters inside the body exposed to electromagnetic radiation and to relate them to prescriptions stated in human exposure guidelines and standards. Also

the configuration of the human body in the vicinity of the antenna does not present any obvious and accurate symmetry, we speculate a reduction of the 3D problem, to a 2D idealized model, based on axial symmetry around the antenna longitudinal axis [8, 9]. This approximation proves to be satisfactory for global estimates, like specific energy absorption rate, or power deposition in a tissue layer, both rated to the total power absorbed by the body, or to the power emitted by the antenna. The approximation is also favored by the fact that the penetration depth of the electromagnetic radiation in biological tissue at microwave frequencies is small ( $< 30$  mm) and the electromagnetic phenomena are superficial. In order to compare the quantitative results obtained with the simplified 2D models with the more realistic ones obtained with 3D models one have to take into account *the power transfer factor* (defined above), whose value is *one* for the 2D models (due to the axial symmetry) and *smaller than one* for the realistic 3D models (generally dependent of the distance between the antenna and the body).

#### 4. Electromagnetic problem formulation

The numerical computation used for the 2D FEM model is based on the FEM-LAB software [10], the *Electromagnetics Module*, in the *axisymmetric transversal magnetic (TM) waves* application mode, *time-harmonic* submode. The wave equations are applied for dispersive media, characterized by the complex electric permittivity  $\underline{\varepsilon}$

$$\nabla \times \left( \frac{1}{\mu_0} \nabla \times \underline{\mathbf{E}} \right) - \omega^2 \underline{\varepsilon} \underline{\mathbf{E}} = 0, \quad \nabla \times \left( \frac{1}{\underline{\varepsilon}} \nabla \times \underline{\mathbf{H}} \right) - \omega^2 \mu_0 \underline{\mathbf{H}} = 0, \quad (1)$$

where the unknown field variables, in the cylindrical coordinate system and in complex form are the electric and the magnetic field strengths:

$$\underline{\mathbf{H}}(r, z, t) = \underline{H}_\varphi(r, z) \mathbf{e}_\varphi e^{i\omega t}, \quad \underline{\mathbf{E}}(r, z, t) = (\underline{E}_r(r, z) \mathbf{e}_r + \underline{E}_z(r, z) \mathbf{e}_z) e^{i\omega t} \quad (2)$$

The computational domain (Fig. 2) is limited with *low-reflecting* boundary conditions

$$\mathbf{n} \times \left( \frac{\underline{\varepsilon}}{\mu_0} \right)^{1/2} \underline{\mathbf{E}} - H_\varphi = -2H_{\varphi 0}, \text{ where } H_{\varphi 0} = 0, \quad (3)$$

and the boundary on the  $(Oz)$  axis satisfies *axial symmetry* conditions

$$E_r = 0, \quad \frac{\partial E_z}{\partial r} = 0, \quad \frac{\partial H_\varphi}{\partial r} = 0. \quad (4)$$

The radiation source is introduced through a nonhomogeneous *magnetic field* boundary condition, simulating the antenna. The magnetic field condition is adjusted for each model, so that the emitted power is constant (1 W) in all studied cases. The FEMLAB linear stationary solver is based on Gaussian elimination. The FEM mesh is composed of triangular elements (Delaunay mesh with Lagrange quadratic elements),

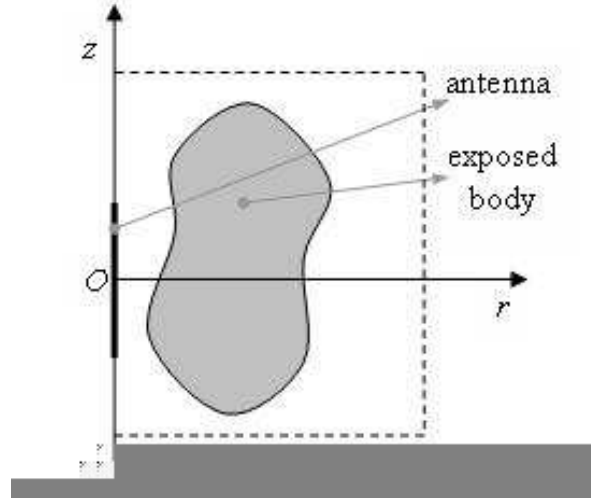


Fig. 2. The general structure of the computational domain based on axial symmetry.

and two accuracy tests were performed to settle its parameters: the constant radiated power and an energetic balance (the radiated power compared with the sum of the power absorbed in the body and the power radiated in the antenna far field). The optimal mesh in our examples is settled to approx. 150 000 elements.

## 5. Equivalent dielectric properties

The anatomical structures presented above are further reduced to more simplified models, having the same external shape and dimensions and an inner homogeneous structure. The electric properties of the reduced model ( $\sigma_{equiv}$  respectively  $\varepsilon_{equiv}$ ) are computed with the 2D FEM model described earlier, by energy based equivalence, considering that the total absorbed power and total electric energy have the same values in the heterogeneous (composed by  $i$  different subdomaines) and equivalent homogeneous models [9]:

$$\int_i \sigma_i (E_i)^2 dv = \sigma_{equiv} \int_i (E_i)^2 dv, \quad (5)$$

$$\int_i \frac{1}{2} \varepsilon_i (E_i)^2 dv = \frac{1}{2} \varepsilon_{equiv} \int_i (E_i)^2 dv. \quad (6)$$

The method presented above is applied to compute the equivalent dielectric properties of the following reduced models, derived from **model A\_6** and **model B\_4** presented above:



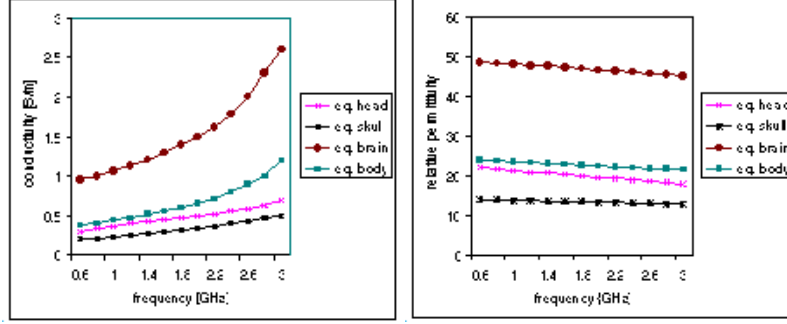


Fig. 3. Frequency dependence of the equivalent dielectric properties (conductivity and relative permittivity), in the GSM frequency range.

- **model A\_2** – *reduced heterogeneous head model with two subdomains: equivalent skull* (skin+fat+bone) and **equivalent brain** (dura+csf+brain);
- **model A\_1** – *reduced homogeneous head model: equivalent head* tissue;
- **model B\_2** – *reduced heterogeneous trunk model: skin* (same properties and thickness as in model B) and **equivalent body** tissue (fat+muscle+bone).
- **model B\_1** – *reduced homogeneous trunk model: equivalent body* tissue.

Fig. 3 presents the frequency dependence of the mentioned equivalent dielectric properties computed in the GSM frequency range.

The equivalent properties are useful in the design of experimental mannequins. Tissue equivalent “phantoms” are used instead of real bodies in the experimental dosimetry [7]. Miniature isotropic  $E$ -field sensors are commonly used as implantable probes. The sensor is immersed into tissue equivalent liquid, and the internal electric field in the phantom is measured; the  $SAR$  is then calculated from internal  $E$ -field. A typical phantom designed for the certification of communication equipment is described in [7] (i.e., the Specific Anthropomorphic Mannequin – SAM) and it consists of a 2 mm polyurethane shell ( $\sigma_{shell} = 0.0012$  S/m respectively  $\varepsilon_{shell} = 5$ ), filled with **simulant tissue solution** ( $\sigma_{simulant} = 0.7$  S/m,  $\varepsilon_{simulant} = 48$ , at 0.835 GHz and  $\sigma_{simulant} = 1.7$  S/m,  $\varepsilon_{simulant} = 41$  at 1.9 GHz). As one could see, the mentioned values of the **simulant tissue solution** are comparable to the **equivalent brain** in Fig. 3.

## 6. Electric field strength and penetration depth in the exposed body

We computed the  $E$ -field strength (rms-values) distribution inside the two anatomical structures exposed in the near field of the antenna at different frequencies; the antenna is placed at 0.01 m distance from the body surface. In the case of the head (models of type A) the antenna is a center fed, half wavelength dipole and in the case of the trunk (models of type B) the antenna is a lower end fed, quarter

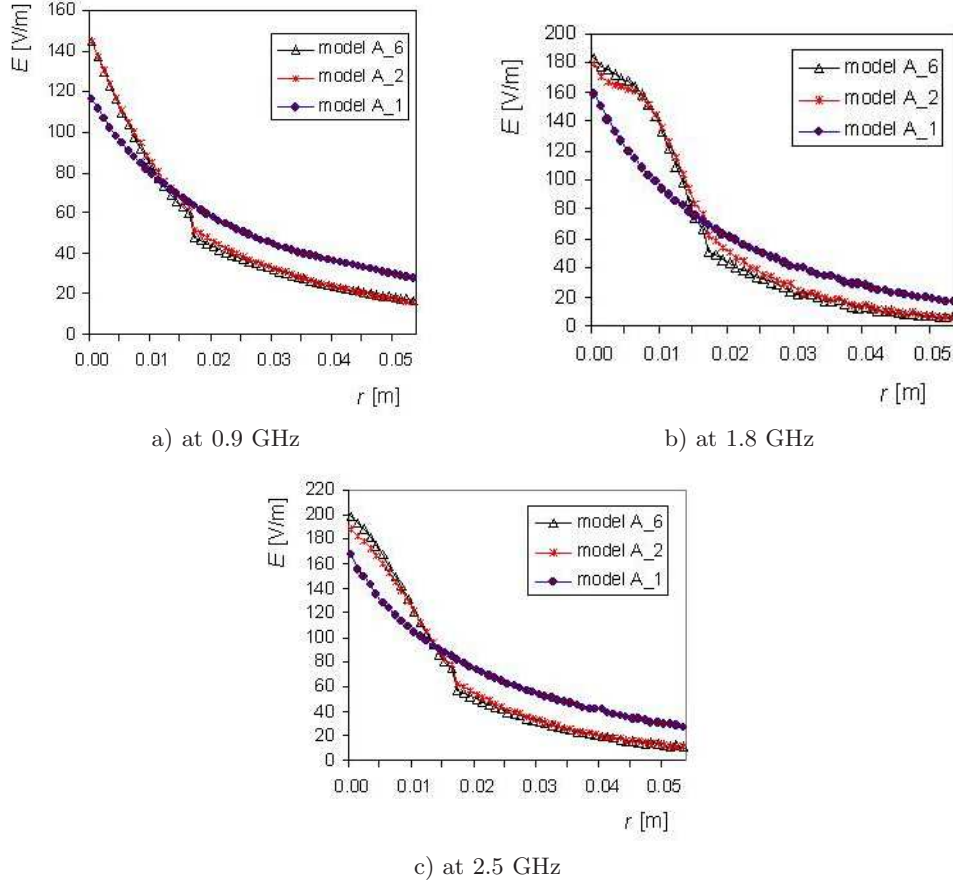


Fig. 4.  $E$  distribution (rms values) versus distance  $r$ , measured from the surface of the skin – head models.

wavelength monopole. The antenna emitted power is set at 1 W in all cases in order to express the  $E$ -field as rated (per power) values, giving the possibility to better analyze them and to scale them for any other value of the emitted power. Electric field penetration in the exposed head (the models of type A) is presented in Fig. 4, at the main frequencies in the MW considered range (0.9 GHz, 1.8 GHz and 2.5 GHz). The presented distributions of the electric field strength are computed on the axes of maximal values.

In a 3D numerical model, the degree of heterogeneity is significant for the complexity of the model and computational resources; thin layers (like skin, fat, dura and csf) should be covered with a very dense FEM mesh. Consequently, any possibility to simplify the structure is appreciated. The results presented in Fig. 4 support the good agreement among equivalent models. This is a good reason to use the values

determined for equivalent dielectric properties in the design of 3D FEM models or in experimental phantoms for the human body.

The reduced heterogeneous structures (**models A.2 and B.2**) prove to achieve the most desirable compromise between the accuracy of anatomical representation and the economy of computational resources. The *E*-field *penetration depth* was estimated in the GSM frequency range for the presented models. Figures 5 and 6 show that the frequency dependence of the penetration depth for the heterogeneous models (**A.6** and **A.2**, respectively **B.4** and **B.2**) is very similar, while the same function for the homogeneous model (**A.1**, respectively **B.1**) has a noticeable different distribution.

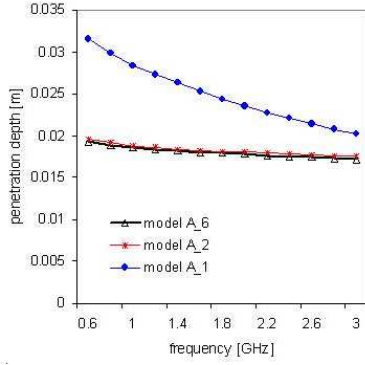


Fig. 5. Frequency dependence of the penetration depth, for type A (head) models.

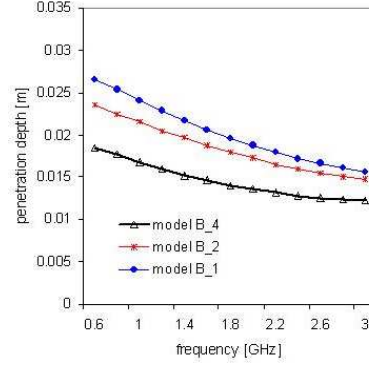
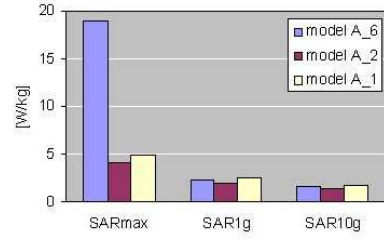


Fig. 6. Frequency dependence of the penetration depth, for type B (trunk) models.

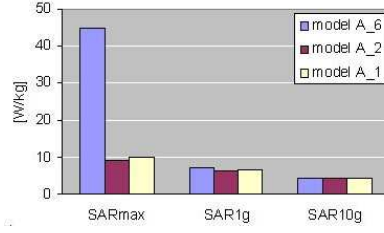
In the heterogeneous models, the presence of the superficial layers (skin fat and bone) has a screening effect for the electric field. The skin has a relatively high permittivity and concentrates the electric field and the absorbed power at the surface of the body; one could see in figure 4 the high initial *E* values. The fat and the bone, with their lower permittivities act like a barrier for *E*-field penetration. These observations support the lower level of the penetration depth for heterogeneous models and the almost constant value regardless the frequency.

## 7. SAR evaluation

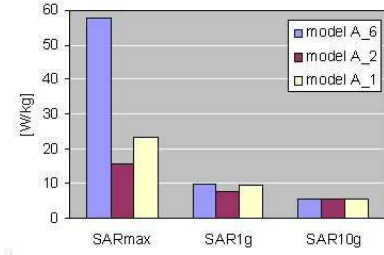
We mentioned in Section 2 that the main dosimetric quantity in microwave exposure of living bodies is the specific energy absorption rate (*SAR*), considered as the *basic restriction* by the most referred standards. It is defined as the absorbed power per unit mass at infinitesimal volume of tissue ( $SAR = \sigma E^2 / \rho$  [W/kg]sg, where *E* is the rms value of the electric field strength,  $\sigma$  is the electric conductivity and  $\rho$  is the mass density of the tissue). *SAR* distribution depends on several factors: the incident field parameters (near or far field), geometric parameters (shape and structure) of the exposed body, physical properties of the tissues (as lossy dielectrics), ground/screen/reflector effects of other objects in the field near the body.



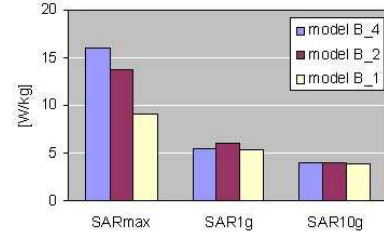
a) at 0.9 GHz.



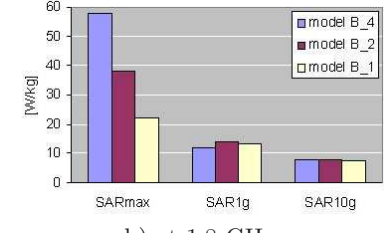
b) at 1.8 GHz.



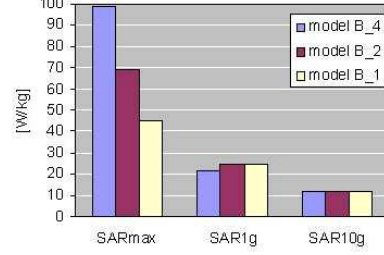
c) at 2.5 GHz.

Fig. 7. *SAR* estimates for the type A (head) models (at 1 W emitted power).

a) at 0.9 GHz.



b) at 1.8 GHz.



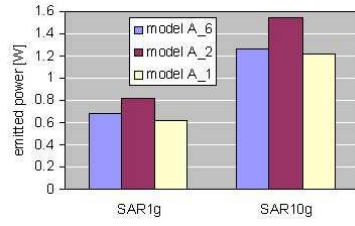
c) at 2.5 GHz.

Fig. 8. *SAR* estimates for the type B (trunk) models (at 1W emitted power).

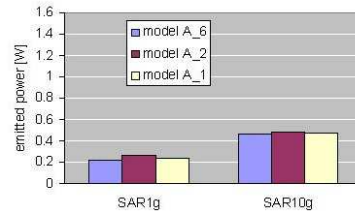
In this section we discuss the significance of the mass-normalized *SAR* over 1 g versus 10 g of tissue, which represents, for an average biological tissue with the mass density of  $1 \text{ kg/m}^3$ , a volume associated to a cube with the edge of 0.01 m, respectively 0.022 m. The standards suggest that the cube volume should include the maximal local *SAR* values in the exposed area. Because the *SAR*, like the *E*-field in the exposed region is maximal at the surface of the body, the volume is selected with one face on the body surface (in practical cases the curvature could be neglected). Figures 7 and 8 display, for three significant frequencies and for the models presented in this case study, the following *SAR* quantities: the local maximal *SAR* (*SAR*<sub>max</sub>) and the averaged *SAR* values for 1 g and 10 g of tissue (*SAR*<sub>1g</sub>, *SAR*<sub>10g</sub>). The emitted radiation power is 1 W in all studied cases and the exposed body is placed at the same distance (0.01 m) of the antenna.

One could observe several characteristics:

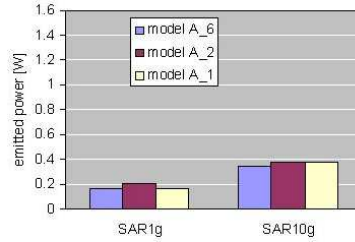
*SAR*<sub>max</sub> is highly dependent on the model heterogeneity; the layered models



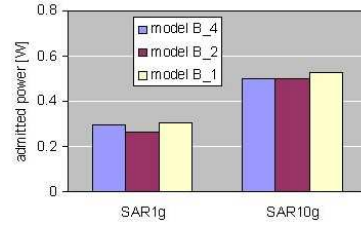
a) at 0.9 GHz.



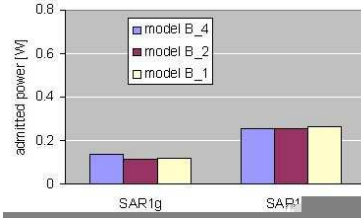
b) at 1.8 GHz.



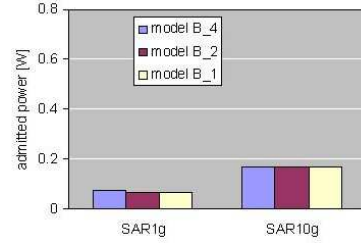
c) at 2.5 GHz.



a) at 0.9 GHz



b) at 1.8 GHz



c) at 2.5 GHz

Fig. 9. Antenna emitted power to produce SAR 1 g and SAR 10 g admitted values in the type A (head) models.

Fig. 10. Antenna emitted power to produce SAR 1 g and SAR 10 g admitted values in the type B (trunk) models.

A\_6 and B\_4 concentrate a large amount of the total absorbed power at the surface of the body because the skin (the peripheral layer which is very thin) has a higher dielectric permittivity than fat and bone; however, in all models, *SAR* distribution is highly focused in the area proximal to the antenna, and decreases rapidly inside the body (as the penetration depth and electric field distribution show).

On the contrary, the averaged values, both for 1 g and 10 g of tissue, are similar for all the compared models; they are insignificantly affected by the heterogeneity. This is because the volume of integration is not negligible in size and its characteristic dimension is larger than the thickness of the tissue layers.

For the same exposure conditions, *SAR*10 g is considerably smaller than *SAR*1 g because *SAR* distribution is highly nonuniform and decreases rapidly with the distance from the peak. Following the restrictions stated by the exposure standards, one could see the opposite relation: the limit imposed by ICNIRP for 10 g (2 W/kg) is more permissive (higher) than the limit imposed by IEEE/ANSI for 1 g (1.6 W/kg).

A more confusing situation appears when we try to solve the “inverse problem”, that is to estimate the admissible power of the radiation source, that produces the SAR1g, respectively the SAR10 g permitted by the standards. Figures 9 and 10 present the values of the emitted power able to produce SAR1g = 1.6 W/kg and SAR10 g = 2 W/kg, for each of the models considered in our case study.

The three compared models give similar values for the maximal admitted antenna power, because the heterogeneity of the model significantly affects local SAR values, but seems to be less important for spatial averaged values, especially at higher frequencies (the depth of penetration decreases with the frequency rise). However, the controversy between the two referred standards is evident and confirmed in all considered examples: the limit values derived from the ICNIRP guidelines [1] are twice more permissive than the ANSI/IEEE standard [2].

## 8. Conclusions

The construction of the simplified 2D models used in this study arises from the necessity to evaluate dosimetric parameters in layered structures like anatomical tissues when exposed to microwave radiation in wireless communications. Compared with more sophisticated models, the 2D models demonstrate advantages in economy of resources, accessibility and rapidity, while the results are sufficiently accurate for global estimates and for comparison with experimental *SAR* and *E* distributions from measurements on phantom human models. The method of equivalence between the heterogeneous anatomical structures and the homogeneous equivalent domains could be applied in different configurations. The results are useful for the optimal design of 3D models. This work presents the electric field distribution inside models representative for parts of the human body (head and trunk) in different exposure conditions. The penetration depth and the specific energy absorption rate are also computed. A critical study for the evaluation of *SAR* shows some controversies produced by important differences between the most known and referred human exposure international standards; this situation is quite confusing for manufacturers and for end-users of wireless devices. The normalization method for *SAR* limit evaluation should be reconsidered and made more appropriate to the structure of the exposed body; we consider that the characteristic dimension of the integration volume should be made smaller for a more accurate estimate both in numerical and experimental models. Besides, the averaged value on a volume in the shape of a cube, over tissues with different physical properties seems to have a poor physical significance. A more localized *SAR* evaluation could be important both for the assessment of thermal and non-thermal biological effects.

It is the role of international standardization authorities, International Electrical Commission (IEC), European Committee for Standardization in Electrotechnic (CENELEC), The Institute of Electrical and Electronic Engineers (IEEE) to synthesize reliable research and to edit technical standards for the design, manufacture and conditions of use of the electric and electronic equipment.

**Acknowledgment.** The work presented is part of the grant CNCSIS 357/2005.

## References

- [1] International Commission on Non-Ionizing Radiation Protection (ICNIRP), *Guidelines for Limiting Exposure to Time-varying Electric, Magnetic and Electromagnetic Fields (up to 300 GHz)*, Health Phys., **74** (1998), No. 4, pp. 494–522.
- [2] ANSI/IEEE, *Safety Levels with respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz*, IEEE Standard C95.1-1999.
- [3] Australian Radiation Protection and Nuclear Safety Agency (ARPANSA), *Maximum Exposure Levels to Radiofrequency Fields – 3 kHz to 300 GHz*, Radiation Protection Series Publication, no. 3, 2002.
- [4] Council of the European Union, *Council Recommendation of 12 July 1999 on the Limitation of Exposure of the General Public to Electromagnetic Fields (0 Hz to 300 GHz)*, Official Journal of the European Communities L199/519/EC, 59-70 (1999).
- [5] CENELEC, *Limitation of Human Exposure to Electromagnetic Fields from Devices Operating in the Frequency Range 0 Hz to 10 GHz, used in Electronic Article Surveillance (EAS), Radio Frequency Identification (RFID) and Similar Applications*, EN 50364:2002 [http://europa.eu.int/comm/health/ph\\_determinants/environment/EMF/emf\\_en.htm](http://europa.eu.int/comm/health/ph_determinants/environment/EMF/emf_en.htm)
- [6] Gabriel C., Gabriel S., *Dielectric Properties of Body Tissues at RF and Microwave Frequencies*, Report for Armstrong Laboratory (AFMC), Occupational and Environmental Health Directorate - Radiofrequency Radiation Division, USA, 2002.
- [7] CTIA – Certification Program Test Plan for Mobile Station Over the Air Performance Cellular Telecommunications & Internet Association Method of Measurement for Radiated RF Power and Receiver Performance, Revision 2.0, March 2003
- [8] Morega Mihaela, Machedon Alina, *Numerical Models of Biological Tissues for Applications in Microwave Dosimetry*, in *Proceedings Third International Workshop on Mathematical Modelling of Environmental and Life Sciences Problems*, Constanța, 2004, Editura Academiei Române, pp.254–263.
- [9] Morega Mihaela, Machedon Alina, *EMF Penetration in Biological Tissue when Exposed in the Near Field of a Mobile Phone Antenna*, 4<sup>th</sup> International Symposium on Advanced Topics in Electrical Engineering, ATEE-2004, Bucharest, CD-ROM, ISBN 973-7728-31-9.
- [10] FEMLAB 3.1, User's Guide and Electromagnetics Module, COMSOL AB, 2004.





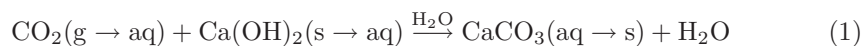
## Lower bounds on the weak solution of a moving-boundary problem describing the carbonation penetration in concrete

Adrian Muntean<sup>\*‡</sup> and Michael Böhm

In this note, we discuss a 1D partly dissipative moving sharp-interface reaction-diffusion system modeling concrete carbonation. We point out a way to obtain non-trivial lower bounds on the concentrations of the chemically active species. A FEM approximation of the solution illustrates numerically the behavior of the concentration profiles and interface position. The lower bounds of  $\text{Ca}(\text{OH})_2(\text{aq})$  concentration obtained numerically are compared with the theoretical lower bounds.

### 1. Introduction

This research is motivated by the macroscopic modeling of the concrete carbonation process by moving the interface where the carbonation reaction



is localized. The interface, which we denote by  $\Gamma(t)$ , advances from the outside boundary inwards the material. It separates the region  $\Omega_1(t)$ , in which all available  $\text{Ca}(\text{OH})_2(\text{aq})$  has been depleted, from the yet unreacted part  $\Omega_2(t)$ . In front of the advancing interface the concentration of freely diffusing  $\text{CO}_2$  molecules is zero. Figure 1 illustrates the situation. The understanding of the process and the ability to calculate, and therefore, to predict the penetration of aggressive carbon dioxide from air inward through the unsaturated porous concrete matrix towards the steel reinforcement is

---

<sup>\*</sup> Centre for Industrial Mathematics, Faculty of Mathematics, University of Bremen, Postfach 330440, 28334 Bremen, Germany, e-mail: [muntean](mailto:muntean), [mbohm@math.uni-bremen.de](mailto:mbohm@math.uni-bremen.de)

<sup>‡</sup> This work was completed with the partial support of DFG SPP1122 *Prediction of the Course of Physicochemical Damage Processes Involving Mineral Materials*.

essential to forecast the penetration of a complex spectrum of really corrosive species like chlorides or sulfates. We present in Section 2 a coupled PDE-ODE model for the simultaneous determination of concentration profiles of the active species and of the position of the carbonation interface. Section 3 contains the weak formulation of the problem and the main results. In Section 4, we point out a way to obtain non-trivial lower bounds for  $\text{Ca}(\text{OH})_2(\text{aq})$  concentration. We use a FEM approximation of the solution to illustrate numerically the behavior of the concentration profiles and interface position. Finally, we compare the numerical lower bounds of  $\text{Ca}(\text{OH})_2(\text{aq})$  with the theoretical estimates. We report on these aspects in Section 5.

## 2. The Sharp-Interface Carbonation Model $P_\Gamma$

The evolution of the interface between carbonated and uncarbonated parts in concrete materials can be modeled (cf. [1, 5, 7]) by the following 1D moving-boundary problem: Determine the concentrations<sup>1</sup>  $\bar{u}_i(x, t)$ ,  $i \in \mathcal{I}$  and the interface position  $s(t)$  which satisfy for all  $t \in S_T$  the equations:

$$\left\{ \begin{array}{l} (\phi\phi_w \bar{u}_i)_{,t} + (-D_i \nu_{i2} \phi\phi_w \bar{u}_{i,x})_x = +f_{i, \text{Henry}}, \quad x \in \Omega_1(t), \quad i \in \{1, 2\}, \\ (\phi\phi_w \bar{u}_3)_{,t} + (-D_3 \phi\phi_w \bar{u}_{3,x})_x = +f_{\text{Diss}}, \quad x \in \Omega_2(t), \\ (\phi\phi_w \bar{u}_4)_{,t} = +f_{\text{Prec}} + f_{\text{Reac}\Gamma}, \quad x \in \Gamma(t), \\ (\phi\bar{u}_5)_{,t} + (-D_5 \phi\bar{u}_{5,x})_x = +f_{\text{Reac}\Gamma}, \quad x \in \Omega_1(t), \\ (\phi\bar{u}_6)_{,t} + (-D_6 \phi\bar{u}_{6,x})_x = 0, \quad x \in \Omega_2(t). \end{array} \right. \quad (2)$$

The initial and boundary conditions are  $\phi\phi_w \nu_{i2} \bar{u}_i(x, 0) = \hat{u}_{i0}$ ,  $i \in \mathcal{I}$ ,  $x \in \Omega(0)$ ,  $\phi\phi_w \nu_{i2} \bar{u}_i(0, t) = \lambda_i$ ,  $i \in \mathcal{I}_1$ ,  $\bar{u}_{i,x}(L, t) = 0$ ,  $i \in \mathcal{I}_2$ ,  $x \in \Omega_2(t)$ , where  $t \in S_T$ . Specific to our problem, we impose the following interface conditions

$$\left\{ \begin{array}{l} [j_1 \cdot n]_{\Gamma(t)} = -\tilde{\eta}_\Gamma(s(t), t) + s'(t)[\phi\phi_w \bar{u}_1]_{\Gamma(t)}, \\ [j_i \cdot n]_{\Gamma(t)} = s'(t)[\phi\phi_w \nu_{i2} \bar{u}_i]_{\Gamma(t)}, \quad i \in \{2, 5, 6\}, \\ [j_3 \cdot n]_{\Gamma(t)} = -\tilde{\eta}_\Gamma(s(t), t) + s'(t)[\phi\phi_w \bar{u}_3]_{\Gamma(t)}, \end{array} \right. \quad (3)$$

$$s'(t) = \alpha \frac{\tilde{\eta}_\Gamma(s(t), t)}{\phi\phi_w \bar{u}_3(s(t), t)}, \quad s(0) = s_0, \quad (4)$$

where  $\nu_{i2} := 1$  ( $i \in \mathcal{I} - \{2\}$ ),  $\nu_{22} := \frac{\phi_a}{\phi_w}$ ,  $j_i := -D_i \nu_{i\ell} \phi\phi_w \bar{u}_i$  ( $i, \ell \in \mathcal{I}_1 \cup \mathcal{I}_2$ ) are the corresponding diffusive fluxes, and  $\alpha > 0$ . Here  $D_i$ ,  $L$ , and  $s_0$  are strictly positive constants,  $\lambda_i$  are prescribed in agreement with the environmental conditions to which

<sup>1</sup>We involve the following mass concentrations:  $\bar{u}_1 := [\text{CO}_2(\text{aq})]$ ,  $\bar{u}_2 := [\text{CO}_2(\text{g})]$ ,  $\bar{u}_4 := [\text{CaCO}_3(\text{aq})]$ , and  $\bar{u}_5 := [\text{H}_2\text{O}]$  for the species present in the region  $\Omega_1(t) := [0, s(t)]$ ;  $\bar{u}_3 := [\text{Ca}(\text{OH})_2(\text{aq})]$  and  $\bar{u}_6 := [\text{H}_2\text{O}]$  for those in  $\Omega_2(t) := [s(t), L]$ . Here  $t \in S_T := ]0, T[$ ,  $T \in ]0, \infty[$ . We use the set of indices  $\mathcal{I} := \mathcal{I}_1 \cup \{4\} \cup \mathcal{I}_2$ .  $\mathcal{I}_1 := \{1, 2, 5\}$  points out the active concentrations in  $\Omega_1(t)$ , and  $\mathcal{I}_2 := \{3, 6\}$  refers to the active concentrations present in  $\Omega_2(t)$ . We assume that  $\text{CaCO}_3(\text{aq})$  is not transported in  $\Omega := \Omega_1(t) \cup \Gamma(t) \cup \Omega_2(t)$ , therefore the only partly dissipative character of the model.

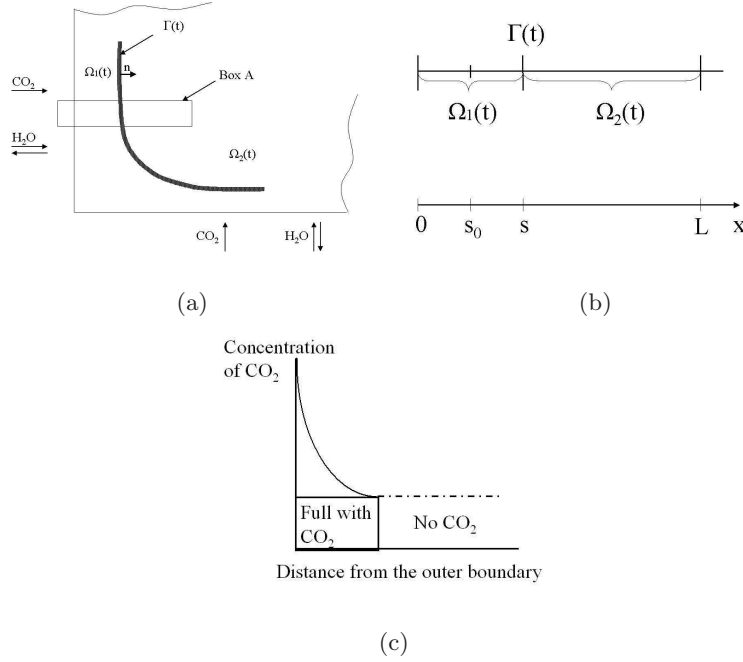


Fig. 1. (a) Basic geometry for  $P_T$  model. (b) Schematic 1D geometry. (c) Definition of the reaction interface, see also Figure 1 in [3].

$\Omega$  is exposed [2]. The initial conditions  $\hat{u}_{i0} > 0$  are determined by the chemistry of the cement. The hardened mixture of aggregate, cement, and water (i.e. the concrete) imposes ranges for the porosity  $\phi > 0$  and also for the water and air fractions,  $\phi_w > 0$  and  $\phi_a > 0$ , [2]. For a derivation of (4), see [1, 7]. The terms  $f_i$  represent (see Section 3) the r.h.s. of the  $i$ th equation in (2), where  $f_{i, Henry}$ ,  $f_{Diss}$ ,  $f_{Prec}$ , and  $f_{Reac\Gamma}$  are sources or sinks by Henry-like interfacial transfer mechanisms, dissolution, precipitation, and carbonation reactions. Typical examples are (cf. [1, 7, 4, 8]):

$$\begin{cases} f_{i, Henry} := (-1)^i P_i (\phi \phi_w \bar{u}_1 - Q_i \phi \phi_a \bar{u}_2) & (P_i > 0, Q_i > 0), \quad i \in \{1, 2\}, \\ f_{Diss} := S_{3, diss} (u_{3, eq} - \phi \phi_w \bar{u}_3), & S_{3, diss} > 0, \quad f_{Prec} := 0, \quad f_{Reac\Gamma} := \tilde{\eta}_\Gamma. \end{cases} \quad (5)$$

Let  $\bar{u}$  denote the vector of concentrations  $(\bar{u}_1, \dots, \bar{u}_6)^t$  and  $M_\Lambda$  be the set of parameters  $\Lambda := (\Lambda_1, \dots, \Lambda_m)^t$  that are needed to describe the reaction rate. For our purposes, it suffices to assume that  $M_\Lambda$  is a non-empty compact subset of  $\mathbb{R}_+^m$ . We introduce the function

$$\bar{\eta}_\Gamma : \mathbb{R}^6 \times M_\Lambda \rightarrow \mathbb{R}_+ \text{ by } \bar{\eta}_\Gamma(\bar{u}(x, t), \Lambda) := k \phi \phi_w \bar{u}_1^{\bar{p}}(x, t) \bar{u}_3^{\bar{q}}(x, t), \quad x = s(t). \quad (6)$$

Here  $m := 3$  and  $\Lambda := \{\bar{p}, \bar{q}, k \phi \phi_w\} \in \mathbb{R}_+^3$ . We define the rate of reaction (1)  $\tilde{\eta}_\Gamma(s(t), t)$ , which arises in (3) and (4), by

$$\tilde{\eta}_\Gamma(s(t), t) := \bar{\eta}_\Gamma(\bar{u}(s(t), t), \Lambda). \quad (7)$$

Equations (2)–(7) define the model  $P_\Gamma$ .

### 3. Notation. Weak Formulation. Main Results

For each  $i \in \mathcal{I}_1 \cup \mathcal{I}_2$ , we denote  $H_i := L^2(a, b)$ , where we set  $[a, b] := [0, 1]$  for  $i \in \mathcal{I}_1$  and  $[a, b] := [1, 2]$  for  $i \in \mathcal{I}_2$ . Moreover,  $\mathbb{H} := \prod_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} H_i$ ,  $V_i = \{u \in H^1(a, b) : u_i(a) = 0\}$ ,  $i \in \mathcal{I}_1$ ,  $V_i := H^1(a, b)$ ,  $i \in \mathcal{I}_2$ , and  $\mathbb{V} = \prod_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} V_i$ , [7]. In addition,  $\|\cdot\| := \|\cdot\|_{L^2(a, b)}$  and  $\|\cdot\| := \|\cdot\|_{H^1(a, b)}$ . If  $(X_i : i \in \mathcal{I})$  is a sequence of given sets  $X_i$ , then  $X^{|\mathcal{I}_1 \cup \mathcal{I}_2|}$  denotes the product  $\prod_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} X_i := X_1 \times X_2 \times X_3 \times X_5 \times X_6$ . Note that sometimes  $u(1)$  and  $u_{,y}(1)$  replace  $u(1, t)$  and  $u_{,y}(1, t)$ .

We re-formulate the model  $P_\Gamma$  in terms of macroscopic quantities by performing the transformation of all concentrations into volume-based concentrations via  $\hat{u}_i := \phi\phi_w \bar{u}_i$ ,  $i \in \{1, 3, 4\}$ ,  $\hat{u}_2 := \phi\phi_a \bar{u}_2$ ,  $\hat{u}_i := \phi \bar{u}_i$ ,  $i \in \{5, 6\}$ . We employ the Landau transformations  $(x, t) \in [0, s(t)] \times \bar{S}_T \mapsto (y, \tau) \in [a, b] \times \bar{S}_T$ ,  $y = \frac{x}{s(t)}$  and  $\tau = t$ , for  $i \in \mathcal{I}_1$ ,  $(x, t) \in [s(t), L] \times \bar{S}_T \mapsto (y, \tau) \in [a, b] \times \bar{S}_T$ ,  $y = a + \frac{x-s(t)}{L-s(t)}$  and  $\tau = t$ , for  $i \in \mathcal{I}_2$  to map  $(P_\Gamma)$  onto a region with fixed boundaries. We re-label  $\tau$  by  $t$  and introduce the new concentrations, which act in the auxiliary  $y$ - $t$  plane, by  $u_i(y, t) := \hat{u}_i(x, t) - \lambda_i(t)$  for all  $y \in [a, b]$  and  $t \in S_T$ . Thus, the model equations are reduced to

$$\begin{aligned} (u_i + \lambda_i)_{,t} - \frac{1}{s^2(t)}(D_i u_{i,y})_{,y} &= f_i(u + \lambda) + y \frac{s'(t)}{s(t)} u_{i,y}, \quad i \in \mathcal{I}_1, \quad (8) \\ (u_i + \lambda_i)_{,t} - \frac{1}{(L - s(t))^2}(D_i u_{i,y})_{,y} &= f_i(u + \lambda) + (2 - y) \frac{s'(t)}{L - s(t)} u_{i,y}, \quad i \in \mathcal{I}_2, \end{aligned}$$

where  $u$  is the vectors of concentrations  $(u_1, u_2, u_3, u_5, u_6)^t$  and  $\lambda := (\lambda_1, \lambda_2, \lambda_3, \lambda_5, \lambda_6)^t$  represents the boundary data. The transformed initial, boundary, and interface conditions are

$$u_i(y, 0) = 0, \quad i \in \mathcal{I}_1 \cup \mathcal{I}_2, \quad u_i(a, t) = 0, \quad i \in \mathcal{I}_1, \quad u_{i,y}(b, t) = 0, \quad i \in \mathcal{I}_2, \quad (9)$$

$$\begin{aligned} \frac{-1}{s(t)} D_1 u_{1,y}(1) &= \eta_\Gamma(1, t) + s'(t)(u_1(1) + \lambda_1), \quad \frac{-1}{s(t)} D_2 u_{2,y}(1) = s'(t)(u_2(1) + \lambda_2), \\ \frac{-1}{L - s(t)} D_3 u_{3,y}(1) &= \eta_\Gamma(1, t) - s'(t)(u_3(1) + \lambda_3), \end{aligned} \quad (10)$$

$$\frac{1}{s(t)} D_5 u_{5,y}(1) - \frac{1}{L - s(t)} D_6 u_{6,y}(1) = s'(t)(u_5(1) + \lambda_5 - u_6(1) - \lambda_6), \quad (11)$$

where  $\eta_\Gamma(1, t)$  represents the reaction rate that acts in the  $y$ - $t$  plane. This is defined by

$$\eta_\Gamma(1, t) := \bar{\eta}_\Gamma(\bar{u}(ys(t), t) + \lambda(t), \Lambda) \quad (12)$$

for a given  $\Lambda \in M_\Lambda$ . Finally, two ODE's

$$s'(t) = \eta_\Gamma(1, t) \text{ and } \hat{u}_4'(s(t), t) = f_4(\hat{u}(s(t), t)) \text{ a.e. } t \in S_T, \quad (13)$$

complete the model formulation, where  $\hat{u} := (\hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{u}_5, \hat{u}_6)^t$ . We also assume

$$s(0) = s_0 > 0, \hat{u}_4(s_0, 0) = \hat{u}_{40} \geq 0. \quad (14)$$

(8)–(14) are the transformed equations of  $(P_\Gamma)$ . Note that the model cannot be complete without the law defining  $s'$ , see [7]. Let  $\varphi := (\varphi_1, \varphi_2, \varphi_3, \varphi_5, \varphi_6)^t \in \mathbb{V}$  be an arbitrary test function, and take  $t \in S_T$ . In order to write the weak formulation of (8)–(14) in a compact form, we introduce the auxiliary notation:

$$\begin{cases} a(s, u, \varphi) &:= \frac{1}{s^2} \sum_{i \in \mathcal{I}_1} (D_i u_{i,y}, \varphi_{i,y}) + \frac{1}{(L-s)^2} \sum_{i \in \mathcal{I}_2} (D_i u_{i,y}, \varphi_{i,y}), \\ b_f(u, \varphi) &:= \sum_{i \in \mathcal{I}} (f_i(u), \varphi_i), \\ e(s, s', u, \varphi) &:= \frac{1}{s} \sum_{i \in \mathcal{I}_1} g_i(s, s', u(1)) \varphi_i(1) + \frac{1}{L-s} \sum_{i \in \mathcal{I}_2} g_i(s, s', u(1)) \varphi_i(1), \\ h(s, s', u, y, \varphi) &:= \frac{s'}{s} \sum_{i \in \mathcal{I}_1} (y u_{i,y}, \varphi_i) + \frac{s'}{L-s} \sum_{i \in \mathcal{I}_2} ((2-y) u_{i,y}, \varphi_i), \end{cases}$$

for any  $u \in \mathbb{V}$  and  $\lambda \in W^{1,2}(S_T)^{|\mathcal{I}_1 \cup \mathcal{I}_2|}$ . Moreover,  $v_4(t) := \hat{u}_4(s(t), t)$  for  $t \in S_T$ . For our concrete application (see (5) and (6)), the interface terms  $g_i (i \in \mathcal{I}_1 \cup \mathcal{I}_2)$  are given by

$$\begin{cases} g_1(s, s', u) := \eta_\Gamma(1, t) + s'(t)u_1(1), & g_3(s, s', u) := -\eta_\Gamma(1, t) + s'(t)u_3(1), \\ g_2(s, s', u) := s'(t)u_2(1), & g_5(s, s', u) := g_6(s, s', u) = 0, \end{cases} \quad (15)$$

whereas the volume terms  $f_i (i \in \mathcal{I})$  are defined as

$$\begin{cases} f_1(u) := P_1(Q_1 u_2 - u_1), & f_4(\hat{u}) := +\tilde{\eta}_\Gamma(s(t), t), \\ f_2(u) := -P_2(Q_2 u_2 - u_1), & f_5(u + \lambda) := +\eta_\Gamma(1, t), \\ f_3(u) := S_{3,diss}(u_3 - u_{3,eq}), & f_6(u) := 0. \end{cases} \quad (16)$$

Set  $M_{\eta_\Gamma} := \sup_{u(y,t) \in \mathcal{K}} \{\eta_\Gamma(1, t) : y \in [a, b], t \in S_T\}$  ( $\mathcal{K} := [0, k_1] \times [0, k_3]$ ), where

$$\begin{cases} k_i := \max\{u_{i0}(y) + \lambda_i(t), \lambda_i(t) : y \in [a, b], t \in \bar{S}_T\}, & i = 1, 2, 3, \\ k_4 := \max\{\hat{u}_{40}(x) + M_{\eta_\Gamma} T : x \in [0, s(t)], t \in \bar{S}_T\}, \\ k_j := \max\{u_{j0}(y) + \lambda_j(t) + M_{\eta_\Gamma} T : y \in [a, b], t \in \bar{S}_T\}, & j = 5, 6. \end{cases} \quad (17)$$

**Definition 1 (LOCAL WEAK SOLUTION).** *We call the triple  $(u, v_4, s)$  a local weak solution to problem  $(P_\Gamma)$  if there is a  $S_\delta := ]0, \delta[$  with  $\delta \in ]0, T]$  such that*

$$v_4 \in W^{1,4}(S_\delta), \quad s \in W^{1,4}(S_\delta), \quad (18)$$

$$u \in W_2^1(S_\delta; \mathbb{V}, \mathbb{H}) \cap [\bar{S}_\delta \mapsto L^\infty(a, b)]^{|\mathcal{I}_1 \cup \mathcal{I}_2|} \cap L^\infty(S_\delta; C^{0, \frac{1}{2}-}([a, b])^{|\mathcal{I}_1 \cup \mathcal{I}_2|}), \quad (19)$$

and

$$\begin{cases} (u'(t), \varphi) + a(s, u, \varphi) + e(s, s', u, \varphi) = b_f(u(t) + \lambda(t), \varphi) + \\ + h(s, s', u, y, \varphi) - (\lambda'(t), \varphi) \quad \text{for all } \varphi \in \mathbb{V}, \text{ a.e. } t \in S_\delta, \\ s'(t) = \eta_\Gamma(1, t), \quad \hat{u}'_4(s(t), t) = f_4(\hat{u}(s(t), t)) \text{ a.e. } t \in S_\delta, \\ u(0) = u_0 \in \mathbb{H}, s(0) = s_0, \hat{u}_4(s(0), 0) = \hat{u}_{40}. \end{cases} \quad (20)$$

The only assumptions that we need for the structure of the reaction rate and the model parameters are the following:

(A1) There exists a positive constant  $C_\eta = C_\eta(\Lambda, u_0, u, \lambda, T)$  such that

$$\bar{\eta}_\Gamma(\bar{u}(s(t), t), \Lambda) \leq C_\eta \bar{u}(s(t), t) \text{ for all } t \in S_T.$$

(A2) There exists a constant  $c_g = c_g(s, s', C_\eta)$  such that

$$|e(s, s', u, \varphi)| \leq c_g(u(1), \varphi(1)) \quad \text{for all } \varphi \in \mathbb{V}.$$

(B) The reaction rate  $\eta_\Gamma$  (defined by (12)) is locally Lipschitz with respect to all variables. More precisely, let  $(u^{(i)}, \hat{u}_4^{(i)}, s^{(i)})$ , where  $i \in \{1, 2\}$ , be two solutions corresponding to the sets of data  $\mathcal{D}_i := (u_0^{(i)}, \lambda^{(i)}, \dots, \Lambda^{(i)})^t$ . Set  $\Delta u := u^{(2)} - u^{(1)}$  and  $\Delta \Lambda := \Lambda^{(2)} - \Lambda^{(1)}$ . The Lipschitz condition on  $\Delta \eta_\Gamma := \Delta \bar{\eta}_\Gamma = \bar{\eta}_\Gamma(\bar{u}^{(2)}, \Lambda^{(2)}) - \bar{\eta}_\Gamma(\bar{u}^{(1)}, \Lambda^{(1)})$  reads: There exists a positive constant  $c_L = c_L(\mathcal{D}_1, \mathcal{D}_2)$  such that the inequality  $|\Delta \eta_\Gamma| \leq c_L(|\Delta u| + |\Delta \Lambda|)$  holds pointwise. For a special choice of  $\bar{\eta}_\Gamma$ ,  $\Lambda$ , and hence  $c_L$ , see (6).

(C1)  $k_3 \leq \min_{[1,2] \times \bar{S}_T} \{|u_{3,eq}(y, t)| : y \in [1, 2], t \in S_T\}$ ;

(C2)  $P_1 Q_1 k_2 \leq P_1 k_1 \leq P_2 Q_2 k_2$ ;

(C3)  $Q_2 > Q_1$ .

**Theorem 1** (LOCAL EXISTENCE AND UNIQUENESS, [7]). *Assume the hypotheses (A1)–(C2) and let the following conditions (21)–(25) be satisfied:*

$$\lambda \in W^{1,2}(S_T)^{|\mathcal{I}_1 \cup \mathcal{I}_2|}, \lambda(t) \geq 0 \text{ a.e. } t \in \bar{S}_T, \quad (21)$$

$$u_0 \in L^\infty(a, b)^{|\mathcal{I}_1 \cup \mathcal{I}_2|}, u_0(y) + \lambda(0) \geq 0 \text{ a.e. } y \in [a, b], \quad (22)$$

$$\hat{u}_{40} \in L^\infty(0, s_0), \hat{u}_4(x, 0) \geq 0 \text{ a.e. } x \in [0, s_0], \quad (23)$$

$$\min\left\{\min_{[1,2] \times S_T} \{u_{3,eq}(y, t)\}, S_{3,diss}, P_1, Q_1, P_2, Q_2\right\} > 0, \quad (24)$$

$$0 < s_0 \leq s(t) \leq L_0 < L \text{ for all } t \in \bar{S}_T. \quad (25)$$

Then the following assertions hold:

(a) There exists a  $\delta \in ]0, T[$  such that the problem  $(P_\Gamma)$  admits a unique local solution on  $S_\delta$  in the sense of Definition 1;

(b)  $0 \leq u_i(y, t) + \lambda_i(t) \leq k_i$  a.e.  $y \in [a, b], i \in \mathcal{I}_1 \cup \mathcal{I}_2$  for all  $t \in S_\delta$ , and  $0 \leq \hat{u}_4(x, t) \leq k_4$  a.e.  $x \in [0, s(t)]$  for all  $t \in S_\delta$ ;

(c)  $v_4, s \in W^{1,\infty}(S_\delta)$ .

The main result of this note is:

**Theorem 2** (STRICT LOWER BOUNDS). *Assume that the hypotheses of Theorem 1 hold. Additionally, if (C3) holds and the initial and boundary data are strictly positive, then there exists a range of parameters such that the positivity estimates stated in Theorem 1 (b) are strict for all times.*

#### 4. Sketch of the Proof of Theorem 2

Within this section, we sketch the basic ideas on which the proof of Theorem 2 relies. For the complete proof, we refer to [7]. The main ingredients are the positivity and maximum estimates provided by Theorem 1. The idea of the proof is a refined version of the arguments used in [7] to show the positivity of concentrations. We focus on getting lower bounds for the  $\text{Ca}(\text{OH})_2(\text{aq})$  concentration and only tersely suggest how the lower bounds for the other concentrations are obtained: We choose in the weak formulation (20) the test function

$$\varphi_i = \begin{cases} -[u_3 - u_3^* \gamma_3(t)]^-, & \text{for } i = 3 \\ 0, & \text{otherwise} \end{cases} \in \mathbb{V}_i \text{ for all } i \in \mathcal{I}. \quad (26)$$

In (26) the function  $\gamma_3 \in C^1(\bar{S}_\delta)$  has to be determined such that  $\gamma_3'(t) \leq 0$  for all  $t \in S_\delta$  and  $\gamma_3(0) = 1$ . On this way, we obtain the following identity

$$\begin{aligned} ((u_3 + \lambda_3)_{,t}, \varphi_3) + \frac{D_3}{(L-s)^2} \|\varphi_3\|^2 &= \frac{1}{L-s} (-\eta_\Gamma(u(1) + \lambda), \varphi_3(1)) + \\ &+ s'(u_3(1) + \lambda_3, \varphi_3(1)) + S_{3,diss}(u_3 + \lambda_3 - u_{3,eq}, \varphi_3) + \\ &+ \frac{s'}{L-s} ((2-y)\varphi_{3,y}, \varphi_3). \end{aligned} \quad (27)$$

We collect some of the terms in (27), add the positive term  $-(\zeta u_3^* \gamma_3(t) - \theta u_3^*, \varphi_3)$  to its right-hand side, and select  $u_3^*$  in the interval  $]0, \min_{S_\delta} \lambda_3(t)[$ . The constants  $\zeta > 0$  and  $\theta > 0$  are chosen such that

$$\zeta \gamma_3(t) - \theta > 0 \text{ for all } t \in [0, \infty[ \text{ and } \lim_{t \rightarrow \infty} (\zeta \gamma_3(t) - \theta) > 0. \quad (28)$$

It yields the inequality

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} |\varphi_3|^2 + \frac{D_3}{(L-s)^2} \|\varphi_3\|^2 &\leq \frac{1}{L-s} (-\eta_\Gamma(u(1) + \lambda) + \\ &+ s'(u_3(1) + \lambda_3, \varphi_3(1)) + S_{3,diss} |\varphi_3|^2 + S_{3,diss} (\lambda_3 - u_{3,eq} + u_3^* \gamma_3(t), \varphi_3) - \\ &- (u_3^* \gamma_3'(t) + \lambda_3', \varphi_3) - (\zeta u_3^* \gamma_3(t) - \theta u_3^*, \varphi_3) + \frac{s'}{L-s} ((2-y)\varphi_{3,y}, \varphi_3). \end{aligned} \quad (29)$$

We state the following auxiliary result:

**Lemma 1.** *Let  $\lambda_3 \in W^{1,2}(S_\delta)$  (together with the respective zero extension to  $]0, \infty[$ ) and set  $\sigma := -S_{3,diss} + \zeta$ ,  $\rho := \frac{1}{u_3^*} (\lambda_3' - S_{3,diss} \lambda_3 + S_{3,diss} u_{3,eq})$ ,  $\chi := \rho - \theta$ , where  $\zeta$  and  $\theta$  are constants satisfying the restrictions (28). The following statements hold:*

(i) *If  $\lambda_3 = \text{const.}$ , and if*

$$\zeta \in \left] \max\{S_{3,diss}, \frac{\sigma\theta}{\theta - \rho}\}, +\infty \right[ \text{ and } \theta \in ]\rho, \sigma + \rho[, \quad (30)$$

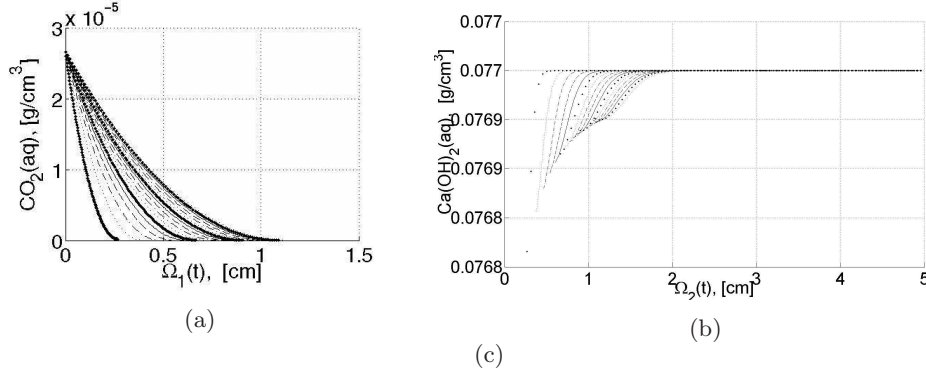


Fig. 2. (a)+(b)  $\text{CO}_2(\text{aq})$  and  $\text{Ca}(\text{OH})_2(\text{aq})$  profiles vs. space. Each curve refers to time  $t = i$  years,  $i \in \{1, \dots, 18\}$ . (c): Interface position against the exp. points “o” (see [2]) after 18 years of exposure to  $\text{CO}_2$  attack.

then the function

$$\gamma_3(t) = -\frac{\chi}{\sigma} + \frac{\sigma + \chi}{\sigma} e^{-\sigma t} \text{ for all } t \in [0, \infty[ \quad (31)$$

is the unique positive solution of the problem

$$\gamma_3'(t) + \sigma \gamma_3(t) + \chi(t) = 0, \quad \gamma_3'(t) < 0 \text{ for all } t \in ]0, \infty[ \text{ provided } \gamma_3(0) = 1. \quad (32)$$

(ii) If  $\lambda_3 \neq \text{const.}$ ,  $\rho(t) > 0$  for all  $t \in ]0, \infty[$ ,  $\zeta$  as it is given in (28), and

$$\sigma - \sigma \int_0^t \chi(\tau) e^{\sigma \tau} d\tau + \chi(t) e^{\sigma t} > 0 \text{ for all } t \in [0, \infty[, \quad (33)$$

then

$$\gamma_3(t) = \left( 1 - \int_0^t \chi(\tau) e^{\sigma \tau} d\tau \right) e^{-\sigma t} \text{ for all } t \in S_\delta \quad (34)$$

is the unique positive solution of (32).

*Proof of Lemma 1.* It is important to note that our choice of (31) (or (34)) relies on  $\rho > 0$ ,  $\sigma > 0$ ,  $\chi < 0$ , (28), (30), and (33). These estimates provide a *strictly positive* and *bounded* function  $\gamma_3$ . The statements (i) and (ii) follow by straightforward verification.  $\square$

Now, with  $\gamma_3(t)$  as in Lemma 1 we force the fourth, fifth and sixth term from the right-hand side of (29) to vanish. Combining Young's inequality and the interpolation inequality (i.e.  $|\varphi_3|_\infty \leq \hat{c}|\varphi_3|^{1-\theta} \|\varphi_3\|^\theta \leq \hat{c}(\xi \|\varphi_3\| + c_\xi |\varphi_3|)$ , where  $\hat{c} = \hat{c}(\theta) > 0$  with  $\theta \in [\frac{1}{2}, 1[$ , and  $\xi, c_\xi$  belong to a compact subset of  $\mathbb{R}_+$ ) we obtain:  $\frac{1}{2} \frac{d}{dt} |\varphi_3|^2 + \frac{D_3}{(L-s)^2} \|\varphi_3\|^2 \leq \frac{s'}{L-s} ((2-y)\varphi_{3,y}, \varphi_3) \leq \frac{\xi}{2} \frac{\|\varphi_3\|^2}{(L-s)^2} + \frac{c_\xi}{2} \hat{c}^{\frac{2}{1-\theta}} |s'|^{\frac{1}{1-\theta}} (L-s)^{\frac{2\theta-1}{1-\theta}} |\varphi_3|^2$ , where



we select  $\xi \in ]0, 2D_3]$ . Since  $\varphi_3(0) = 0$ , we can use Gronwall's inequality to conclude that  $u_3 \geq u_3^* \gamma_3(t)$  for all  $t \in S_\delta$ .

Making use of test functions similar to that in (26), we can find non-trivial lower bounds for all active concentrations, see [7] for details.  $\square$

## 5. Numerical Illustration

We focus on the typical behavior of concentrations and interface penetration in a real-world situation, namely a 18 years old concrete wall made of the cement PZ35F is exposed to  $\text{CO}_2$  attack, see Table 3.1 in [2]. The weak formulation (20) allows us

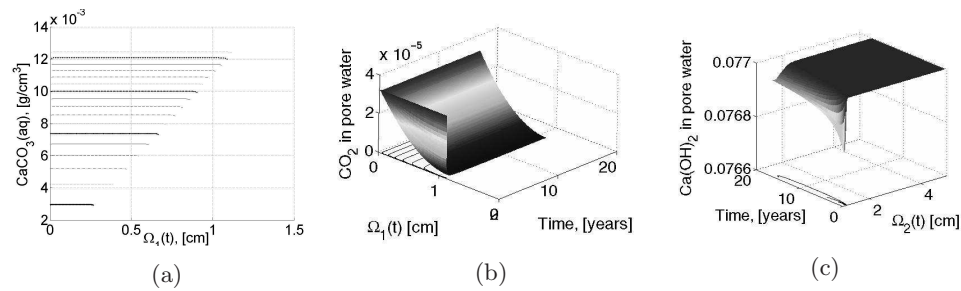


Fig. 3. (a)  $\text{CaCO}_3(\text{aq})$  profiles vs. space. Each curve refers to time  $t = i$  years,  $i \in \{1, \dots, 18\}$ . (b)+(c)  $\text{CO}_2(\text{aq})$  and  $\text{Ca(OH)}_2(\text{aq})$  vs. time and space.

to approximate the underlying moving-boundary problem by using the finite element method. The examples shown in Figures 2–4 are obtained with a uniform 1D Galerkin scheme. Extensive numerical simulations of carbonation scenarios and details on the numerical scheme can be found in [1, 5, 7]. Observe that steep concentration gradients arise near  $\Gamma(t)$  and the calculated interface location is in the experimental range. Figure 4 shows a typical example in which numerical and theoretical lower bounds of  $\text{Ca(OH)}_2(\text{aq})$  concentration are compared. For the chosen parameter set (in which we set  $u_3^* = \frac{1}{100}$  and  $\gamma_3(t)$  cf. (31)), the theoretical bounds underestimate the numerical prediction.

## References

- [1] M. Böhm, J. Kropp, A. Muntean, *On a prediction model for concrete carbonation based on moving interfaces – Interface concentrated reactions*. Berichte aus der Technomathematik 03-03, University of Bremen, 2003.
- [2] D. Bunte, *Zum Karbonatisierungsbedingten Verlust der Dauerhaftigkeit von Außenbauteilen aus Stahlbeton*. Dissertation, Braunschweig, 1994.

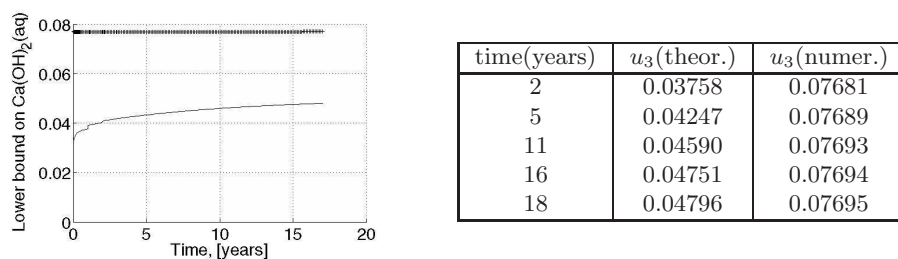


Fig. 4. Theoretical vs. numerical lower bounds of  $\text{Ca}(\text{OH})_2(\text{aq})$  concentration.

- [3] J. Crank, *Chemical and biological problems*. MBP in Heat Flow and Diffusion, Eds. Ockendon, J.R. et al., 62–70, Oxford, 1974.
- [4] Logan, J. D. *Transport Modeling in Hydrogeochemical Systems*. Springer-Verlag, Berlin, 2001.
- [5] A. Muntean, M. Böhm, *On a prediction model for the service life of concrete structures based on moving interfaces*. In Proceedings of ICLODC, Eds. Stangenberg, F. et al., 209–218, Bochum, 2004.
- [6] A. Muntean, M. Böhm, *Dynamics of a moving reaction interface in a concrete wall*, to appear in Free and Moving Boundary Problems. Theory and Applications, Editors J. F. Rodrigues et al., Birkhäuser, 2006.
- [7] A. Muntean, *A Moving-boundary problem: modeling, analysis and simulation of concrete carbonation*, Dissertation, University of Bremen, 2006.
- [8] Ortoleva, P. J. *Geochemical Self-Organization*. OUP, Oxford, 1994.

## Discretization Techniques and Numerical Treatment For First Kind Integral Equations

Elena Pelican<sup>\*‡</sup> and Elena Băutu<sup>\*‡</sup>

Many real-world problems are modeled by first kind integral equations (e.g., backwards heat equation, inverse scattering problems, the hanging cable, geological prospecting, computerized tomography, electric potential problems, diffusion and biochemical reactions, etc). In this paper we realize a comparative study for solving this type of equations, with continuous kernels. As discretization techniques, we use the collocation method (the classical version, and the extended one, proposed by one of the authors in a previous paper) and the quadrature method. We use Kovarik-like algorithms as numerical solvers for the associated linear systems; the experiments present systematic tests. We compare the above mentioned methods for certain types of such equations.

### 1. Introduction

Let  $K : L^2([0, 1]) \longrightarrow L^2([0, 1])$  be the integral operator

$$Kx(t) = \int_0^1 k(t, s)x(s)ds, \quad (1)$$

with continuous kernel  $k : [0, 1] \times [0, 1] \longrightarrow \mathbb{R}$ . An inverse problem for first kind integral equation (denoted from now on by FKIE) is: for a given right hand side  $y \in L^2([0, 1])$ , find  $x \in L^2([0, 1])$  such that

$$Kx(t) = y(t), \quad \forall t \in [0, 1]. \quad (2)$$

The drawback of the problem (2) is its ill-posed nature. Even if we know or can prove that (2) has solution, this one is not stable (with respect to small changes in  $y$ ).

---

<sup>\*</sup> “Ovidius” University of Constanța, Faculty of Mathematics and Computer Science, Romania, e-mail: [epelican@univ-ovidius.ro](mailto:epelican@univ-ovidius.ro), [erogojina@univ-ovidius.ro](mailto:erogojina@univ-ovidius.ro)

<sup>‡</sup> The paper was supported by the PNCDI INFOSOC Grant 131/2004.

And it is the stability issue that is of main concern because  $y$  is often a measured quantity and therefore subject to errors. As in most cases  $y \notin R(K)$  (where by  $R(K)$  we denoted the range of  $K$ ), the equation (2) has no longer solution. Thus, if in addition, we suppose that  $y \in D(K^+)$ , where by  $D(K^+) = R(K) \oplus R(K)^\perp$  we denoted the domain for the Moore-Penrose pseudoinverse of the linear compact operator  $K$  from (2) (see e.g. [1]), we can reformulate (2) as the least-squares problem: find  $\bar{x} \in L^2([0, 1])$  such that

$$\| K\bar{x} - y \|_{L^2([0,1])} = \min!, \quad (3)$$

where  $\| f \|_{L^2([0,1])} = (\int_0^1 (f(t))^2 dt)^{\frac{1}{2}}$ . It is well known that, if (3) holds, then the problem (4) has a minimal norm solution,  $x_{LS}$ , given by

$$x_{LS} = K^+ y. \quad (4)$$

This solution also satisfies (in classical sense) the associated normal equation  $K^* Kx = K^* y$ , where  $K^*$  is the adjoint of  $K$  defined by  $K^* z(\tau) = \int_0^1 k(s, \tau) z(s) ds$ ,  $\tau \in [0, 1]$ . In what follows, we shall shortly present some methods for approximating  $x_{LS}$  given by (5).

## 2. Discretization Techniques

In this section we shall present the following discretization techniques for FKIE: the collocation method, Nystrom's quadrature formula, and Landweber-Friedman's iterations.

### 2.1. Collocation Method

For  $n \geq 2$  arbitrary fixed and  $T_n = \{t_1, \dots, t_n\}$  the set of (collocation) points in  $[0, 1]$  ( $0 \leq t_1 < t_2 < \dots < t_n \leq 1$ ), we consider the collocation discretization of (2): find  $x \in L^2([0, 1])$  such that

$$Kx(t_i) = y(t_i), \quad \forall i = 1, \dots, n. \quad (5)$$

If  $t_i \in T_n$  we define  $k_{t_i} : [0, 1] \longrightarrow \mathbb{R}$  and  $\tilde{y}_i$  by  $k_{t_i}(s) = k(t_i, s)$ ,  $\forall s \in [0, 1]$ ,  $\tilde{y}_i = y(t_i)$ ,  $i = 1, \dots, n$ . Then, the equation (7) can be written as

$$C_n x = \tilde{y}, \quad (6)$$

where  $\tilde{y} \in \mathbb{R}^n$  and  $C_n : L^2 \longrightarrow \mathbb{R}^n$  are defined by  $C_n z = (\langle k_{t_1}, z \rangle, \dots, \langle k_{t_n}, z \rangle)^t$ ,  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)^t$ . If

$$y \in R(K), \quad (7)$$

let  $x^{LS}$  be the minimal norm least-squares solution of (2) and  $x_n^{LS}$  the similar one for (7) (or (9)), given by

$$x^{LS} = K^+ y, \quad x_n^{LS} = C_n^+ \tilde{y}. \quad (8)$$

**Assumption CW.** It exists a sequence of positive integers  $0 < n_1 < n_2 < \dots < n_p < n_{p+1} < \dots$  such that  $\dim(Y_{n_p}) < \dim(Y_{n_{p+1}})$ ,  $\forall p \geq 1$ , with  $Y_n = \text{span}\{k_t, t \in T_n\}$ .

**Remark 1.** The above assumption **CW** tells us that the number of linearly independent functions  $k_t$  in the subspaces  $Y_n$  tends to infinity together with  $n$ , but not all the functions in each  $Y_n$  are linearly independent, as in the original assumption (see [4]).

Let  $\Delta_n$  be defined  $\Delta_n = \sup_{t \in [0,1]} \left( \inf_{t_i \in T_n} |t - t_i| \right)$ . The following result is proved in [5] (Theorem 4).

**Theorem 1.** Under the above assumption **CW**, if (7) holds, and  $\lim_{n \rightarrow \infty} \Delta_n = 0$ , then

$$\lim_{n \rightarrow \infty} \|x_n^{LS} - x^{LS}\| = 0. \quad (9)$$

In [5] it is proven that  $x_n^{LS}$  can be computed as

$$x_n^{LS}(t) = \sum_{j=1}^n \alpha_j k(s_j, t), \quad t \in [0, 1], \quad (10)$$

where  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$  is the minimal norm solution of the system

$$A_n \alpha = b_n \quad (11)$$

and the entries for matrix  $A_n$  and vector  $b_n$  are given by

$$(A_n)_{ij} = \int_0^1 k(s_i, t) k(s_j, t) dt, \quad (b_n)_i = y(s_i), \quad i, j = 1, \dots, n.$$

For the case  $y \in R(K) \oplus R(K)^\perp$ , it is considered instead of (2), its normal equation

$$\tilde{Q}x = w, \quad (12)$$

where  $\tilde{Q} = K^*K$ ,  $w = K^*y$ . Because of the equality (see [1])  $\tilde{Q}^+w = K^+y$ , it results that the equations (2) and (20) have the same minimal norm solution  $x^{LS}$  given by (12). Then, we replace (4) by the problem: find  $x \in L^2([0, 1])$  such that

$$\sum_{i=1}^n \left( \tilde{Q}x(t_i) - w(t_i) \right)^2 = \min! \quad (13)$$

In this case, under a similar assumption as **CW**, **Theorem 1** still holds (see [5]), where  $x_n^{LS}$  is the minimal norm solution for (20).

Also,  $x_n^{LS}$  can be computed as

$$x_n^{LS}(t) = \sum_{j=1}^n \alpha_j \tilde{Q}(s_j, t), \quad t \in [0, 1], \quad (14)$$

where  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$  is the minimal norm solution of the (consistent) system  $Q_n \alpha = \tilde{w}$  and, in this case, the entries for matrix  $Q_n$  and vector  $\tilde{w}$  are given by

$$(Q_n)_{ij} = \int_0^1 \tilde{Q}(s_i, t) \tilde{Q}(s_j, t) dt, \quad \tilde{w} = (w(t_1), \dots, w(t_n)), \quad i, j = 1, \dots, n.$$

## 2.2. Nystrom's Quadrature Formula

The idea of this method is to approximate the integral operator from (1) with the sequence of integral operators as follows

$$(K_n x)(s) := \sum_{k=1}^n \alpha_k^{(n)} k(s, t_k^{(n)}) x(t_k^{(n)}), \quad s \in [0, 1]. \quad (15)$$

The solution for (1) will be approximated by the sequence of solutions for

$$K_n x_n = y. \quad (16)$$

The main results for this method are the following two theorems (for proof, see [4]).

**Theorem 2.** *Let  $x_n$  be the solution for  $\sum_{k=1}^n \alpha_k k(s, t_k) x_n(t_k) = y(s)$ , with  $s \in [0, 1]$ . Then the values  $x_j^{(n)} = x_n(t_j)$ ,  $j = 1, \dots, n$ , verify the linear system*

$$\sum_{k=1}^n \alpha_k k(t_j, t_k) x_k^{(n)} = y(t_j), \quad j = 1, 2, \dots, n. \quad (17)$$

**Theorem 3.** *The sequence  $\{K_n\}_{n \geq 1}$  is pointwise convergent,  $K_n x \rightarrow Kx$ , as  $n \rightarrow \infty$  for any  $x \in C([0, 1])$ , but does not converge in norm.*

**Remark 2.** *In practice, we shall approximate the solution for (2) with the solution for the system*

$$K_n x_n = y_n, \quad (18)$$

where

$$(K_n)_{ij} = \alpha_k k(t_j, t_k), \quad (x_n)_k = x_n(t_k^{(n)}), \quad (y_n)_j = y(t_j^n). \quad (19)$$

*If the problem (2) is not consistent, then instead of (18), we consider a least-squares formulation  $\|K_n x_n - y_n\| = \min!$*

## 2.3. Landweber-Friedman's iterations

The Landweber-Friedman method is an iterative scheme for approximating the solution for a FKIE. It is defined as: for  $x_0 \in L^2([0, 1])$ , and  $k = 0, 1, \dots$

$$x_k = x_{k-1} + \omega K^*(y - Kx_{k-1}), \quad 0 < \omega < 2/\|K^*K\|. \quad (20)$$

In 1951, Landweber proved that  $x_k \rightarrow K^+ y$ ,  $k \rightarrow \infty$ , for  $y \in R(K)$ . But  $\|x_k\| \rightarrow \infty$  for  $y \notin R(K)$ . Thus, the area of application for this method is restricted to consistent equations of the form (2).

In practice, this method can be applied by approximating the integrals from (20) using a quadrature formula (e.g., the Simpson composite rule).

### 3. Kovarik-like Algorithms for Symmetric Matrices

With well-posed problems, better results are obtained as we refine the discretization. However, for a FKIE, refining the discretization causes the discrete problem to more mirror the ill-posed nature of continuous problem. In fact, MATLAB computations of the condition number of the matrices for the linear systems obtained using the types of discretization from the previous section are confirming this. All the above mentioned matrices are rank-deficient, very ill-conditioned, and symmetric. Thus, using a classical direct or iterative method to solve these systems is not a good idea.

A class of iterative solvers for relatively dense symmetric linear systems are the Kovarik-like approximative orthogonalization algorithms (see [3]). We shall briefly describe them in what follows.

**Algorithm KOBS.** Let  $A_0 = A$  a symmetric matrix. For  $k = 0, 1, \dots$  do:  
 $K_k = (I - A_k)(I + A_k)^{-1}$ ,  $A_{k+1} = (I + K_k)A_k$ .

**Theorem 4.** *If none of the eigenvalues of  $A$  is in the set*

$$E = \left\{ -\frac{1}{\alpha_j}, j \in \mathbb{N} \mid \alpha_0 = 1, \alpha_{j+1} = 2\alpha_j + 1 \right\},$$

*then the sequence  $(A_k)_{k \geq 0}$  generated as above is well defined, convergent, and  $\lim_{k \rightarrow \infty} A_k = A^+ A$ .*

**Remark 3.** *Actually, this algorithm acts as a preconditioner for the matrix  $A$ . So, it can be combined with any other solver.*

In order to avoid the computation of the inverse at each step of the previous algorithm, we shall use a modified version of that one. The inverse  $(I + A_k)^{-1}$  will be approximated by ( $q \geq 1$  arbitrary fixed)  $S(A_k; q) = \sum_{i=0}^q a_i (-A_k)^i$ , with  $a_0 = 0, a_{j+1} = \frac{2j+1}{2j+2} \cdot a_j, j > 0$  (see [3]).

**Algorithm MKOBS.** Let  $A_0 = A$  a symmetric matrix with  $\sigma(A) \subset [0, 1]$ . We construct the sequence  $(A_k)_{k \geq 0}$ ,  $(K_k)_{k \geq 0}$  via:

$$K_k = (I - A_k)S(A_k; n_k); \quad A_{k+1} = (I + K_k)A_k. \quad (21)$$

**Remark 4.** *In order to minimize the computational effort per iteration, in applications we choose  $n_k = 1, \forall k \geq 1$ . The algorithm becomes*

$$K_k = (I - A_k) \left( I - \frac{1}{2} A_k \right); \quad A_{k+1} = (I + K_k) A_k. \quad (22)$$

For solving the linear least-squares problem of the form

$$\|Ax - b\| = \min! \quad (23)$$

the following right hand side (rhs, for short) version of algorithm **MKOBS** was proposed in [3].

**Algorithm MKOBS-rhs.** Let  $A_0 = A$ ,  $b^0 = b$ ; for  $k = 0, 1, \dots$  do

$$K_k = (I - A_k)S(A_k; n_k), \quad A_{k+1} = (I + K_k)A_k, \quad b^{k+1} = (I + K_k)b^k. \quad (24)$$

In [3] are proved the following results.

**Theorem 5.** (i) *If the problem (2) is consistent, then the sequence  $(b^k)_{k \geq 0}$  is convergent and*

$$\lim_{k \rightarrow \infty} b^k = A^+ b = x_{LS}. \quad (25)$$

(ii) *If the problem (2) is not consistent, then the sequence  $(A_k b^k)_{k \geq 0}$  is convergent and*

$$\lim_{k \rightarrow \infty} A_k b^k = A^+ b = x_{LS}. \quad (26)$$

*In this case,  $\lim_{k \rightarrow \infty} \|b^k\| = \infty$ .*

**Remark 5.** *The last relationship can generate problems. That's why, in practice, it is used a modified version of MKOBS-rhs algorithm.*

**Algorithm MKOBS-rhs-1.**

$$K_k = (I - K_k)(I - \frac{1}{2}A_k), \quad A_{k+1} = (I + K_k)A_k, \quad \alpha^{(k+1)} = (I + K_k)^2 \alpha^{(k)}. \quad (27)$$

**Remark 6.** *The above algorithm MKOBS-rhs-1 has the same convergence behaviour as described in Theorem 5.*

## 4. Numerical Experiments

The experiments were ran with different settings and convergence results are presented in Table 1. The convergence speed with respect to the number of iterations is very high (around 16 iterations), although it does not improve significantly with the increase in the number of collocation points. The tests were conducted on two directions; first, we tested the algorithms presented earlier on a problem with a known solution, and second, on a problem where we do not know of the existence of a solution, moreover the problem may be inconsistent.

### 4.1. Numerical experiments – The problem

Consider the FKIE (1)–(2) with the kernel

$$k(s, t) = \frac{1}{\sqrt{(1 + (s - t)^2)^3}}, \quad s, t \in [0, 1].$$

The problem is a simplified version of a problem arising from the field of electrical potential generated by a known electric field. It was specifically chosen as a model problem to test the algorithms presented, since it is a symmetric function, thus being appropriate to all three discretization methods described.

For  $y(s) = \sin(\arctan(1-s)) - \sin(\arctan(-s))$ , a solution is  $x(t) = 1$  ( $y \in R(K)$ ) – denote this case as problem (Pa).

For  $y(s) = s$ , there is no known solution for the equation, thus the problem may be inconsistent – denote this case as problem (Pb). The solution for (Pa) with K OBS



Table 1

Pa: KOBS and MKOBS maximum admissible error  $10^{-6}$ 

Collocation points	KOBS iterations	MKOBS iterations
10	18	18
50	16	16
100	16	16
200	15	15

and MKOBS are very similar (see Figure 1) and at the same time very close to the known solution  $x(t) = 1$ . The kernel is symmetric, so Nystrom method can be used for discretization; the solution found in 21 iterations with a maximum allowed error of  $10^{-3}$  is presented in Figure 2. If MKOBS is left to run more than 21 iterations in order to obtain a smaller residual, the effect is quite opposite: the error grows rapidly and the algorithm diverges. The solution found by the Landweber Friedman iterative method yields a good result in maximum 50 iterations, although it's shape doesn't resemble that of the solutions found with (M)KOBS. Also, for a greater number of iterations the method diverges.

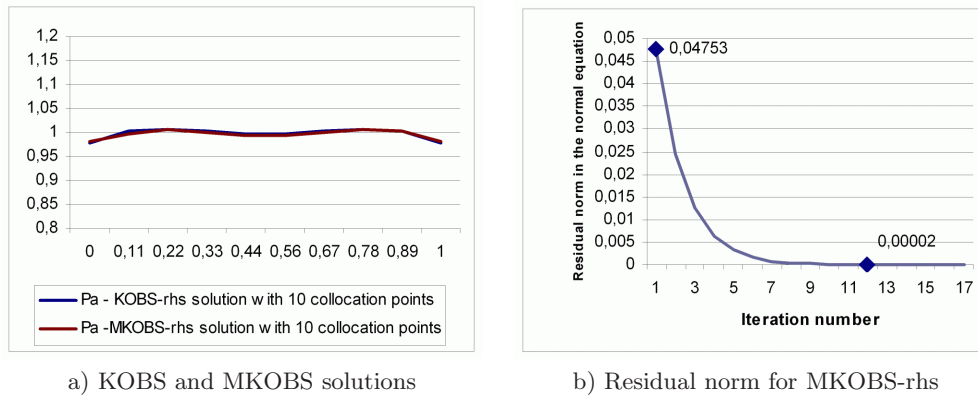


Fig. 1. Pa: KOBS and MKOBS with collocation discretization with 10 points.

With respect to problem (Pb), the same tests were ran. Both collocation method and Nystrom discretization techniques yielded similar results in terms of the shape of the solution. The residual norm of the solution found, when collocation discretization was used, was  $10^3$  times bigger than the residual norm of the solution following Nystrom quadrature method as a discretization technique. Landweber Friedman iterative method does not reach a satisfactory result, thus confirming what has been proven in theory that it is only suited for the consistent case.

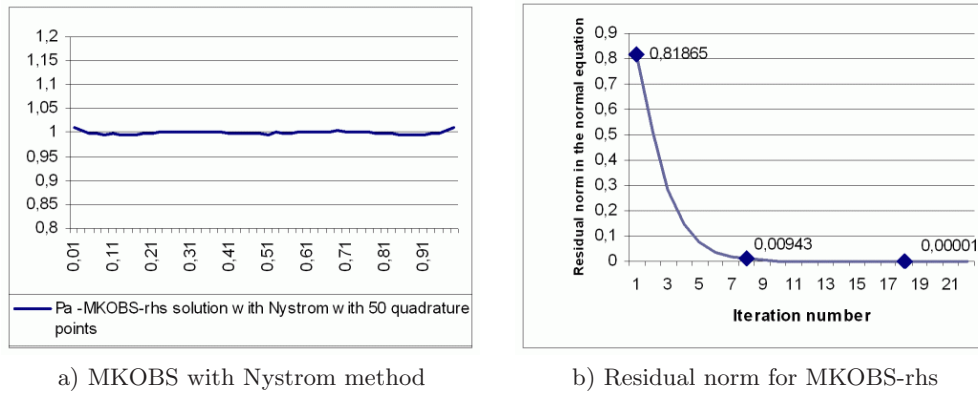


Fig. 2.  $Pa$ : MKOBS with Nystrom method, 21 iterations, residual  $10^{-3}$ .

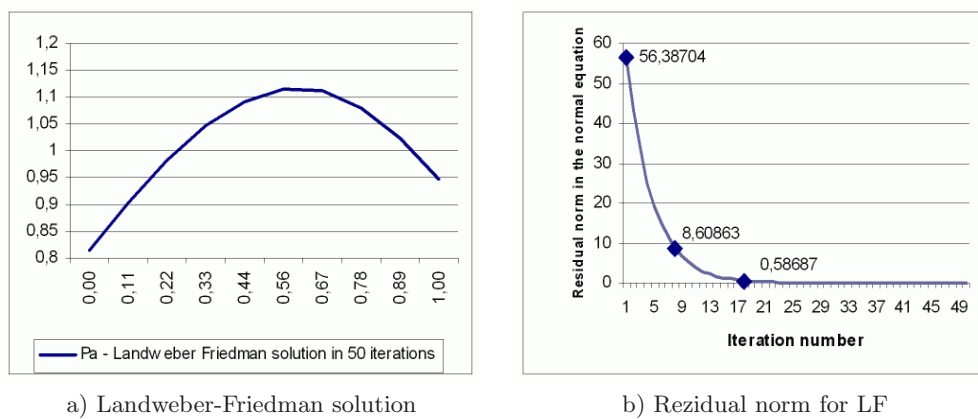
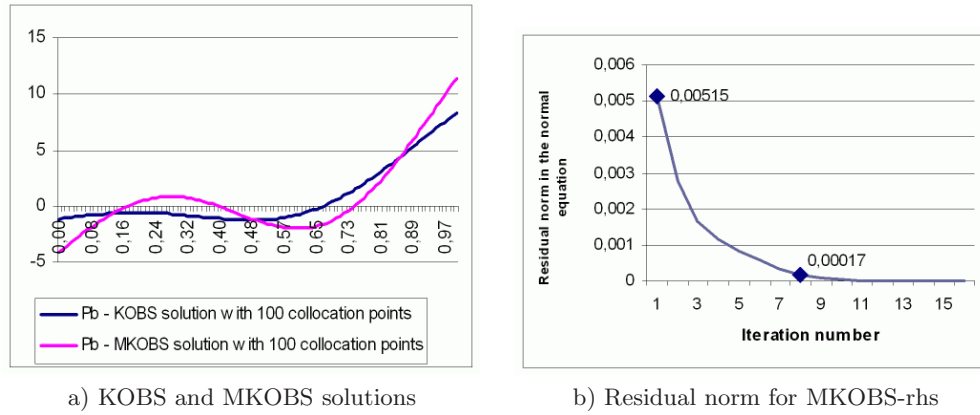


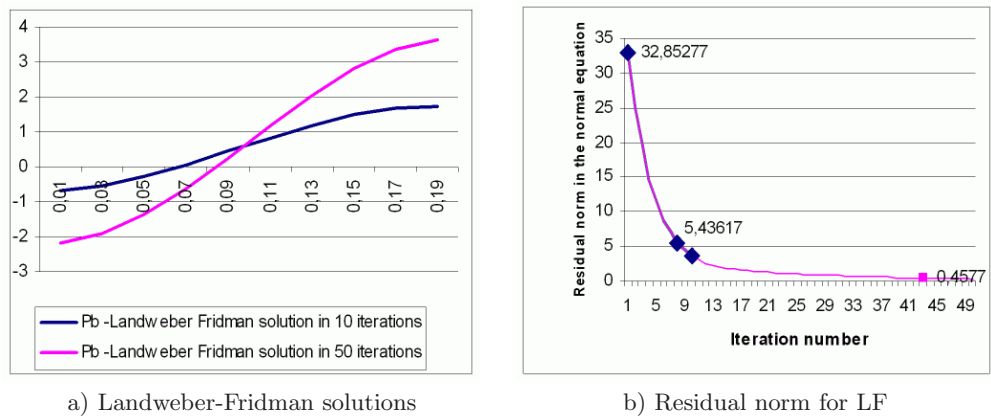
Fig. 3.  $Pa$ : Landweber-Friedman in 50 iterations.



a) KOBS and MKOBS solutions

b) Residual norm for MKOBS-rhs

Fig. 4. Pb: KOBS and MKOBS using collocation discretization (100 points).



a) Landweber-Fridman solutions

b) Residual norm for LF

Fig. 5. Pb: Landweber-Fridman solutions.

## References

- [1] T.L. Boullion and P.L. Odell, *Generalized inverse matrices*. Willey-Interscience,
- [2] R. Kress, *Linear Integral Equations*, Springer-Verlag, 1989.
- [3] M. Mohr, C. Popa, U. Rude, *A Kovarik Type Algorithm without Matrix Inversion for the Numerical Solution of Symmetric Least-Squares Problems*, Lehrstuhlbericht 05-2, 2005
- [4] M.Z. Nashed, G. Wahba, *Convergence rates of approximate least-squares solutions of linear integral and operator equation of the first kind* Math. of Comp., **28(125)** (1974).
- [5] E. Pelican, *Extensions of Projection and Collocation Methods for First Kind Integral Equations*, submitted to Rev. Roumaine Math. Pures Appl.

## A fast approximation for discrete Laplacian

Constantin Popa<sup>\*‡</sup> and Tudor Udrescu<sup>\*‡</sup>

The discrete Laplacian is used as preconditioner in many partial differential equations modelling real world problems. Consequently, a fast approximation for the discrete Laplacian is essential. According to this, in the present paper we describe two multigrid approaches based on the original formulation by Braess [1]. The first one deals with Braess' original method, whereas the second algorithm is an adaptation of the full multigrid method (see e.g. [2]). Numerical experiments presented for 2D Poisson equation, illustrate the robustness and the efficiency of our approaches compared to classical solvers.

### 1. Fast approximation of the discrete Laplacian

Many important classes of partial differential equations (PDEs, for short) give rise, after finite element or finite differences approximations to big, sparse, ill-conditioned (possible non-symmetric) linear systems of equations of the form

$$Ax = b. \quad (1)$$

Among the most important and efficient solvers for (1) are the Conjugate Gradient algorithms (CG, for short, see [5]), which for non-symmetric matrices  $A$  becomes the CGN method, i.e. the CG algorithm applied to the normal equation of (1),  $A^t Ax = A^t b$ . The CGN method has an “error reduction factor” of the form (see [5])

$$\| b - Au^k \| \leq 2 \left( \frac{k_2(A) - 1}{k_2(A) + 1} \right)^k \| b - Au^0 \|, \quad (2)$$

where  $k_2(A)$  is the spectral condition number of  $A$ . The bigger the value of  $k_2(A)$ , closer to 1 will be the error reduction factor in (2) and the convergence of the CGN

---

<sup>\*</sup> “Ovidius” University of Constanța, Faculty of Mathematics and Computer Science, Romania, e-mail: [cpopa@univ-ovidius.ro](mailto:cpopa@univ-ovidius.ro) and [tudrescu@univ-ovidius.ro](mailto:tudrescu@univ-ovidius.ro)

<sup>‡</sup> Supported from PNCDI INFOSOC Grant 131/2004.

method will slow down. In order to eliminate this bad aspect, we can perform a preconditioning to the initial system (1). One possibility in this sense is to use a symmetric and positive definite matrix (SPD, for short)  $P$  with a decomposition of the form (e.g. Cholesky)

$$P = CC^t. \quad (3)$$

The preconditioned version of (1) can be written as (see e.g. [7], [8])

$$(C^{-1}AC^{-t})(C^tx) = C^{-1}b \Leftrightarrow \hat{A}\hat{x} = \hat{b}, \quad (4)$$

with

$$\hat{A} = C^{-1}AC^{-t}, \quad \hat{x} = C^tx, \quad \hat{b} = C^{-1}b. \quad (5)$$

Then, the Preconditioned CGN algorithm (PCGN, for short), i.e. CGN applied to the preconditioned system (4)–(5) (see [5] for details) has a similar error reduction formula with (2) holds, but with  $k_2(\hat{A})$  instead of  $k_2(A)$ . If  $k_2(\hat{A}) \ll k_2(A)$ , (e.g. for “mesh independent” preconditioning) then the convergence of PCGN is much faster than that of CGN, but, it requires in each iteration two inversions of the matrix  $P$ . If this is not done in an efficient way the “much better” behaviour of PCGN remains only at a theoretical level and the method is useless in practice.

**Remark 1.** *Even though a Cholesky-like decomposition of  $P$  as in (3) is available, for big dimensions of (1) the forward and backward substitutions required by  $P^{-1} = C^{-t}C^{-1}$  can be too costly.*

This is the reason for which we must look for much faster algorithms. In this paper we shall present two of them for the case  $P = \Delta_h$ , with  $\Delta_h$  the 5-point stencil discretization of the 2D Laplacian, i.e.

$$\Delta_h = \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_h. \quad (6)$$

There are many important cases in which the theory behind preconditioning (4)–(5) suggests the use of  $\Delta_h$ ; for example, if we consider the 2D convection-diffusion problem

$$\begin{cases} -\Delta u + \alpha \cdot \frac{\partial u}{\partial x} + \beta \cdot \frac{\partial u}{\partial y} = f, & \text{in } \Omega \subset \mathbb{R}^2 \\ u = g, & \text{on } \partial\Omega \end{cases} \quad (7)$$

discretized with centered finite differences, then the symmetric part of the discretization matrix  $A$ ,

$$M = \frac{1}{2}(A + A^t), \quad (8)$$

recommended as a preconditioner in [4], is exactly  $\Delta_h$  from (6). Another example is when we use the variational finite element algorithm (VFEM, for short). If we suppose that  $\Omega$  is a polygonal domain with a triangulation as in Figure 1 then, the classical VFEM with piecewise linear basis functions in  $H_0^1(\Omega)$  gives us that the Gram matrix associated to this basis in the  $H_0^1$ -norm is exactly  $\Delta_h$  from (6). This Gram matrix is recommended as an efficient preconditioner in (4) in the papers [7], [8] (see also the references therein).

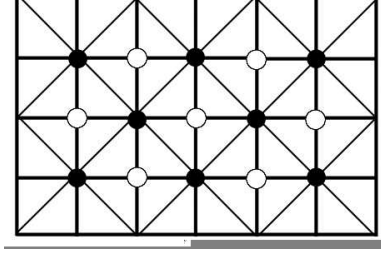


Fig. 1. Triangulation of a polygonal domain.

## 2. Braess multigrid algorithm

D. Braess proposed in [1] an interesting and efficient multigrid algorithm (MG, for short), based on the “red-black” version of the classical Gauss-Seidel iteration (RBGS, for short), which we shall briefly describe in what follows. By again referring to Figure 1, the points marked with  $\bullet$  will be the “black” ones (B), the other “red” (R). With respect to this triangulation for  $\Omega$ , we consider the Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega = (0, 1)^2 \\ u = 0, & \text{on } \partial\Omega \end{cases} \quad (9)$$

With piecewise linear basis functions VFEM we obtain the associated linear system ( $h$  is the mesh size).

$$A^h x^h = b^h, \quad (10)$$

where

$$(A^h)_{ij} = \begin{cases} 1, & \text{if } i = j \\ -\frac{1}{4}, & \text{if } i \neq j \text{ and } p_i, p_j \text{ are adjacent in } \Omega_h, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

i.e. exactly  $\Delta_h$  from (6), up to a multiplicative factor. If  $\Omega_H$  denotes the set of all black points, then the RBGS iteration can be splitted into 2 “disjoint” steps as follows (see [1])

$$(G_h^I x^h)_i = \begin{cases} x_i^h, & p_i \in \Omega_H \\ \frac{1}{4} \sum_j x_j^h + b_i^h, & p_i \notin \Omega_H \end{cases} \quad (12)$$

$$(G_h^{II} x^h)_i = \begin{cases} \frac{1}{4} \sum_j x_j^h + b_i^h, & p_i \in \Omega_H \\ x_i^h, & p_i \notin \Omega_H \end{cases} \quad (13)$$

**Remark 2.** We have to observe that, if  $N$  is the total number of points in  $\Omega_h$  (i.e. the dimension of  $A_h$ ), then each step of the RBGS from (12)–(13) requires approximately  $\frac{5 \cdot N}{2}$  flops (because of the 5-diagonal structure of  $A_h$  in (11)). This computational effort fits well into the general theory about efficient MG algorithm (see [2]).

The 2-grid algorithm proposed by Braess in [1] can be written as follows:

$$\begin{aligned}
&\textbf{Step 1 } x^{h,k,1} = (G_h^{II} \circ G_h^I)(x^{h,k,0}) \\
&\textbf{Step 2 } x^{h,k,2} = G_h^I(x^{h,k,0}) \\
&\textbf{Step 3 } \text{compute the residual } d^h = b^h - A_h \cdot x^{h,k,2} \\
&\quad \text{restrict } d_h \text{ to } \Omega_H \text{ and solve the coarse grid equation} \\
&\quad \begin{cases} y_i = \frac{1}{4} \sum_{j \in H} y_j + d_i, & i \in \Omega_H \\ y_i = 0, & i \in \Omega_h \setminus \Omega_H \end{cases} \quad (14) \\
&\textbf{Step 4 } \text{correction } x^{h,k,3} = x^{h,k,2} + y \\
&\textbf{Step 5 } x^{h,k,4} = G_h^I(x^{h,k,3}) \text{ and set } x^{h,k+1,0} = x^{h,k,4}
\end{aligned}$$

**Remark 3.** We have to observe that, without the coarse grid solution in **Step 3**, the other steps of (14) are just the two parts of *RGBS* (12), (13), which gives a  $\mathcal{O}(N)$  computational effort.

Starting from (14), in the same paper, Braess describes the general multigrid algorithm (more than two levels, denoted by *BMG*). Thus, if  $q = 0, 1, \dots, q_{\max}$  are successive discretization levels as in Figure 1,  $\Omega_q$ ,  $A^q x^q = b^q$ , the corresponding grid and discrete systems, respectively, and

$$h_{q-1} = \sqrt{2}h_q, \quad (15)$$

the associated mesh sizes, the  $(q+1)$ -grid *BMG* algorithm can be written as follows:

$$\begin{aligned}
&\textbf{Step 1 } x^{q,k,1} = (G_q^{II} \circ G_q^I)(x^{q,k,0}) \\
&\quad x^{q,k,2} = G_q^I x^{q,k,1} \quad (16) \\
&\textbf{Step 2 } \text{compute } d^q = b^q - A^q x^{q,k,2}; \\
&\quad \text{let } y^{q-1} \text{ be the solution of} \\
&\quad A^{q-1} y^{q-1} = d^{q-1} \quad (*)
\end{aligned}$$

1. if  $q = 1$  then  $(*)$  is solved exactly;
2. if  $q > 1$  then  $\mu$  iterations ( $\mu = 1, 2, 3$ ) of the  $q-1$  algorithm are performed for  $(*)$  with  $y^{q-1,0,0} = 0$ .

For  $\mu = 1$  in (17) we get the V-cycle *MG* algorithm shown in Figure 2, for  $q_{\max} = 3$ . For  $\mu = 2$  and 3 we get the W-cycle type algorithm described in Figure 3 and Figure 4 (also for  $q_{\max} = 3$ ). The following convergence result is proved in [1].

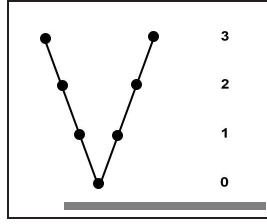
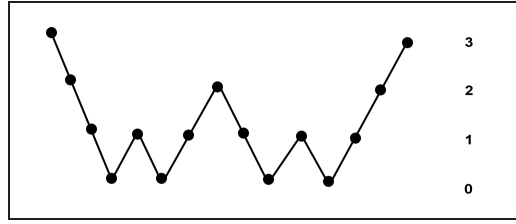
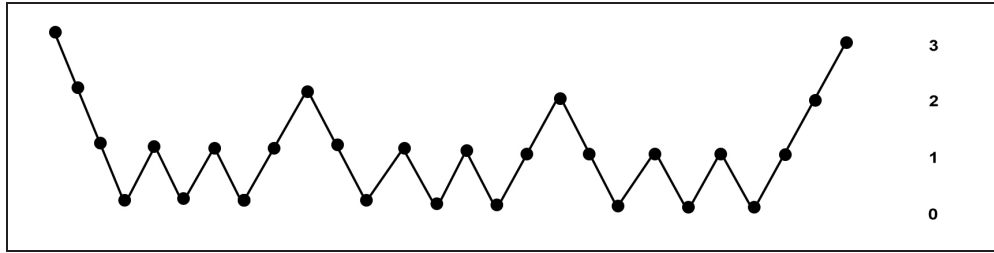
**Theorem 1.** The error reduction factor per iteration in the  $(q+1)$ -grid *BMG* algorithm (17) and with respect to the energy norm, i.e.

$$\|x^{q,k+1,0} - x^q\|_{A^q} \leq \delta_q \|x^{q,k,0} - x^q\|_{A^q}, \quad (17)$$

is given by the recursion

$$\delta_0 = 0, \quad \delta_q = \frac{1}{2}(1 + \delta_{q-1}^\mu), \quad q = 1, \dots, q_{\max}. \quad (18)$$



Fig. 2. V-cycle MG algorithm with  $\mu = 1$ .Fig. 3. W-cycle MG algorithm with  $\mu = 2$ .Fig. 4. W-cycle MG algorithm with  $\mu = 3$ .

**Remark 4.** Some values of  $\delta_{q_{\max}}$  for different combinations  $(\mu, q_{\max})$  are described in Table 1 below.

Table 1  
Error reduction factors  $\delta_{q_{\max}}$

		$q_{\max}$								
		0	1	2	3	4	5	6	7	8
$\mu$	1	0	0.5	0.750	0.875	0.938	0.969	0.985	$\approx 1$	$\approx 1$
	2	0	0.5	0.625	0.696	0.742	0.776	0.801	0.821	0.837
	3	0	0.5	0.563	0.589	0.602	0.610	0.614	0.616	0.617

### 3. The Full Multigrid Algorithm

The Full Multigrid Algorithm (FMG, for short) was first proposed by A. Brandt in [2]. We shall briefly describe its main ideas by using the particular case with 3 grids presented in Figure 5 below. The algorithms in the square brackets are classical

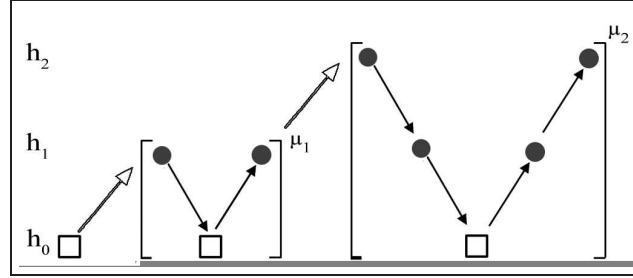


Fig. 5. The Full Multigrid Algorithm.

V-cycle MG methods. We start with an approximation  $x^0$  of the solution on the coarsest grid  $\Omega_{h_0}$ ; we interpolate  $x^0$  to  $\tilde{x}^1$  on  $\Omega_{h_1}$ ; then we perform an  $(h_1, h_0)$  V-cycle  $\mu_1$ -times on  $\tilde{x}^1$  and get  $x^1$ ; we interpolate  $x^1$  to  $\tilde{x}^2$  on  $\Omega_{h_2}$  (the finest level in our case); then we perform an  $(h_2, h_1, h_0)$  V-cycle  $\mu_2$ -times on  $\tilde{x}^2$  and get  $x^2$ . According to Brandt, if the interpolation of the coarse grid solution to finer ones and the numbers of V-cycle sweeps  $\mu_1$  and  $\mu_2$  are chosen in an appropriate way we get for  $x^2$  the same order of approximation as the order of truncation error given by the discretization scheme used, i.e. (in our case see [3])

$$\|x^2 - u^2\| = \mathcal{O}(h_2^2), \quad (19)$$

for an approximate norm  $\|\cdot\|$  (where  $u^2$  is the exact discrete solution on  $\Omega_{h_2}$ ). Starting from the above mentioned basic ideas of the FMG algorithm, we constructed the following one which uses the elements from Section 2. For this, let  $h_0, h_1, \dots, h_q$  ( $h_{q-1} = \sqrt{2}h_q$ ) be  $(q+1)$ -grids,  $\Omega_q$  as in Section 2;  $A^k$  and  $b^k$ ,  $k = 0, \dots, q$  will be the matrices and right hand side of the associated linear systems on  $\Omega_{h_k}$ .

**Note 1.** The matrices  $A^k$  have the same elements as in (11) (thus we don't need to construct them), only their structure is determined by the connections between the  $\Omega_{h_k}$  points.

**Note 2.** The right hand sides  $b^k$  can be easily obtained from  $b^{q_{\max}}$  by using the connections of the finite element basis functions between two successive levels.

**Note 3.** For interpolating a solution  $x^{q-1} = (x_i^{q-1})_i$  from  $\Omega_{h_{q-1}}$  to  $x^q$  from  $\Omega_{h_q}$  we do the following two steps.

**IS1.** Compute  $(\tilde{x}_i^q)_i$  on  $\Omega_{h_q}$  by

$$\tilde{x}_i^q = \begin{cases} x_i^{q-1}, & i \in \Omega_{h_{q-1}} \\ 0, & i \in \Omega_{h_q} \setminus \Omega_{h_{q-1}} \end{cases} \quad (20)$$

**IS2.** Compute  $x^q$  on  $\Omega_{h_q}$  by

$$x^q = G_q^I(\tilde{x}^q), \quad (21)$$

with  $G^I$  as in (12).  $\Omega_h = \Omega_q$  and  $\Omega_H = \Omega_{q-1}$ .

Now we can describe the FMG version of Braess BMG algorithm (BFMG, for short).

**Step 1** Solve (exactly)  $A^0 x^0 = b^0$  for  $x^0$  on  $\Omega_{h_0}$ ; Set  $k = 0$ ;

**Step 2 (i)** if  $k = q$  then GO TO Step 3

**(ii)** if  $k < q_{\max}$  then

1. interpolate  $x^k$  to  $x^{k+1}$  as in (20)-(21)
2. set  $\tilde{x}^{k+1} = x^{k+1}$
3. perform  $\mu_{k+1}$  sweeps of a V-cycle on  $\tilde{x}^{k+1}$  according to the grids  $h_0, \dots, h_{k+1}$  and get  $x^{k+1}$
4. set  $k = k + 1$  and GO TO (i)

**Step 3** The approximate solution is  $x^q$ .

**Note.** In each of the above V-cycles the system on the coarsest grid  $\Omega_{h_0}$  is solved exactly.

**Proposition 1.** Let  $k \in \{1, \dots, q\}$  arbitrary fixed,  $h_{k-1}, h_k$  two consecutive levels,  $x^{k-1}$  the exact solution of  $A^{k-1} x^{k-1} = b^{k-1}$  and  $x^k$  on level  $h_k$  obtained as in (20)-(21). Then,

$$\max_{p_i \in \Omega_k} |x_i^k - u_i^k| = \mathcal{O}(h_{k-1}^2). \quad (22)$$

*Proof.* Because  $x^{k-1}$  is the exact solution on  $\Omega_{k-1}$  we have

$$x_i^{k-1} = u_i^{k-1}, \quad \forall p_i \in \Omega_{k-1}. \quad (23)$$

Let now  $u^{ex} : \Omega \rightarrow \mathbb{R}$  be the exact (unique) solution of our continuous problem (9). From the theory of the approximation error in FEM method (see e.g. [3]) we get

$$u_i^{ex} = u_i^{k-1} + c_i^{k-1} \cdot h_{k-1}^2 = u_i^k + c_i^k \cdot h_k^2, \quad (24)$$

where  $c_i^{k-1}$  and  $c_i^k$  are uniformly bounded independently on  $h$  and  $i$ , i.e

$$m \leq c_i^{k-1} \leq M; \quad m \leq c_i^k \leq M, \quad \forall i, k. \quad (25)$$

Moreover, from the definition of the exact solution  $u_k$  on  $\Omega_k$ , and the construction of  $x^k$  in (20)-(21) we have that (by also using (23))

$$\begin{aligned} x_i^k &= b_i^k + \frac{1}{4} \sum_{j \in N(i)} x_j^k = b_i^k + \frac{1}{4} \sum_{j \in N(i)} x_j^{k-1} = \\ &= b_i^k + \frac{1}{4} \sum_{j \in N(i)} u_j^{k-1}, \quad p_i \in \Omega_k \setminus \Omega_{k-1}, \end{aligned} \quad (26)$$

$$u_i^k = b_i^k + \frac{1}{4} \sum_{j \in N(i)} u_j^k, \quad p_i \in \Omega_{k-1}. \quad (27)$$

Then, from (24)–(27) we get, for any  $p_i \in \Omega_k \setminus \Omega_{k-1}$ ,

$$u_i^k - x_i^k = \frac{1}{4} \sum_{j \in N(i)} (u_j^k - u_j^{k-1}) = \mathcal{O}(h_{k-1}^2), \quad (28)$$

which gives us (22) and completes the proof.  $\square$

We are now able to prove the approximation property of the above described BFMG algorithm. For this, let's suppose we have a sequence of consecutive grids  $q = 0, 1, \dots, q_{\max}$  as in Figure 5 satisfying (15). We consider a BFMG algorithm as described before, in which we solve exactly the systems on the coarser grid  $q = 0$  and perform  $\mu_q \geq 1$  iterations of the corresponding V-cycles algorithms  $q = 1, \dots, q_{\max}$  (i.e.  $\mu_1 \geq 1$  iterations for the V-cycle with the grids  $(h_0, h_1)$ ;  $\mu_2 \geq 1$  iterations for the V-cycle with the grids  $(h_0, h_1, h_2)$ ,  $\dots$ ,  $\mu_{q_{\max}} \geq 1$  iterations with the final complete V-cycle, i.e. with all the grids  $(h_0, h_1, \dots, h_{q_{\max}})$ ).

**Theorem 2.** *Let  $x^{q_{\max}}$  be the final vector on  $\Omega_{q_{\max}}$  generated with the above described BFMG algorithm. Then there exists integers*

$$\mu_1 \geq 1, \mu_2 \geq 1, \dots, \mu_{q_{\max}} \geq 1 \quad (29)$$

such that

$$\max_{p_i \in \Omega_{q_{\max}}} |x_i^{q_{\max}} - u_i^{q_{\max}}| = \mathcal{O}(h_{q_{\max}}^2), \quad (30)$$

where  $u^{q_{\max}}$  is the exact discrete solution on  $\Omega_{q_{\max}}$ .

*Proof.* The proof is almost obvious according to Proposition 1. Indeed, for two consecutive grids  $(h_{q-1}, h_q)$ ,  $q = 1, \dots, q_{\max}$  if the solution on grid  $\Omega_{h_{q-1}}$ ,  $x^{q-1}$  satisfy

$$\max_{p_i \in \Omega_{q-1}} |x_i^{q-1} - u_i^{q-1}| = \mathcal{O}(h_{q-1}^2) \quad (31)$$

it results from Proposition 1 that for the approximation  $x^q$  on  $\Omega_{h_q}$  we have

$$\max_{p_i \in \Omega_q} |x_i^q - u_i^q| = \mathcal{O}(h_{q-1}^2) = 2\mathcal{O}(h_q^2) \quad (32)$$

Thus, what we need is to eliminate the factor 2 from (32) with  $\mu_q$  V-cycles on the grids  $(h_0, \dots, h_q)$ . Theoretically this number depends on the number of levels (see Table 1), but it exists and it can be theoretically determined according to (18) and Table 1. This completes the proof.  $\square$

**Remark 5.** *Although in theory the above values  $\mu_q$  can be big (for a big number of levels), in practical applications only some of them must be taken equal with 2, the other ones being 1 (see Section 4 of the paper). This confirms the considerations made by Brandt in [2].*

#### 4. Numerical experiments

We considered in our experiments the 2D Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega = (0, 1)^2 \\ u = 0, & \text{on } \partial\Omega \end{cases} \quad (33)$$

with the exact solution  $u_h^{ex} = e^{xy} \sin(\pi x) \sin(\pi y)$ . We applied the BMG algorithm described in Section 2 on a workstation equipped with an Intel Pentium 4 Processor with a clock speed of 3 Ghz and 1 GB RAM of DDR memory. The application was coded in the Java language, and we used the Java Runtime Environment Version 5.0 to run all tests. The results obtained for different levels of grid coarseness and  $\mu$  iterations of the BMG algorithm are shown in Table 2. In comparison, performing the same test using only Gauss-Seidel iterations, takes 1840 seconds on the same platform! Additionally, we present in Table 3 the results obtained with the BFMG

*Table 2*  
Computational times for BMG  
( $n = 128$ )

Algorithm type	Time [s]		
	$\mu = 1$	$\mu = 2$	$\mu = 3$
5-Grid	13.88	11.49	17.78
6-Grid	52.51	10.09	18.66
7-Grid	50.84	10.50	23.29
8-Grid	$\infty$	10.69	30.47
9-Grid	$\infty$	11.52	40.07

*Table 3*  
Computational times for BFMG

Algorithm type	Time [s]		
	$n = 32$	$n = 64$	$n = 128$
2-Grid	1.36	-	-
3-Grid	0.42	10.245	-
4-Grid	0.29	2.293	-
5-Grid	0.23	1.202	13.81
6-Grid	0.22	0.971	5.61
7-Grid	0.21	0.951	4.48
8-Grid	0.21	0.912	4.32
9-Grid	0.20	0.911	4.43

algorithm on grids with different number of levels and using three mesh sizes (32, 64 and 128). If the number of levels is relatively small, the computational effort is large, due to the cost of solving directly the system on the coarsest grid. In our experiments we have used for  $\mu_i, i = 0, \dots, q_{\max}$  the value 1, with one exception in each case –  $\mu_{q_{\max}-1} = 2$ . For assesment of the performance of the BMG algorithm, we also present the 3D plot of the relative errors between the solution computed at each iteration of the algorithm and the solution computed directly by solving the system  $A_h u_{direct} = b^h$ . Figures 6, 7, 8, 9, 10 and 11 illustrate the the relative errors for the case of the 5-grid BMG algorithm with a mesh size of 32 and  $\mu = 2$ . The numerical experiments presented clearly illustrate the robustness and efficiency of our approaches compared to classical solvers.

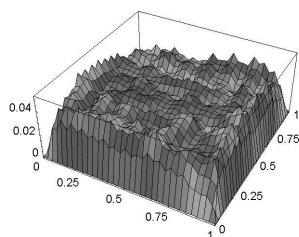


Fig. 6. First iteration.

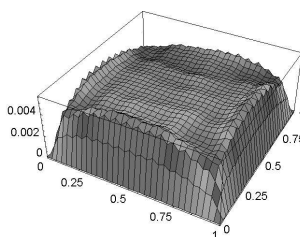


Fig. 7. Second iteration.

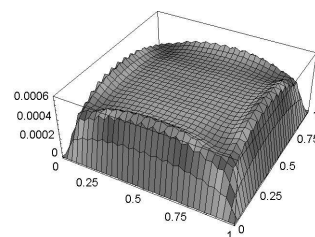


Fig. 8. Third iteration.

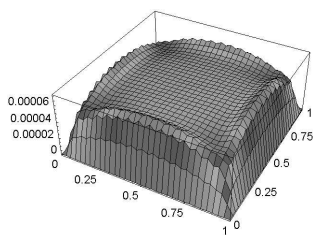


Fig. 9. Fourth iteration.

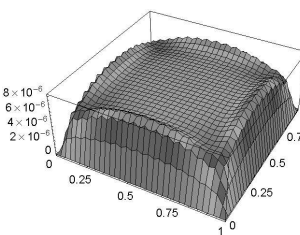


Fig. 10. Fifth iteration.

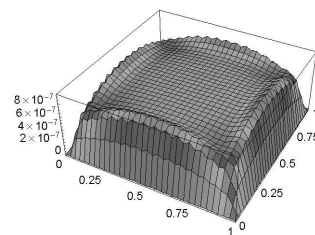


Fig. 11. Sixth iteration.

## References

- [1] Braess, D. *The contraction number of a multigrid method for solving the Poisson equation*, Numerische Mathematik, **37** (1981), 387–404.
- [2] Brandt, A. *Multi-level adaptive solutions to boundary-value problems*, Math. of Comput., **31** (1977), 138, 333–390.
- [3] Ciarlet, Ph. G. *The finite element method for elliptic problems*, North-Holland, New York, 1979.
- [4] Elman, H. C. and Schultz, M. H. *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., **23**(1) (1986), 44–57.
- [5] Golub, G.H. and Van Loan, C.F. *Matrix computations*, The Johns Hopkins University Press, Baltimore, 1996.
- [6] Hackbusch, W. *Elliptic differential equations. Theory and numerical treatment* Springer-Verlag, Berlin, 1987.
- [7] Popa, C. *Mesh independence of the condition number of discrete Galerkin systems by preconditioning*, Intern. J. Comp. Math., **51** (1994), 127–132.
- [8] Popa, C. *Preconditioning conjugate gradient method for non-symmetric systems*, Intern. J. Comp. Math., **58** (1995), 117–133.

## Gibbs regularized tomographic image reconstruction with DW algorithm based on generalized oblique projections

Constantin Popa\* and Rafal Zdunek\*\*

In our previous paper [4] we considered the Diagonal Weighting algorithm (DW) for consistent linear least-squares formulations coming from the field of image reconstruction and processing. There, we proposed a new construction for the Sparsity Pattern Oriented (SPO) family of matrices in the case of Component Averaging (CAV) version of the DW method. This new choice was compared with the classical one proposed in the paper by Censor and Byrne for some image reconstruction model problems from borehole tomography. But, unfortunately in practical applications the corresponding mathematical least-squares formulation is inconsistent and ill-conditioned. In order to cover also this general case, in the present paper we propose a Tikhonov-like regularization for this inconsistent case together with a regularized version of the CAV method. The regularization method is based on the Gibbs prior from statistical image reconstruction with Green and Gaussian potential functions. Numerical experiments are also presented for geophysical imaging in borehole tomography.

### 1. The classical CAV algorithm

In this section we shall very briefly replay the constructions and considerations from [4]. Let  $A$  be an  $m \times n$  (sparse) real matrix,  $b \in \mathbb{R}^m$ ,  $a_i = (a_{i1}, \dots, a_{in})^t \in \mathbb{R}^n$

---

\* “Ovidius” University, Faculty of Mathematics and Computer Science, Constanța, Romania, e-mail:cpopa@univ-ovidius.ro; for this author the paper was supported by the PNCDI INFOSOC Grant 131/2004.

\*\* Institute of Telecommunications, Teleinformatics and Acoustics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, e-mail: Rafal.Zdunek@pwr.wroc.pl

the  $i$ -th row of  $A$  and  $b_i \in R$  the  $i$ -th component of  $b$ . We shall denote by  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$  the Euclidean scalar product and the associated norm, respectively. With these notations we shall define the hyperplane  $H_i = \{x \in \mathbb{R}^n, \langle x, a_i \rangle = b_i\}$  and subspace  $S_i = \{x \in \mathbb{R}^n, \langle x, a_i \rangle = 0\}$  associated to the  $i$ -th equation of the linear system

$$Ax = b. \quad (1)$$

If  $G$  is a diagonal positive semi-definite  $n \times n$  matrix  $G = \text{diag}(g_1, g_2, \dots, g_n)$ ,  $g_j \geq 0$ ,  $\forall j = 1, \dots, n$  we shall denote by  $G^{-1}$  its Moore-Penrose pseudoinverse

$$(G^{-1})_{ij} = \begin{cases} \frac{1}{g_j}, & \text{if } i = j, g_j \neq 0 \\ 0, & \text{else} \end{cases} \quad (2)$$

and by  $\langle \cdot, \cdot \rangle_G$ ,  $\| \cdot \|_G$  the scalar semi-product and the associated semi-norm defined by

$$\langle x, y \rangle_G = \sum_{j=1}^n g_j x_j y_j, \quad \|x\|_G^2 = \langle x, x \rangle_G, \quad \forall x, y \in \mathbb{R}^n. \quad (3)$$

With these notations we can define the **generalized oblique projection** of a point  $x \in \mathbb{R}^n$  onto  $H_i$  or  $S_i$  with respect to  $G$  by

$$P_{H_i}^G(x) = x + \frac{b_i - \langle x, a_i \rangle}{\langle a_i, a_i \rangle_{G^{-1}}} G^{-1} a_i, \quad P_{S_i}^G(x) = x - \frac{\langle x, a_i \rangle}{\langle a_i, a_i \rangle_{G^{-1}}} G^{-1} a_i. \quad (4)$$

A family  $\{G_i\}_{i=1, \dots, m}$  of real diagonal  $n \times n$  matrices such that

$$G_i = \text{diag}(g_{i1}, g_{i2}, \dots, g_{in}), \quad g_{ij} \geq 0, \quad \sum_{i=1}^m G_i = I \quad (5)$$

(with  $I$  the unit matrix) will be called **sparsity pattern oriented (SPO)**, for short) with respect to the matrix  $A$  if, for every  $i = 1, \dots, m, j = 1, \dots, n$ , we have  $g_{ij} = 0$  if and only if  $a_{ij} = 0$ . Then, the **Diagonal Weighting** algorithm (**DW**) for the system (1) can be written as: let  $x^0 \in \mathbb{R}^n$  be the initial approximation and for  $k = 0, 1, \dots$  do

$$x^{k+1} = x^k + \lambda_k \sum_{i=1}^m G_i (P_{H_i}^{G_i}(x^k) - x^k), \quad (6)$$

where  $\lambda_k \in (0, 2)$  are relaxation parameters. If for  $j = 1, \dots, n$  the numbers  $s_j \in \{0, 1, \dots, m\}$  are defined by

$$s_j = \text{card}(\{i \in \{1, \dots, m\}, a_{ij} \neq 0\}) \quad (7)$$

and the elements of  $G_i$  by

$$g_{ij} = \begin{cases} \frac{1}{s_j}, & \text{if } a_{ij} \neq 0 \\ 0, & \text{if } a_{ij} = 0 \end{cases} \quad (8)$$



then the method (7) will be called the **Component Averaging** algorithm (**CAV**, for short; see [2]). In [4] we proposed the following different construction for the above SPO family:

$$g_{ij} = \frac{|a_{ij}|}{\sum_{k=1}^n |a_{kj}|}. \quad (9)$$

**Remark 1.** In [2] it is proved that, for consistent systems as (1) and  $\lambda_k = 1$ ,  $\forall k \geq 0$ , then for  $x^0 = 0$  the sequence  $(x^k)_{k \geq 0}$  generated with the DW algorithm (6) converges to the minimal norm solution  $x_{LS}$  of (1).

## 2. The regularized formulation

As we have already mentioned before, in practical applications the problem (1) becomes inconsistent and has to be reformulated in the least-squares sense

$$\|Ax^* - b\| = \min! \quad (10)$$

Keeping the notation  $x_{LS}$  for its minimal norm solution, because of the ill-conditioning of  $A$  this can be much different than the exact image  $x_{EX}$ . In order to eliminate this difficulties we consider the regularized weighted least-squares version of the problem (10): find  $x^* \in \mathbb{R}^n$  such that

$$\min \Psi(x^*) = \min_{x \in \mathbb{R}^n} \Psi(x), \quad \Psi(x) = \|Ax - b\|_{\Sigma^{-1}}^2 + \beta R(x), \quad (11)$$

where  $\Sigma$  is a symmetric and positive definite  $m \times m$  matrix which attributes weights to data,  $\beta$  is a regularization parameter, and  $R(x)$  is functional that measures the roughness in the image. For constructing  $R$  we start from observations (according to Gaussian distribution)

$$p(b|x) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2} \|Ax - b\|_{\Sigma^{-1}}^2\right) \quad (12)$$

and combine them with the image (w.r.t Gibbs prior from statistical image reconstruction)

$$p(x) = \frac{1}{Z} \exp\left(\beta \sum_{j=1}^n \sum_{k \in N_j} w_{jk} V(x_j - x_k, \delta)\right), \quad (13)$$

where  $Z$  – the partition function – is supposed to be constant. Then, following the Bayes theorem (see [1]) we get

$$\max_x \left( p(b|x) = \frac{p(b|x)p(x)}{p(b)} \right) = \min_x (-2 \log p(x|b)). \quad (14)$$

In this way the above regularized formulation (11) becomes

$$\min \Psi(x^*) = \min_{x \in \mathbb{R}^n} \{\Psi(x) = -2 \log p(x|b)\}, \quad (15)$$

with

$$\Psi(x) = \|Ax - b\|_{\Sigma^{-1}}^2 + \beta \sum_{j=1}^n \sum_{k \in N_j} w_{jk} V(x_j - x_k, \delta). \quad (16)$$

As potential function  $V$  in (16) we consider the following two variants (see [5] for more variants):

**Green**

$$V^{(\text{Green})}(x_j - x_k, \delta) = \delta \log \left[ \cosh \left( \frac{x_j - x_k}{\delta} \right) \right] \quad (17)$$

**Gaussian**

$$V^{(\text{Gaussian})}(x_j - x_k, \delta) = \left( \frac{x_j - x_k}{\delta} \right)^2. \quad (18)$$

If we define by  $U(x)$  the function

$$U(x) = \sum_{j=1}^n \sum_{k \in N_j} w_{jk} V(x_j - x_k, \delta), \quad (19)$$

then the regularized version of CAV that we proposed, RCAF for short (also according to the arguments in [3]), is the following

$$x^{k+1} = x^k + \lambda_k \sum_{i=1}^m G_i \left( P_{H_i}^{G_i}(x^k) - x^k \right) - 2\beta \nabla U(x^k). \quad (20)$$

**Remark 2.** In the Gaussian approach (17), the expression of  $U(x)$  becomes

$$U(x) = \frac{1}{\delta^2} x^T \left( I - \frac{W}{4} \right) x, \quad (21)$$

where  $I \in \mathbb{R}^{n \times n}$  is a unit matrix, and

$$W = (w_{jk}), \quad w_{jk} = \begin{cases} 1, & \text{for } k \in \{N, E, W, S\}_j \\ 0, & \text{else} \end{cases} \quad (22)$$

**Remark 3.** The Hessian of the minimization functional  $\Psi(x)$  from (16), for the Gaussian potential function (18) is given by  $H = \nabla^2 \Psi(x) = A^T \Sigma^{-1} A + \frac{2\beta}{\delta^2} \left( I - \frac{W}{4} \right)$ . By construction the matrix  $I - \frac{W}{4}$  is symmetric and irreducible diagonally dominant, thus invertible. From its symmetry and Gershgorin's theorem it then result that it is also positive definite. This fact, together with the positive definiteness of the matrix  $\Sigma$  tells us that the Hessian  $H$  is symmetric and positive definite, thus the functional  $\Psi$  is strictly convex. This means that the regularized problem (15)–(16) has a unique solution which satisfies the “normal equation”  $\nabla \Psi = 0$ . This is an argument for considering the regularized CAV step as in (20) (see some details in [3]).

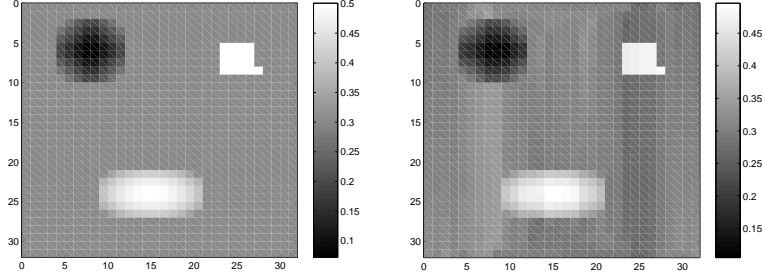


Fig. 1. Original image (left); minimal-norm least-squares solution (right).

### 3. Numerical Results

The proposed method is tested on the data from borehole tomography. In this application, the system of linear equations (1) is inconsistent and rank-deficient. Thus, the minimal-norm least-squares solution is in a certain distance to the true solution. This is shown in Fig. 1, where the left image presents the true solution ( $x_{EX}$ ), and the right one  $x_{LS}$ . In the tests, we use the noise-free and noisy data generated from the true image. The noisy data were perturbed with a zero-mean Gaussian noise with  $SNR = 30$  dB. The images reconstructed with the **CAV**, Gaussian regularized **CAV**, and Gibbs regularized **CAV** are illustrated in Figs. 2–6. The distance and relative errors between the current reconstruction  $x^{(k)}$  and the true image  $x_{EX}$  are presented in Figs. 7–10.

### 4. Conclusions

Fig. 2 shows that the reconstruction with **CAV** for noise-free data is convergent to  $x_{LS}$ , whereas for noisy data the best results are obtained for 50 iterations, and more iterations gradually degrades the image. This is also well visible in Figs. 7–10. The proposed regularization robustly stabilizes the reconstruction (see Figs. 3–6). Carefully adjusted regularization parameter assures the monotonic convergence, which is shown in Figs. 7–10 (parameter  $\beta$  or  $\gamma$ ). The use of the Green function in (16) gives better results than the Gaussian function (see Figs. 3–10) but this is achieved with higher complexity. For the Gaussian function the total energy function can be expressed in a very simplified form (see (21)).

Summing up, the Gibbs regularized **CAV** is a very robust algorithm especially in application to noisy data in borehole tomography. The use of the Green function to penalize the reconstruction gives slightly better results than the typical Gaussian function, but this entails an increase in an overall computational cost.

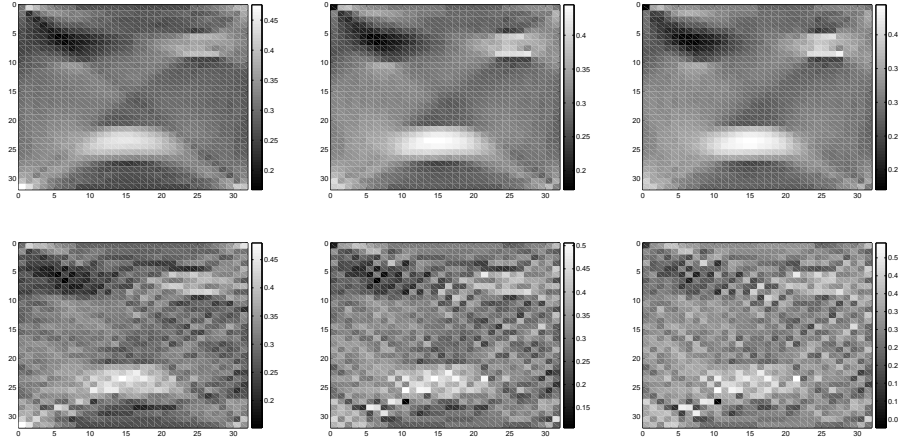


Fig. 2. Images reconstructed from noise-free (upper) and noisy (bottom) data with **CAV** at 50, 150 and 250 iterations, respectively.

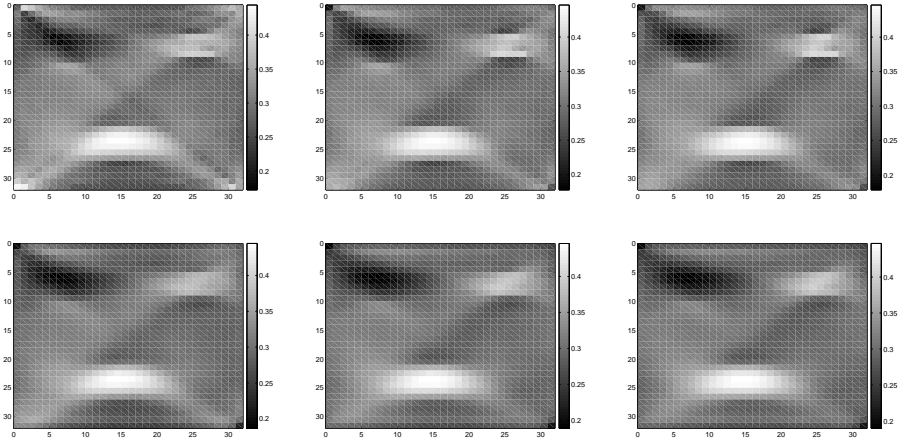


Fig. 3. Images reconstructed from noise-free data with Gaussian regularized **CAV** for  $\gamma = 0.01$  (upper),  $\gamma = 0.1$  (bottom), at 50, 150 and 250 iterations, respectively.

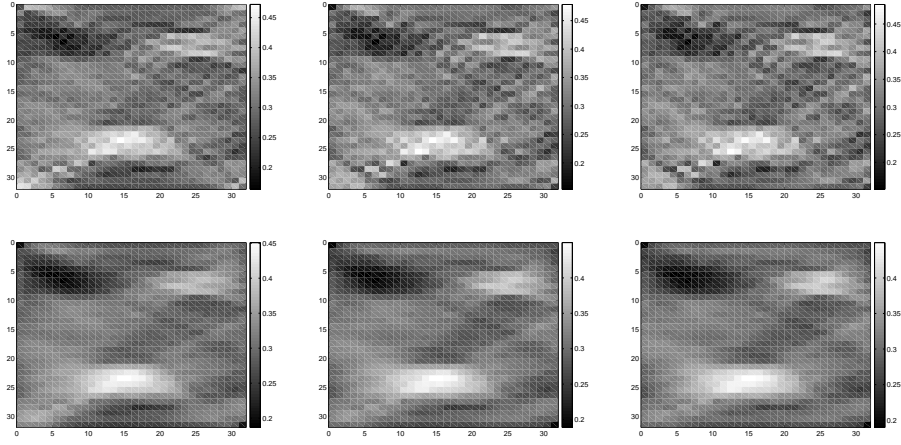


Fig. 4. Images reconstructed from noisy data with Gaussian regularized **CAV** for  $\gamma = 0.01$  (upper),  $\gamma = 0.1$  (bottom), at 50, 150 and 250 iterations, respectively.

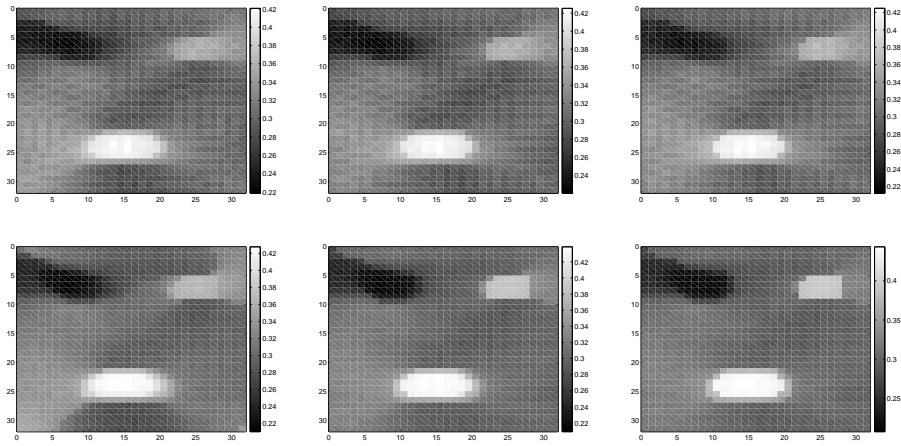


Fig. 5. Images reconstructed from noise-free data with Green regularized **CAV** for  $\beta = 10^{-3}$  (upper),  $\beta = 5 \times 10^{-4}$  (bottom), at 50, 150 and 250 iterations, respectively.

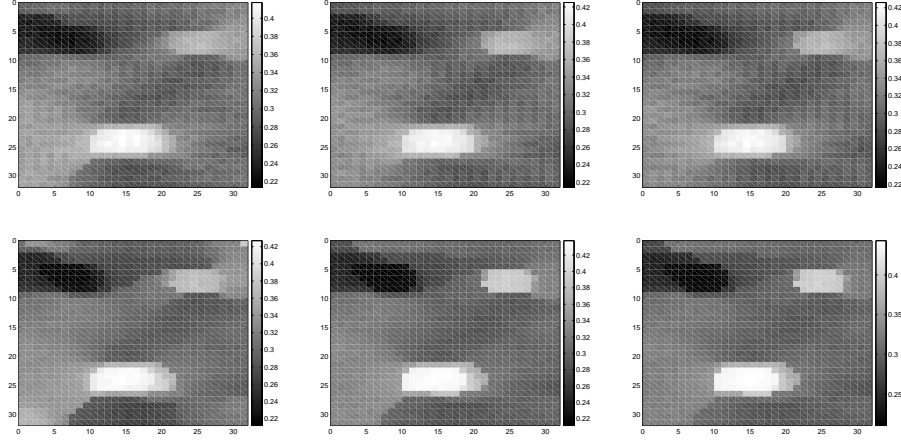


Fig. 6. Images reconstructed from noisy data with Green regularized **CAV** for  $\beta = 10^{-3}$  (upper),  $\beta = 5 \times 10^{-4}$  (bottom), at 50, 150 and 250 iterations, respectively.

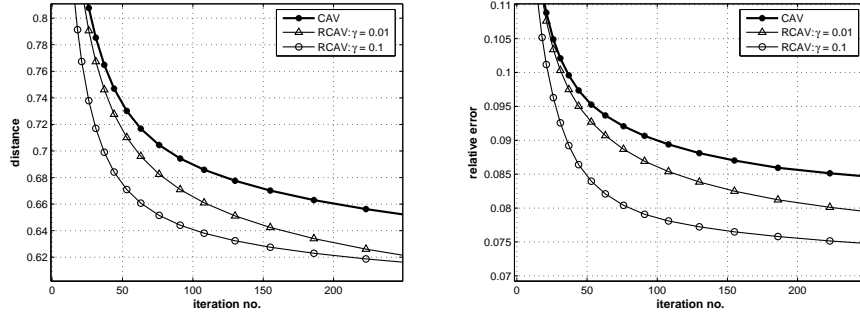


Fig. 7. Distance (left) and relative errors (right) between the true image and  $x^{(k)}$  obtained with Gaussian regularized **CAV** from noise-free data.

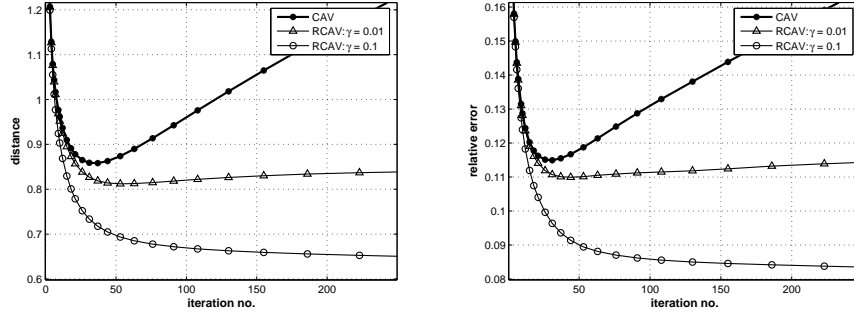


Fig. 8. Distance (left) and relative errors (right) between the true image and  $x^{(k)}$  obtained with Gaussian regularized **CAV** from noisy data.

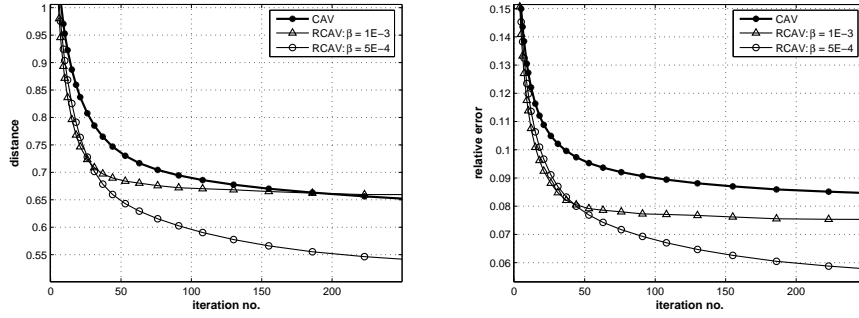


Fig. 9. Distance (left) and relative errors (right) between the true image and  $x^{(k)}$  obtained with Green regularized **CAV** from noise-free data.

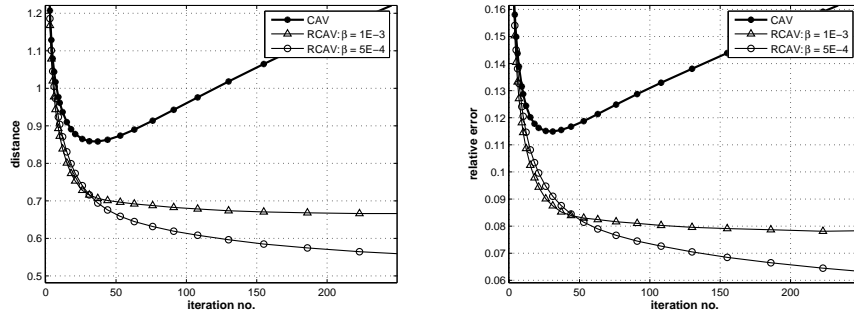


Fig. 10. Distance (left) and relative errors (right) between the true image and  $x^{(k)}$  obtained with Green regularized **CAV** from noisy data.

## References

- [1] H. Akaike, *Likelihood and Bayes Procedure*, In Bayesian Statistics, J. M. Bernardo, et al. (Eds.), University Press, Valencia, 143/166 (1980).
- [2] Y. Censor, D. Gordon, R. Gordon, *Component averaging: an efficient iterative parallel algorithm for large and sparse unstructured problems*, Parallel Computing, **27** (2001), pp. 777–808.
- [3] M. Jiang, G. Wang, *Convergence studies on iterative algorithms for image reconstruction*, IEEE Transactions on Medical Imaging, **22(5)** (2003), 569–579.
- [4] C. Popa, R. Zdunek, *New generalized oblique projections in DW algorithm with application to borehole tomography*, in: *Proceedings of The Third Workshop on Mathematical Modelling of Environmental and Life Sciences problems*, May 27–30, 2005 Constanța, Romania, pp. 231–241, Editura Academiei Române, București, 2004.
- [5] C. Popa, R. Zdunek, *Penalized Least-Squares Image Reconstruction for Borehole Tomography*. Proceedings of ALGORITMY 2005 Conference, Vysoké Tatry–Podbanske, Slovakia, March 13–18, 2005, pp. 260–269.



## **Post–Synaptic Nicotinic Currents Triggered by the Acetylcholine Distribution within the Synaptic Cleft**

**Anca Popescu\* and Alexandru Morega\*\***

A mathematical model that describes the postsynaptic nicotinic currents out of the acetylcholine distributions within the synaptic cleft is proposed. The model describes the most important steps of synaptic transmission: neurotransmitter release from presynaptic vesicles, its diffusion in the synaptic cleft, receptor-neurotransmitter coupling, and the induced post-synaptic current inflow. Relying on previous results on acetylcholine distribution in the synaptic space, the current work is concerned with the postsynaptic currents that convect through open nicotinic acetylcholine receptors, which act as ionic channels. The nicotinic currents are the outcome of a deterministic model that is based on the intra-synaptic cleft acetylcholine space-time distribution, and on the nicotinic receptors opening dynamics. The mathematical model is solved numerically, using the FEMLAB 3.1 software package by a Galerkin finite element method.

### **1. Introduction**

Synaptic transmission, or the transmission of the information between neurons in the nervous system, occurs through several mechanisms: neurotransmitter release from the presynaptic neuron, as a result of neurotransmitter binding to the receptor, and inward ionic current flow through open receptors leading to postsynaptic membrane depolarization [1], [2], [4], [5]. In order to model synaptic transmission, all these important steps have to be considered. Previous models focus either on neuro-

---

\* Department of Bioengineering and Biotechnology, “Politehnica” University of Bucharest, Romania and Department of Biophysics, “Carol Davilla” University of Medicine and Pharmacy Bucharest, Romania.

\*\* Department of Bioengineering and Biotechnology, Department of Electrical Engineering, “Politehnica” University of Bucharest and “Gheorghe Miho–Caius Iacob” Institute of Statistical Mathematics and Applied Mathematics, Romanian Academy.

transmitter release and diffusion [9], [11], [12], or on electrical phenomena occurring at postsynaptic level [9], [8]. The model proposed in this paper is concerned with computing postsynaptic currents out of neurotransmitter distribution in the synaptic cleft, in the particular case of a nicotinic synapse, where the neurotransmitter is acetylcholine (ACh) that acts on a specific class of ACh receptors: nicotinic, ionotropic receptors, which – as a result of their stimulation – open ionic channels.

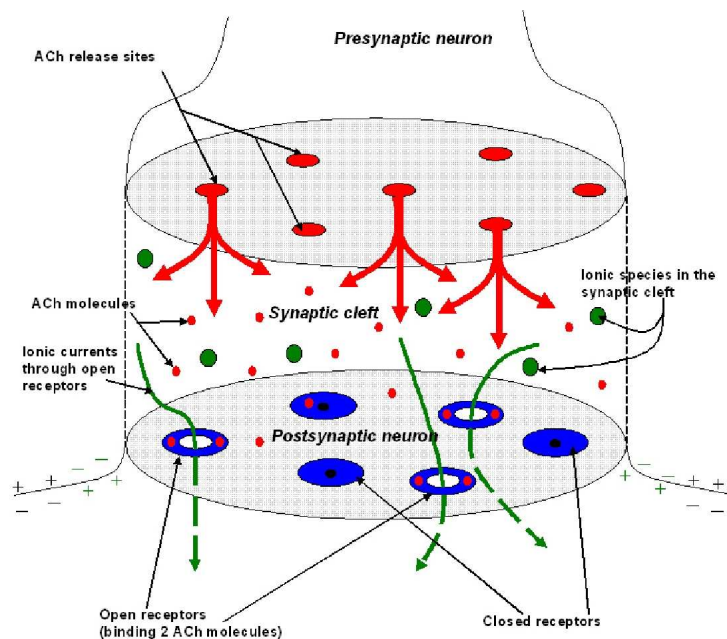


Fig. 1. The nicotinic synapse.

Since nicotinic synapses are widely distributed in the human nervous system (both in central nervous system and at the neuromuscular junction) and since there are several debilitating diseases affecting the nicotinic synapse, a global model of this type of synapse is necessary in order to better understand particular synaptic physiology and pathology. Further more, such a model would offer the possibility of pre-clinical testing the effects produced by drugs that act on the nicotinic synapses.

By adjusting the numerical parameters that occur in the model, the model can be easily modified for any type of chemical synapse (involving ionotropic postsynaptic receptors).

## 2. Physical and Mathematical Models

The neuro-muscular junction is a particular type of synapse using acetylcholine (ACh) as neurotransmitter, which binds to specific receptors called nicotinic receptors

[4], [5]. Figure 1 shows a sketch of the synaptic domain and of the processes involved in neurotransmission, that our model accounts for.

### 2.1. ACh Release and Diffusion

ACh concentration diffusion is modeled on a 2D geometry that assumes Cartesian symmetry for the synaptic space (Fig. 2). The hatched squares in the synaptic space sketch correspond to the enzyme ACh-esterase.

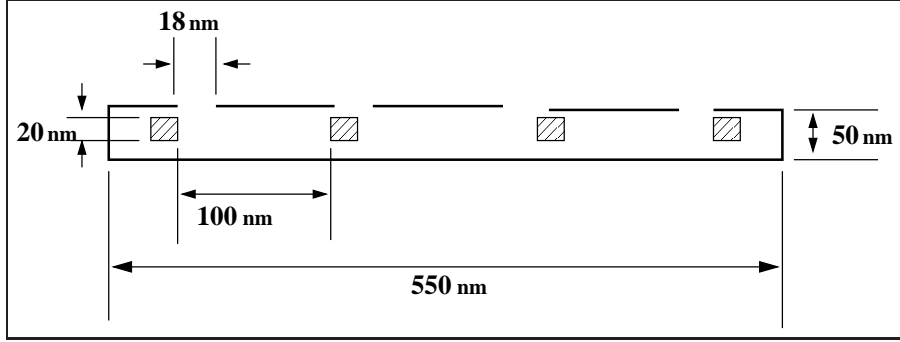


Fig. 2. 2D synaptic geometry used for modeling ACh concentration diffusion. The overall width of the synapse is  $30 \times 50$  nm [1], [4], [5].

Our model accounts for the degradation of ACh by the ACh-esterase enzyme (the hatched squares in the synaptic sketch, Fig. 2). The concentration of ACh is given by [9], [11], [12]

$$\frac{dC}{dt} = \nabla \cdot (D \nabla C), \quad (1)$$

where  $C$  is the ACh concentration and  $D$  is the diffusivity of ACh - in this specific environment,  $D = 4 \cdot 10^{-4} \mu\text{m}^2/\mu\text{s}$ ,  $C(0) = 0$ .

The boundary conditions [9], [11], [12] are given by

$$\mathbf{n} \cdot (D \nabla C) = Q(t), \quad (2)$$

on the ACh release ports

$$\mathbf{n} \cdot (D \nabla C) = -kC, \quad (3)$$

on the ACh-esterase ports

$$\mathbf{n} \cdot (D \nabla C) = 0,$$

elsewhere.

Here  $k = 2 \cdot 10^{-3} \mu\text{m}/\mu\text{s}$  is the activity constant of the ACh-esterase enzyme [11] and  $Q(t)$  is the ACh flux through the release ports [9]

$$Q(t) = \frac{S_0 C_0}{l\tau \left(1 - e^{-\frac{4t}{\tau}}\right)} e^{-\frac{t}{\tau}}, \quad (4)$$

where  $S_0 = 65 \text{ nm}^2$  is the average, maximum (initial) cross-section area of the pre-synaptic vesicle,  $C_0 = 0.3 \text{ molecule/nm}^3$  is the initial ACh concentration,  $l = 18 \text{ nm}$  is the diameter of the release port, and  $T = 6.25 \mu\text{s}$  is the characteristic time of the diffusion process. For  $\tau$  (the characteristic time of the ACh release process) we use here  $\tau \approx T$  (the values should vary according to different ACh release kinetics).

The total amount of ACh in the synaptic place is computed by

$$C_A(t) = \int_{\text{synapse}} C(t) dS, \quad (5)$$

where  $C_A$  is the total number of ACh molecules in the synaptic space,  $d$  is the depth of the synaptic space,  $d = 10 \mu\text{m}$ , and  $n_x = 30$  is the average number of geometric periods in our simplified model of the synaptic space.

The mathematical model was solved numerically by the finite element method using FEMLAB 3.1i [3].

## 2.2. Receptor dynamics

ACh receptors present two binding sites for ACh molecules. Only the double-ligated receptor opens, permitting the inward current flow. There are several intermediate states of the receptor between the non-ligated-closed-receptor and the double-ligated-open-receptor [5], [6]. The transition rates between intermediate states depend on the available amount of ACh. Our model is based on a simplifying hypothesis : it considers only two different states of the receptor – open and closed (Fig. 3). The transition rates between these two states are chosen such as to account for the intermediate steps:  $k_+$  (transition closed-open receptor) =  $0.04 \text{ ms}^{-1}$ , and  $k_-$  (transition open-closed) =  $0.02 \text{ ms}^{-1}$  [6].

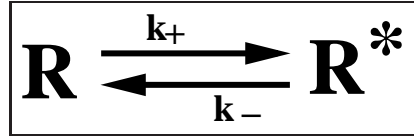


Fig. 3. Receptor dynamics – simplified model.  $R$ –closed nicotinic receptors,  $R^*$  – open nicotinic receptors.

The mathematical model is made of the following Cauchy-type problem

$$\frac{d^* R^*(t)}{dt} = k_+ R(t) - k_- R^*(t), R^*(0) = 0, \quad (6)$$

where

$$R^*(t) + R(t) = R_{\text{total}}(t), \quad (7)$$

$R^*(t)$  is the number of open synaptic receptors,  $R(t)$  is the number of closed receptors, and  $R_{\text{total}} = 10^5$  is the total number of postsynaptic receptors. All these processes occur only if there is enough free ACh in the synaptic space. Considering that the

amount of free ACh equals the difference between the amount of total synaptic ACh and the amount of bound (to receptors) ACh, and the every open receptor is bound to 2 ACh molecules, this condition can be written as

$$C_A(t) - 2R^*(t) > 0, \quad (8)$$

### Non-dimensional equations

Considering the following changes of variables

$$N = \frac{R_{\text{tot}} - R}{R_{\text{tot}}}, \tau = \frac{t}{T}, \quad (9)$$

where

$$T = \frac{1}{k_+ - k_-}, \quad (10)$$

the non-dimensional form of problem (6) is

$$\frac{dN}{d\tau} + N = \frac{k_+}{k_+ + k_-}, N(0) = 0. \quad (11)$$

The solution to this problem poses no difficulties: it was computed numerically by a finite differences backward Euler technique using a constant time-step algorithm implemented in MATLAB [7]. Accuracy test consisted of using several values for the time step and checking that the results do not change.

### 2.3. The Postsynaptic Voltage

The postsynaptic voltage was computed based on a equivalent electrical circuit of the membrane shown in Fig. 4 [9], [10].

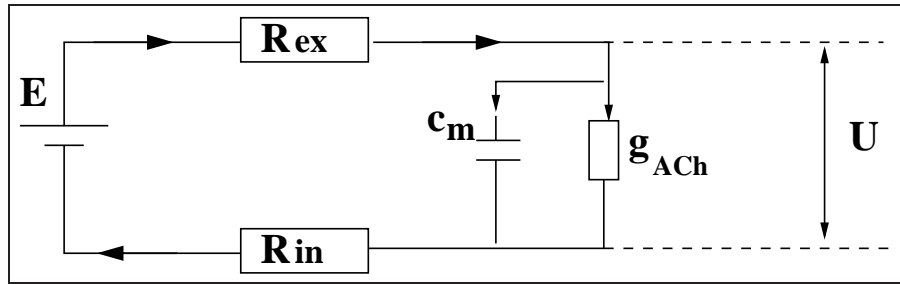


Fig. 4. Equivalent electrical circuit of the membrane.  $E$  – membrane-specific equivalent e.m.f. (due to ionic pumps in the membrane);  $R_{ex}$  – resistance of the extra-cellular space;  $R_{in}$  – resistance of the intracellular space,  $C_m$  – membrane capacity;  $g_{ACh}$  – membrane conductance (due to open nicotinic receptors),  $U$  – postsynaptic voltage.

The postsynaptic voltage is obtained by solving, again, a Cauchy-type problem

$$C_m \frac{dU}{dt} + U \left[ g_{ACh}(t) + \frac{1}{R_{ex}} \right] = \frac{E}{R_x}, \quad U(0) = E, \quad (12)$$

where

$$g_{ACh}(t) = \gamma R^*(t), \quad (13)$$

where  $E = -70$  mV is membrane-specific equivalent e.m.f.,  $R_{ex} = 5 \cdot 10^4 \Omega$  is the resistance of the extra-cellular space,  $R_{in}$  (approx. 0) is the resistance of the intracellular space,  $C_m = 4 \cdot 10^{-12}$  F is the membrane capacity,  $g_{ACh}$  is the membrane conductance (due to open nicotinic receptors),  $\gamma = 20$  pS is the single-channel conductance (for 1 single nicotinic channel), and  $U$  is the postsynaptic voltage.

The non-dimensional form of problem (12) is

$$\frac{d\tilde{U}}{d\tau} + \tilde{U} [\gamma R_{ex} R_{tot} N(\tau) + 1] = 1, \quad \tilde{U}(0) = 1, \quad (14)$$

with

$$\tilde{U} = \frac{U}{E}, \quad \tau = \frac{t}{T}, \quad \text{and} \quad T = \frac{1}{R_{ex} C_m}. \quad (15)$$

In this case too, the solution poses no difficulties: it was computed numerically by the same finite differences backward Euler technique using a constant time-step algorithm implemented in MATLAB [7]. The same type of accuracy test with respect to the time step was performed.

### 3. Results

The number of ACh molecules in the synaptic space during the first  $6.25 \mu s$  ( $\tilde{T} = 1$ ) and  $62.5 \mu s$  ( $\tilde{T} = 10$ ) is shown in Fig. 5.

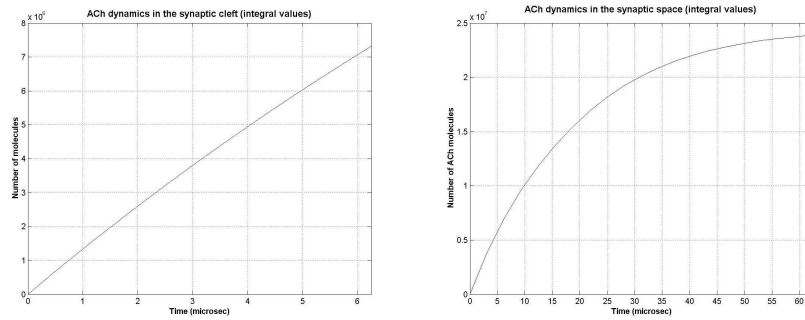


Fig. 5. ACh diffusion dynamics.

Figure 6 presents the receptor opening dynamics.

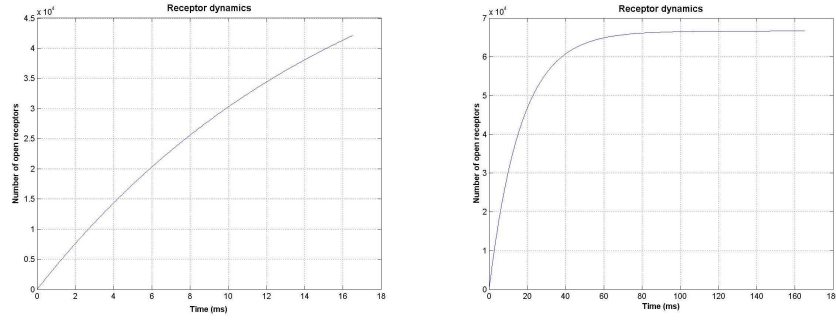


Fig. 6. Receptor dynamics.

We also investigated how the opening kinetics of the postsynaptic receptors influences the triggering profiles: Figure 7 shows the results, for different opening/closing rates.

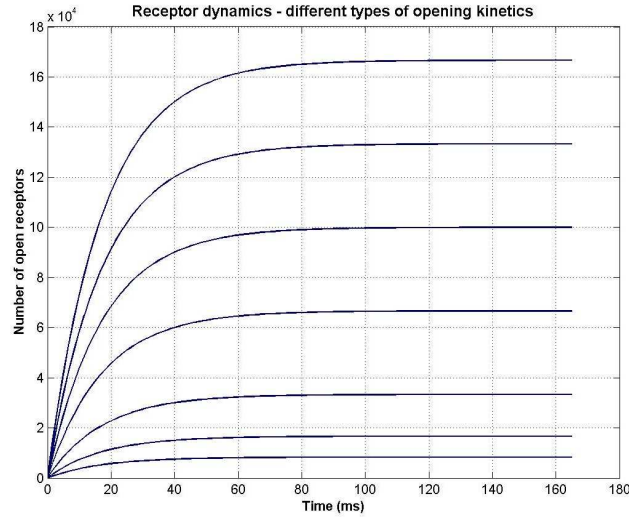
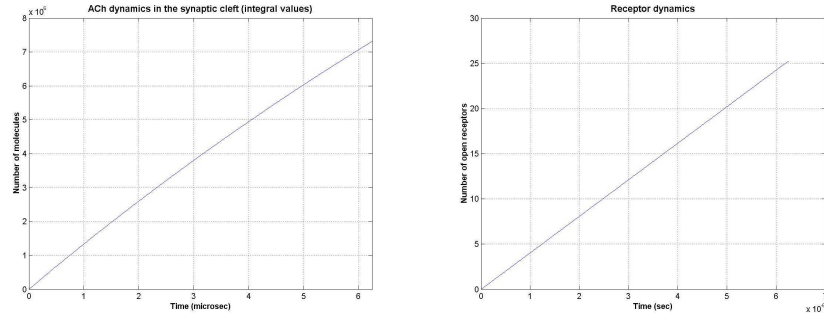


Fig. 7. Receptors triggering profiles for different opening kinetics.

$$\text{From top: } \frac{k_+}{k_+ + k_-} = \frac{5}{3}; \frac{4}{3}; \frac{3}{3}; \frac{2}{3}; \frac{1}{3}; \frac{0.5}{3}; \frac{0.25}{3}.$$

While ACh concentration rapidly reaches high values, only a small number of receptors open in the same period of time ( $6.25\mu s$  from the moment  $t = 0$ ). Figure 8 compares the ACh and receptor profiles in the first  $6.25\mu s$  after the initiation of the ACh release process.

Depolarization of the membrane occurs rapidly after the first receptor channels

Fig. 8. ACh *vs* receptor dynamics.

open, and reaches a positive value. Either different opening kinetics for the receptors or different values from the extra-cellular resistance or single-channel conductance slightly modify the plateau. Figure 9 shows the postsynaptic membrane voltage computed from our model, for  $\frac{k_+}{k_+ + k_-} = \frac{2}{3}$  and the values for the extra-cellular resistance and single-channel conductance mentioned in the text above.

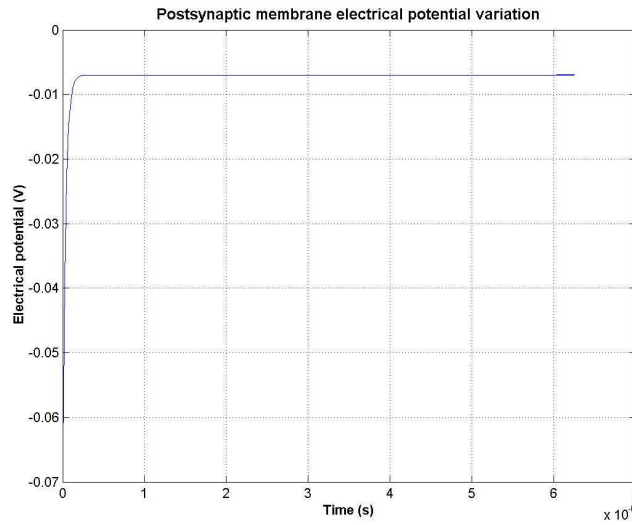


Fig. 9. Postsynaptic membrane voltage.



#### 4. Discussion and Conclusions

Apparently, cholinergic synaptic transmission involves 3 important processes, each with its own time constant:

- $T_{ACh} = 6.25 \mu s$  – for ACh diffusion;
- $T_r = 16.5 ms$  – for receptor opening;
- $T_m = 0.25 \mu s$  – for membrane depolarization.

Although the fastest process is depolarization of the membrane, the leading process (i.e. triggers all the other processes) is the ACh release and diffusion. The receptors need a longer time to reach a plateau. In our model, the full saturation of receptors (100% of the receptors open) cannot be achieved, because of the opening kinetics of the receptors  $\left(\frac{k_+}{k_+ + k_-} = \frac{2}{3}\right)$ . In a model that considers  $k_- = 0$  (Fig. 7 the third curve from the top), saturation is achieved after approximately 80 ms. The first two curves in the same figure are merely theoretical – they suppose a negative value for  $k_-$ , which is not possible from a biological point of view.

The concentration *vs* open-receptor profiles show that is enough ACh to open the postsynaptic receptors in the first phase: ACh concentration grows much faster than the number of open receptors. The influence of the ACh concentration on receptor kinetics becomes important after a much longer time (not in the time-domain of our simulation). Depolarization of the membrane occurs rapidly after receptor-opening. The characteristic peak of postsynaptic membrane depolarization is not explained in our model because of several aspects that our model does not account for: different release kinetics for ACh (simulating trains of impulses, which covers better for the physiological situation), the “refractory period” of the receptors - receptors close after a certain time, and enter a refractory state, where they stay closed, independently from the ACh molecules still existing in the cleft; the propagation of depolarization on the postsynaptic membrane. Further simulations should also consider the influence of ACh on the receptor opening profiles –  $k_+$  and  $k_-$  should vary with the concentration, but effect of a burst of impulses arriving at pre-synaptic level, the saturation of receptors.

#### References

- [1] Bennett M.R., *Synaptic transmission at single buttons in sympathetic ganglia*, News Physiol. Sci., **15** (2000), pp. 98–101.
- [2] Castonguay A. et al., *Differential regulation of transmitter release by presynaptic and glial  $Ca^{2+}$  internal stores at the neuromuscular synapse*, J. Neurosci., **21** (2001), No. 6, pp. 1911–1922.
- [3] COMSOL AB – FEMLAB 3.1*i*.

- [4] Dimoftache C., Sonia Herman, *Principii de biofizică umană*, Editura Universitara “Carol Davilla”, Bucureşti, 2003.
- [5] Guyton A.C., *Textbook of medical physiology*, 8<sup>th</sup> Edition, W.B. Saunders Comp., 1991.
- [6] Keleshian A.M. et al., *Evidence for cooperativity between nicotinic acetylcholine receptors in patch-clamp records*, Biophys. J., **78** (2000), pp. 1–12.
- [7] MathWorks – MATLAB 6.5
- [8] Popescu, Anca, Al. Morega, *Current and Voltage Distribution Within the Synaptic Cleft – A Modelling Study*, in: *Proceedings of the Third Workshop on Mathematical Modelling of Environmental and Life Sciences Problems*, Constanţa, România, Editura Academiei Române, 2005.
- [9] Popescu, Anca and Al. Morega, *Current and Voltage Distribution Within the Synaptic Cleft*, in: *Proceedings of the Second Workshop on Mathematical Modelling of Environmental and Life Sciences Problems*, Bucharest, Romania, Editura Academiei Române, Bucureşti, 2005.
- [10] Savtchenko, L.P. et al., *Effect of voltage drop within the synaptic cleft on the current and voltage generated at a single synapse*, Biophysics. J., **78** (2000), No.3, pp. 1119–1125.
- [11] Smart J.L. et al., *Analysis of synaptic transmission in the neuromuscular junction using a continuum finite element model*, Biophysics. J., **75** (1998), pp. 1679–1688.
- [12] Tai, K. et al., *Finite element simulations of acetylcholine diffusion in the neuromuscular junction*, Biophysics, J., **64** (2003), pp. 2234–2241.

## Effect of tricyclic antidepressants on the frog epithelium

Corina Prica<sup>\*†</sup>, Emil Neaga<sup>\*</sup>, Beatrice Macri<sup>\*</sup>, Dumitru  
Popescu<sup>\*‡</sup> and Maria Luiza Flonta<sup>\*</sup>

Our study was undertaken with the aim of testing the action of tricyclic antidepressants, amitriptyline (AMT) imipramine (IMI) and desipramine (DES) on the epithelium sodium channel (ENaC) using the voltage-clamp technique on the frog epithelium (*Rana ridibunda*). The epithelium sodium channel belongs to Deg/ENaC family, like ASICs and many other putative members in the brain. We have studied the effect of amitriptyline, imipramine and his metabolite desipramine on the short-circuit current ( $I_{sc}$ ) and transmembrane conductance ( $G_t$ ) on the range of concentration of 1  $\mu$ M to 200 M. We observed that all three antidepressant have a dual effect on short-circuit current  $I_{sc}$ , and on the conductance, increasing these two parameters on the range of concentration 1-50  $\mu$ M and reducing them at the concentration 100-200  $\mu$ M. These results suggest once again, multiple effects of tricyclic antidepressant on different ionic channels and receptor: blocks sodium current sensitive at TTX from neurons DRG and cardiac muscle, interact with the channels and carriers of  $K^+$  and act on the cholinergic receptors, inhibits the recapture of serotonin and norepinephrine. All these effects of tricyclic antidepressants could be utilized to explain the effect of antidepressive and all secondary effects which appear during the treatment.

### 1. Introduction

Amiloride sensitive  $Na^+$  channels are essential control elements for the regulation of  $Na^+$  transport into cells and across epithelia.

---

<sup>\*</sup> Faculty of Biology, University of Bucharest, Splaiul Independenței, 91-95, 050095, Bucharest, Romania.

<sup>‡</sup> “Gheorghe Mihoc–Caius Iacob” Institute of Statistical Mathematics and Applied Mathematics, Calea 13 Septembrie 13, 050711, Bucharest, Romania.

<sup>†</sup> e-mail: corina.prica@gmail.com, phone 0040-21-318 15 69, fax 0040-21-318 15 73.

Sodium is the most abundant ion in the extracellular fluid and total body sodium is an important determinant for the regulation of the extracellular fluid volume. The kidney of land vertebrates possesses mechanism that regulate salt balance with the tendency to retain salt and water in the body in order to preserve the volume of the extra cellular fluid and to maintain adequate blood circulation. The epithelial sodium channel (ENaC) in the renal distal part of the nephron is an important component in the control of sodium balance. ENaC is located in the apical or the outward-facing membrane of many other salt-reabsorbing epithelia and facilitates  $\text{Na}^+$  movement across this membrane as a first step in the process of  $\text{Na}^+$  transport.

Epithelial tissues play a critical role in controlling the whole body internal environment, by providing gas exchange, solute and water uptake and contain the routes for secretion and excretion. Typically, they are arranged in sheets consisting of one or more layers of communicating cells that realize the transport function.

A characteristic of all epithelial is their polarity: the apical (luminal) membrane is morphologically and functionally distinct from the basal and lateral membrane.

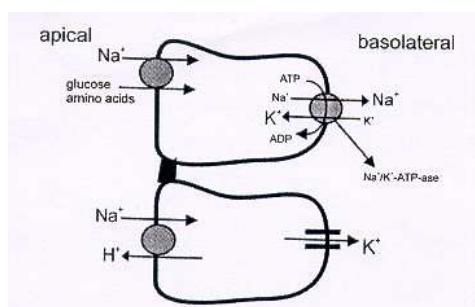
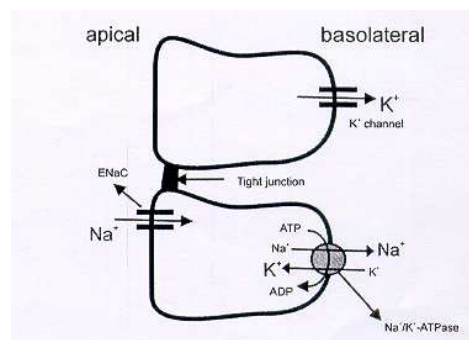
Specialized junctions that connect the cells to each other make up the transition between the apical and lateral membrane at the luminal border. The apical border can be exposed to air (epidermis, tracheal and corneal epithelium) or it can face the lumen of the cavity of the organ that contains solutions with an ionic and solute composition that differs strongly from the body fluids (gut, bladder, small intestine and kidney tubule). The basal and lateral sides called the basolateral domains have specialized regions for cell-cell adhesions and for signal transduction.

According to the permeability properties of the tight junctions, epithelia are classified in tight and leaky.

Leaky epithelia (small intestine, gall bladder and proximal kidney tubule) are characterized by a high permeability of the paracellular pathway. Their apical membranes mainly possess co-transporter systems, such as  $\text{Na}^+$ /glucose and  $\text{Na}^+$ / $\text{Po}_4^{3-}$  and antiporter systems as, for instance,  $\text{Na}^+$ / $\text{H}^+$  that provide  $\text{Na}^+$  uptake.

$\text{Na}^+$  movement across the tight epithelia such as urinary bladder, distal kidney tubule and large intestine the paracellular pathway has a low permeability. Apical  $\text{Na}^+$  uptake occurs through  $\text{Na}^+$  selective channels. The two steps involved in  $\text{Na}^+$  transport across tight epithelia are illustrated in Fig. 2.

ENaC is involved in regulation of  $\text{Na}^+$  uptake by facilitating the entry of ENaC ions into the cell driven by an electrochemical gradient. This gradient is generated by  $\text{Na}^+/\text{K}^+-\text{ATPase}$ . Which extrudes three  $\text{Na}^+$  ions from the cell in the exchange for two  $\text{K}^+$  ions and thus maintaining cell concentration of  $\text{Na}^+$  low and  $\text{K}^+$  high. The electrochemical  $\text{K}^+$  gradient across the  $\text{Na}^+/\text{K}^+-\text{ATPase}$  are responsible for maintaining the intracellular negative potential that facilitates  $\text{Na}^+$  uptake at the apical border. It is generally accepted that under most circumstances that  $\text{Na}^+/\text{K}^+-\text{ATPase}$  only works at about one-third of this maximal capacity. However, the uptake of  $\text{Na}^+$  is the rate-limiting step for the  $\text{Na}^+$  reabsorption and the represents the principal target of transport regulation.

Fig. 1.  $\text{Na}^+$  transport across leaky epithelia.Fig. 2.  $\text{Na}^+$  transport across tight epithelia.

## 2. Aims

It is generally accepted that there is a relationship between chronic pain and depression [8]. In this sense, it is known that serotonin, noradrenaline [1] and opioids [5] are involved in both nociceptive and depressive disorders, as well as in the mechanism of action underlying the antinociceptive [13, 14] and antidepressant effects of antidepressants [2]. In clinical practice, antidepressants – usually tricyclics – are widely used in several painful conditions, as well as opioids and other analgesics and have been proven to be effective in the management of pain of diverse aetiology [15, 7, 10]. On the other hand, it has been previously reported that opiate analgesics with monoaminergic reuptake inhibitory properties may induce an antidepressant-like effect in mice [11] accounting for the interrelationship between these two entities.

In spite of the wide use of antidepressants in painful disorders, the nature of antidepressant-induced analgesia remains to be elucidated. It has been suggested that antidepressant drugs have specific analgesic properties, and a body of clinical [8, 9] and experimental [12, 14, 4] types of evidence together seems to demonstrate that the analgesic may be independent of the antidepressant effects.

However, the relationships between antidepressant and analgesic effects of these drugs have not yet been fully elucidated [6].

The analgesic mechanism of amitriptyline, imipramine and desipramine is unclear. One explanation is that it blocks Na channels, reducing in this way spontaneous activity in nerve fibres, which cause pain. A new class of ion channels known as the Deg/ENaC (degenerins-epithelial Na channel) family which includes ASIC, DRASIC mediate pain induced by acidosis in damaged or inflamed tissue. The same family includes also EnaC, which can be easily studied in the amphibian epithelial model.

This study was undertaken with the aim of testing the actions of amitriptyline, imipramine and desipramine on native epithelial Na channels from frog skin. We assumed that, having a common protein structure, characterization of the amitriptyline-ENaC interaction could help to elucidate the analgesic mechanism of this antidepressant.

### 3. Materials and Methods

#### 3.1. Mounting of the epithelium

**Preparation of the frog skin.** Frog (*Rana ridibunda*), weighting around 50 g, were kept at 17° C with free access to tap water. The abdominal skin of doubly pithed frog was dissected and mounted in an Ussing-type Lucite chamber. The chamber ensured negligible edge damage and allowed continuous perfusion with fresh solutions of both the mucosal and serosal sides of the epithelium, at a rate of 5 mL min<sup>-1</sup>. The tissue area in contact with the bathing media was 1 cm<sup>2</sup>.

In order to perform measurements of the transepithelial current and voltage under many different circumstances, i.e. with varying solutions on the outer and inner side of the tissue, we mounted the preparation in a chamber. The tissue was placed as a flat sheet between two fluids filled compartments. This technique is often attributed to H.H. Ussing (1949). The “Ussing chamber” we used is shown in Fig. 3.

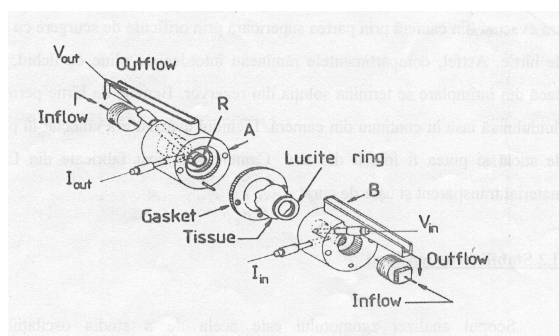


Fig. 3. The “Ussing chamber”.

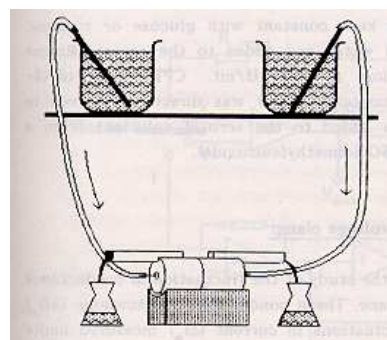


Fig. 4. Design of Ussing chamber.

These compartments were filled with Ringer solution to equal levels, this maintaining the pressure difference across the tissue equal to zero. When the two chamber halves were put together with screws, the edges of the tissue became compressed between the rim (R) on the left and the right compartment. The pressure that was then extending on the tissue's edge could be adjusted by changing the thickness of the gasket, which was put between the two chamber halves. In this way edge damage of the tissue was minimized. A thin film of the silicon vacuum grease on the rim of each chamber half was used to seal the chamber and also the further reduce edges damage. The skin area exposed to the bathing solution was 0,5 cm<sup>2</sup>. Both chamber compartments were continuously perfused during the experiment with a solution flow of 4–6 ml/min. The incoming fluid next to directed towards the tissue face. In this way a layer of stagnant fluid next to the tissue was prevented. The design of the chamber allowed for rapid change of solution without of voltage clamping, Fig. 4.

Each chamber halve had its own reservoir, which consisted of the beaker filled with a Ringer solution, was placed on the higher level than the chamber, so that the solution flow was maintained by gravity.

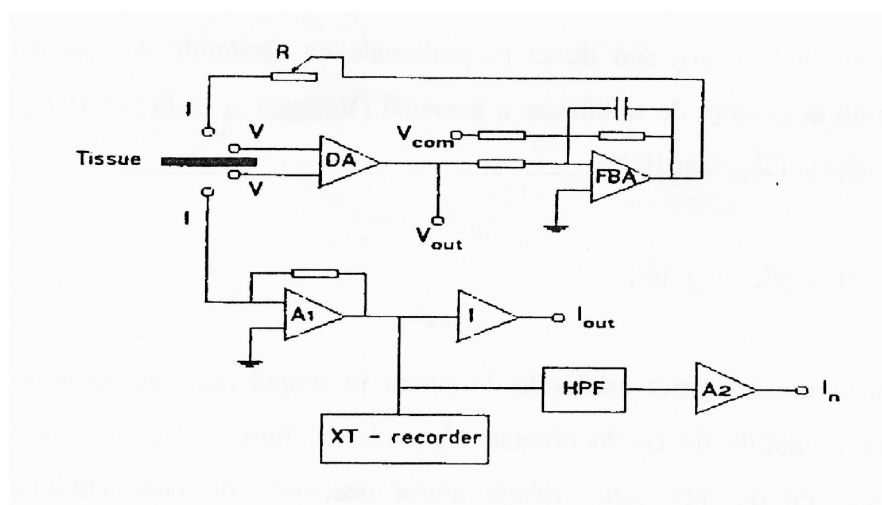


Fig. 5. The voltage clamp circuit.

One end of the tubing was connected with a metal tube that was placed into the beaker. The other end of the tubing was connected with the chamber. The solution could easily be changed by moving the metal end of the tubing from one reservoir into another. The fluid left the chamber via the upper side, through spillways that contained paper strips. In this way the compartments always remained filled with solution, also when the (by accident) one of the reservoirs became empty. The paper strips allowed the fluid to leave the chamber continuously. Via the strips the fluid was collected in the glass beakers. It could be used again by emptying the beakers in the reservoirs.

### 3.2. The voltage clamp

The purpose of noise analysis is the study of the fluctuation in conductance of an ionic pathway in a cell membrane. This conductance fluctuation ( $G$ ) are directly proportional to the fluctuation in current ( $I$ ) measured under voltage clamp conditions.

So, one can record current fluctuation in order to obtain information about the conductance fluctuation. Indeed, the shape of their spectra will be identical. On the other hand, the relation between conductance fluctuation and fluctuation in voltage measured under current clamp conditions is more complicated. The relation between the power spectral density of the voltage and conductance fluctuation will depend on the frequency dependence of the impedance of the membrane.

For this reason, we performed the experiments under voltage clamp conditions. This yields information about the number of channels, the single channel current and the probability that the channel is in a specific state. Prerequisite for the measurements of the membrane noise is a low-noise measuring circuit. A schematic plot of the



voltage clamp circuit we used is shown in Fig. 5. The voltage clamp is a feedback circuit that keeps the voltage across the epithelium constant by varying the current flow through the epithelium. In order to avoid large junction potentials, the connection between the voltage input and current output of the clamp was made by two current (I) and two-voltage (V) electrodes, figure 5. They consisted of Ag/AgCl/3M KCl electrodes which were in contact with the solutions in the “Ussing chamber” through agar bridges ( 3 gr agar per 100 ml 1M KCl ). The two voltage electrodes placed to each side of the epithelium allow the recording of the transepithelial potential (PD) and two current electrodes allow injection of the clamp current. Voltage clamp of PD to zero allows to assess  $I_{scNa}$  transport ( $I_{Na}$ ) was defined as the amiloride-sensitive component of  $I_{sc}$ . An automatic electronic circuit clamps the spontaneous potential difference across the epithelium to zero ( voltage clamping to short-circuit state). The voltage clamp is an automatic fed-back device intended to impose a constant voltage across the epithelium. The input stage of the clamp consisted of the low-noise differential amplifier ( DA ) with a gain of 100x. The feedback amplifier (FBA) converts the voltage offset at the input to current. Moreover, the FBA is made as summing amplifier so that a clamp potential between  $-100$  to  $+100$  mV can be imposed to the epithelium. Via an external input (Vcom) this potential could be varied under computer control to record current-voltage relations.

The epithelium can be represented as a RC-network. The impedance of this circuit varies among the different tissues. This RC circuit influence the frequency characteristics of the voltage clamp. The frequency response of the voltage clamp circuit could be varied by changing the capacitance in the FBA and by varying the resistance of the potentiometer ( R ) which was installed in the feedback circuit. When the feedback loop was closed this resistance was  $470 \Omega$  . Concomitantly the resistance was gradually decreased to make the response of the voltage clamp faster. An optimal adjustment of the feedback capacitance and of R was obtained by inspection of the transepithelial current and the voltage response to train of square wave pulses with a frequency of 200 Hz which was applied at Vcom. The feedback current was measured with a current to voltage converter. The output signal of this amplifier was connected to three different setups.

1. The transepithelial current was continuously recorded on a XT-recorder. The transepithelial conductance was obtained from the current deflections caused by the application of 10 mV transepithelial potential pulses at  $V_{com}$ .

2. After inversion (I), the signal was displayed on the liquid-crystal display and could be sampled under computer control.

3. The output signal was connected via a high pass filter ( HPF ) with adjustable cut-off frequency to a voltage amplifier with a total gain of 1000. The output of this amplifier (  $I_n$  ) was sent to a computer for analysis.

### 3.3. Noise system

With noise analysis, we intend to analyse the random fluctuation in current associated with ionic movement through biological membranes. The thermal or Brownian movement to which the charge carriers are subjected produces what is called ther-



mal or Johnson-Nyquist noise. In non-equilibrium conditions (net ionic current flow), other contributions besides the thermal noise exist, together denoted as excess noise. In the accessible ms to s time range, distinctive features (e.g. Lorentzian) in the spectral analysis (power spectrum) of the current fluctuation originate from stochastic changes in the conductive pathways for ion movements through the membrane. The underlying processes for this "Lorentzian noise" are thought to consist of probabilistic opening and closing events of ion channels. When recorded, in conditions where the voltage across the membrane is maintained constant, the fluctuation in current is directly proportional to the conductance fluctuation. Information about the number of channels in the ensemble and about the single channel current can be obtained from the analysis of the noise caused by random blockage of ion channels by reversible blockers. Fluctuation of current caused by random interaction of the blocker with the channel give rise to Lorentzian noise. Its relaxation time varies with blocker concentration which enables the calculation of the rate constants of the blocker and the single channel parameters.

The power density spectrum of fluctuations in current that passes through gated channels has a Lorentzian shape. To resolve the entire frequency range, we measured current fluctuations in a frequency range between 0.2 Hz and 1 kHz. To guarantee a sufficiently fine spectral resolution, and to keep the measuring time short, we analyzed the current-noise signal in two different frequency ranges: a low and a high frequency range. The 1000x amplified fluctuations in current (In from clamp) were sent to two separate modules (designated HS and LS in Fig. 12) each consisting of a pre-filter and an anti-aliasing filter. The fundamental frequency was fixed in most experiments to 0.5 Hz.

The output of the pre-filter and the anti-aliasing filter was amplified (1-20x) to optimize the analog-to-digital conversion in the computer. The low and the high signal were sampled consecutively by a multiplexer and sent via a sample and hold system (S&H) to an A/D converter. After digitising, data were converted from the time to the frequency domain using an FFT (Fast Fourier Transform) software routine and a double logarithmic plot of the power density  $S(f)$  as the function of the frequency ( $f$ ) was displayed on the computer's monitor. An individual spectrum was obtained from a series of 50 periods of data acquisition (sweeps). Each sweep consisted of a sequential digitization of the low (4096 points/2 sec) and then the high (4096 points/0.25 sec) analog signals. The duration of the sweep was therefore 2.25 sec and approximately 3 minutes were required to obtain and process the 50 sweep for a single spectrum which was displayed immediately on the computer's monitor. The FFT output frequencies for the LS ranged from 0.5 to 256 Hz and for the HS from 4 to 2048 Hz. In order to obtain spectra that can easily be analyzed by a computer, the number of frequency points was reduced. Each octave was divided in 8 intervals and the average value of LS and HS frequency was calculated for each interval. This resulted in 8 values per octave that yielded 74 spectral data points in the power density spectrum. The data were saved on disk for further analysis.

#### 4. Fitting procedure

Theoretical calculations on the stochastic conductance ( or current ) fluctuation between multiple conductance states of an ionic channel predict PDS that contain a number of Lorentzian curves equal to the number of conductance states minus one. The Lorentzian function :

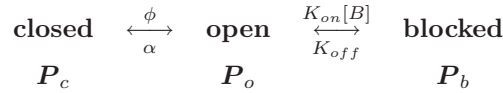
$$S(f) = \frac{S_0}{1 + \left(\frac{f}{f_c}\right)^2}$$

is characterized by two parameters: the plateau value  $S_0$  represents the noise power at zero Hz and the corner frequency  $f_c$  is the frequency where the noise power has dropped to its half-maximal value (  $S_0/2$  ).

In the fitting procedure of  $1/f$  excess noise of the preparation must be taken into account. Instrumental noise becomes dominant at higher frequencies and is excluded from the fit by selecting a suitable frequency range.

#### 5. Three state model

In the presence of the blocker the channels are distributed among three states:



Where  $\alpha$  and  $\beta$  are the rates of the open-close reaction. The parameters  $K_{on}$  and  $K_{off}$  are association and the dissociation rates of the blocker, respectively. The probability of the channel of being closed is  $P_c$ , whereas of the blocked state is  $P_b$ . In the absence of the blocker (B) the open probability is :

$$P_o = \frac{\alpha}{\alpha + \beta}.$$

The power density is now equal to the sum of two Lorentzian functions:

$$S(f) = \frac{S_o^{(1)}}{1 + \left[\frac{f}{f^{(1)}}\right]^2} + \frac{S_o^{(2)}}{1 + \left[\frac{f}{f^{(2)}}\right]^2}.$$

Thus is the case of three states “open”, “closed” or “blocked” the power density spectrum of the fluctuation in current is the sum of two Lorentzian functions. The corner frequency and plateau value of the Lorentzian components will depend on the concentration of the blocker. Patch clamp experiments univocally indicate that ENaC fluctuates between open and closed times with long mean open and closed times and would give rise to a Lorentzian with a corner frequency smaller than or in the range of 0.1 Hz. Therefore, in the frequency range of our analysis, expose of the epithelium to a blocker, give rise to a single Lorentzian in the PDS. From the blocker-induced noise, we determined  $P_o$  and  $P_b$ .

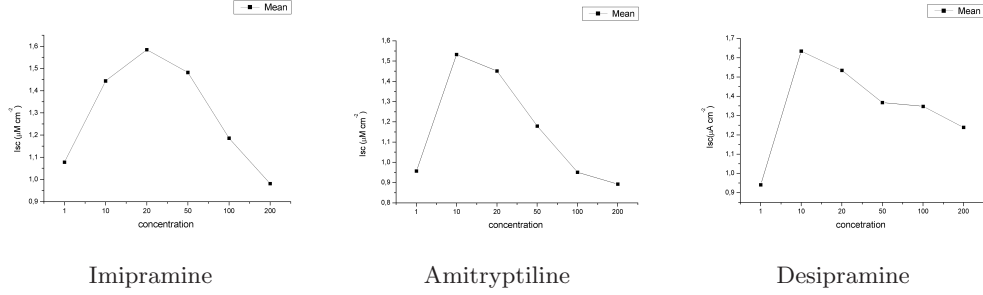


Fig. 6. The effect of imipramine, amitriptyline and desipramine on short-circuit current  $I_{sc}$  in Ringer Cl solution.

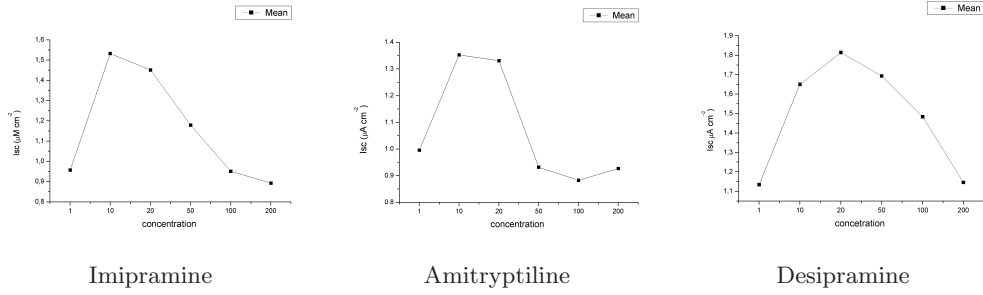


Fig. 7. The effect of imipramine, amitriptyline and desipramine on short-circuit current  $I_{sc}$  in Ringer  $SO_4$  solution.

The two Lorentzian parameters, the low-frequency plateau ( $S_o$ ) and the corner frequency ( $f_c$ ) were determined by non-linear curve fitting of the spectra presuming a first order interaction between channel and blocker. The on- ( $K_{on}$ ) and off ( $K_{off}$ ) rate constant of the blocking reaction were calculated from:

$$2\pi f_c = K_{on}[B] + K_{off},$$

where  $[B]$  is the blocker concentration. Using the values of the Na current and the Lorentzian parameters ( $f_c$  and  $S_o$ ) obtained from noise analysis, we estimated  $i_{Na}$  si  $N_o$ .

$$i_{Na} = \frac{S_o (2\pi f_c)^2}{4I_{Na} K_{01} [B]}, \quad N_0 = \frac{I_{Na}}{i_{Na}}.$$

## 6. Results

This study is a continuation of a research started with amitriptyline. We effected a curve dose-effect (1–200  $\mu M$ ) and we have found out that the tricyclic

antidepressants have the higher effect at a concentration of 10  $\mu\text{M}$ . We used two different solutions of Ringer; Ringer  $\text{Cl}^-$  and Ringer  $\text{SO}_4$  to see if there is a different effect of chloride ions.

Graphs 1 to 11, represent the effect of short-circuit current  $I_{sc}$  when we applied the antidepressants on the apical membrane of the epithelium. The current increases progressively until a concentration of 50  $\mu\text{M}$  and at higher concentrations appears an inhibitory effect. The effect is similar in Ringer Cl or Ringer SO solutions, which exclude a specific effect of tricyclic antidepressant on chloride ions conductance, which demonstrates that the flow of sodium is modulated by tricyclic antidepressant.

We observed that all three antidepressants have a dual effect on short-circuit current  $I_{sc}$ , and on the conductance, increasing these two parameters on the range of concentration 1–50  $\mu\text{M}$  and reducing them at the concentration 100–200  $\mu\text{M}$ . The inhibitory effect of tricyclic antidepressants on sodium ion flow at higher concentrations can be explained by the phenomenon of feed-back inhibition of ENaC at higher concentrations on Na intracellular. These results suggest once again, multiple effects of tricyclic antidepressant on different ionic channels and receptors: blocks sodium current sensitive to TTX from neurons DRG and cardiac muscle, interacts with the channels and carriers of  $\text{K}^+$  and acts on the cholinergic receptors, inhibits the recapture of serotonin and norepinephrine.

All these effects of tricyclic antidepressant could be utilized to explain the effect of antidepressive and analgesic and all secondary effects which appear during the treatment. But we must mention that in our experimental studies the concentration that we used is higher than the plasma concentration which is 1  $\mu\text{M}$ . This suggests the existence of receptors with higher affinity responsible for the therapeutic effects of tricyclic antidepressants.

We also observed that at a concentration of 10  $\mu\text{M}$  the amitriptyline has a higher effect comparative with imipramine and desipramine (graph 12).

## References

- [1] Basbaum, A.I., Fields, H.L., *Endogenous pain control systems: brainstem spinal pathways and endorphin circuitry*, Annu. Rev. Neurosci., **7** (1984), pp. 309–338.
- [2] Carlsson, A., Corrodi, H., Fuxe, K., Hokfelt, T., *Effects of some antidepressant drugs on the depletion of intraneuronal brain catechol-amine stores caused by 4- $\alpha$ -dimethyl-meta-tyramine*, Eur. J. Pharmacol., **5** (1969), pp. 367–373.
- [3] Casas, J., Gibert-Rahola, J., Chover, A.J., Mico, J.A., *Test-dependent relationship of the antidepressant and analgesic effects of amitriptyline*. Methods Find. Exp. Clin. Pharmacol., **17** (1995), pp. 583–588.
- [4] Casas, J., Gibert-Rahola, J., Valverde, O., Tejedor-Real, P., Romero, P., Mico, J.A., *Antidepressants, is there a correlation between their analgesic and antidepressant effect?*, Eur. Neuropsychopharmacol., **3** (1993), pp. 343–344.

- [5] De Gandarias, J.M., Irazusta, J., Varona, A., Gil, J., Fernandez, D., Casis, L., *Effect of imipramine on enkephalin-degrading peptidases*, Eur. Neuropsychopharmacol., **9** (1999), pp. 493–499.
- [6] O'Malley, P.G., Balden, E., Tomkins, G., Santoro, J., Kroenke, K., Jackson, J.L., *Treatment of fibromyalgia with antidepressants. A meta-analysis*, J. Gen. Intern. Med., **15** (2000), pp. 659–666.
- [7] Onghena, P., Van Houdenhove, B., *Antidepressant-induced analgesia in chronic non-malignant pain: a meta-analysis of 39 placebo-controlled studies*, Pain, **49** (1992), pp. 205–219.
- [8] Magni, G., Conlon, P., Arsie, D., *Tricyclic antidepressants in the treatment of cancer pain: a review*, Pharmacopsychiatry, **20** (1987), 160–164.
- [9] Magni, G., *The use of antidepressants in the treatment of chronic pain. A review of the current evidence*, Drugs, **42** (1991), pp. 730–74.
- [10] McQuay, H.J., Tramer, M., Nye, B.A., Carroll, D., Wiffen, P.J., Moore R.A., *A systematic review of antidepressants in neuropathic pain*, Pain, **68** (1996), pp. 217–227.
- [11] Rojas-Corrales, M.O., Gibert-Rahola, J., Mico, J.A., *Tramadol induces antidepressant-type effects in mice*, Life Sci., **63** (1998), pp. PL175–PL180.
- [12] Spiegel, K., Kalb, R., Pasternak, G.W., *Analgesic activity of tricyclic antidepressants*, Ann. Neurol. **13** (1983), pp. 462–65.
- [13] Schmauss, C., Emrich, H.M., *Narcotic antagonist and opioid treatment in psychiatry*, in: Rogers, R.J., Cooper, S.J. (Eds.), *Enkephalins, Opiates and Behavioral Processes*, Wiley, New Delhi, 1988, pp. 327–351.
- [14] Valverde, O., Mico, J.A., Maldonado, R., Mellado, M.L., Gibert-Rahola, J., *Participation of opioid and monoaminergic mechanisms on the antinociceptive effect induced by tricyclic antidepressants in two behavioral pain test in mice*, Prog. Neuropsychopharmacol. Biol. Psychiatry, **18** (1994), pp. 1073–1092.
- [15] Tura, B., Tura, S.M., *The analgesic effect of tricyclic antidepressants*, Brain Res., **518** (1990), pp. 19–22.
- [16] Walsh, T.D., *Antidepressants in chronic pain*, Clin. Neuropharmacol., **6** (1983), pp. 271–295.



## On the Solvability of Navier-Stokes Equations

Cristina Sburlan\*

In this paper we investigate the solvability of Navier-Stokes system, in case of an incompressible dynamical viscous fluid. The system can be reduced to an evolution problem in which are involved a linear monotone operator and a compact nonlinear one. We will use functional and topological methods to study the problem, such as Fourier method and the topological degree.

### 1. Introduction

The solvability of the Navier-Stokes system represents an actual problem.

We consider the steady state flow equation as a bifurcation problem with respect to the dynamical viscosity, and this approach permits us to use the powerful topological degree methods to study this problem.

Then, in the next section, we consider the linearized Navier-Stokes system (Stokes equation) for the flow of incompressible fluids and we succeed to find the weak solution of this evolution problem as a Fourier series.

### 2. Coincidence Degree for Steady State Flow of Incompressible Fluids

First we present the theoretical approach to the coincidence degree, and then we apply this theory to the steady state flow of incompressible fluids. Let  $H$  be a real Hilbert space and consider  $\mathcal{L} : D(\mathcal{L}) \subseteq H \rightarrow H$  a linear and maximal monotone

---

\* Faculty of Mathematics and Computer Science, “Ovidius” University of Constanța, Romania, e-mail: [c\\_sburlan@univ-ovidius.ro](mailto:c_sburlan@univ-ovidius.ro)

operator and  $\mathcal{N} : D(\mathcal{N}) \subseteq H \rightarrow H$  a (nonlinear) compact one. We can write

$$\lambda\mathcal{L} + \mathcal{N} = I + \lambda\mathcal{L} - I + \mathcal{N} = (I + \lambda\mathcal{L})(I - (I + \lambda\mathcal{L})^{-1}(I - \mathcal{N})),$$

so we have the equivalence

$$\lambda\mathcal{L}y + \mathcal{N}(y) = 0 \Leftrightarrow y - (I + \lambda\mathcal{L})^{-1}(I - \mathcal{N})(y) = 0 \quad (1)$$

Denote by  $\mathcal{M}(\lambda)$  the compact operator  $(I + \lambda\mathcal{L})^{-1}(I - \mathcal{N})$ . For  $D \subset H$  open bounded set such that  $\lambda\mathcal{L}y + \mathcal{N}(y) \neq 0, \forall y \in \partial D$ , we define the coincidence degree of  $(\mathcal{L}, \mathcal{N})$  relatively to  $D$  by

$$d_\lambda((\mathcal{L}, \mathcal{N}), D) := d_{LS}(I - \mathcal{M}(\lambda), D, 0), \quad (2)$$

( $d_{LS}$  stands for Leray-Schauder degree). This coincidence degree has all the properties of Leray-Schauder degree, such that

**(P1)** (The solution property). *If  $d_\lambda((\mathcal{L}, \mathcal{N}), D) \neq 0$ , then equation (1) has at least one solution in  $D$ .*

**(P2)** (The homotopy invariance). *Let  $(\lambda_t)_{t \in [0,1]} \subset (0, +\infty)$  be a continuous deformation such that equation  $\lambda_t\mathcal{L}y + \mathcal{N}(y) = 0$  has no solutions  $y \in \partial D$  for all  $t \in [0, 1]$ . Then  $d_{\lambda_t}((\mathcal{L}, \mathcal{N}), D)$  is independent of  $t \in [0, 1]$ .*

Denote by  $C(\mathcal{L})$  the set of all characteristic values of  $\mathcal{L}$ . We have the following theorem (see [3] and [5]).

**Theorem 1.** *Suppose that there exists  $k > 0$  such that*

$$\|\mathcal{N}(y)\| \leq k \|y\|, \quad y \in H.$$

*If  $\lambda_0 \in C(\mathcal{L})$  is such that  $\text{dist}(\lambda_0, C(\mathcal{L}) \setminus \{\lambda_0\}) > 2k$ , then equation (1) has at least one bifurcation point with  $\lambda \in (\lambda_0 - k, \lambda_0 + k)$ .*

We set  $S = \{(\lambda, u) \in \Lambda \times H \mid \lambda\mathcal{L}u + \mathcal{N}(u) = 0\}$  and  $S' = S \cup \{C(\mathcal{L}) \times \{0\}\}$ . Due to classical bifurcation Rabinowitz theorem we state:

**Theorem 2.** *Under the hypotheses of Theorem 1, the maximal connected set  $S_{\lambda_0} \subset S'$  intersecting  $(\lambda_0 - k, \lambda_0 + k) \times H$  is either unbounded in  $\Lambda \times H$  or it contains a finite number of points  $(\lambda_j, 0)$  with  $\lambda_j \in \Lambda$ . Moreover, the number of these points having odd algebraic multiplicity – including  $(\lambda_0, 0)$  – is even.*

**Remark 1.** *The above results remain true if, instead the condition from Theorem 1,  $\mathcal{N}$  satisfies the following condition*

$$\exists k > 0 \text{ such that } \|\mathcal{N}(y)\| \leq k \|y\|^2, \quad y \in H.$$

Let now  $\Omega \subseteq \mathbb{R}^N$ , ( $N = 2$  or  $3$ ), be a domain with smooth boundary  $\partial\Omega$  and denote by  $Q := \Omega \times (0, +\infty)$  and by  $\Sigma := \partial\Omega \times (0, +\infty)$ .

We will study the case of the *steady state flow*

$$u_t = 0 \Leftrightarrow u = \text{const. (in } t) = u_0(x)$$



regarding the flow equation as the eigenvalue equation

$$(u \cdot \nabla)u(x, t) - \nu \Delta u(x, t) + \nabla p(x, t) = f(x, t) \quad (3)$$

with respect to the eigenvalue parameter  $\nu$  – the *dynamical viscosity*, and the boundary conditions are given by

$$u = 0 \text{ on } \Sigma. \quad (4)$$

If  $\Omega$  is unbounded, we consider moreover that

$$u \rightarrow 0 \text{ for } |x| \rightarrow +\infty.$$

We study the incompressible fluids, so we also have the condition

$$\nabla \cdot u = 0. \quad (5)$$

Suppose that  $\Omega$  is pathwise connected. The space  $(L^p(\Omega))^N$  can be decomposed as follows

$$(L^p(\Omega))^N = H_p(\Omega) \oplus \{\nabla g | g \in W^{1,p}(\Omega)\}, \quad 1 < p < +\infty,$$

where  $X = H_p(\Omega) := \overline{\{u \in (C_0^\infty(\Omega))^N | \nabla \cdot u = 0\}}^{\|\cdot\|_{L^p(\Omega)}}$ .

Hence each  $u \in (L^p(\Omega))^N$  can be uniquely represented as

$$u = u_1 + u_2, \quad u_1 \in H_p(\Omega), \quad u_2 \in \{\nabla g | g \in W^{1,p}(\Omega)\}.$$

Denote by  $Pu = u_1$  the projection from  $(L^p(\Omega))^N$  onto its divergence free part  $H_p(\Omega)$ , called the *Helmholtz projection*. It is a bounded linear operator and it preserves the regularity, i.e.,  $P((W^{m,p}(\Omega))^N) \subset (W^{m,p}(\Omega))^N$  with continuous embedding. Moreover,  $H_p(\Omega) = P((L^p(\Omega))^N)$  is a reflexive, separable Banach space (see [10]).

Because  $P\nabla p = 0$ , applying  $P$  to the steady state flow equation we obtain

$$P(u \cdot \nabla)u - \nu P\Delta u = Pf, \quad P(\nabla \cdot u) = 0. \quad (6)$$

Let  $E := \{y \in H_p(\Omega) | y \in (W_0^{1,p}(\Omega))^N\}$  be a subspace of  $H_p(\Omega)$ . Denote by  $A \in L(E, E^*)$  the *Stokes operator*

$$\langle Ay, w \rangle := \sum_{i=1}^N \int_{\Omega} \nabla y_i \cdot \nabla w_i dx, \quad \forall y, \forall w \in E$$

and define the threeilinear form

$$b(y, z, w) := \sum_{i,j=1}^N \int_{\Omega} y_i D_i z_j w_j dx, \quad \forall y, \forall z, \forall w \in E$$

which determines the nonlinear operator  $C : E \rightarrow E^*$

$$\langle C(y), w \rangle := b(y, y, w), \quad \forall y, \forall w \in E.$$

Then we can reformulate (6) as the eigenvalue problem

$$\nu Ay + C(y) = Pf, \quad (7)$$

where  $A$  is symmetric, i.e.  $\langle Ay, w \rangle = \langle y, Aw \rangle$ , and strictly monotone because

$$\langle Ay, y \rangle = \sum_{i=1}^N \int_{\Omega} \nabla y_i \cdot \nabla y_i dx = \int_{\Omega} |\nabla y|^2 dx > 0, \quad \forall y \in E.$$

Since  $b$  is threelinear on  $E$  we have that there exists  $c > 0$  such that

$$\|C(y)\| \leq c \|y\|^2, \quad \forall y \in E, \quad (8)$$

and we deduce that there exists  $k > 0$  such that

$$\|(C - Pf)(y)\| \leq k \|y\|^2, \quad y \in E.$$

If  $\Omega$  is a bounded set, for  $p = 2$  we have that  $E$  and  $X$  are Hilbert spaces and the embedding  $E \hookrightarrow X$  is compact ( $2 \leq N \leq 3$ ) (from Sobolev-Kondrashov theorem). Then the operator  $\tilde{C} : E \subset X \rightarrow X$ ,  $\tilde{C} = I \circ (C - Pf)$ , is compact (as the composition of a compact operator with a continuous one). Similarly, the operator  $\tilde{A} = I \circ A$  is compact, too. Equation (7) becomes

$$\nu \tilde{A}y + \tilde{C}(y) = 0$$

and we can apply to this equation the above theory.

If  $\Omega$  is a complementary set (therefore unbounded) of a compact domain  $\mathcal{P}$ , i.e.,  $\Omega = \mathbf{R}^N - \mathcal{P}$ , the operators  $A$  and  $C$  are not compact because the Sobolev-Kondrashov theorem does not work in this case. We solve this problem using the following theorem (see [4]):

**Theorem 3 (BROWDER-TON).** *Let  $X$  be a reflexive separable Banach space and  $S \subset X$  be a countable set. Then there exist a separable Hilbert space  $W$  and a compact one-to-one linear operator  $\psi : W \rightarrow X$  such that  $S \subset \psi(W)$  and  $\psi(W)$  is dense in  $X$ .*

When  $X$  is the reflexive separable Banach space  $H_p(\Omega)$ , let  $\psi : W \rightarrow X$  be the above compact one-to-one linear operator and  $\varphi : X^* \rightarrow W$  be the adjoint operator defined by

$$(\varphi(v), w) = \langle v, \psi(w) \rangle, \quad v \in X^*, \quad w \in W,$$

where  $(\cdot, \cdot)$  is the scalar product in  $W$ .

Then, denoting by  $\mathcal{L} := \varphi A \psi : W \rightarrow W$  (linear, maximal monotone and compact operator) and  $\mathcal{N} := \varphi(C - Pf)\psi : W \rightarrow W$  (nonlinear compact operator), problem (7) is equivalent with

$$\nu \mathcal{L}y + \mathcal{N}(y) = 0, \quad y \in W. \quad (9)$$

Now the above theory is applicable to this equation, so writing it as a eigenvalue problem and using the coincidence degree we obtain the bifurcation points.

### 3. Fourier Method for Stokes equation

In this section we try to find the weak solution of the linearized Navier-Stokes system as a Fourier series.

Consider the Navier-Stokes system, for the flow of an incompressible fluid:

$$(\nabla \cdot u)(x, t) = 0 \quad (10)$$

$$u_t(x, t) + (u \cdot \nabla)u(x, t) - \nu \Delta u(x, t) + \nabla p(x, t) = f(x, t), \quad (x, t) \in Q \quad (11)$$

$$u = 0 \text{ on } \Sigma, \quad (12)$$

where  $\Omega \subseteq \mathbf{R}^N$ ,  $N = 2$  or  $3$ , is a bounded domain with smooth boundary  $\partial\Omega$  and  $Q := \Omega \times (0, +\infty)$ ,  $\Sigma := \partial\Omega \times (0, +\infty)$ .

Suppose that the body forces are of potential type, *i.e.*

$$f(x, t) = \nabla_x V(x, t)$$

and denote by  $q := V - p$ , where  $p$  is the (unknown) pressure in fluid. Then we can write (11) under the form:

$$u_t(x, t) + (u \cdot \nabla)u(x, t) - \nu \Delta u(x, t) = \nabla q(x, t).$$

We study the case of a dynamical viscous fluid, with big  $\nu$ . In this case we have

$$\nu \Delta u(x, t) \gg (u \cdot \nabla)u(x, t)$$

and we can approximate the above equation with:

$$u_t - \nu \Delta u(x, t) = \nabla q(x, t). \quad (13)$$

Let  $X := \{y \in (L^2(\Omega))^N; \nabla \cdot y = 0, y \cdot n = 0 \text{ on } \partial\Omega\}$  be the Hilbert space of “incompressible fluids”,  $E := \{y \in (H_0^1(\Omega))^N; \nabla \cdot y = 0\}$  be subspace of  $X$  and let  $P : (L^2(\Omega))^N \rightarrow X$  be Leray projector.

Denote again by  $A \in L(E, E)$  the Stokes operator:

$$(Ay, w) = \sum_{i=1}^N \int_{\Omega} \nabla y_i \cdot \nabla w_i dx, \quad \forall y, w \in E.$$

Then we can reformulate problem (13) as an evolution equation

$$\frac{dy}{dt} + \nu Ay = P(q),$$

where  $A$  is symmetric, *i.e.*  $(Ay, w) = (y, Aw)$ , and strongly monotone, because  $(Ay, y) \geq \|y\|^2$ .

For solving this problem, we will apply the Fourier method developed in [6]. Here,  $E$  is a Hilbert space with respect to the energetic inner product:

$$(y, w)_E := (Ay, w), \forall y, w \in E,$$

and the embedding  $E \hookrightarrow X$  is compact. Identifying  $X$  with its dual  $X^*$ , we have

$$E \hookrightarrow X \hookrightarrow E^*.$$

Consider the duality map  $J : E \rightarrow E^*$ ,

$$\langle Jy, w \rangle := (y, w)_E, \forall y, w \in E$$

which is a linear homeomorphism with  $\|Jy\|_{E^*} = \|y\|_E, \forall y \in E$  (see [11]), and it is an extension of  $A$ , i.e.  $Jy = Ay, \forall y \in E$ . Now consider the Friedrichs extension  $B : D(B) \subseteq X \rightarrow X$  of the operator  $A$ ,

$$By := Jy, \forall y \in D(B),$$

where  $D(B) := \{y \in E \mid Jy \in X\}$ . The Friedrichs extension is maximal monotone ( $D(B)$  is dense in  $X$ ),  $B$  is closed, self-adjoint and strongly monotone, i.e.  $(By, y) \geq c\|y\|^2, \forall y \in D(B)$ , and the inverse operator  $B^{-1} : X \rightarrow X$  is linear, continuous, self-adjoint and compact. Because the embedding  $E \hookrightarrow X$  is compact, we have the following result (see [4]).

**Theorem 4.** *There exist the sequences  $\{e_n\} \subset E$  and  $\{\lambda_n\} \subset (0, +\infty)$  that are eigensolutions of  $B$ , i.e.,*

$$(Be_n, w) = \lambda_n (e_n, w), \forall w \in X, n \in \mathbb{N}$$

such that

- 1°.  $\{e_n\}$  is an orthonormal basis in  $E$ ;
- 2°.  $\{\sqrt{\lambda_n}e_n\}$  is an orthonormal basis in  $X$ ;
- 3°.  $\{\lambda_n e_n\}$  is an orthonormal basis in  $E^*$ ;
- 4°.  $\{\lambda_n\}$  is a divergent sequence increasing to  $+\infty$ .

Consider the Cauchy problem in  $X$ :

$$\begin{cases} y'(t) + By(t) = P(q)(t), & t \in (0, T) \\ y(0) = y_0, \end{cases} \quad (14)$$

where  $0 < T < \infty$ , and  $y_0 \in X$ . We have also  $P(q) \in L^2((0, T); X)$ .

We will find the solution  $y(t)$  as a Fourier series in  $X$  of the form

$$y(t) = \sum_{n \geq 1} b_n(t) e_n. \quad (15)$$

Formally, we arrive to the scalar Cauchy problem:

$$\begin{cases} b'_n(t) + \lambda_n b_n(t) = P_{q,n}(t) \\ b_n(0) = y_{0n}, \end{cases} \quad (16)$$

where  $y_{0n}$  and  $P_{q,n}(t)$  are the Fourier coefficients  $y_{0n} := \langle y_0, e_n \rangle_E = \lambda_n(y_0, e_n)$ ,  $P_{q,n}(t) := \langle P(q)(t), e_n \rangle_E = \lambda_n(P(q)(t), e_n)$ , a.a.  $t \in (0, T)$ . After solving (16) we obtain

$$b_n(t) = y_{0n}e^{-\nu\lambda_n t} + \int_0^t P_{q,n}(s)e^{\nu\lambda_n(s-t)}ds \quad (17)$$

and it is true the result

**Proposition 1.** *The function  $y(t)$  given by (15) and (17) belongs to the space  $C((0, T); E) \cap H^1((0, T); E^*)$  and it is the unique weak solution of problem (14), namely:*

$$y(0) = y_0 \text{ and}$$

$$(y'(t), w) + (By(t), w) = (P(q)(t), w), \forall w \in E.$$

Furthermore, if  $P(q) \in H^1((0, T); X)$ , then  $y \in C((0, T); E) \cap C^1((0, T); X)$  and this is the classical solution of problem (14).

*Sketch of proof.* We have that:

$$b'_n(t) = -\nu\lambda_n y_{0n}e^{-\nu\lambda_n t} + P_{q,n}(t) + \nu\lambda_n \int_0^t P'_{q,n}(s)e^{\nu\lambda_n(s-t)}ds$$

and we prove the estimations:

$$\begin{aligned} |b_n(t)|^2 &\leq 2(y_{0n}^2 e^{-2\nu\lambda_n t} + \left( \int_0^t |P_{q,n}(s)e^{\nu\lambda_n(s-t)}| ds \right)^2) \leq \\ &\leq 2 \left( y_{0n}^2 + T \int_0^T |P_{q,n}(s)|^2 ds \right), \\ \frac{|b'_n(t)|^2}{\lambda_n^2} &\leq 3 \left( \nu^2 y_{0n}^2 + \frac{|P_{q,n}(t)|^2}{\lambda_n^2} + \frac{\nu^2 T}{\lambda_1^2} \int_0^T |P'_{q,n}(t)|^2 ds \right). \end{aligned}$$

One can show the convergence of the corresponding Fourier series by using the Cauchy's uniformly convergence criteria.  $\square$

**Remark 2.** *With slight modifications, these results are also true in the case when  $\Omega$  is the complementary set of a compact one in  $\mathbb{R}^N$ .*

#### 4. Conclusions

We have developed a mathematical theory writing the steady state flow of incompressible fluids as an eigenvalue problem in abstract spaces, and we deduce not only the existence of solutions, but also the branch structure of solutions and the bifurcation points, which is perfectly concordant with the physical meaning of the problem. Also, using Fourier method for Stokes equation we deduce not only the existence and uniqueness of the weak solution, but we determine explicitly this solution and its properties.

## References

- [1] P. Constantin, C. Foias, *Navier-Stokes Equations*, Chicago Lectures in Mathematics series, 1988.
- [2] D. Fujiwara, H. Morimoto, *An  $L_r$ -theorem of the Helmholtz decomposition vector fields*, J. Fac. Sci. Univ. Tokio, Sect. IA Math., **(24)** (1977), pp. 685–700.
- [3] C. Mortici, S. Sburlan, *A Coincidence Degree for Bifurcation Problems*, Nonlinear Analysis **(53)**, 2003, pp. 715–721.
- [4] S. Sburlan, G. Moroşanu, *Monotonicity Methods for PDE's*, MB-11 / PAMM, Budapest, 1999.
- [5] C. Sburlan, S. Sburlan, *Topological Degree Approach to Steady State Flow*, in Analysis and Optimization of Differential Systems, Ed. Kluwer Academic Publishers, 2002, pp. 369–374.
- [6] S. Sburlan, C. Sburlan, *Fourier Method for Evolution Problems*, PAMM, Budapest, 2002.
- [7] S. Sburlan, *Gradul topologic*, Editura Academiei Române, Bucureşti, 1983.
- [8] H. Sohr, *The Navier-Stokes Equations. An Elementary Functional Analytic Approach*, Birkhauser Verlag, 2001.
- [9] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland Publishing Company, 1977.
- [10] W. von Wahl, *The Equations of Navier-Stokes and Abstract Parabolic Equations*, Friedr. Vieweg&Sohn, Braunschweig/Wiesbaden, 1985.
- [11] E. Zeidler, *Applied Functional Analysis*, Springer-Verlag, Berlin, 1995.

## **The Influence of Initial Fields on the Propagation of Attenuated Waves along an Edge of a Cubic Crystal**

**Olivian Simionescu-Panait\***

This paper investigates the conditions of propagation of attenuated waves in cubic crystals subject to initial deformation and electric fields. The analysis is extended to all symmetry classes belonging to the cubic system, exhibiting, or not, the piezoelectric effect. We derive the velocities of propagation and the attenuation coefficients in closed-form, and we analyze the influence of the initial fields on the wave polarization in the case of propagation along a cube edge. In this particular case, for special choices of the initial electric field, we derive approximate expressions for the solutions. We perform a parametric study on the influence of the initial mechanical and electric fields on wave velocities and attenuation coefficients.

### **1. Introduction**

Last years, problems related to electroelastic materials subject to incremental fields superposed on initial mechanical and electric fields have attracted considerable attention, due their complexity and to multiple applications (see, for example, [2] and [18]). The basic equations of the theory of piezoelectric bodies subject to infinitesimal deformations and fields superposed on initial mechanical and electric fields, were established by Eringen and Maugin in the well-known monograph [4]. In [17] E. Soós obtained the governing equations in an alternate and simpler way.

In paper [1] the fundamental equations for piezoelectric crystals subject to initial fields have been established and important results concerning the dynamic and static local stability conditions of such media were obtained. In particular, the problem of plane wave propagation in piezoelectric crystals subject to initial fields was

---

\* “Gheorghe Mihoc–Caius Iacob” Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Calea 13 Septembrie, No. 13, 050711 Bucharest, Romania, e-mail: o\_simionescu@yahoo.com

considered. In order to clarify the complicated aspects regarding the influence of the initial fields on plane wave propagation in piezoelectric crystals, for various symmetry classes, Soós and Simionescu studied in [15] the case of 6-mm type crystals. This case is important, due to its complexity from theoretical point of view, and its practical applications. In [8, 9] we obtained for 6-mm type crystals the plane wave velocities in closed form, analyzed the directions of polarization, defined new electro-mechanical coupling coefficients, and demonstrated the influence of initial fields on the shape of slowness surfaces. In [10] we studied the electrostrictive effect on plane wave propagation in isotropic solids subject to initial fields. In [11] we investigated the conditions of propagation of plane waves in cubic crystals subject to initial deformations and electric fields. We developed the previous results studying the problem of attenuated wave propagation in isotropic solids and cubic crystals subject to initial electro-mechanical fields (see [12, 13, 14]). Our results generalize, in a significant manner, those presented in classical works such as [5, 7, 16].

In this paper we investigate the conditions for attenuated wave propagation in cubic crystals subject to initial deformation and electric fields. The crystal is supposed to be a linear elastic dielectric, the analysis being extended to all the symmetry classes belonging to the cubic system, with or without the piezoelectric effect. We derive the velocities of propagation and the coefficients of attenuation in closed-form, and analyze the influence of the initial fields on the wave polarization in the case of propagation along a cube edge. In this particular case, supposing special forms of the initial electric field, we obtain approximate expressions of the obtained solutions in order to compare them with classical solutions. Finally, we perform a parametric study on the influence of the initial mechanical and electric fields on wave velocities and attenuation coefficients.

## 2. Fundamental equations

The basic equations of piezoelectric bodies for infinitesimal deformations and fields superposed on initial deformations and electric fields were given by Eringen and Maugin in their monograph [4]. An alternate derivation of these equations was obtained by Baesu, Fortuné and Soós in [1].

We assume the material to be an elastic dielectric, which is nonmagnetizable and conducts neither heat, nor electricity. We shall use the quasi-electrostatic approximation of the equations of balance. Furthermore, we assume that the elastic dielectric is linear and homogeneous, that the initial homogeneous deformations are infinitesimal and that the initial homogeneous electric field has small intensity. To describe this situation we use three different configurations : the *reference configuration*  $B_R$  in which at time  $t = 0$  the body is undeformed and free of all fields; the *initial configuration*  $\overset{\circ}{B}$  in which the body is deformed statically and carries the initial fields; the *present (current) configuration*  $B_t$  obtained from  $\overset{\circ}{B}$  by applying time dependent incremental deformations and fields. In what follows, all the fields related to the initial configuration  $\overset{\circ}{B}$  will be denoted by a superposed “o”.



In this case the *homogeneous field equations* take the following form:

$$\begin{aligned} \overset{\circ}{\rho} \ddot{\mathbf{u}} &= \text{div } \boldsymbol{\Sigma}, \text{div } \boldsymbol{\Delta} = 0, \\ \text{rot } \mathbf{e} &= 0 \Leftrightarrow \mathbf{e} = -\text{grad } \varphi, \end{aligned} \quad (1)$$

where  $\overset{\circ}{\rho}$  is the mass density,  $\mathbf{u}$  is the incremental displacement from  $\overset{\circ}{B}$  to  $B_t$ ,  $\boldsymbol{\Sigma}$  is the incremental mechanical nominal stress tensor,  $\boldsymbol{\Delta}$  is the incremental electric displacement vector,  $\mathbf{e}$  is the incremental electric field and  $\varphi$  is the incremental electric potential. All incremental fields involved into the above equations depend on the spatial variable  $\mathbf{x}$  and on the time  $t$ .

We have the following incremental constitutive equations:

$$\begin{aligned} \Sigma_{kl} &= \overset{\circ}{\Omega}_{klmn} u_{m,n} + \overset{\circ}{\Lambda}_{mkl} \varphi_{,m} + d_{klmn} \dot{u}_{m,n}, \\ \Delta_k &= \overset{\circ}{\Lambda}_{kmn} u_{n,m} + \overset{\circ}{\epsilon}_{kl} e_l = \overset{\circ}{\Lambda}_{kmn} u_{n,m} - \overset{\circ}{\epsilon}_{kl} \varphi_{,l}. \end{aligned} \quad (2)$$

In these equations  $\overset{\circ}{\Omega}_{klmn}$  are the components of the instantaneous elasticity tensor,  $d_{klmn}$  are the components of attenuation (damping) tensor,  $\overset{\circ}{\Lambda}_{kmn}$  are the components of the instantaneous coupling tensor and  $\overset{\circ}{\epsilon}_{kl}$  are the components of the instantaneous dielectric tensor. The instantaneous coefficients can be expressed in terms of the classical moduli of the material and on the initial applied fields as follows:

$$\begin{aligned} \overset{\circ}{\Omega}_{klmn} &= \overset{\circ}{\Omega}_{nmlk} = c_{klmn} + \overset{\circ}{S}_{kn} \delta_{lm} - e_{kmn} \overset{\circ}{E}_l - e_{nkl} \overset{\circ}{E}_m - \eta_{kn} \overset{\circ}{E}_l \overset{\circ}{E}_m, \\ \overset{\circ}{\Lambda}_{mkl} &= e_{mkl} + \eta_{mk} \overset{\circ}{E}_l, \quad \overset{\circ}{\epsilon}_{kl} = \overset{\circ}{\epsilon}_{lk} = \epsilon_{kl} = \delta_{kl} + \eta_{kl}, \end{aligned} \quad (3)$$

where  $c_{klmn}$  are the components of the constant elasticity tensor,  $e_{kmn}$  are the components of the constant piezoelectric tensor,  $\epsilon_{kl}$  are the components of the constant dielectric tensor,  $\overset{\circ}{E}_i$  are the components of the initial applied electric field and  $\overset{\circ}{S}_{kn}$  are the components of the initial applied symmetric (Cauchy) stress tensor.

It is important to observe that the previous material moduli have the following symmetry properties:

$$\begin{aligned} c_{klmn} &= c_{lkmn} = c_{klnm} = c_{mnkl}, \quad e_{mkl} = e_{mlk}, \\ d_{klmn} &= d_{lkmn} = d_{klnm} = d_{mnkl}, \quad \epsilon_{kl} = \epsilon_{lk}. \end{aligned}$$

Hence, in general there are 21 independent elastic coefficients  $c_{klmn}$ , as well as 21 independent attenuation components  $d_{klmn}$ , 18 independent piezoelectric coefficients  $e_{klm}$  and 6 independent dielectric coefficients  $\epsilon_{kl}$ . From the relations (3) we see that  $\overset{\circ}{\Omega}_{klmn}$  is not symmetric in indices  $(k, l)$  and  $(m, n)$  and  $\overset{\circ}{\Lambda}_{mkl}$  is not symmetric in indices  $(k, l)$ . It follows that, generally, there are 45 independent instantaneous

elastic moduli  $\overset{\circ}{\Omega}_{klmn}$ , 27 independent instantaneous coupling moduli  $\overset{\circ}{\Lambda}_{mkl}$  and 6 independent instantaneous dielectric moduli  $\overset{\circ}{\epsilon}_{kl}$ .

The main goal of this work is to study the conditions for propagation of incremental mechanical attenuated waves in an unbounded three dimensional material described by the previous constitutive equations. Therefore, we suppose that the displacement vector and the electric potential have the following form:

$$\begin{aligned} \mathbf{u}(\mathbf{x}, t) &= \mathbf{a} \exp(-\boldsymbol{\alpha} \cdot \mathbf{x}) \exp[i(\omega t - \mathbf{p} \cdot \mathbf{x})], \\ \varphi(\mathbf{x}, t) &= \bar{a} \exp[i(\omega t - \mathbf{p} \cdot \mathbf{x})]. \end{aligned} \quad (4)$$

Here  $\mathbf{a}$  and  $\bar{a}$  are constants, characterizing the *amplitude* of the wave,  $\mathbf{p} = p \mathbf{n}$  (with  $\mathbf{n}^2 = 1$ ) is a constant vector,  $p$  representing the *wave number* and  $\mathbf{n}$  denoting the *direction of propagation* of the wave,  $\boldsymbol{\alpha} = \alpha \mathbf{n}$  (with  $\alpha$  defining the *attenuation coefficient*). Here  $\omega$  is the *frequency* of the wave. The *velocity of propagation* of the wave is defined by  $v = \omega/p$ . The validity of the hypothesis saying that the direction of propagation coincides with the direction of attenuation was analyzed in monograph [6].

Introducing these forms of  $\mathbf{u}$  and  $\varphi$  into the field equations (1) and taking into account the constitutive equations (2), (3) we obtain the *condition of propagation* of attenuated waves:

$$\overset{\circ}{Q} \mathbf{a} = \overset{\circ}{\rho} \omega^2 \mathbf{a} \quad (5)$$

with the components of the *instantaneous acoustic tensor*  $\overset{\circ}{Q}$  having the following form:

$$\begin{aligned} \overset{\circ}{Q}_{lm} = & \overset{\circ}{\Omega}_{klmn} (p_k - i\alpha_k)(p_n - i\alpha_n) + \frac{(\overset{\circ}{\Lambda}_{uvl} p_u p_v) [\overset{\circ}{\Lambda}_{rsm} (p_r - i\alpha_r)(p_s - i\alpha_s)]}{\overset{\circ}{\epsilon}_{ij} p_i p_j} + \\ & + i\omega d_{klmn} (p_k - i\alpha_k)(p_n - i\alpha_n). \end{aligned} \quad (6)$$

It is evident, that for the problem of attenuated wave propagation these components are *complex numbers* and that the tensor  $\overset{\circ}{Q}$  is *not symmetric*. Consequently, the arguments used in [1] and [3] to derive the condition of propagation, supposing the symmetry and the positive definiteness of the acoustic tensor, are no longer valid here, for the general formulation.

In this paper we deal with the problem of propagation of attenuated waves along an edge of a cubic crystal subject to initial electro-mechanical fields. For particular directions of propagation and attenuation we shall obtain the *phase velocities*, the *attenuation coefficients* and we shall study the corresponding *polarization*.

### 3. Attenuated wave propagation along an edge of a cubic crystal subject to initial electro-mechanical fields

It is known that, in the case of a cubic crystal, the *elasticity tensor* contains three independent constants (see [16], or [7]). Using Voigt's convention we have:

$$\mathbf{c} = \begin{pmatrix} c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{12} & 0 & 0 & 0 \\ c_{12} & c_{12} & c_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{44} \end{pmatrix}. \quad (7)$$

Among the five symmetry classes belonging to the cubic system, only  $\bar{4}3m$  and 23 classes exhibit the piezoelectric effect, for the others (i.e.  $m\bar{3}m, 432, m\bar{3}$ ) the piezoelectric effect is absent.

Similarly, the *attenuation tensor* possesses three independent coefficients:

$$\mathbf{d} = \begin{pmatrix} d_{11} & d_{12} & d_{12} & 0 & 0 & 0 \\ d_{12} & d_{11} & d_{12} & 0 & 0 & 0 \\ d_{12} & d_{12} & d_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & d_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{44} \end{pmatrix}. \quad (8)$$

In case of symmetry classes  $\bar{4}3m$  and 23, the *piezoelectric tensor* contains one constant:

$$\mathbf{e} = \begin{pmatrix} 0 & 0 & 0 & e_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & e_{14} \end{pmatrix}, \quad (9)$$

while the *dielectric tensor* has one constant, for all five symmetry classes:

$$\boldsymbol{\eta} = \begin{pmatrix} \eta & 0 & 0 \\ 0 & \eta & 0 \\ 0 & 0 & \eta \end{pmatrix}. \quad (10)$$

To study the attenuated wave propagation along the  $[001]$  axis, we shall assume that the direction of propagation coincides with  $x_3$  axis (i.e.  $n_3 = 1, n_1 = n_2 = 0$ ).

It follows that the acoustic tensor  $\overset{\circ}{\mathbf{Q}}$  has the following components:

$$\overset{\circ}{Q}_{lm} = [\overset{\circ}{\Omega}_{3lm3} + \frac{\overset{\circ}{\Lambda}_{33l}\overset{\circ}{\Lambda}_{33m}}{\overset{\circ}{\epsilon}_{33}} + i\omega d_{3lm3}](p - i\alpha)^2 = \overset{\circ}{Q}'_{lm} (p - i\alpha)^2. \quad (11)$$

If we denote by  $V = \frac{\overset{\circ}{\rho} \omega^2}{(p - i\alpha)^2}$ , the condition of propagation (5) will take the form of an *eigenvector problem*:

$$\overset{\circ}{\mathbf{Q}}' \mathbf{a} = V \mathbf{a}, \quad (12)$$

or:

$$(\overset{\circ}{Q}'_{lm} - V\delta_{lm})a_m = 0, \quad l = \overline{1,3}. \quad (13)$$

This problem is usually associated to the following *eigenvalue problem* (*characteristic equation*):

$$\det(\overset{\circ}{Q}'_{lm} - V\delta_{lm}) = 0. \quad (14)$$

Note that, for this particular direction of propagation the acoustic tensor becomes symmetric,  $\overset{\circ}{Q}'_{lm} = \overset{\circ}{Q}'_{ml}$ .

Thus we obtain the components of the tensor  $\overset{\circ}{Q}'$  in the form:

$$\begin{aligned} \overset{\circ}{Q}'_{11} = a &= c_{44} + i\omega d_{44} + \overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_1^2, \quad \overset{\circ}{Q}'_{12} = \overset{\circ}{Q}'_{21} = b = -\frac{\eta}{1+\eta} \overset{\circ}{E}_1 \overset{\circ}{E}_2, \\ \overset{\circ}{Q}'_{13} = \overset{\circ}{Q}'_{31} = c &= -\frac{\eta}{1+\eta} \overset{\circ}{E}_1 \overset{\circ}{E}_3, \quad \overset{\circ}{Q}'_{22} = d = c_{44} + i\omega d_{44} + \overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_2^2, \\ \overset{\circ}{Q}'_{23} = \overset{\circ}{Q}'_{32} = e &= -\frac{\eta}{1+\eta} \overset{\circ}{E}_2 \overset{\circ}{E}_3, \quad \overset{\circ}{Q}'_{33} = f = c_{11} + i\omega d_{11} + \overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_3^2. \end{aligned} \quad (15)$$

From the analysis of the form of the previous coefficients, we can easily observe that the piezoelectric effect is absent, for this direction of propagation, even if the crystal is piezoelectric active.

With this notation, the condition of propagation (13) becomes:

$$\begin{cases} a a_1 + b a_2 + c a_3 = V a_1 \\ b a_1 + d a_2 + e a_3 = V a_2 \\ c a_1 + e a_2 + f a_3 = V a_3, \end{cases} \quad (16)$$

while, the characteristic equation (14) has the form:

$$F(V) = \begin{vmatrix} a - V & b & c \\ b & d - V & e \\ c & e & f - V \end{vmatrix} = 0. \quad (17)$$

#### 4. Analysis of particular cases

In order to obtain the phase velocities and the attenuation coefficients in closed form, in what follows we shall present two important particular cases:

##### 4.1. Longitudinal initial electric field ( $\overset{\circ}{E}_1 = \overset{\circ}{E}_2 = 0, \overset{\circ}{E}_3 \neq 0$ )

This case can be defined as an *electro-acoustic Pockels effect* (see [4] for the analogous electro-optical effect). Here, the expressions  $b, c, e$  being zero, the charac-

teristic equation (17) has the following roots:

$$\begin{aligned} V_1 = V_2 &= c_{44} + i\omega d_{44} + \overset{\circ}{S}_{33}, \\ V_3 &= c_{11} + i\omega d_{11} + \overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_3^2. \end{aligned} \quad (18)$$

As regards the *polarization* of the obtained waves, using the condition of propagation (16) in this particular case, we can easily see that  $V_3$  corresponds to a *longitudinal wave* with electrostrictive effect, while  $V_1 = V_2$  are linked to *transverse waves*, arbitrarily polarized.

To find the *phase velocities* and *attenuation coefficients* related to the previous roots, we shall denote by:

$$V_3 = A_L + iB_L = \frac{\overset{\circ}{\rho} \omega^2}{(p_L - i\alpha_L)^2}, \quad (19)$$

$$A_L = c_{11} + \overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_3^2, \quad B_L = \omega d_{11}.$$

It yields a phase velocity  $v_L$ , in the form:

$$v_L^2 = \frac{\omega^2}{p_L^2} = \frac{2(A_L^2 + B_L^2)}{\overset{\circ}{\rho} (\sqrt{A_L^2 + B_L^2} + A_L)}, \quad (20)$$

and an attenuation coefficient  $\alpha_L$ , given by the relation:

$$\alpha_L^2 = \frac{\overset{\circ}{\rho} \omega^2}{2} \cdot \frac{\sqrt{A_L^2 + B_L^2} - A_L}{A_L^2 + B_L^2}. \quad (21)$$

We can conclude that the displacement vector, in this particular case, has the form  $\mathbf{u}_L = (0, 0, u_3^L)$ , with:

$$u_3^L(x_3, t) = a_3 \exp(-\alpha_L x_3) \exp \left[ i\omega \left( t - \frac{x_3}{v_L} \right) \right]. \quad (22)$$

We easily observe that the attenuation affects the phase velocity  $v_L$  (by  $\omega$ ), and the amplitude of the longitudinal wave (by  $\alpha_L$ ). Moreover, the *electrostrictive effect* is represented by the term  $-\frac{\eta}{1+\eta} \overset{\circ}{E}_3^2$ .

To obtain an *approximate solution* of this problem, we shall denote by  $\bar{\epsilon} = \omega d_{11}/c_{11}$  a non-dimensional number. Supposing that  $\bar{\epsilon} \ll 1$ , we shall approximate the expression (20) and (21) for phase velocity and attenuation coefficient, to first order in  $\bar{\epsilon}$ .

Neglecting the terms containing powers of order greater than  $\bar{\epsilon}$ , we derive the approximate form of the phase velocity:

$$v_L \simeq v_L^* \sqrt{1 + \psi}, \quad v_L^* = \sqrt{c_{11}/\overset{\circ}{\rho}}, \quad \psi = \frac{\overset{\circ}{S}_{33} - \frac{\eta}{1+\eta} \overset{\circ}{E}_3^2}{c_{11}}. \quad (23)$$

Here  $v_L^*$  is the longitudinal velocity in the classical case, without initial fields, and  $\psi$  is a non-dimensional number describing the influence of the initial fields.

In a similar way, we can derive an approximate form of the attenuation coefficient:

$$\alpha_L \simeq \alpha_L^* \cdot \frac{1}{(1+\psi)^{3/2}}, \quad \alpha_L^* = \frac{\tau \omega^2}{2v_L^*}, \quad \tau = d_{11}/c_{11}. \quad (24)$$

Here  $\alpha_L^*$  is the attenuation coefficient in the case without initial fields, as defined in [7].

Applying the same procedure, as in the case of the longitudinal wave, we find the phase velocity and attenuation coefficients for the transverse waves. So, using the notation:

$$V_1 = V_2 = A_T + iB_T = \frac{\overset{\circ}{\rho} \omega^2}{(p_T - i\alpha_T)^2}, \quad (25)$$

$$A_T = c_{44} + \overset{\circ}{S}_{33}, \quad B_T = \omega d_{44},$$

we obtain the phase velocity  $v_T$  in the form:

$$v_T^2 = \frac{\omega^2}{p_T^2} = \frac{2(A_T^2 + B_T^2)}{\overset{\circ}{\rho} (\sqrt{A_T^2 + B_T^2} + A_T)}, \quad (26)$$

and the attenuation coefficient  $\alpha_T$ , as:

$$\alpha_T^2 = \frac{\overset{\circ}{\rho} \omega^2}{2} \cdot \frac{\sqrt{A_T^2 + B_T^2} - A_T}{A_T^2 + B_T^2}. \quad (27)$$

We can conclude that the displacement vector, in this case, has the form  $\mathbf{u}_T = (u_1^T, u_2^T, 0)$ , with:

$$u_k^T(x_3, t) = a_k \exp(-\alpha_T x_3) \exp \left[ i\omega \left( t - \frac{x_3}{v_T} \right) \right], \quad k = 1; 2. \quad (28)$$

We observe that the phase velocity  $v_T$  depends on  $\omega$  and the amplitude is affected by  $\alpha_T$ . Similar approximate forms for the phase velocity and attenuation coefficient can be obtained in this case, too.

#### 4.2. Transverse initial electric field ( $\overset{\circ}{E}_1 \neq 0, \overset{\circ}{E}_2 \neq 0, \overset{\circ}{E}_3 = 0$ )

This case can be defined as an *electro-acoustic Kerr effect* (see [4] for the analogous electro-optical effect).

In this case, the coefficients  $c$  and  $e$  being zero, the characteristic equation (17) has the following three roots:

$$V_1' = c_{44} + \overset{\circ}{S}_{33} + i\omega d_{44}, \quad V_2' = c_{44} + \overset{\circ}{S}_{33} + i\omega d_{44} - \frac{\eta}{1+\eta} (\overset{\circ}{E}_1^2 + \overset{\circ}{E}_2^2), \quad (29)$$

$$V_3' = c_{11} + \overset{\circ}{S}_{33} + i\omega d_{11}.$$

As regards the *polarization* of the obtained waves, using the condition of propagation (16) in this particular case, we can easily see that  $V'_3$  corresponds to a *longitudinal wave*, while  $V'_1$  is linked to a *transverse wave*, whose polarization direction is fixed by the initial electric field. Indeed, in this case, the system (16) reduces to the equation  $\overset{\circ}{E}_1 a_1 + \overset{\circ}{E}_2 a_2 = 0$ .

$V'_2$  corresponds to a *transverse wave*, with a direction of polarization fixed by the initial electric field, given by the equation  $\overset{\circ}{E}_2 a_1 - \overset{\circ}{E}_1 a_2 = 0$ , normal to the preceding direction.

To find the *phase velocities* and *attenuation coefficients* related to the previous roots, we shall proceed as in the case with longitudinal initial electric field. So, using the notation:

$$V'_3 = A'_L + iB'_L = \frac{\overset{\circ}{\rho} \omega^2}{(p_L - i\alpha_L)^2}, \quad A'_L = c_{11} + \overset{\circ}{S}_{33}, \quad B'_L = \omega d_{11}, \quad (30)$$

we obtain a phase velocity  $v_L$  in the form:

$$v_L^2 = \frac{\omega^2}{p_L^2} = \frac{2(A_L'^2 + B_L'^2)}{\overset{\circ}{\rho} (\sqrt{A_L'^2 + B_L'^2} + A'_L)}, \quad (31)$$

and an attenuation coefficient  $\alpha_L$ :

$$\alpha_L^2 = \frac{\overset{\circ}{\rho} \omega^2}{2} \cdot \frac{\sqrt{A_L'^2 + B_L'^2} - A'_L}{A_L'^2 + B_L'^2}. \quad (32)$$

We conclude that the displacement vector, in this case, has the form  $\mathbf{u}_L = (0, 0, u_3^L)$ , with:

$$u_3^L(x_3, t) = a_3 \exp(-\alpha_L x_3) \exp \left[ i\omega \left( t - \frac{x_3}{v_L} \right) \right]. \quad (33)$$

We observe that the attenuation affects the phase velocity  $v_L$  (by  $\omega$ ), and the amplitude of the longitudinal wave (by  $\alpha_L$ ). In this case, the electrostrictive effect is absent.

Applying the same procedure, we find the phase velocity and attenuation coefficient for the transverse waves. So, letting:

$$V'_1 = A_{T_1} + iB_{T_1} = \frac{\overset{\circ}{\rho} \omega^2}{(p_{T_1} - i\alpha_{T_1})^2}, \quad A_{T_1} = c_{44} + \overset{\circ}{S}_{33}, \quad B_{T_1} = \omega d_{44}, \quad (34)$$

we obtain the phase velocity  $v_{T_1}$  in the form:

$$v_{T_1}^2 = \frac{\omega^2}{p_{T_1}^2} = \frac{2(A_{T_1}^2 + B_{T_1}^2)}{\overset{\circ}{\rho} (\sqrt{A_{T_1}^2 + B_{T_1}^2} + A_{T_1})}, \quad (35)$$

and attenuation coefficient  $\alpha_{T_1}$ :

$$\alpha_{T_1}^2 = \frac{\overset{\circ}{\rho} \omega^2}{2} \cdot \frac{\sqrt{A_{T_1}^2 + B_{T_1}^2} - A_{T_1}}{A_{T_1}^2 + B_{T_1}^2}. \quad (36)$$

We can see that the displacement vector, in this case, has the form  $\mathbf{u}_{T_1} = (u_1^{T_1}, u_2^{T_1}, 0)$ , where:

$$u_k^{T_1}(x_3, t) = a_k \exp(-\alpha_{T_1} x_3) \exp \left[ i\omega \left( t - \frac{x_3}{v_{T_1}} \right) \right], \quad k = 1; 2. \quad (37)$$

Similar approximate forms for the phase velocity and attenuation coefficient can be obtained in this case. This transverse wave is attenuated, has the polarization fixed by the initial electric field, and is not affected by the electrostrictive effect.

Finally, on letting:

$$V_2' = A_{T_2} + iB_{T_2} = \frac{\overset{\circ}{\rho} \omega^2}{(p_{T_2} - i\alpha_{T_2})^2}, \quad (38)$$

$$A_{T_2} = c_{44} + \overset{\circ}{S}_{33} - \frac{\eta}{1 + \eta} (\overset{\circ}{E}_1^2 + \overset{\circ}{E}_2^2), \quad B_{T_2} = \omega d_{44},$$

we obtain the phase velocity  $v_{T_2}$  in the form:

$$v_{T_2}^2 = \frac{\omega^2}{p_{T_2}^2} = \frac{2(A_{T_2}^2 + B_{T_2}^2)}{\overset{\circ}{\rho} (\sqrt{A_{T_2}^2 + B_{T_2}^2} + A_{T_2})}, \quad (39)$$

and an attenuation coefficient  $\alpha_{T_2}$ :

$$\alpha_{T_2}^2 = \frac{\overset{\circ}{\rho} \omega^2}{2} \cdot \frac{\sqrt{A_{T_2}^2 + B_{T_2}^2} - A_{T_2}}{A_{T_2}^2 + B_{T_2}^2}. \quad (40)$$

We conclude that the displacement vector, in this case, has the form  $\mathbf{u}_{T_2} = (u_1^{T_2}, u_2^{T_2}, 0)$ , with:

$$u_k^{T_2}(x_3, t) = a_k \exp(-\alpha_{T_2} x_3) \exp \left[ i\omega \left( t - \frac{x_3}{v_{T_2}} \right) \right], \quad k = 1; 2. \quad (41)$$

Similar approximate forms for the phase velocity and attenuation coefficient can be obtained in this case. This transverse wave is attenuated, has the polarization fixed by the initial electric field, and is affected by the electrostrictive effect.

### 4.3. Parametric study

Here we analyze the influence of the initial mechanical and electric fields on wave velocities and attenuation coefficients, corresponding to a longitudinal initial electric field.



In Table 1 we compute the rapport between longitudinal wave velocities in the case without initial fields, and with initial strain fields of order 1%, 2% and 5%, resp. of attenuation coefficients (see formulae (23) and (24)). The influence of initial electric field on wave velocities and attenuation coefficients is very weak, even if its intensity is important:  $\overset{\circ}{E}_1 = 10^3 \sqrt{Pa} = 10^8 \text{ V/m}$ . The superior value corresponds to a traction stress  $\overset{\circ}{S}_{11}$ , while the inferior value is related to a compression stress  $\overset{\circ}{S}_{11}$ , that generate the initial strain fields. One can observe important differences between these values, due to the initial strain fields.

*Table 1*  
The influence of initial strain fields on wave velocities and attenuation coefficients for longitudinal initial electric field

	0% initial strain field	1%	2%	5%
$v_L/v_L^*$	1	1.005/0.995	1.010/0.990	1.025/0.975
$\alpha_L/\alpha_L^*$	1	0.985/1.015	0.971/1.031	0.929/1.080

## References

- [1] E. Baesu, D. Fortuné and E. Soós, *Incremental behaviour of hyperelastic dielectrics and piezoelectric crystals*, J. Appl. Math. Phys. (ZAMP), **54**, 160–178 (2003).
- [2] J.C. Baumhauer and H.F. Tiersten, *Nonlinear electrostatics equations for small fields superimposed on a bias*, J. Acoust. Soc. Amer., **54**, 1017–1034 (1973).
- [3] N.D. Cristescu, E.M. Crăciun and E. Soós, *Mechanics of elastic composites*, Chapman & Hall/CRC, Boca Raton (2004).
- [4] A.C. Eringen and G.A. Maugin, *Electrodynamics of continua*, vol. I, Springer, New York (1990).
- [5] F.I. Fedorov, *Theory of elastic waves in crystals*, Plenum Press, New York (1968).
- [6] L. Landau and E. Lifchitz, *Électrodynamique des milieux continus*, MIR, Moscou (1969).
- [7] D. Royer and E. Dieulesaint, *Elastic waves in solids*, vol. I – *Free and guided propagation*, Springer, Berlin (2000).
- [8] O. Simionescu-Panait, *The influence of initial fields on wave propagation in piezoelectric crystals*, Int. J. Appl. Electromagnetics and Mechanics, **12**, 241–252 (2000).
- [9] O. Simionescu-Panait, *Progressive wave propagation in the meridian plane of a 6mm-type piezoelectric crystal subject to initial fields*, Math. and Mech. Solids, **6**, 661–670 (2001).

- [10] O. Simionescu-Panait, *The electrostrictive effect on wave propagation in isotropic solids subject to initial fields*, Mech. Res. Comm., **28**, 685–691 (2001).
- [11] O. Simionescu-Panait, *Wave propagation in cubic crystals subject to initial mechanical and electric fields*, J. Appl. Math. Phys. (ZAMP), **53**, 1038–1051 (2002).
- [12] O. Simionescu-Panait, *Propagation of attenuated waves in isotropic solids subject to initial electro-mechanical fields*, Proc. of Symp. *New Trends in Continuum Mechanics*, 267–275, Ed. Theta, Bucharest (2005) (in press).
- [13] O. Simionescu Panait, *Propagation of attenuated waves along an edge of a cubic crystal subject to initial electro-mechanical fields*, Math. and Mech. Solids, **10** (2005) (in press).
- [14] O. Simionescu-Panait, *Attenuated wave propagation on a face of a cubic crystal subject to initial electro-mechanical fields*, Int. J. Appl. Electromagnetics and Mechanics (2005) (in press).
- [15] O. Simionescu and E. Soós, *Wave propagation in piezoelectric crystals subjected to initial deformations and electric fields*, Math. and Mech. Solids, **6**, 437–446 (2001).
- [16] I.I. Sirotnin and M.P. Shaskolskaya, *Crystal physics*, Nauka, Moscow (1975) (in Russian).
- [17] E. Soós, *Stability, resonance and stress concentration in prestressed piezoelectric crystals containing a crack*, Int. J. Engn. Sci., **34**, 1647–1673 (1996).
- [18] H.F. Tiersten, *On the accurate description of piezoelectric resonators subject to biasing deformations*, Int. J. Engn. Sci., **33**, 2239–2259 (1995).

## Model for molecular dynamics simulation of radiation-induced defect formation in fcc metals

Daniel Șopu<sup>\*‡</sup>, Bogdan Nicolescu<sup>\*\*‡</sup>, and M.A. Gîrțu<sup>\*\*\*‡</sup>

We present here a model that describes the radiation-induced defect formation during prolonged irradiation of nuclear materials using molecular dynamics simulation methods. The purpose of the simulation is to study the formation of vacancies and interstitial defects in the face centered cubic metals and to set the ground for comparing and contrasting the localization of the heat spikes and of the resulting molten region in metallic targets. The interactions between the high energy ion and the atoms in the target are modelled by the Ziegler-Biersack-Littmark potential, which gives a good fit to reasonably accurate quantum calculations of interatomic potentials in the repulsive region, while the equilibration within the system can be described by means of the Morse potential. The energetic ion (with initial kinetic energy lower than 5 keV) impinging on a dense solid produces a sequence of atomic collisions that cause structural defects and can even lead to a local melting of the crystal, followed by quenching into a strongly disordered phase.

### 1. Introduction

The degradation of the physical properties of the metal alloys used in pressure vessels in nuclear power plants as well as in metal coatings for fusion-based alternative

---

\* Department of Physics, “Ovidius” University of Constanța, Romania

\*\* Department of Mathematics, “Ovidius” University of Constanța, Romania.

\*\*\* Department of Physics, “Ovidius” University of Constanța, Romania,

e-mail: [girtu@univ-ovidius.ro](mailto:girtu@univ-ovidius.ro)

‡ Supported in part by the Romanian Ministry of Education and Research through a PNCDI-INFOSOC grant 131/20.08.2004 and a CNCSIS grant 881/2004, as well as by CopyRo – The Romanian Society for Reproduction Rights, through the research grant 1636/2004.

energy sources can be modeled at the shortest length and time scales (nanometers and picoseconds) using molecular dynamics simulations [15]. The energetic ion impinging on a dense solid produces a sequence of atomic collisions that cause structural defects, such as vacancies and interstitial defects and can even lead to a local melting of the crystal, followed by quenching into a strongly disordered phase [2].

The study of the formation of such defects in the low energy range (namely below about 5 keV), where elastic collisions dominate the slowing-down process, computer simulation methods have been commonly used [2]. Binary collision approximation methods provide a fairly efficient means for calculating ion ranges, but molecular dynamics (MD) methods, although require larger computational efforts in terms of memory and computer time, describe the interactions involved in radiation damage much more realistically [15].

The slowing down of ions in solid materials is conventionally interpreted to be due to two separate processes, electronic and nuclear slowing down (stopping) [21]. Defining the stopping power as the energy loss per unit distance, it was shown [21] that electronic slowing down dominates the stopping of the impinging ion at high ion energies (higher than roughly 1 keV/amu). However, when the ion has slowed down sufficiently, nuclear slowing down will always sooner or later becomes significant. The maximum of the nuclear stopping curve typically occurs at energies around 1 keV/amu while the minimum at about 100 keV/amu. (It should be noted, however, that for very light ions slowing down in heavy materials, the nuclear stopping is weaker than the electronic at all energies.) Therefore, at keV energies the electronic slowing down does not significantly contribute to the production of lattice defects. Defect production is chiefly caused by nuclear stopping, i.e. elastic collisions between a recoiling ion and atoms in the medium.

When an energetic projectile ion collides with a target atom in a crystal lattice and gives enough energy to it, the lattice atom will collide with other lattice atoms, resulting in a large number of successive collisions. All such atomic collision processes initiated by a single ion are called a collision cascade. A collision cascade is a complex process that can be divided into three phases [5]. The initial stage, during which atoms collide strongly, is called the collisional phase, and typically lasts about 0.1–1 ps. In the second phase, the high kinetic energy of the atoms affected by the collision processes decreases by dissipation in the crystal by means of heat conduction. This phase is called the thermal spike, and lasts roughly 1 ns. During the last phase, after the cooling down, the crystal is usually left with a large quantity of defects. The defects can vary from vacancies and interstitial atoms to complex interstitial-dislocation loops and volume defects [11, 16]. If the lattice temperature is high enough, many of these defects will relax by thermally activated migration [9]. This is the so called relaxation phase of the collision cascade.

The type of damage produced during ion irradiation may be very complex and varies a great deal for different ion types, sample materials and implantation conditions. However, some general characteristics can be identified for irradiation of metals. For instance, at a certain irradiation intensity (roughly  $10^{14}$  ions/cm<sup>2</sup> for some metals) the sample may become amorphous [16].

## 2. Model

In molecular dynamics simulations the time evolution of a system of atoms is calculated by solving the equations of motion numerically [1, 10]. Since the movement of each individual atom involved in a collisions cascade can be followed in MD simulations, they offer the most realistic way of examining defect formation during ion implantation. One of the very first uses of molecular dynamics methods was in fact simulation of collision sequences in metals [6, 8]; since then MD simulations have been used to study a large variety of phenomena in collision cascades [16].

The molecular dynamics simulation process starts by calculating the force acting on each atom in the system. The equations of motion for the system are solved using some suitable algorithm [3, 17]. The solution yields the change in the atom positions, velocities and accelerations over a finite time step  $\Delta t$ . After the atoms have been moved the simulation continues by recalculating the forces in the new positions. The atoms that are included in the calculation are usually placed in an face-centered cubic lattice within an orthogonal simulation cell.

The forces governing the simulation can be obtained from classical or quantum mechanical calculations. Although promising advances have recently been made using tight-binding molecular dynamics methods, quantum molecular dynamics methods are still far too time-consuming to allow simulation of full collision cascades [7]. Therefore classical MD simulations have to be used in the foreseeable future for descriptions of energetic collision cascades.

In classical MD simulations the interaction between atoms in the sample are described with an interatomic potential  $V(r)$ , generally assumed to depend only on the distance  $r$  between two atoms. Probably the most common choices for the interatomic potential are the Lennard-Jones [12] and Morse [14] potentials, the former more suitable for closed shell systems such as the noble gases while the latter more appropriate for metals. The expressions for these potentials are:

$$V_{LJ}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \quad (1)$$

and

$$V_M(r) = D e^{a(1-r/r_e)} (1 - e^{a(1-r/r_e)}) \quad (2)$$

for the Lennard-Jones and Morse cases, respectively. In these equations  $r$  is the distance between the atoms. Also,  $\sigma$  and  $\epsilon$  are the characteristic length and energy scales of the Lennard-Jones interaction, while  $r_e$  is the equilibrium bond distance,  $D$  is the depth of the potential energy well function, and  $a$  controls the width of the potential.

The type of interaction occurring between the projectile ion and the neutral atoms in the target depends on the energy of the incoming ion. At energies larger than 1 keV the slowing down is mostly due to electrons, while at lower energies the slowing down is mostly nuclear. Consequently, the interactions have to be considered carefully. Collisions of the impinging particle with the recoil atoms can be described by repulsive interactions, which at small distances can be regarded as essentially

Coulombic. At greater distances, the electron clouds screen the nuclei from each other and the repulsive potential can be described by multiplying the Coulombic repulsion between nuclei with a screening function,

$$V_{ZBL}(r) = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2}{r} \varphi(r), \quad (3)$$

where  $\varphi(r)$  goes to unity as the distance  $r$  between the nuclei vanishes, and  $Z_1$  and  $Z_2$  are the charges of the interacting nuclei.

A large number of different repulsive potentials and screening functions have been proposed over the years, some determined semi-empirically, others from theoretical calculations. A much used repulsive potential is the one given by Ziegler, Biersack and Littmark (ZBL) [21]. It has been constructed by fitting a universal screening function to theoretically obtained potentials calculated for a large variety of atom pairs. The ZBL screening function has the form

$$\varphi(r) = 0.1818e^{-3.2r/\bar{r}} + 0.5099e^{-0.9423r/\bar{r}} + 0.2802e^{-0.4029r/\bar{r}} + 0.02817e^{-0.2016r/\bar{r}}, \quad (4)$$

where

$$\bar{r} = \frac{0.8854}{Z_1^{0.23} + Z_2^{0.23}} a_B \quad (5)$$

( $a_B$  being the Bohr atomic radius, equal to 0.529 Å). It can be verified that at small distances  $\varphi(r)$  goes to unity, while as the distance increases the screening is more effective and  $\varphi(r)$  vanishes.

### 3. Simulation-algorithm, parameters and initial state preparation

In molecular dynamic simulations the time evolution of a system of atoms is calculated by solving the equations of motion numerically. In the Newtonian formalism the force acting on an atom is calculated based on the sum of the contributions of all other particles, taking into account the gradient of the interatomic potential [1]. After the force calculation, the equations of motion for the system are solved using an integration algorithm, to provide the new positions and velocities for each particle at the next moment. The time step,  $\Delta t$ , is appropriately chosen to optimize a compromise between accuracy and duration of the simulation. The process is repeated by calculating the forces in the new positions and integrating to find the new positions and velocities.

Our simulation algorithm is based on Newtons classical laws of motion. The equations of motion for the typical equilibrium simulation, performed at constant energy (in the microcanonical ensemble), are given simply by Newtons second law. The most common integration algorithm is the so called velocity Verlet algorithm [19], whose defining relations are:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \mathbf{a}(t)(\Delta t)^2/2, \quad (6)$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + [\mathbf{a}(t) + \mathbf{a}(t + \Delta t)](\Delta t)/2. \quad (7)$$

The positions and velocities at a given moment are determined based on the positions and velocities at the previous moment and the accelerations at both the previous and the present time.

In more sophisticated simulations at constant temperature and pressure, the equations of motion are more complicated, originating, however, from the same Newton's law. The temperature and the pressure of the system are kept constant for instance by scaling the velocities and positions of all the particles, respectively, following a certain equilibration time frame. A typical approach is the one proposed by Berendsen [4], which introduces the scaling factors

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau_T} \left( \frac{T_d}{T} - 1 \right)} \quad (8)$$

and

$$\mu = \sqrt[3]{1 - \frac{\beta \Delta t}{\tau_P} (P_d - P)}, \quad (9)$$

where the temperature and pressure equilibration times ( $\tau_T$  and  $\tau_P$ , respectively) are usually  $> 100\Delta t$ . Here,  $T_d$  and  $P_d$  are the desired temperature and pressure, while  $\beta$  is the isothermal compressibility (inverse bulk modulus).

The infinite range of the potential implies that every particle interacts with all the other particles. As the simulated systems grow larger and larger the number of force computations grows with the square of the number of particles, leading to long computer times for each run. A partial solution to this problem is the potential truncation, suggested by the rapid decrease of the strength of the interaction at large distances [1]. Typical cutoff radii are  $r_c = 2.5\sigma$  for the Lennard-Jones,  $r_c = 1.8r_e$ , about 5 Å for the Morse, and around 3 Å for the ZBL potential.

A second step in reducing computational time is the use of a neighbor list [18]. The neighbor list keeps track of the particles located just outside the cutoff radius of a certain particle, to minimize the search for interaction candidates. Given the small number of particles used in this work the use of neighbor lists is not justified. However, for future simulations, with larger number of particles the use of a neighbor list is likely to reduce significantly the computation time.

The most important criterion for selecting the minimum size of the simulation cell during a recoil event calculation is that all atoms within the cutoff distance of the recoil atom must be present at all times during the simulation. Therefore, a simulation cell with a side length of 10–15 Å is large enough to contain all atoms that, at a given moment, interact with the recoil atom. This typically amounts to a cell containing 50–100 atoms. A length of 10–15 Å, however, cannot contain the entire path of an implanted ion in the keV energy range, as it may move several hundreds or thousands of angstroms in the implanted sample. Therefore, a mechanism for ensuring that the recoil atom is always surrounded by lattice atoms is needed. We investigate here collision cascades of low energy ions (lower than 1 keV), and we used a lattice of 1 000 such atoms.

As the number of particles used in the simulations is relatively small a real system can be simulated only using periodic boundary conditions [1]. The boundary

conditions apply only along the two transverse directions with respect to the incoming ion and insure that i) once a particle escapes through a wall it has to enter the system from the opposite side, and ii) when distances between particles are evaluated, one has to choose between the particle or its image in the adjacent box to determine the total force on the respective particle. To prevent unphysical double counting (of both the particle and its image) one has to impose the condition that the size of the box is larger than twice the cutoff distance,  $r_c$ . One possible drawback of using conventional periodic boundary conditions is the risk for the recoil atom to move in a simulation cell damaged by its own previous motion. This is another reason to choose a larger system.

The time step  $\Delta t$  is usually fixed in simulations of systems in thermal equilibrium. In collision cascade calculations the initial time step must be very short. Therefore, using a fixed time throughout the simulation is very ineffective. Instead, we choose the length of the time step dynamically to allow a good compromise between the speed and the accuracy of the simulation. The time step is made inversely proportional to the velocity of the fastest moving atom. Thus, the time step becomes longer as the recoil atom slows down, significantly reducing the calculation time. After the collisional phase the time step is held constant again, equal to 2 fs.

#### 4. Results and discussion

The simulations were performed on a personal computer with an Intel Celeron processor running at 2.4 GHz clock speed, using an in house program written in C based on the MDRANGE program [13].

The equilibrium state of the system is prepared by equilibrating 1000 copper atoms located on a fcc crystalline lattice. In the initial displacement calculation, periodic boundary conditions and a constant time step of 2 fs are used. The atoms are given initial velocities in random directions following a Maxwell distribution corresponding to a given initial temperature. The final (desired) temperature is set for 300 K. The simulation is carried out until the average temperature averaged over the last 2000 time steps yields the desired temperature  $T$  (within the error bounds). Figure 1 shows the time evolution of the system temperature during the equilibration stage, starting from 450 K and ending with fluctuations around 300 K. The characteristic temperature equilibration time was  $\tau_T = 100\Delta t = 200$  fs.

During the preparation of the equilibrium state the pressure of the system is also monitored (see Fig. 2). The cell size is scaled such that the system returns to atmospheric pressure, as during the first steps of the equilibration process it departs from the initial value of atmospheric pressure. The characteristic pressure equilibration time was  $\tau_P = 100\Delta t = 200$  fs.

The time evolution of order parameter of the fcc lattice is shown in Fig. 3. As the initial positions of all atoms were on a perfect lattice, the order parameter starts at unity and decreases due to displacements from the lattice values.

The initial state for the collision simulations is shown in Fig. 4 (left). The system is made up of one incoming copper ion of various kinetic energies, initially located at



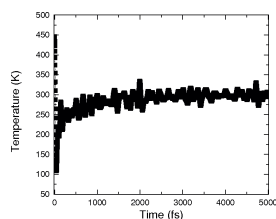


Fig. 1. Temperature versus time during the preparation of the initial state for the collision simulation.

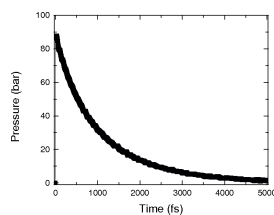


Fig. 2. Pressure versus time during the preparation of the initial state for the collision simulation.

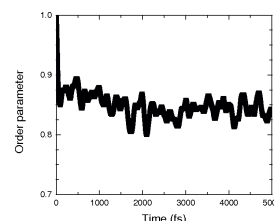


Fig. 3. Order parameter versus time during the preparation of the initial state for the collision simulation.

10 Å away from the target, impinging centrally, at normal incidence, on 1000 copper atoms. The target was previously equilibrated for 5000 fs (2500 simulation steps). The grayscale coding shown in Fig. 4, ranging from high energies (black and dark grey) to room temperature kinetic energies (light grey), is valid for all figures in this work.

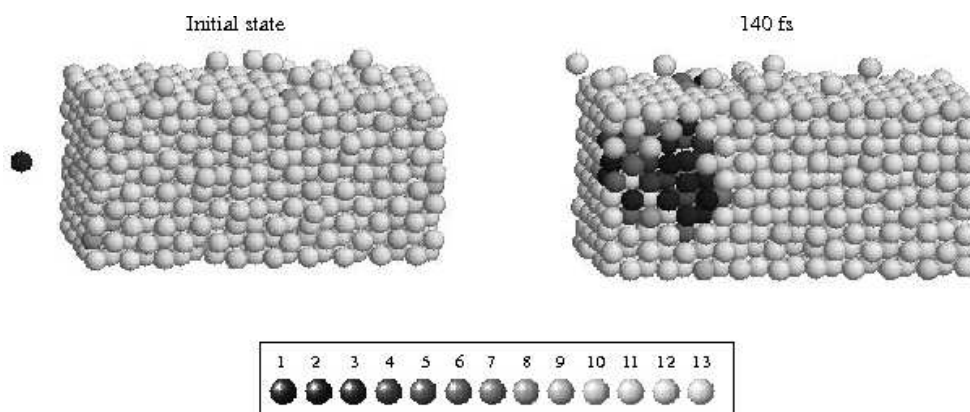


Fig. 4. The initial state of the collision simulation (left), the result of the collision with a 500 eV particle, after 140 fs from the contact (right), and the grayscale energy coding (bottom): 1)  $> 25$  eV, 2) 25–5 eV, 3) 5–1.2 eV, 4) 1.2–0.7 eV, 5) 0.7–0.5 eV, 6) 0.5–0.35 eV, 7) 0.35–0.2 eV, 8) 0.2–0.14 eV, 9) 0.14–0.1 eV, 10) 0.1–0.06 eV, 11) 0.06–0.03 eV, 12) 0.03–0.005 eV, 13)  $< 0.005$  eV. This color coding is valid for all figures in this work.

Simulations were performed at various initial energies of the incoming particle: 10, 20, 50, 100, 200, 500, and 1000 eV. We show here details of our simulations for two of the most interesting cases, namely at 50 eV (Fig. 5) and 500 eV (Fig. 6).

Displayed in Fig. 5 are two-dimensional images of the system at various times during the collision process. It can be seen that the projectile particle is rapidly slowed down in the target and that the damage is limited in the system. The collision

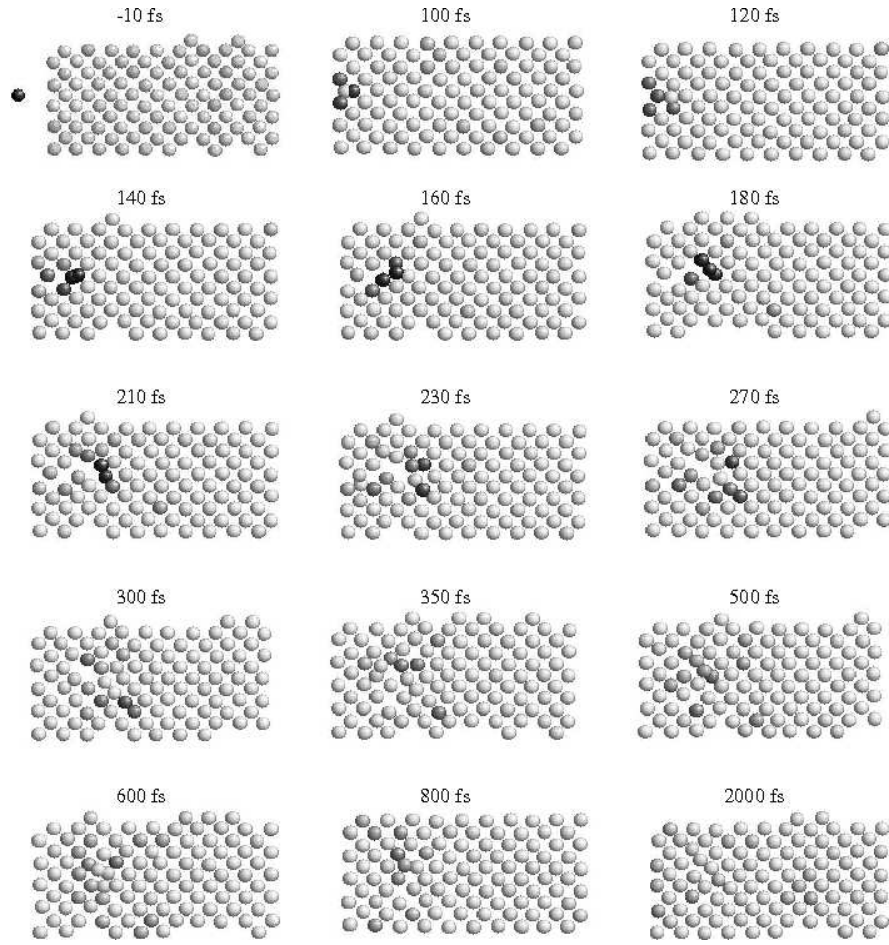


Fig. 5. Two dimensional images of the collision processes at various times in the case of an 50 eV incoming particle.

phase lasts less than 200 fs, being followed by a thermal spike during which the heat is dissipated in roughly 1 000 fs. The relaxation phase lasts more than 2 000 fs, the defects, not very numerous, migrating through the lattice.

In the case of a 500 eV projectile, the damaging effects of the collision are more spectacular. As shown in Fig. 6, the collision phase lasts less than 200 fs. The energy of the incoming particle is transferred to the target atoms causing major defect formation. The thermal spike allows the heat dissipation in roughly 1 000 fs. Again, the relaxation phase lasts more than 2 000 fs, the defects, much more numerous this time, migrating through the lattice. It can be seen that the left part of the system was melted during the collision, being left in an roughly amorphous state, while the right

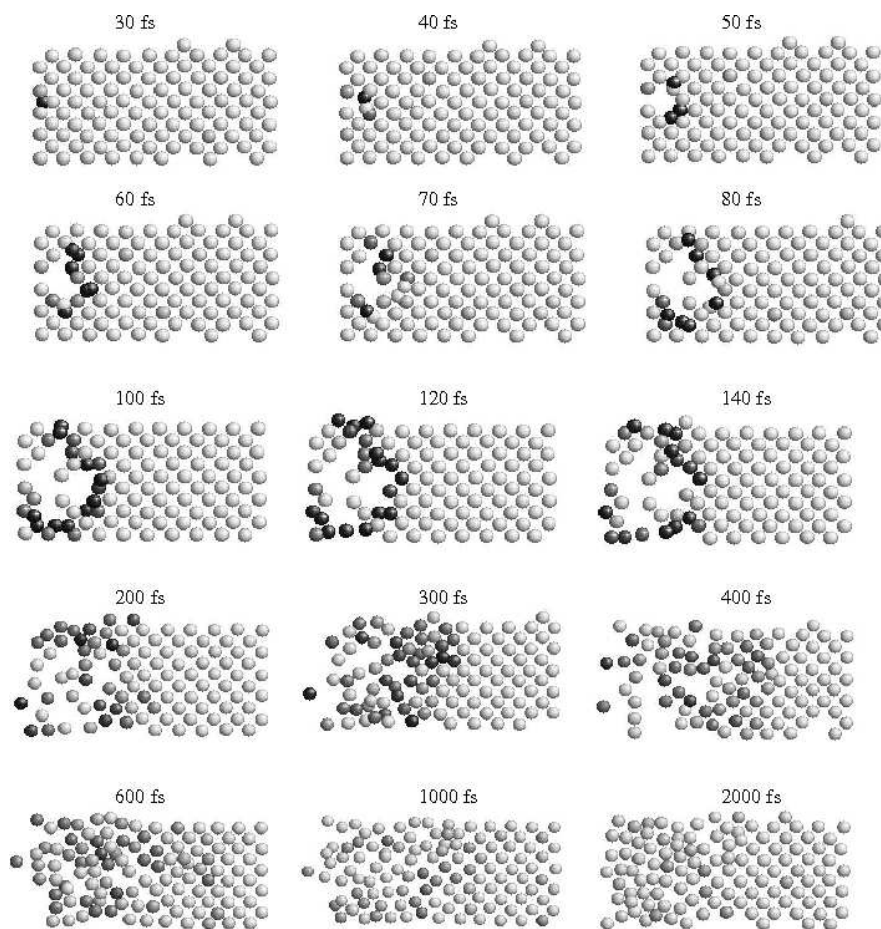


Fig. 6. Two dimensional images of the collision processes at various times in the case of an 500 eV incoming particle.

part was less affected and preserved its crystallinity. The displacements caused by the recoil atom lead to various defects, most often, Frenkel pairs (an interstitial atom, leaving behind a vacant lattice site). The number of defects increases, as expected with the kinetic energy of the incoming particle. For energies of 1000 eV the particles breaks through the target, which clearly shows that a larger simulation cell with a larger number of atoms is needed to describe the collision properly.

The number of Frankel pairs varies strongly between individual recoil events, which suggests that if one wishes to make quantitative conclusions on defect production from MD simulations of collision cascades, it is essential to simulate a large number of events in order to obtain a statistically significant average of the number of defects produced.

## 5. Conclusions

We reported studies of the interaction involving low energy ions impinging on fcc metal surfaces aimed at revealing the damage caused by the recoil events. We used molecular dynamics simulations to describe the radiation-induced defect formation in copper. We found that, as expected, the number of defects increases with the kinetic energy of the incoming particle, for energies larger than 1 000 eV the simulation cell being insufficient. We found that to study quantitatively the defect production, it is imperative that a large number of collision events are simulated for a proper averaging of the results.

## References

- [1] M.P. Allen and D.J. Tildesley, *Computer simulation of liquids*, Clarendon, Oxford, 1987.
- [2] Averback, R.S. and T. Diaz de la Rubia, *Displacement Damage in Irradiated Metals and Semiconductors.*, in *Solid State Physics*, **51**, edited by H. Ehrenfest and F. Spaepen, Academic Press, New York, 1997.
- [3] D. Beeman, *Some Multistep Methods for Use in Molecular Dynamics Calculations*, J. Comp. Phys., **20** (1976), 130.
- [4] A. Berendsen, *Some Multistep Methods for Use in Molecular Dynamics Calculations*, J. Chem. Phys., **81** (1984), 3684.
- [5] W. Bolse, *Ion-beam induced atomic transport through bi-layer interfaces of low- and medium-Z metals and their nitrides*, Mat. Sci. Eng. Rep., **R12** (1994), 53.
- [6] C. Erginsoy, G. H. Vineyard, and A. Englert, *Dynamics of Radiation Damage in a Body-Centered Cubic Lattice*, Phys. Rev., **133** (1964), 595.
- [7] W. M. C. Foulkes and R. Haydock, *Tight-binding models and density-functional theory*, Phys. Rev. B, **39** (1989), 12520.
- [8] J. B. Gibson, A. N. Goland, M. Milgram, and G. H. Vineyard, *Dynamics of Radiation Damage*, Phys. Rev., **120** (1960), 1229.
- [9] H. L. Heinisch, B. N. Singh, and T. Diaz de la Rubia, *Calibrating a multi-model approach to defect production in high-energy collision cascades*, J. Nucl. Mat., **212-215** (1994), 127.
- [10] D. W. Heermann, *Computer Simulation Methods in Theoretical Physics*, Springer, Berlin, 1986.
- [11] P. Jung, *Atomic displacement functions of cubic metals*, J. Nucl. Mat., **117** (1983), 70.
- [12] J.E. Lennard-Jones, Proc. Camb. Phil. Soc., **27** (1931), 469.

- [13] K. Nordlund *et al.*, *The MDRANGE program, V1.0-V1.83b*, [http://beam.helsinki.fi/knordlun/mdh/mdh\\_program.html](http://beam.helsinki.fi/knordlun/mdh/mdh_program.html), August 2002.
- [14] P.M. Morse, *Diatomic molecules according to the wave mechanics. II Vibrational levels*, Phys. Rev., **34** (1929), 57.
- [15] K. Nordlund, M. Ghaly, R.S. Averback, M. Caturla, T. Diaz de la Rubia and J. Tarus, *Defect production in collision cascades in elemental semiconductors and fcc metals*, Phys. Rev. B, **57** (1998), 7556.
- [16] T. Diaz de la Rubia and M. W. Guinan, *New Mechanism of Defect Production in Metals: A Molecular-Dynamics Study of Interstitial-Dislocation-Loop Formation at High-Energy Displacement Cascades*, Phys. Rev. Lett., **66** (1991), 2766.
- [17] R. Smith and D. E. Harrison, Jr., *Algorithms for molecular dynamics simulations of keV particle bombardment*, Computers in Physics Sep/Oct 1989 (1989), 68.
- [18] S.M. Thompson, *Use of neighbor lists in molecular dynamics*, CCP5 Quaterly, **8** (1983), 20.
- [19] L. Verlet, *Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Phys. Rev. **159** (1967), 98.
- [20] L. Verlet, *Computer experiments on classical fluids. II. Equilibrium correlation function*, Phys. Rev., **165** (1968), 201.
- [21] J. F. Ziegler, J. P. Biersack, and U. Littmark, *The Stopping and Range of Ions in Matter*, Pergamon, New York, 1985.