“OVIDIUS” UNIVERSITY OF CONSTANTZA
UNIVERSITATEA „OVIDIUS” CONSTANŢA

“OVIDIUS” UNIVERSITY ANNALS - CONSTANTZA
Year XI
(2009)

Series: CIVIL ENGINEERING

SPECIAL ISSUE DEDICATED TO THE 5-TH CONFERENCE
“DYNAMICAL SYSTEMS AND APPLICATIONS”

ANALELE UNIVERSITĂŢII „OVIDIUS”CONSTANŢA
ANUL XI
(2009)

Seria: CONSTRUCŢII

VOLUM SPECIAL DEDICAT CELEI DE A 5-A CONFERINŢE
“DYNAMICAL SYSTEMS AND APPLICATIONS”

Ovidius University Press
2009

“OVIDIUS” UNIVERSITY OF CONSTANTZA
UNIVERSITATEA „OVIDIUS” CONSTANŢA

“OVIDIUS” UNIVERSITY ANNALS - CONSTANTZA
Year XI
(2009)

Series: CIVIL ENGINEERING

SPECIAL ISSUE DEDICATED TO THE 5-TH CONFERENCE
“DYNAMICAL SYSTEMS AND APPLICATIONS”

ANALELE UNIVERSITĂŢII „OVIDIUS”CONSTANŢA
ANUL XI
(2009)

Seria: CONSTRUCŢII

VOLUM SPECIAL DEDICAT CELEI DE A 5-A CONFERINŢE
“DYNAMICAL SYSTEMS AND APPLICATIONS”

Ovidius University Press
2009

"OVIDIUS" UNIVERSITY ANNALS - CONSTANTZA – SERIES: CIVIL ENGINEERING
ANALELE UNIVERSITĂŢII „OVIDIUS" CONSTANŢA – SERIA: CONSTRUCŢII

DESK EDITORS
Lucica ROŞU, Elena Pelican, Alina Barbulescu
Mail address: "OVIDIUS" University, Faculty of Civil Engineering,
124, Mamaia Blvd., 900527, RO., Constantza, Romania
E-mail: lrosu@yahoo.fr;

ORDERING INFORMATION
The journal may be obtained by ordering at the "OVIDIUS" University, or on exchange basis with similar Romanian or foreign institutions.
Revista poate fi procurată prin comandă la Universitatea „OVIDIUS", sau prin schimb de publicaţii cu instituţii similare din ţară şi străinătate.
124, Mamaia Blvd., 900527 - Constantza, Romania

ISSN 1584-5990

TABLE OF CONTENTS

ISSN-1584-5990

B. Mechanics

C. Dynamical Systems

D. Numerical Methods

E. Differential and Integral Operators & Equations

F. Computer Science

SECTION A

APPLICATIONS IN EARTH SCIENCES & ENGINEERING

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Dynamic Simulation of the Multicomponent Distillation Colums Using CHEMCAD

Suleyman Karacan
Ankara University, Faculty of Engineering
E-mail:karacan@eng.ankara.edu.tr

Abstract

The aim of this study is to compare the dynamic behaviour of a plate with packed distillation column. Stagewise and Plug-Flow mathematical models have been developed for the system. Aromatic compounds are yielded from naphtha reforming in a petrochemical plant, and the products are separated to distill Benzene, Toluene and p-Xylene (BTX) mixture. Finally, the dynamic simulation result of the two types of column are compared. Numerical simulations have been done by CHEMCAD software. A comparison of different models is presented.

Keywords: Multicomponent distillation column, Mathematical model, BTX

1 Introduction

The multicomponent distillation operation is multivariable and highly non-linear process, that presents a very complex dynamics. It constitutes therefore a very serious control problem, been at the same time very widely used in process plant [1]. Distillation is energy-intensive separation process in which a liquid or vapor mixture of two or more substances is separated into its component fractions of desired purity by the application and removal of heat. It can be classified into two ways namely binary distillation for separation of mixture of two substances and multi-component distillation for separation of mixture of more than two substances [2].

Both plate and packed columns are widely used in chemical industries. There are many texts and papers which compare the steady state advantages and disadvantages of these two types of columns. But there are few papers which compare the dynamic behaviour of plate and packed columns [3-5]. However, if for any reason one type of column should be replaced by another type, information on dynamic behaviour of both columns is necessary because in the new case all of the regulating elements and set points should be recognized and readjusted.

ISSN-1584-5990

Computer-aided design and simulation of multicomponent multiphase processes is conventionally performed by equilibrium stage models. This approach is not completely justified because actual equipment rarely, if ever, operates at equilibrium. Moreover, in the enthalpy and K-values calculations, the use of tray efficiency leads to inconsistencies that cause most of computational difficulties [6-8]. Transfer-based models allow to overcome these and other drawbacks. Furthermore they are very powerful in correctly solving design and scale-up problems; in fact, their behavior depends on structural equipment parameters and then equipment configuration has an exact correspondence within the model [9-11].

In this study, I try to show the most important differences between the dynamic simulation of the continuous BTX(Benzene, Toluene,p-Xylenes) mixture fraction system by using a plate and then a packed distillation column. The column models may consist of the different configurations of mass and energy, thermodynamics and hydraulic equations. To find the comparable models for plate and packed columns, I review again the principal differences between plate and packed columns. The structural design procedure is applied to find optimum operating parameters of fractionation process in a naphtha reforming plant. For equilibrium computation and the design of operational variables, the commercial software CHEMCAD [12] is implemented. The performance of models is compared with each other to examine energy saving of the column.

2 Mathematical Models

2.1 Stagewise Model

The dynamic mathematical model of the column was written for n equilibrium stages in Figure 1 and used throughout the column for numerical solution. The mass balance equations of traditional multicomponent stagewise model [4] are written separately for each phase. To determine the liquid compositions for stages, unsteady state mass balance is as follow:

$$(dMx)_n/dt = (Vy)_{n-1} + (Lx)_{n+1} - (Vy)_n - (Lx)_n \tag{1}$$

Where M is the amount of liquid accumulated in the nth plate. Total material balance around the same stage is:

$$d(M_n)/dt = V_{n-1} + L_{n+1} - V_n - L_n \tag{2}$$

The liquid flow rate is function of liquid accumulation in the stage:

$$d(L_n/dt) = (1/\tau_h)(dM_n/dt) \tag{3}$$

Component molar accumulation is written as:

$$M_{n,i} = x_{n,i}M_n \tag{4}$$

In the nth stage, energy balance is:

$$d(Mh_L)_n/dt = (Lh_L)_{n+1} + (VH_V)_{n-1} - (Lh_L)_n - (VH_V)_n \tag{5}$$

The thermodynamic equilibrium at the vapor-liquid interface is usually described as follows:

$$y_{n,i} = K_n x^*_{n,i} \tag{6}$$

Where vapor-liquid equilibrium constants Kn are determined from the selected thermodynamic models, such as UNIQUAC or NRTL, and the extended Antonie equation for the vapor pressure [13]

2.2 Plug-Flow Model

This model involves the two-film theory of mass transfer using the average vapor phase mass fluxes of each component. In the two-film model, it is assumed that all of the resistance to mass transfer is concentrated in thin films adjacent to the interface and that transfer occurs within these films by unsteady state molecular diffusion alone. Multicomponent balance equation as follows:

In the liquid phase:

$$\varphi_L \frac{\partial(Lx_i^B)}{\partial t} = \frac{\partial(Lx_i^B)}{\partial l} + (N^B_{L,i} a^I A_c); i = 1, 2, ..., nc \tag{7}$$

In the vapor phase:

$$\varphi_V \frac{\partial(Vy_i^B)}{\partial t} = \frac{\partial(Vy_i^B)}{\partial l} - (N^B_{L,i} a^I A_c); i = 1, 2, ..., nc \tag{8}$$

The interfacial mass transfer rates are calculated based on the Maxwell-Stefan equations to account for diffusional interactions and thermodynamic non-idealities.

To relate the multicomponent mass transfer rates to binary mass transfer coefficients, the method of Krishna and Standart [14] is used. The diffusional fluxes can be calculated from The volumetric liquid hold-up depends on the vapor and liquid flows and is empirical correlations.

The energy balances written for continuous systems are as follows: In the liquid phase:

$$\varphi \frac{\partial(h_L^B)}{\partial t} = -\frac{\partial(Lh_L^B)}{\partial l} + (Q_L^B a^I A_c) \tag{9}$$

In the vapor phase:

$$\varphi \frac{\partial(H_V^B)}{\partial t} = \frac{\partial(VH_V^B)}{\partial l} - (Q_V^B a^I A_c) \tag{10}$$

3 Case Study

The BTX mixture separation system (Figure 1) is considered as a case study. The column contains 30 theoretical stages (total condenser+reboiler). The feed stage number is the 12-th for the stage and packed column.

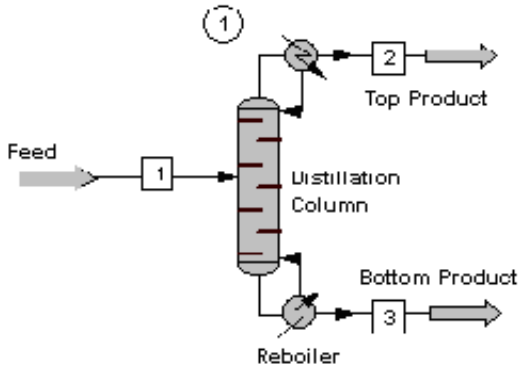


Figure 1: The flow diagram of BTX separation which is applied in the case study

3.1 Numerical Simulation Approach

Steady-state and dynamic simulation of these models was done by CHEMCAD [12]. Todays chemical processing industry (CPI) faces numerous challenges: rising fuel and feedstock costs, reduced engineering staff, shorter product life cycles, increased global competition, and increased regulation. These challenges require that CPI companies seek out and use the best tools to increase productivity and improve engineering decisions. CHEMCAD is a powerful and flexible chemical process simulation environment.

At zero time, the distillation column is empty and dry. So the mass and energy balance equations become trivial. To solve this problem and initialize the dynamic simulation, a steady state simulation is provided. After the steady-state simulation the column is moderately irrigated and the dynamic simulation can be initialized.

4 Results and Discussions

4.1 Steady-State Results

In this part, steady-state solution of two different models is shown and compared with each other. Operating conditions of the distillation column are obtained. Reboiler heat duty and reflux ratio are main control variables of the distillation unit. In the usual operation the reboiler heat duty is determined by the prior

plant, so the reflux ratio is the key control variable. The product purity should be maintained within constraint which can be accomplished by proper reflux ratio. The normal operating range of reflux ratio of the distillation column to product high purity benzene is around 5. The top product purity of benzene should be at least 99.9 percentage , but the top product purity of toluene and p-xylene is not important because the product is mainly composed of three components.

Figures 2 and 3 show the relation between mole fraction of the top product and changes of reflux ratio and rebolier heat duty in each column.

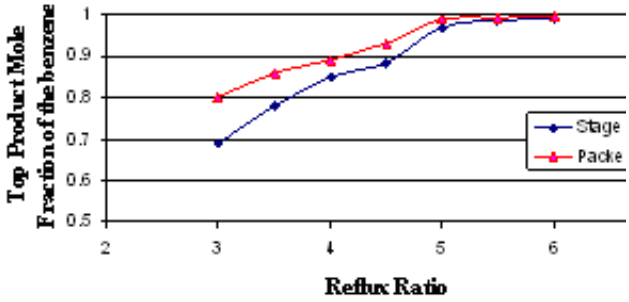


Figure 2: Comparison of steady-state mole fraction for different reflux ratio

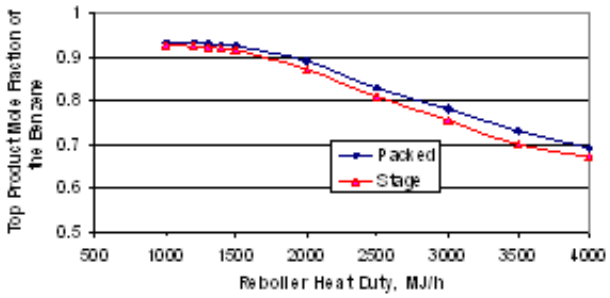


Figure 3: Comparison of steady-state mole fraction for different reboiler heat duty

I have found that at least over 5, a high purity of benzene (mole fraction of 0.99) is produced. Packed column performance is better than stage column. Top product mole fraction of the benzene of packed column is higher than that of stage column at different reboiler duty.

Figure 4 shows liquid mole fraction of the components in each stage. As mole fraction of benzene at top stage (stage number 1) is about 0.99, those of the toluene and p-xylene is about zero.

Figure 5 shows temperature profile in each stage for different columns. The temperatures of the stage column are bigger than those of the packed column,

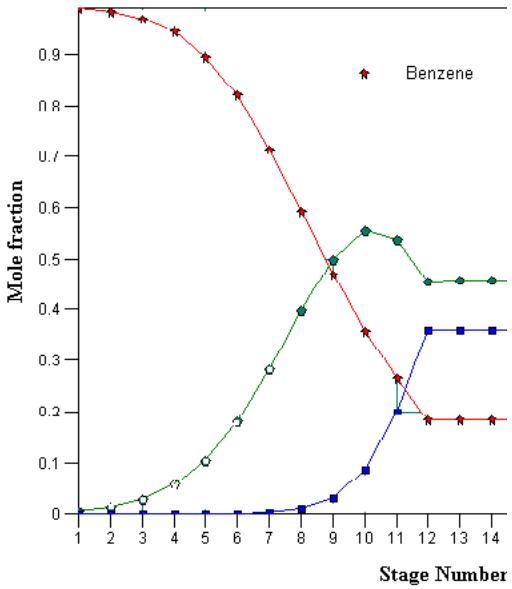


Figure 4: Liquid mole fraction vs. stage

because liquid mole fractions of the packed column are higher than those of the stage column.

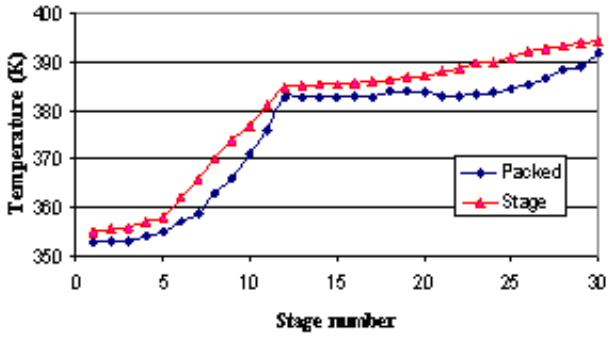


Figure 5: Temperature profile of the column

4.2 Dynamic Results

Dynamic simulation carried out under unsteady-state conditions are reported for stage and packed distillation column. Figures 6 and 7 show top product mole fraction of benzene for stage and packed column, respectively. A negative load disturbance of the feed mole fraction of the benzene from 0.32 to 0.20 is introduced. As shown in the Figures packed column simulation are faster than

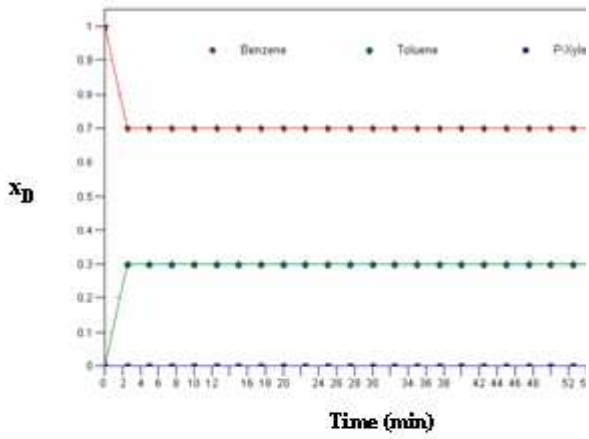


Figure 6: Dynamic response of top product mole fraction of components to step change feed mole fraction for stage column

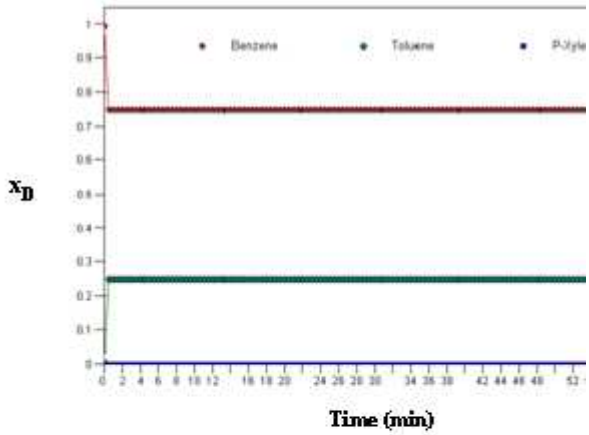


Figure 7: Dynamic response of top product mole fraction of components to step change feed mole fraction for packed column

stage column and reach to steady-state more quickly.

Nomenclature

A: column cross section area,
a: specific packing surface,
H: vapor enthalpy,
h: liquid enthalpy,
kG: gas phase mass transfer coefficient,
i: component index
L: liquid flow rate,

l: column height,
m: holdup in column,
NL: interfacial molar flux,
x: liquid phase mole fraction
V: vapor flow rate,

References

[1] B. Jensy, M.Kenneth, *Process Control: Structures and Applications*,, (1988).

[2] C.C. Yu , M.L. Luyben, *Interpretation of temperature control for ternary distillation*,Industrial & Engineering Chemistry Research, 44(2005), 8277-8290.

[3] A.Gorak, IChemE Symp. Series,104(1987), A413-A424.

[4] M.Alpbaz, S. Karacan, Y.Cabbar and H. Hapoglu, *Application of Model Predictive Control and Dynamic Analysis to A Pilot Distillation Column and Experimental Verification*, Chemical Engineering Journal, 88(2002), 163-17.

[5] Carla Da Porto and Deborha Decorti, *Effect of cooling conditions on separation of volatile compounds in grappa using tray and packed columns without reflux*, International Journal of Food Science and Technology, 43(2008), 638-643.

[6] A. Higler, R.Chande, R. Taylor, R. Baur,R.Krishna, *Nonequilibrium modeling of three-phase distillation*, Computers and Chemical Engineering, 28(2004), 2021-2036.

[7] P. A. Springer, M., R. Baur, & R.Krishna,*Composition trajectories for heterogeneous azeotropic distillation in a bubble-cap tray column: Influence of mass transfer*, Chemical Engineering Research & Design, 81(2003), 413-426.

[8] M. Ottenbacher and H. Hasse,*Continuous Three-Phase Distillation A Process for Separating Thermally Unstable Substances*,Chemical Engineering Research and Design, 85(2007) 144-148.

[9] S. Karacan, Y. Cabbar, M. Alpbaz, H.Hapoglu,*The Steady-State and Dynamic Analysis of Packed Distillation Column Based on Partial Differential Approach*, Chemical Engineering and Processing, 37(1998), 379-388.

[10] G. Pagani, A. DArminio Monforte, G. Bianchi,*Transfer-based models implementation in an equation oriented package*,Computers and Chemical Engineering, 25(2001), 1493-151.

[11] Y.Ishiia, Fred D. Ottob,*An enhanced pseudo-binary-mixture (PBM) algorithm for multistage separation processes*, Computers and Chemical Engineering, 28(2004), 2553-2567.

[12] Chemstations, (2008), CHEMCAD Version 6.1 User Guide

[13] R.C. Reid, J.M. Prausnitz, B.E. Poling,*The properties of gases and liquids*, fourth ed., (1987) McGraw Hill, USA.

[14] R. Krishna, G.L. Standart,*Mass and energy transfer in multicomponent systems*,Chem. Eng. Commun. 3(1979), 201-275.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Nonlinear Multivariable Control of a Multicomponent Continuous Packed Distillation Column Using Artificial Neural Networks

Fatma Varol and Suleyman Karacan
Ankara University, Faculty of Engineering
E-mail:karacan@eng.ankara.edu.tr

Abstract

Distillation of a multicomponent alcohol mixture was researched in a laboratory scale continuous packed distillation column. In this work, multi-input multi-output (MIMO) control of top and bottom product temperature of the column were made theoretically and experimentally. Model predictive control (MPC) algorithm based on artificial neural networks (ANN) was used for control studies. Dynamic experiments were made to system identification by selecting reflux ratio and reboiler heat duty as input variables, top and bottom product temperatures as output variables. For control algorithm, ANN model of process was proposed and was trained with backpropagation algorithm. By basing on this model, MPC algorithm was obtained. In experimental control, model predictive control (MPC) algorithm based on artificial neural networks (ANN) was written at Visual Basic language. This algorithm was added to software. In control studies, the top and bottom product temperatures were controlled by giving positive and negative step effects to set points of the top and bottom product temperatures and the concentration of feed mixture.

Keywords: Artificial neural networks, Multicomponent distillation column, MIMO control

1 Introduction

The multicomponent distillation operation is multivariable and highly non-linear process, that presents a very complex dynamics. It constitutes therefore a very serious control problem, been at the same time very widely used in process plant [1]. Distillation is energy-intensive separation process in which a liquid or vapor mixture of two or more substances is separated into its component fractions of desired purity by the application and removal of heat. It can be classified into two ways namely binary distillation for separation of mixture of two substances

ISSN-1584-5990

and multi-component distillation for separation of mixture of more than two substances [2]. In certain control applications, there are situations when some of the parameters cannot be measured economically online. This is because either the instrumentation is very costly or measurement introduces lags, which makes it impossible to design an effective control scheme. In such situations from the secondary measurements, inference is made for the desired parameters, artificial neural network (ANN) best suits for this application. Artificial neural networks have been extensively studied to face a great variety of problems. One of them has been to approximate relationships between multiple inputs and outputs [3]. The neural network is trained to develop these relationships with data obtained from the process under analysis. ANN has also received considerable attention in multivariable control system applications using even predictive effects [4]. Over the last twenty years, many papers and applications of model-predictive control (MPC) have appeared in the open literature [5-6]. MPC has been successfully applied in chemical process industries. The MPC algorithm has many attractive features such as dead time compensation, multivariable control and handling of system constraints. The nonlinear modeling capability of NN is well documented [7,8]. Modeling methods employing neural networks (NNs) are utilised to provide viable process models [9]. This is due to their ability to approximate virtually any arbitrary mapping between a known input and output space. In this study, a nonlinear MPC strategy based on artificial NNs is presented for the control of a chemical plant. Multi-layer feedforward neural networks are the most extensively utilized. The neural-network based predictive control (NNPC) strategies have been found to be effective in controlling a wide class of nonlinear processes in the past [l0-11]. In the NNPC, the neural network will be used as the prediction model of the nonlinear plant and the system performance is greatly dependent on the online optimization procedure. In this study, a nonlinear MPC strategy based on artificial NNs is presented for the control of a multicomponent packed distillation column.

In this work, multi-input multi-output (MIMO) control of top and bottom product temperature of the column were made. Model predictive control (MPC) algorithm based on artificial neural networks (ANN) was used for control studies. Dynamic experiments were made to system identification by selecting reflux ratio and reboiler heat duty as input variables, top and bottom product temperatures as output variables. For control algorithm, ANN model of process was proposed and was trained with backpropagation algorithm. In control studies, the top and bottom product temperatures were controlled by giving positive and negative step effects to set points of the top and bottom product temperatures and the concentration of feed mixture.

2 Nonlinear model predictive control based on ANNs

A combination of multiple ANNs is used to model an M input N-output nonlinear dynamic system. The proposed system consists of a two-dimensional array of NN blocks. Each block consists of a one step- ahead predictive neural model, NNj, which is identified to represent each output yj of the MIMO system. Therefore, each block represents a multiple-input single-output (MISO) subsection of the whole MIMO system. All blocks in the jth row utilise the same model as NNj. These models are employed to predict the future outputs of the output yj over the prediction horizon of P time steps. The neural models are multilayer feed-forward NNs containing one hidden layer. The hidden layer contains 10 neurons. The activation function used for the neurons in the hidden layer is a Hyperbolic Tangent function. A linear activation function is used for the single output node of each network. Figure 1 shows the details of a typical NN block used in this system. As shown in this figure, past and current samples of each process input ui, and past and current output samples of the process output yj are used as inputs to the network. At time k, the input vector $Uj(k)$ to the block $NNj(1)$ is defined as:

$$\mathbf{U_j(k)} = [u_1(k-d_{1j}), ..., u_1(k-d_{1j}-n_{u1}), u_2(k-d_{2j}), ...,$$

$$u_2(k-d_{2j}-n_{u2}), ..., u_{ni}(k-d_{nij}), ..., u_{ni}(k-d_{nij}-n_{uni}), yp_j(k), ..., ym_j(k-n_{yj})] \tag{1}$$

where $ym_j = y_j$ is the jth measured output of the plant, d_{ij} is the time delay between the ith input and jth output and the ns are integers indicating the model orders. The input vector to the block $NN_j(in)$ is given by $\mathbf{U}_j(k+in-1)$. The future outputs in this vector are supplied by the preceding blocks.

A correction term d_j is added to the model output ys_j to obtain the predicted output yp_j.The correction term dj counts for the difference between the measured plant output and the model output. Each predicted disturbance $d_j(k+in)$ for any future time $k+in$ is assumed to be equal to the present $d_j(k)$.

An attempt was made to identify and use a large one-step-ahead predictor for the MIMO process at each prediction step. However, finding such a large model required further training time and effort. In addition, it was not possible to complete the training to a reasonably small error. Similar results were obtained by using a large P-step-ahead predictor for the MIMO process.

3 Nonlinear optimization

The task of the nonlinear optimiser is to calculate the present and future control actions which minimise the performance index at time k:

$$E = \sum_{j=1}^{N} \sum_{i_n=1+d_{ij}}^{P} [yp_j(k+i_n - ysp_j(k+i_n)]^2/2 \tag{2}$$

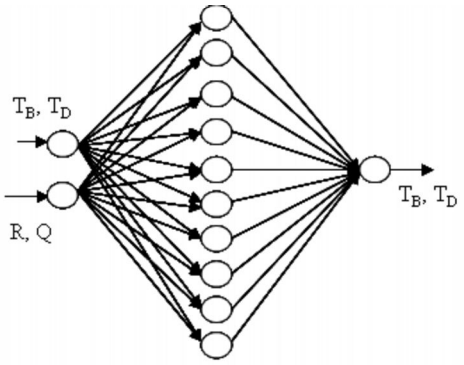


Figure 1: Typical NN block for distillation column

where: $yp_j(k+i_n), i_n = 1+d_{ij}, ..., P$ and $ysp_j(k+i_n), i_n = 1+d_{ij}, ..., P$ are the predicted and desired trajectories, respectively, of the j–th controlled variable. At the current time, the i–th present and future manipulated inputs $u_i(kd_{ij}+i_n-1), i_n = d_{ij}+1, ..., d_{ij}+M; d_{ij}+M \leq P$ are calculated repeatedly as:

$$(u_i(k^{'}))_{new} = (u_i(k^{'}))_{old} - \eta(\partial E/\partial u_i(k^{'})) \tag{3}$$

where $k = k - d_{ij} + i_n - 1$, and η is the step size of the steepest descent method. According to Equation 2, the gradients of the objective function with respect to the manipulated variables can be obtained as:

$$\partial E/\partial u_i(k^{'}) = \sum_{j=1}^{N} \sum_{i_n=1+d_{ij}}^{P} [yp_j(k+i_n) - ysp_j(k+i_n)][\partial yp_j(k+i_n)/\partial u_i(k^{'})] \tag{4}$$

It can be shown that the partial differential of the output yp_j of the NN employed in this work with respect to its ith input u_i can be given by:

$$\partial yp_j/\partial u_i = \sum_{i=1}^{H} w_2(j,i_h)[1 - Q_{hid}(i_h)^2]w_1(i_h,i) \tag{5}$$

where w_1 and w_2 are connection weights in the first and second layers, respectively, $Q_{hid}(ih)$ is the output of the ith neuron in the hidden layer, and H is the number of neurons in the hidden layer.

Equation 5 is computed by the partial differential chain operations applied to the multiple neural network system. Using a high-level pseudo code, these operations are realized in MATLAB.

4 Experimental procedures

Experiments were carried out in a laboratory scale packed column to distil the multicomponent mixture. All experimental equipments were shown in Figure

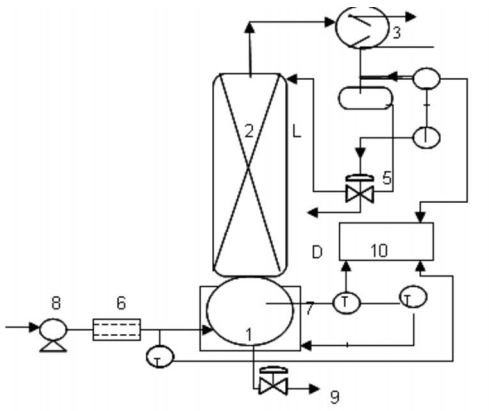


Figure 2: Experimental Equipment

2. In the experiments, overhead product composition and temperature changes with time were observed at steady-state and dynamic conditions..

Mixture including five components, namely methanol-ethanol-n buthanol-isoamine alcohol-anisol, was used as a feed mixture. The column utilized has 1 m packing height. Packing type is raschng ring 20-15 mm sizes. The reboiler was made from 2 L glass container. A peristaltic pump was utilized to feed the relevant liquid into the column. Reflux ratio was adjusted by on-line computer. The system temperatures were measured with three thermocouples. Each thermocouple was connected to a controller module and was transferred to the computer with a Digital/ Digital (D/D) converter. Temperature data measured at each second was recorded. Temperature profiles observed on the computer were recorded and samples were taken regularly from the top and bottom of the column. The samples were analyzed by GC/MS. When the concentrations and temperatures of top and bottom product are constant, the system is said to have reached the steady-state condition.

In experimental control, model predictive control (MPC) algorithm based on artificial neural networks (ANN) was written at Visual Basic language. This algorithm was added to software. In control studies, the top and bottom product temperatures were controlled by giving positive and negative step effects to set points of the top and bottom product temperatures and the concentration of feed mixture.

5 Results and discussions

In this study, the use of ANNs for modeling and control in a packed distillation column is demonstrated. The weights of the networks are estimated off-line and the learning is carried out with inputoutput data provided by suitable open

loop identification experiments. The backward propagation of error signals is used to update the connection weights. Finally, a network is achieved which can predict the output for any input vector. The input neurons transform the input signal and transmit the resulting value to the hidden layer. Each neuron in the hidden layers individually sums the signals they receive together with the weighted signal from bias neuron and transmit the result of each of the neurons in the next layer. Ultimately, the neurons in the output layer receive weighted signals from neurons in the penultimate layer sum the signals and emit the transformed sums as output from the net. The output vector is the temperature of the distillate.

When the multicomponent distillation column works under the steady-state condition at total reflux, reflux ratio and feed flow rate are adjusted to 4 and 0.28 mol/min, respectively, and then 0.45 mol ethanol in mixture is fed to column for continuous operating condition. The system works under this condition and occurrence of the steady-state condition is waited. For control calculations, prediction horizon P is fixed at 10 and control horizon M is varied from 1 to 5. The Run deals with a MIMO ANNMPC simulation for a step increase from 66.0C to 67.0C in the setpoint of the top product temperature (TD) and from 77C to 78C in the setpoint of the bottom product temperature (TB) shown in Figure 3. We get perfect tracking of the specified setpoint change (TD=1, TB=1). Integral Square Error (ISE) values were obtained as follows, 2.03 for output TD and 4.08 for TB. Manipulated variables are plotted in Figure 4, where reflux ratio is used for controlling of the top product temperature and reboiler heat duty is used for controlling the bottom product temperature. The Run 2 gives the results for a step decrease from 66.0C to 65.0C in the setpoint of the top product temperature and from 77C to 76C in the setpoint of the bottom product temperature shown in Figure 5 and 6. ISE values were obtained as follows, 3.84 for output TD and 4.09 for TB. The disturbance rejection capabilities of the MMPC controller were also studied. Feed composition acts as a disturbance to the column. The Run 3 deals with the ANN MPC simulation to positive load effect of feed mole fraction of the process from 0.4556 to 0.5837. At this level of operation controller give similar performance as shown in Figure 7 and 8. ISE values were 0.698 for output TD and 1.216.

The control quality is evaluated by computing the performance criteria for different values of the control horizon M. The integral square error (ISE) was used as the key measure. Control horizon M is varied from 1 to 5. Referring to these values, one can see that the controller with the value of M = 3 has a superior performance. However, in real control, a high value of M can yield more dynamic actions and stability problems can arise. Therefore, a lower value must be chosen for M.

Acknowledgements

The authors are grateful to Ankara University, Research Foundation for its financial support. Ankara, Turkey, Grant no: 2007.0745.003 HPD.

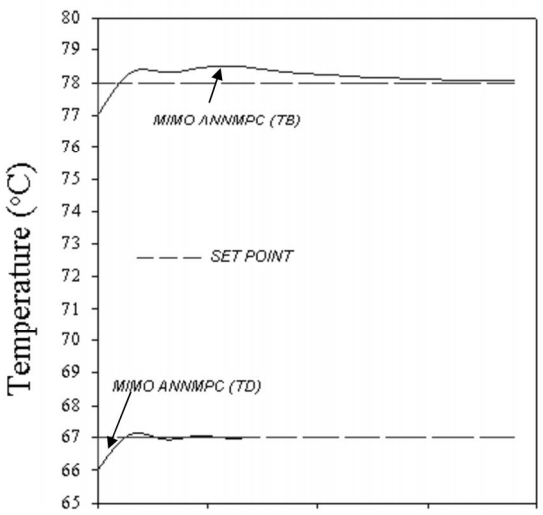


Figure 3: Response of the top and bottom product temperatures using ANN MPC for positive step changes of the product temperatures.

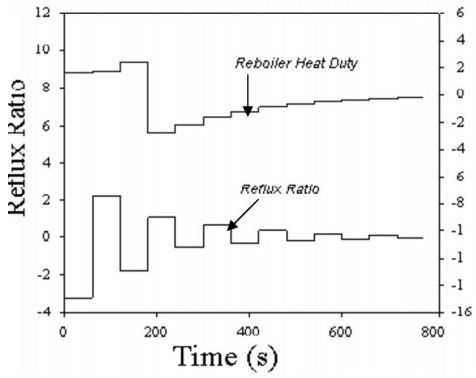


Figure 4: Control variables change as reflux ratio and reboiler heat duty

References

[1] B. Jensy, M.Kenneth, *Process Control: Structures and Applications*, 1988.

[2] M.T. Lin , C.C. Yu , M.L. Luyben, *Interpretation of temperature control for ternary distillation*, Industrial and Engineering Chemistry Research, 44(2005), 8277-8290.

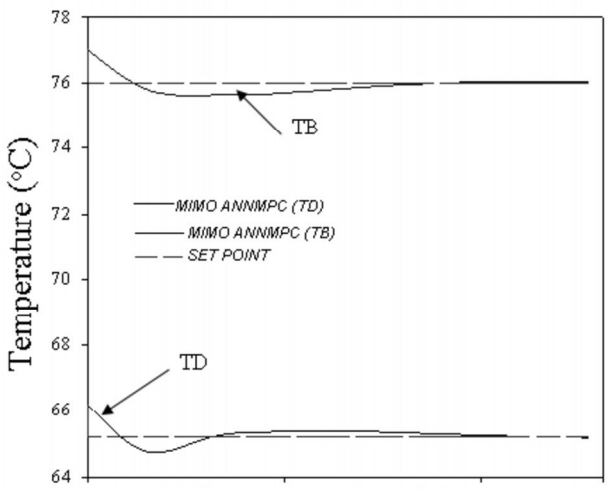


Figure 5: Response of the top and bottom product temperatures using ANN MPC for negative step changes of the product temperatures.

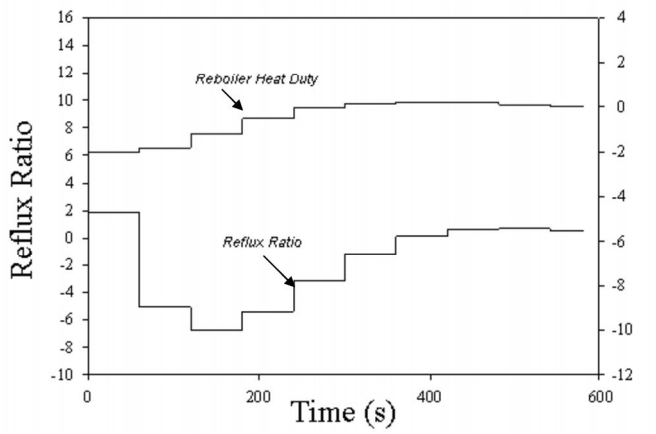


Figure 6: Control variables change as reflux ratio and reboiler heat duty

[3] K.S. Narendra, K. Parthasarathy,*Gradient methods for optimization of dynamical systems containing neural networks*IEEE Trans. Neural Net. 23(1991), 252-262.

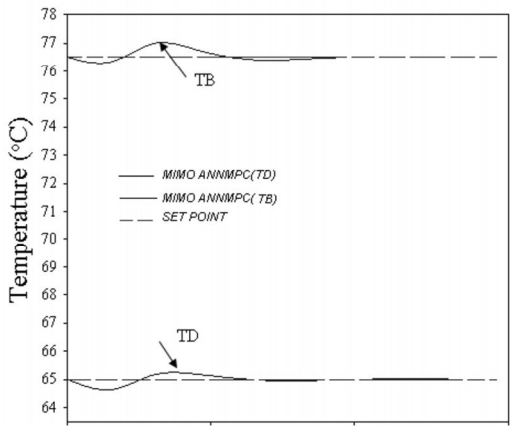


Figure 7: Response of the top and bottom product temperatures using ANN-MPC for positive step changes of the mole fraction of the feed.

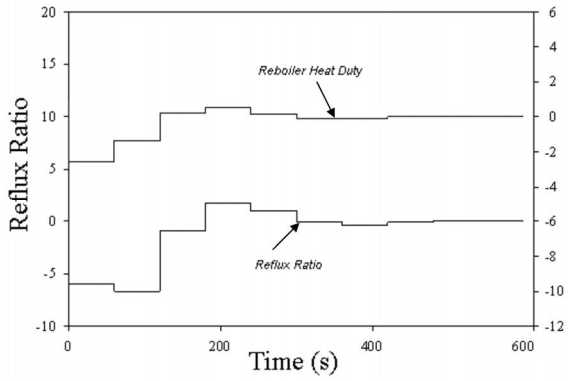


Figure 8: Control variables change as reflux ratio and reboiler heat duty

[4] A. Draeger, S. Engell, H. Ranke, *Model predictive control using neural networks*, IEEE Control Syst.Mag.,15(1995), 61-66.

[5] C. E .Garcia, M. Morari, *model control. 1. A unifying review and some new results*, Ind Eng Chem Proc Des Dev 21(1982), 308-323.

[6] M. A. Henson, ,Nonlinear model predictive control: current status and future directions. Comp. Chem. Eng. 23(1998), 187-202.

[7] K. Hornik, , M. Stinchcombe, H. White,*Multilayer feedforward network are universal approximators*, Neural Networks,2(1989), 359-366.

[8] T.P. Chen, H. Chen,*Approximations of continuous functionals by neural network with application to Dynamics systems*,IEEE Trans. Neural Networks, 4(1993), 910-918.

[9] B. Lennox, G. A. ,Montague, A. M. Frith, C. Gent, V. Beuan,*Industrial application of neural networksan investigation*,J Proc Cont. 11(2001), 497-507.

[10] Y. Zhang, Z.Q. Chen, P. Yang, Z.Yuan,*Multivariable nonlinear proportional– integral–derivative decoupling control based on recurrent neural networks*,Chin. J. Chem. Eng. 12(2004),677- 681.

[11] X. Wang, J. Xiao,*PSO-based model predictive control for nonlinear processes*, In: Lecture Notes in Computer Science 361(2005), 196-203.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Slope Stability Analysis for Earthquakes

Mezouar Khoudir
Technical University of Civil Engineering, Bucharest, Romania

Boukhemacha Mohamed Amine
Technical University of Civil Engineering, Bucharest, Romania

Romeo Ciortan
IPTANA S.A. and Ovidius University of Constanta, Romania

George Paduraru
Faculty of Civil Engineering, Ovidius University of Constanta, Romania

Abstract

The slope stability can be affected by earthquakes with different modes; from exceeded displacement, lose of strength to soil liquefaction. This paper resumes the behavior of slope soil during seismic shaking. The analysis methods for seismic slope stability are grouped into: inertia analysis for those materials that retain their shear strength during the earthquake, and weakening analysis for those materials that will experience a significant reduction in shear strength during the earthquake

Keywords: Slope stability, earthquakes, ground displacement, liquefaction

1 Introduction

Failures occur in both natural and manmade slopes. In our natural environment landslides occur frequently and they are part of the ongoing evolution of landscape. More rarely failures occur in manmade earth slopes which are designed specifically to resist the forces of nature. Landslides have caused great amounts of damage and loss of life throughout history. Examples of catastrophic landslides in the 20th century include the earthquake induced landslides in the Ningxia province of China in 1920 ($>$ 100000 deaths), the earthquake induced landslides in Alaska in 1964 (tremendous damage) and the reservoir induced landslide at Vaiont dam, Italy in 1963 (2300 deaths). Earthquake induced ground shaking is one of the most frequent causes of landslides. This is happening in marginally or moderately stable slopes where earthquake inertial forces may be sufficient to trigger a failure. In the case of weak foundations or

ISSN-1584-5990

embankment fill materials, repeated ground shaking may cause loss of strength of the soil materials (e.g. liquefaction) and subsequent slope failure. The possibility of the occurrence of a landslide where a slope is subject to earthquake loading, depends on numerous factors which include the geometry, the geology of the slope, the soil engineering properties, the ground water regime, the presence of preexisting shear zones, the weather etc. It is not uncommon for a slope to survive stronger earthquake shaking and fail under lower earthquake shaking because some of the above factors were more favorable for the second case. Landslides induced by earthquakes may be classified into three broad categories (Varnes, 1978).

1. Disrupted slides and falls;
2. Coherent slides;
3. Lateral spreads and flows.

In the case of disrupted slides and falls, the soil or rock material in the slide is sheared and distorted in a nearly random manner. The slopes involved are usually steep and failures take place very suddenly. The damages and loss of life from such slides in developed areas may be devastating. Disrupted slides and falls include disrupted soil/rock slides, soil/rock falls and soil/rock avalanches. Coherent slides generally occur at deeper failure surfaces in moderate to steeply sloping ground and they involve rotational and translational failures of coherent soil and/or rock blocks. These failures include rock/soil slumps, rock/soil block slides and slow earth flows. They develop at slow to rapid velocities. Keefer (1984) studied the effect of earthquake Magnitude and epicentral distance on the occurrence of earthquake induced landslides. The study was based on 300 U.S. earthquakes and it showed that for local magnitudes less than 4.0, landslides rarely occur.

2 Seismic slope stability

The slope stability can be affected by earthquakes with different modes, from exceeded displacement, lose of strength to soil liquefaction. Depending on the behavior of the soil during seismic shaking, seismic instabilities may be grouped into two categories:

1. Inertial instabilities;
2. Weakening instabilities.

3 The inertia slope stability analysis

The inertia slope stability analysis is preferred for those materials that retain their shear strength during the earthquake. Examples of these types of soil and rock are as follows:

- Massive crystalline bedrock and sedimentary rock that remains intact during the earthquake, such as earthquake-induced rock block slide;
- Soils that tend to dilate during the seismic shaking, or, for example, dense to very dense granular soil and heavily overconsolidated cohesive soil such as very stiff to hard clays;
- Soils that have a stress-strain curve that does not exhibit a significant reduction in shear strength with strain. Earthquake-induced slope movement in these soils often takes the form of soil slumps or soil block slides;
- Clay that has a low sensitivity;
- Soils located above the groundwater table. These soils often have negative pore water pressure due to capillary action;
- Landslides that have a distinct rupture surface and the shear strength along the rupture surface is equal to the drained residual shear strength $\phi_r^{'}$.

Seismic Coeff	**Remarks**
0.10	Major earthquake,$FS \succ 1.0$, Corps of Eng. Manual (1982)
0.15	Great earthquake, $FS \succ 1.0$, Corps of Eng. Manual (1982)
0.15-0.25	Japan, $FS \succ 1.0$
0.05-0.15	State of California, $FS \succ 1.0$
0.15	Seed (1979), with $FS \succ 1.15$ and a 20% strength reduction
0.33-0.50	Marcuson and Franklin (1983), $FS \succ 1.0$
0.50	Hynes and all(1984), $FS \succ 1.10$ and a 20% strength reduction
0.10	Terzaghi (1950), sever earthquake
0.20	Terzaghi (1950), violent and destructive earthquake
0.50	Terzaghi (1950), catastrophic earthquake

Table 1: Typical seismic coefficients (k_h) and factor of safety (FS) used in practice

3.1 Pseudostatic method

When the dynamic normal and shear stresses on a potential failure surface are superimposed upon the corresponding static stresses, these may produce inertial instability of the slope if the shear stresses exceed the shear strength of the soil. The original application of the pseudostatic method has been credited to Terzaghi (1950).This method ignores the cyclic nature of the earthquake and treats it as if it applied an additional static force upon the slope. In particular, the pseudostatic approach is to apply a lateral force acting through the centroid of the sliding mass, acting in an out-of-slope direction. The pseudostatic lateral force $F_h$is calculated by using equation 1.

$$F_h = ma = \frac{Wa}{g} = \frac{Wa_{max}}{g} = k_h W, \tag{1}$$

where m: total mass of slide material, which is equal to W/g;
W: total weight of slide material;
a: acceleration, which in this case is the maximum horizontal acceleration at ground surface caused by earthquake ($a = a_{max}$);
$a = a_{max}$: the peak ground acceleration;
$a = a_{max}/g = k_h$: seismic coefficient, also known as pseudostatic coefficient.

The only unknowns in the pseudostatic method are the weight of the sliding mass W and the seismic coefficient k_h. Based on the results of subsurface exploration and laboratory testing, the unit weight of the soil or rock can be determined. The other unknown is the seismic coefficient k_h, which is much more difficult to determine. Typical seismic coefficients and factors of safety used in practice today are given in table I.

3.1.1 Wedge method

The simplest type of slope stability analysis is the wedge method. Figure 1 illustrates the free-body diagram for the wedge method. The failure wedge has a planar slip surface inclined at an angle a to the horizontal. Although the failure wedge passes through the toe of the slope, the analysis could also be performed for the case of the planar slip surface intersecting the face of the slope.

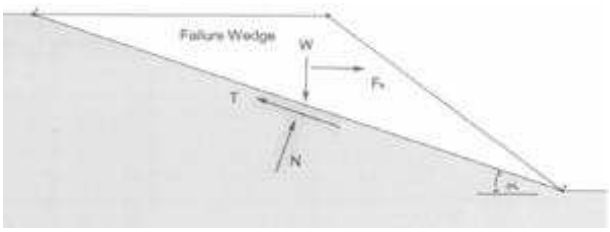


Figure 1: Wedge method, with the forces acting on the wedge

$$T = cL + N \tan \phi \tag{2}$$

or

$$T = s_u L,$$

where
N: normal force acting on the slip surface;
T: shear force acting along the slip surface.

The shear force is also known as the resisting force because it resists failure of the wedge. Based on the Mohr-Coulomb failure law, the shear force is equal to the following:
For a total stress analysis:

For an effective stress analysis:

$$T = c^{'} L + N^{'} \tan \phi^{'}, \tag{3}$$

where:

L: length of the planar slip surface;
c,ϕ : shear strength parameters in terms of a total stress analysis,
s_u: undrained shear strength of the soil (total stress analysis);
$c^{'}$,$\phi^{'}$: shear strength parameters in terms of an effective stress analysis;
$N^{'}$: effective normal force acting on the slip surface.

The assumption in this slope stability analysis is that there will be movement of the wedge in a direction that is parallel to the planar slip surface. Thus the factor of safety of the slope can be derived by summing forces parallel to the slip surface, and it is as follows.

Total stress pseudostatic analysis:

$$FS = \frac{resistingforce}{drivingforces} = \frac{cL + N \tan \phi}{W \sin \alpha + F_h \cos \alpha} = \frac{cL + (W \cos \alpha - F_h \sin \alpha) \tan \phi}{W \sin \alpha + F_h \cos \alpha} \tag{4}$$

Effective stress pseudostatic analysis:

$$FS = \frac{c^{'} L + N^{'} \tan \phi^{'}}{W \sin \alpha + F_h \cos \alpha} = \frac{c^{'} L + (W \cos \alpha - F_h \sin \alpha - uL) \tan \phi^{'}}{W \sin \alpha + F_h \cos \alpha}, \tag{5}$$

where:

u: average pore water pressure along the slip surface.

The total stress pseudostatic analysis is performed in those cases where the total stress parameters of the soil are known. A total stress pseudostatic analysis is often performed for cohesive soil, such as silts and clays.

The effective stress pseudostatic analysis is performed in those cases where the effective stress parameters of the soil are known. Note that in order to use an effective stress analysis equation 5, the pore water pressure u along the slip surface must also be known. The effective stress analysis is often performed for cohesionless soil, such as sands and gravels.

3.1.2 Method of slices

In the method of slices, the failure mass is subdivided into vertical slices and the factor of safety is calculated based on force equilibrium equations. A circular arc slip surface and rotational type of failure mode are often used for the method

of slices, and for homogeneous soil, a circular arc slip surface provides a lower factor of safety than assuming a planar slip surface.

The calculations for the method of slices are similar to those for the wedge-type analysis, except that the resisting and driving forces are calculated for each slice and then summed in order to obtain the factor of safety of the slope. For the ordinary method of slices, the equations used to calculate the factor of safety is identical to equations 4 and 5, with the resisting and driving forces calculated for each slice and then summed to obtain the factor of safety.

The method of slices is not an exact method because there are more unknowns than equilibrium equations. This requires that an assumption be made concerning the interslice forces. Table 2 presents a summary of the assumptions for the various methods.

Type of method of slices	**Assumption concerning interslice forces**	**Reference**
Ordinary method of slices	Resultant of interslice forces is parallel to average inclination of slice	Fellenius (1936)
Bishop simplified method	Resultant of interslice forces is horizontal (no interslice shear forces)	Bishop (1955)
Janbu simplified method	Resultant of interslice forces is horizontal (a correction factor is used to account for interslice shear forces)	Janbu (1968)
Janbu generalized method	Location of interslice normal force is defined by an assumed line of thrust	Janbu (1957)
Spencer method	Resultant of interslice forces is of constant slope throughout the sliding mass	Spencer (1967, 1968)
Morgenstern-Price method	Direction of resultant interslice forces is determined by using a selected function	Morgenstern and Price (1965)

Table 2: Assumptions conserning interslice forces for different method of slices

4 Newmark method

The purpose of the Newmark (1965) method is to estimate the slope deformation for those cases where the pseudostatic factor of safety is less than 1.0 (i.e., the failure condition). The Newmark (1965) method assumes that the slope will deform only during those portions of the earthquake when the out-of-slope earthquake forces cause the pseudostatic factor of safety to drop below 1.0.

When this occurs, the slope will no longer be stable, and it will be accelerated downslope. The longer that the slope is subjected to a pseudostatic factor of safety below 1.0, the greater the slope deformation. On the other hand, if the pseudostatic factor of safety drops below 1.0 for a mere fraction of a second, then the slope deformation will be limited.

Figure 2 can be used to illustrate the basic premise of the Newmark (1965) method. Figure 2a shows the horizontal acceleration of the slope during an earthquake. Those accelerations that plot above the zero line are considered to be out-of-slope accelerations, while those accelerations that plot below the zero line are considered to be into-the-slope accelerations. It is only the out-of-slope accelerations that cause downslope movement, and thus only the acceleration that plots above the zero line is considered in the analysis. In Figure 2, a dashed line has been drawn that corresponds to the horizontal yield acceleration, which is designated a_y. This horizontal yield acceleration a_y is considered to be the horizontal earthquake acceleration that results in a pseudostatic factor of safety that is exactly equal to 1.0. The portions of the two acceleration pulses that plot above ay have been darkened. According to the Newmark (1965) method, it is these darkened portions of the acceleration pulses that will cause lateral movement of the slope.

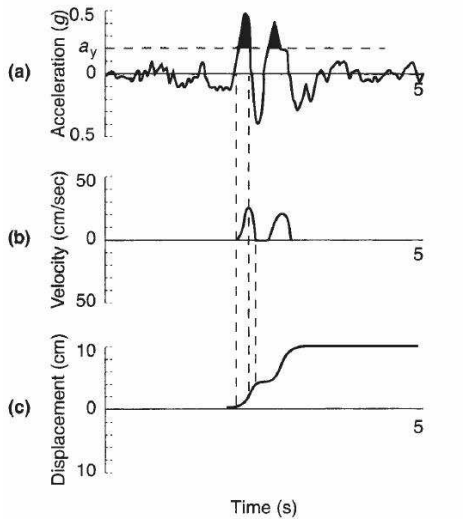


Figure 2: Diagram illustrating the Newmark method. (a) Acceleration versus time; (b) velocity versus time for the darkened portions of the acceleration pulses; (c) the corresponding downslope displacement versus time in response to the velocity pulses. (After Wilson and Keefer 1985).

Newmark (1965) developed an upper bound estimate for earthquake-induced slope displacements using an empirical curve that bounded the results from his analyses of four earthquake time histories. The proposed upper-bound curve is

given by the equation 6.

$$d_{max} = \frac{V^2}{2ga_y}\left(1 - \frac{a_y}{a_{max}}\right)\frac{a_{max}}{a_y} \tag{6}$$

d_{max}: maximum displacement;
V: peak velocity;
a_y: yield acceleration that produces a factor of safety equal to 1.0;
a_{max}: pick ground acceleration.

Many different equations have been developed utilizing the basic Newmark (1965) method as outlined above. One simple equation is given by equation 7([1]).

$$\log d = 0.90 + \log\left[\left(1 - \frac{a_y}{a_{max}}\right)^{2.53}\left(\frac{a_{max}}{a_y}\right)^{-1.09}\right], \tag{7}$$

where d : estimate downslope movement caused by earthquake.

5 The weakening slope stability analysis

The weakening slope stability analysis is preferred for those materials that will experience a significant reduction in shear strength during the earthquake. Examples of these types of soil and rock are as follows:

- Foliated or friable rock that fractures apart during the earthquake, resulting in rockfalls, rock slides, and rock slumps;
- Sensitive clays that lose shear strength during the earthquake.
- Soft clays and organic soils that are overloaded and subjected to plastic flow during the earthquake. The type of slope movement involving these soils, is often termed slow earth flows.
- Loose soils located below the groundwater table and subjected to liquefaction or a substantial increase in excess pore water pressure. There are two cases of weakening slope stability analyses involving the liquefaction of soil:
  - Flow slid
  - Lateral spreading

5.1 Flow slides

Flow slides develop when the static driving forces exceed the shear strength of the soil along the slip surface, and thus the factor of safety is less than 1.0. There are three general types of flow slides:

- Mass liquefaction: This type of flow slide occurs when nearly the entire sloping mass is susceptible to liquefaction. These types of failures often occur to partially or completely submerged slopes, such as shoreline embankments. For design conditions, the first step in the analysis is to determine the factor of safety against liquefaction. If it is determined that the entire sloping mass, or a significant portion of the sloping mass, will be subjected to liquefaction during the design earthquake, then the slope will be susceptible to a flow slide.
- Zonal liquefaction: This second type of flow slide develops because there is a specific zone of liquefaction within the slope. For design conditions, the first step is to determine the location of the zone of soil expected to liquefy during the design earthquake. Then a slope stability analysis is performed by using various circular arc slip surfaces that pass through the zone of expected liquefaction. If the factor of safety of the slope is less than 1.0, then a flow slide is likely to occur during the earthquake.
- Landslide movement caused by liquefaction of soil layers or seams: The third type of slope failure develops because of liquefaction of horizontal soil layers or seams of soil.

For design conditions, it can be difficult to evaluate the possibility of landslide movement due to liquefaction of soil layers or seams. This is because the potentially liquefiable soil layers or seams can be rather thin and may be hard to discover during the subsurface exploration. In addition, when the slope stability analysis is carried out, the slip surface must pass through these horizontal layers or seams of liquefied soil. Thus a slope stability analysis is often performed using a block-type failure mode (rather than using circular arc slip surfaces).

5.1.1 Factor of safety against liquefaction

The factor of safety against liquefaction for slopes is calculated by making an adjustment of that calculated for level-ground sites.

5.1.2 Factor of safety against liquefaction for level-ground sites

The most common type of analysis to determine the liquefaction potential is to use the standard penetration test (SPT) (Seed et al. 1985, Stark and Olson 1995). The analysis is based on the simplified method proposed by Seed and Idriss (1971). The factor of safety against liquefaction (FSL) for level-ground sites is given by the ratio of the cyclic resistance ratio from the SPT, (CRR), by the cyclic resistance stress induced by earthquake, (CSR), as it is given in equation 8.

$$FSL = \frac{CRR}{CSR} \tag{8}$$

The first step in the simplified procedure is to calculate the cyclic stress ratio, also commonly referred to as the seismic stress ratio (SSR) that is caused

by the earthquake. To develop the CSR earthquake equation, it is assumed that there is a level ground surface and a soil column of unit width and length, and that the soil column will move horizontally as a rigid body in response to the maximum horizontal acceleration a_{max} exerted by the earthquake at ground surface. Figure 3 shows a diagram of these assumed conditions.

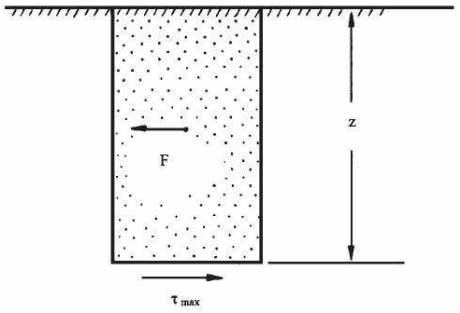


Figure 3: Conditions assumed for the derivation of the CSR earthquake equation

Given these assumptions, the weight W of the soil column is equal to $\gamma_t z$, where γ_t: total unit weight of the soil and z: depth below ground surface. The horizontal earthquake force F acting on the soil column (which has a unit width and length) is given by equation 9.

$$F = ma = \frac{W}{g}a == \frac{\gamma_t z}{g}a_{max} = \sigma_{v0}\frac{a_{max}}{g} \tag{9}$$

σ_{v0} : total vertical stress at bottom of soil column.
By summing forces in the horizontal direction, the force F acting on the rigid soil element is equal to the maximum shear force at the base on the soil element. Since the soil element is assumed to have a unit base width and length, the maximum shear force F is equal to the maximum shear stress τ_{max}.

$$\tau_{max} = \sigma_{v0}\frac{a_{max}}{g} \tag{10}$$

Dividing both sides of the equation by the vertical effective stress $\sigma_{v0}^{'}$ gives:

$$\frac{\tau_{max}}{\sigma_{v0}^{'}} = \frac{\sigma_{v0}}{\sigma_{v0}^{'}}\frac{a_{max}}{g} \tag{11}$$

Since the soil column does not act as a rigid body during the earthquake, but rather the soil is deformable, Seed and Idriss (1971) incorporated a depth reduction factor rd into the right side of equation 12, or:

$$\frac{\tau_{max}}{\sigma_{v0}^{'}} = r_d\frac{\sigma_{v0}}{\sigma_{v0}^{'}}\frac{a_{max}}{g} \tag{12}$$

The r_d values are usually obtained from the curve labeled *Average values by Seed and Idriss (1971)* in figure 4. Another option is to assume a linear

relationship of r_d versus depth and use the following equation (Kayen et al. 1992):

$$r_d = 1 - 0.012z \tag{13}$$

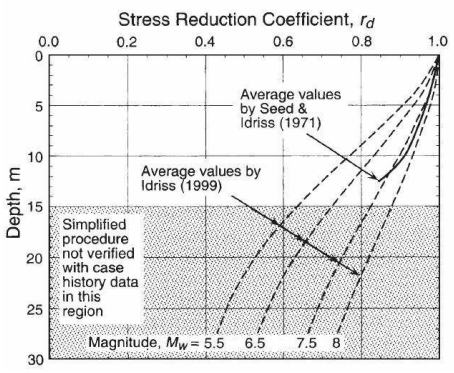


Figure 4: Reduction factor r_d versus depth below level or gently sloping ground surface

For the simplified method, Seed et al. (1975) converted the typical irregular earthquake record to an equivalent series of uniform stress cycles by assuming that the uniform cyclic shear stress amplitude of the earthquake τ_{cyc} is given by equation 14.

$$\tau_{cyc} = 0.63\tau_{max} \tag{14}$$

In the end, the cyclic resistance stress induced by earthquake (CSR), can be calculated using equation 15.

$$CSR = \frac{\tau_{cyc}}{\sigma_{v0}^{'}} = 0.63\frac{\tau_{max}}{\sigma_{v0}^{'}} = 0.63 r_d \frac{\sigma_{v0}}{\sigma_{v0}^{'}} \frac{a_{max}}{g} \tag{15}$$

5.1.3 Factor of safety against liquefaction for slopes

For sloping ground sites, the factor of safety against liquefaction calculated from equation 4 may need to be adjusted. Figure 5 presents a chart that can be used to adjust the factor of safety for sloping ground conditions, (original from Seed and Harder 1990, reproduced from Kramer 1996).

In the horizontal axis is designated α, given by the ratio of the static shear acting on a horizontal plane $\tau_{h.static}$ by the vertical effective stress $\sigma_{v0}^{'}$.

$$\alpha = \frac{\tau_{h.static}}{sigma_{v0}^{'}} \tag{16}$$

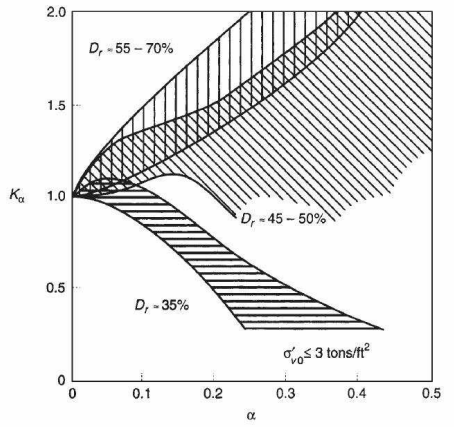


Figure 5: Chart that can be used to adjust the factor of safety against liquefaction for sloping ground, (after Seed and Harder, 1990. Bolton Seed memorial Symposium Proceeding, Vol. 2)

5.1.4 Stability analysis for liquefied soils

The first step in a flow analysis is to determine the factor of safety against liquefaction for the various soil layers that comprise the slope. The factor of safety against liquefaction is based on level-ground assumptions. Then the factor of safety against liquefaction is adjusted for the sloping ground conditions, by using figure 5. If it is determined that the entire sloping mass, or a significant portion of the sloping mass, will be subjected to liquefaction during the design earthquake, then the slope will be susceptible to a flow slide. No further analyses will be required for the "mass liquefaction" case.

For the cases of zonal liquefaction or liquefaction of soil layers or seams, a slope stability analysis is required. To perform a slope stability analysis for soil that is anticipated to liquefy during the earthquake, there are two different approaches: (1) using a pore water pressure ratio equal to 1.0 or (2) using zero shear strength for the liquefiable soil.

5.1.5 Pore water pressure ratio (r_u=1.0)

The first approach is to assume that the pore water pressure ratio of the liquefied soil is equal to 1.0. As previously mentioned, the pore water pressure ratio r_u is defined as $r_u = \frac{u}{\gamma_t h}$, where u: pore water pressure, γ_t : total unit weight of the soil, and h: depth below the ground surface.

At a factor of safety against liquefaction FSL equal to 1.0 (i.e., liquefied soil), $r_u = 1.0$. Using a value of $r_u = 1.0$, then $r_u = 1.0 = \frac{u}{\gamma_t h}$. This means that the pore water pressure u must be equal to the total stress $\sigma = \gamma_t h$ and hence

the effective stress σ' is equal to zero.

5.1.6 Shear strength equal zero for liquefied soils

The second approach is to assume that the liquefied soil has zero shear strength. If a total stress analysis is used, then the liquefied soil layers are assumed to have an undrained shear strength equal to zero ($s_u = 0$). If an effective stress analysis is used, then the effective shear strength parameters are assumed to be equal to zero.

5.1.7 Liquefaction induced lateral spreading

The concept of cyclic mobility is used to describe large-scale lateral spreading of slopes. In this case, the static driving forces do not exceed the shear strength of the soil along the slip surface, and thus the ground is not subjected to a flow slide. Instead, the driving forces only exceed the resisting forces during those portions of the earthquake that impart net inertial forces in the downslope direction. Each cycle of net inertial forces in the downslope direction causes the driving forces to exceed the resisting forces along the slip surface, resulting in progressive and incremental lateral movement. Often the lateral movement and ground surface cracks first develop at the unconfined toe, and then the slope movement and ground cracks progressively move upslope.

5.1.8 Empirical method

A commonly used approach for predicting the amount of horizontal ground displacement resulting from liquefaction-induced lateral spreading is to use the empirical method developed by Bartlett and Youd (1995). As stated in their paper, both U.S. and Japanese case histories of lateral spreading of liquefied sand were used to develop the displacement equations. Based on the regression analysis, two different equations were developed: (1) for lateral spreading toward a free face, such as a riverbank, and (2) for lateral spreading of gently sloping ground where a free face is absent. The equations are as follows:

Lateral spreading toward a free face:

$$\log D_H = -16.366 + 1.178M - 0.927 \log R - 0.013R + 0.657 \log W + 0.348 \log T + \\ + 4.527 \log (100 - F) - 0.922 D_{50} \tag{17}$$

Lateral spreading of gently sloping ground:

$$\log D_H = -15.787 + 1.178M - 0.927 \log R - 0.013R + 0.439 \log S + 0.348 \log T + \\ + 4.527 \log (100 - F) - 0.922 D_{50} \tag{18}$$

D_H: horizontal ground displacement due to lateral spreading, m;
M: earthquake magnitude of the design earthquake;

R: distance to the expected epicenter or nearest fault rupture of the design earthquake, km;
W: free face ratio, expressed as a percentage. The free face ratio is defined as $\frac{100H}{L}$, where H: height of the free face and L: horizontal distance from base of free face to location of the site;
T: cumulative thickness (meters) of the submerged sand layers having $(N_1)_{60} < 15$;
F: fines content of soil comprising layer T, expressed as percentage. The fines content is defined as the percent of soil particles, based on dry weight, that pass the No. 200 sieve;
D_{50}: grain size corresponding to 50 percent fines of soil comprising layer T, mm;
S: slope gradient, expressed as percentage. For example, a 20:1 (horizontal:vertical) slope has an angle of inclination of 2.9 and a slope gradient of 5 percent.

6 Conclusion

The slope stability can be affected by earthquakes by different forms; exceeded displacement, shear strength decreasing. The slope stability analysis for earthquake is done using one of two global types of comportment of the soil under the dynamic action of the earthquake. For the soils presenting a total or a partial liquefaction of the soil, the weakness slope stability is used, and for those soils where the liquefaction is not suspected, the inertial slope stability is used.
Considering the dynamic form of earthquakes, the reduction of shear strength is a natural consequence of the cyclic deformation applied.
The slope stability analysis is a complex problem approached by mathematical and/or experimental methods. The analysis by finite element method is also used, but the most usually methods are those described in this paper.

References

[1] Ambraseys, N. N. and Menu, J. M., *Earthquake-Induced Ground Displacements*, 1988.

[2] Day, R.W., *Geotechnical earthquake engineering handbook*, 2002.

[3] Duncan, J. M., and Wright, S. J., *Soil strength and slope stability*, 2005.

[4] Idriss, I. M., *Presentation Notes: An Update of the Seed-Idriss Simplified Procedure for Evaluating Liquefaction Potential*, 1999.

[5] Kyrou, K., *Seismic lope stability analysis*, 2000.

[6] Lambe, T. W. and Whitman, R. V., *Soil Mechanics*, Wiley, New York, 1969.

[7] Meehan, C. L., A*n experimental Study of the slickensided Surfaces*, PHD thesis. Virginia Polytechnic Institute, 2005.

[8] Marcuson, W. F. and Hynes, M. E., *Stability of slopes and embankments during earthquakes*, 1990.

[9] Newmark, N.M., *Effect of earthquake on dams and embankments*,1965.

[10] Sassa, K., *Prediction of earthquake induced landslides*,1996.

[11] Trandafir, A.C. and Sassa, K., *Newmark deformation analysis of earthquake-induced catastrophic landslides in liquefiable soils*, 2004.

[12] Seed, R. B., L*iquefaction Manual. Course Notes for CE 275: Geotechnical Earthquake Engineering*, College of Engineering, University of California, Berkeley, 1991.

[13] Tika-Vassilikos, T. E et al, *Seismic displacements on shear surfaces in cohesive soils*, 1993.

[14] US Army Engineering Corp, *Slope stability*, Engineering manual, 2003.

[15] VARNES, D.J., *Slope movement types and processes, in Landslides: Analysis and Control*,1978.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Coastal Dynamics and Coastline Management in Mamaia North-Navodari(Romania)

Mezouar Khoudir
Technical University of Civil Engineering, Bucharest, Romania

Boukhemacha Mohamed Amine
Technical University of Civil Engineering, Bucharest, Romania

Romeo Ciortan
IPTANA S.A. and University of Ovidius Constanta, Romania

George Paduraru
Faculty of Civil Engineering, Ovidius University of Constanta, Romania

Abstract

At present, erosion problems exist on important parts of the Romanian coast, to different intensities depending on the zone, but being especially important on some reaches of the south coast, where erosion rates of the order of 8 m/year are locally registered. The beaches of the Northern coast of Mamaia have been in a permanent erosion process that has dramatically increased in the past 10 years. Changes in the littoral dynamics, mainly due to human action, have generated a coastline regression rate, estimated at 1m per year and more. Studies of coastal morphodynamics are becoming increasingly more focused on quantification of relationships between processes, form and function of dynamic landform systems because wave climates (e.g., wave height, wave period, seasonality, cyclical patterns) and sediments (i.e., composition, size, and shape) interact in various ways to collectively produce distinctive types of beaches. In the paper, the Romanian shoreline evolution is analysed. The reasons of seashore degradation and the means for stopping these phenomena and for beach rehabilitation are described. After the failure of many "hard" measures for protecting sandy beaches (like groins, breakwaters or sea-walls) become evident from both the technical and economic point of view, the interest of using "soft" remedial measures like beach nourishment has largely increased.

Keywords: Mamaia beach, Romanian coast, Coastal regression

ISSN-1584-5990

1 Introduction

Studies of coastal morphodynamics are becoming increasingly more focused on quantification of relationships between processes, form and function of dynamic landform systems because wave climates (e.g., wave height, wave period, seasonality, cyclical patterns) and sediments (i.e., composition, size, and shape) interact in various ways to collectively produce distinctive types of beaches. The components of any coastal system will tend towards a dynamic equilibrium state, within which the energy inputs can be dissipated and a balanced state between activity, three-dimensional geometry and sediment transport is maintained with no net outputs to the system (Bruun, P., 1962 and Pethick, J., 1984) However, subsequent changes to this dynamic equilibrium can occur when further change in energy input (e.g. storm activity, sea-level rise), or human interference with either the sediment budget (e.g. coastal defence works) or coastal morphology (e.g. dredging, land claim, re-alignment of coastal defences) occurs. Human attempts at coastal defence, for example, can have serious impacts (both intentional and unintentional) on the input, transport and output of littoral sediments within coherent geomorphological coastal units (known as sediment cells and sub-cells) (Nicholas J. et al.2001). The fundamental implication associated with natural or artificial changes to any coastal system is that the previously existing dynamic equilibrium within the sediment cell will be altered. In many such cases, the energy inputs can no longer be dissipated without net output from the system, thus resulting in changes in the morphology and/or sediment characteristics and/or sediment transport processes elsewhere within the cell. It is, therefore, vitally important that any efforts at shoreline management should be based on a thorough understanding of the coastal geomorphology (e.g. form, function and processes) within relevant sediment cells in order to minimise disruption to the natural coastal system. Furthermore, consideration should also be given to the effect of shoreline management activities, over a range of timescales, on adjacent coastal frontages within the wider-scale geomorphological setting within which they are set (Nicholas J. et al.2001). In the Romanian littoral (Black sea) region of Mamaia , the migration problem from rural to urban areas since 1970s has caused several environmental and coastal problems. Unlawful and unconscious urbanization is one of the most important ones. As the urban population goes up, the need for various coastal structures and recreation areas also increases. These structures have broken the coastal balance and caused generally erosion and sometimes local siltation problems. In a study made for solving this problem, shore-parallel revetments (breakwater) were proposed in the southern region of Mamaia, and it was observed that these structures have diminished the damage to the highway. Especially in recent years, groins and offshore breakwaters have been successfully employed in coastal protection schemes. Especially groins are supposed to be useful in the studied region. The interest of using "soft" remedial measures like beach nourishment has largely increased. Once the beach erosion phenomenon has been defined, by identifying the causes (through a study on morphological processes) and the local interests (evaluating the different aspects related to safety,

recreation, environment and economy), the careful selection of the protective measures to be adopted is of primary interest. Nowadays beach nourishment, eventually integrated with some "supporting structures", is an increasingly important option.

2 The studied region

2.1 Location

The studied area (Mamaia North- Navodari) beach is situated in the south eastern extremity of Romania, near Constanta city, on a narrow sand bar, 250 - 350 m wide, between the Black Sea and Siutghiol. Mamaia is the largest touristic seaside resort of Romania, stretching 8 Km from north to south. It is formed by sandy material that originates from the Danube. Mamaia beach is facing east and is a natural low sandy beach characterised by gentle sloping underwater profile down to - 6 m. The beach consists of alluvial sediments (brought into the Black Sea by the Danube and transported to the beaches by combined wave action and the north to south flowing current along the Romanian coast) and biogenic shells sediments. The sand is fine and has a grey light colour. The site under study is located on the Northern of Mamia beach to Media harbour between the geographical coordinates 4420'N - 2838'15"E and 4415'N - 2837'30"E. Shoreline length of the region is nearly 03 km. General alignment of the shoreline is north to south.(see figure 1)

2.2 Dynamic characteristics

The coastal zone of Mamaia is under seasonal wind regime. The hind cast wind data by the European center for medium weather forecasting (ECMWF) has been analyzed for direction wise frequency distribution of wind speeds (see figure 2). Northerly winds form NNW to NE are predominant in the constanta area, occupying 37 percent of the total occurrences. Strong winds exceeding 10 m/sec appear frequently in the northerly direction.(see figure 3)

The most frequent waves landing on the Mamaia coast come from the north east sector with dominating winds as well as from the north sector.

Maximum wind speed is about 40 m/s. Maximum wave height during these storms is about 9.5 m and about 8 m near the shore. The North-South orientation of the Romanian shore the bathymetric contours determine the asymmetry of wave propagation. Winds from West have a confined fetch and wave crests run parallel to the shoreline because a refraction in the shallow water near the shore. The highest values of the average wave parameters are recorded for waves from the East direction perpendicular to the shore: length (Lm) is about 34 m, height (Hm) about 1.2 m and the period (Tm) about 5 sec (see fig.5)

The predominant wave direction is from NNE to E. which occupy 50 % of the whole waves. Waves from S and SSW are also present with the rate of 15 %

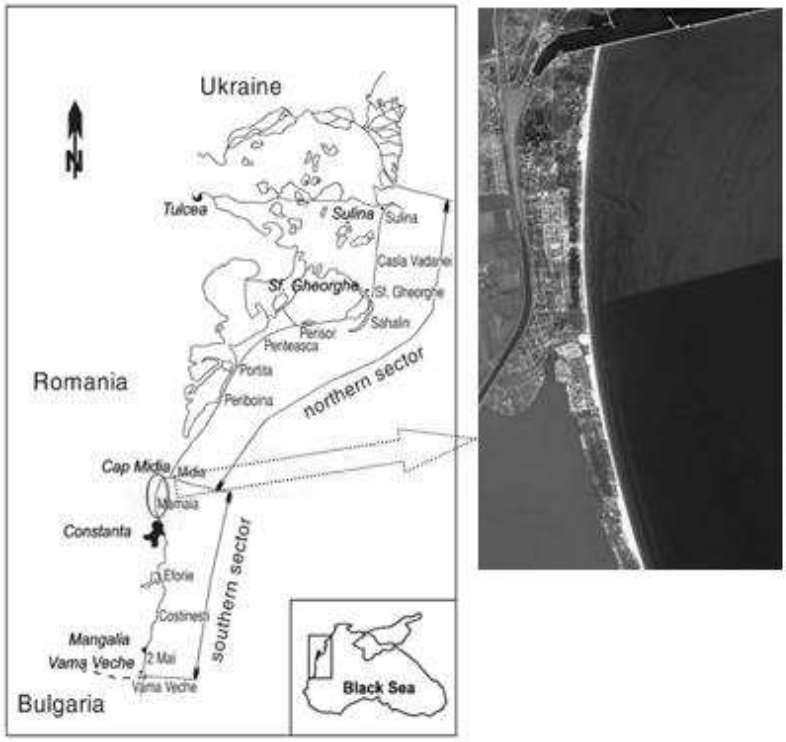


Figure 1: Studied area (Mamaia North- Navodari), Romania.

because the offshore hind cast location is open to the southwest with a certain fetch, but they are small in height and short in period.(see figure 4).

3 Erosion and shoreline evolution

Shore and shoreline evolution both due to natural and human-induced causes or factors can be variable over a wide range of different temporal and/or spatial scales. Our capability to understand and especially predict this variability is still limited. This can lead to misinterpretation of coastal change information, which hampers informed decision making and the subsequent design and implementation of (soft) engineering interventions. Collecting and describing example observations of shore and shoreline variability is one way to support and improve such human intervention. Efforts have been undertaken in the literature to quantify shoreline variability. A particular form of this variability is 'beach mobility', which was defined by (Dolan et al.1978) as the standard deviation of the shoreline relative to its linear trend. It has been suggested that this is a function of the morphodynamic state of the beach (Short and Hesp, 1982): dissipative, intermediate, and reflective beaches correspond to low-moderate,

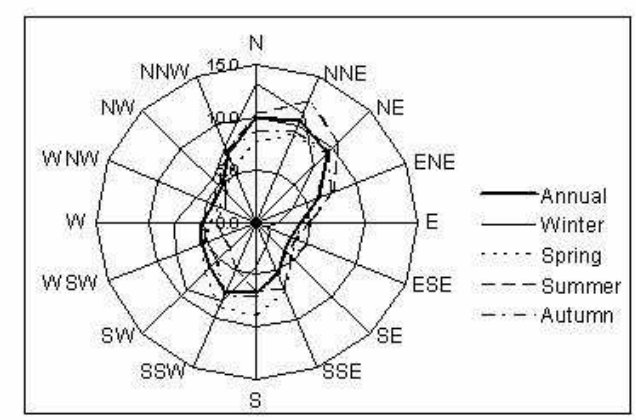


Figure 2: Wind hindcast by the European Centre for Medium-range Weather Forecast (ECMWF) 1991-2000.

moderate-high, and low beach mobility, respectively.

The figures 6 and 7 indicate that the variability of shorelines displays itself differently in space and in time and differently at the low water (LW), high water (HW), and dune foot position. By looking more closely at the causes and effects illustrated by case examples in the following discussion, we may try to understand and thereby quantify the shoreline variability more precisely. As an introduction, Table I and Table II list, respectively, natural and human causes and factors and the resulting typical coastal evolution trends by scale.

While, in principle, all, or almost all, typical evolutions may be associated with the causes and factors behind shore evolution, we have ordered the causes and the evolutions approximately by importance, following (Stive et al.1990). Note that for the evolutions associated with natural causes and factors, the typical evolutions are trends for the larger-scale and for the smaller-scale fluctuations. For the evolutions associated with human causes, these are trends and trend changes, respectively.

However, we would wish to state that generally speaking, the notion of relative importance might be difficult to justify. Dominance will vary from coastal section to coastal section. For instance, in Mamaia, a major long-term interference is sediment starvation due to: the Midia harbour extension dikes (5 Km long) . The dike deflects the longshore sediment drift offshore, to the south-east, bypassing Mamaia beach. The coastal cell of Mamaia beach was transformed in a bay, which almost totally lacks a natural sediment supply. The general decreasing of the sediment supply into the Black Sea as a result of Danube River damming adds to the problem. The development of hydraulic structures on the Danube and its affluents (dams, locks, bottom sills, etc.), and of the works against soil erosion, the alluvial flow at the river mouth was substantially reduced. The jetties for navigation maintenance at Sulina mouth of

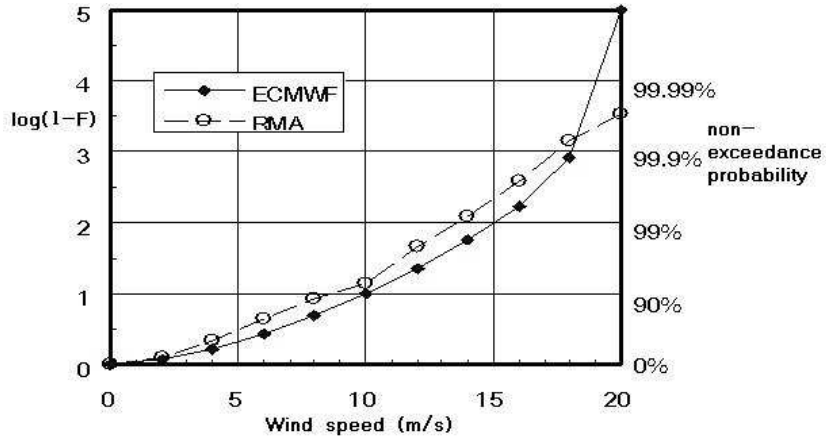


Figure 3: Non-exceedance curves of wind speed at Constanta

the Danube River advanced more than 8 km, determining a constant migration of sediment discharging point to areas of larger depths (under -15 m). At such depths, orbital wave motions are less effective in transporting the sediments to the shore; therefore, the alluviums are transported in open sea. However, this sediment load has an important role in replenishing the coastal sandbars, located southward of Sf. Gheorghe mouth of the Danube River. Transverse coastal structures, like the harbour breakwaters, have a negative effect as concerns the continuity of the littoral transport of sand, in the same way as groins cause upstream accumulation and downstream erosion. Another source of sand for seashore is an organic source, resulted from some species of shells. The water pollution may reduce the number of shells and, as a consequence, the sand supply of the beaches. The unbalance was created by some factors natural and some human activities, and particularly by:

- Global sea-level rise is projected to accelerate two to four fold during the next century, increasing storm surge and shoreline retreat along low-lying, unconsolidated coastal margins. The shoreline is particularly vulnerable to erosion and inundation due to the rapid deterioration of coastal barriers combined with relatively high rates of land subsidence. The local part introduces variations in relative sea-level rise along the Western Black Sea between 2 - 4 mm/yr. Relative sea-level rise is larger in the Constanta (3.3mm/year)than along the remainder of the shores.

- Cases of dune degradation in Romania have been analysed. The main causes for dune degradation in Romania consist of the following: Massive tourist development ;road and boulevard constructions ;dune mining; littoral drift interruption; dune recreational pressure; inadequate waterfront constructions; human trampling; off-road vehicles and parking lots; agricultural practices and afforestation; garbage dumping; water extraction;

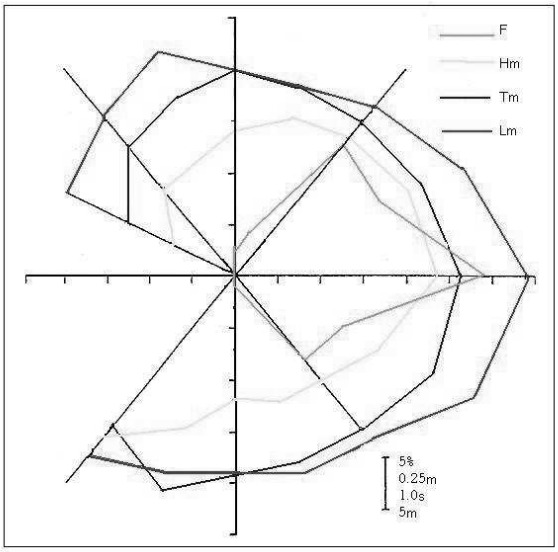


Figure 4: Waves parameters distribution depending on waves direction

civil engineering works.

- Waves are generated by offshore and near-shore winds, which blow over the sea surface and transfer their energy to the water surface. As they move towards the shore, waves break and the turbulent energy released stirs up and moves the sediments deposited on the seabed. The wave energy is a function of the wave heights and the wave periods. As such the breaking wave is the mechanical cause of coastal erosion in most of cases reviewed and in particular on open straight coasts. There is a 50% probability that over one year to have waves higher than 0.2m. The direction in which the waves move is NE-SE. Very close to the shore, the wave's regime is controlled by the beach inclination.

- Winds acts not just as a generator of waves but also as a factor of the landwards move of dunes (Aeolian erosion). This is particularly visible along some sandy coasts. The average wind speed in the NV region is between 6.5 and 5m/s. The main directions of the wind are N, V and S a greater weight having the N-V direction. During the summer months the predominant direction is SSE. The storms have a predominant N direction, with an average wind speed of 9.8m/s, during a period of time of 8 to 22 hours.

- Dams prevent natural sedimentation processes by restraining the flow of riverine fresh water, so reduce sand supply to the coastline and deltas. After the construction of the Portile de Fier I Dam in 1970, the total sed-

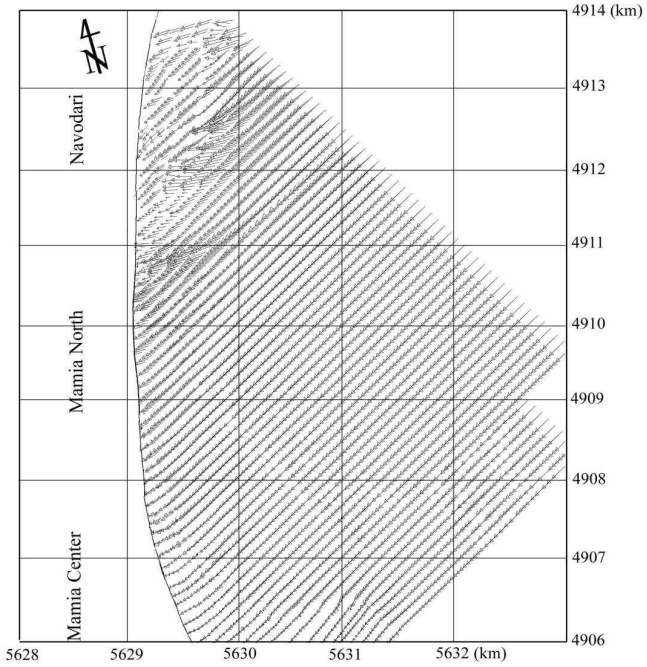


Figure 5: Propagation diagram of NE waves showing the good agreement of the wave rays direction after refraction.

iment discharge has diminished by approximately 30-40% .The sediment retention in the dam lakes is even higher, but the Danube partially compensates for the sedimentary deficit by eroding the riverbed (Panin 1996). the sediment load brought by the Danube into the Black Sea littoral zone has been drastically reduced from approximately 65 x 106 t in 1858 to 38 x106 t in 1988 (Panin 1996; Giosan et al. 1997).

- Harbour and other coastal engineering works: The changes occurred during the last century in the Danube Delta, which have altered the conditions of river drift towards the sea, and particularly regulation of the sulina channel. In relation to this objective, two breakwaters were executed in order to protect the fairway at the outlet into the sea, which gradually acquired a length of 7.5 km, thus relocating offward the drift unloading point, so that the not all the alluviums return to the alongshore drift circuit; these breakwaters and the bar that is formed constitute an obstacle for the northerly drift, particularly that of the Chilia Channel. The sedimentary deficit is even more intense because of the dredging of the Sulina bar and the dumping of this material approximately 4 km south-east of the end of the jetties. Maximum shore retreating distance was 145m in the case of Sulina-Sf. Gheorghe sector and maximum accretion was of 101m

Scale	Natural causes/factors	Typical evolutions
Very long term time scale: centuries to millenia; Space scale: ~ 100 km and more	-'sediment availability' - relative sea-level changes - differential bottom changes - geological setting - long-term climate changes - paleomorphology (inherited morphology)	- (quasi-)linear trends - trend changes (reversal, asymptotic, damping) - fluctuations (from (quasi-) cyclic to noncyclic
Long term time scale: decades to centuries; Space scale: ~ 10–100 km	- relative sea-level changes - regional climate variations - coastal inlet cycles - 'sand waves' - extreme events	- (quasi-)linear trends - fluctuations (from (quasi-) cyclic to noncyclic) - trend changes (reversal, asymptotic, damping)
Middle term Time scale: years to decades; Space scale: ~ 1 –5 km	- wave climate variations - surf zone bar cycles - extreme events	- fluctuations (from (quasi-) cyclic to noncyclic - (quasi-)linear trends - trend changes (reversal, asymptotic, damping)
Short term Time scale: hours to years; Space scale: ~ 10 m–1 km	- wave, tide and surge conditions - seasonal climate variations	- fluctuations (from (quasi-) cyclic to noncyclic - (quasi-)linear trends - trend changes (reversal, asymptotic, damping)

Table 1: Natural causes and factors and associated evolutions for shore and shoreline variability; see text for further explanation (based upon and adapted from (Stolk, 1989 and Stive et al., 1990)

distance near Sulina branch outlet.

The permanent extension of commercial activities led to the necessity of building enlargements to three harbours on the Romanian seashore. The protective sea walls of Midia harbour act against the natural development of the beaches situated to the South of them. The most important is the Midia Harbour jetty, which interrupts the southward transport of the Danubian sediment by the longshore current. Thus, these sediments are either deposited north of the harbour, or redirected offshore. Because of this, the entire southern sector of the Romanian littoral zone is almost completely deprived of Danubian sediments.

The shoreline distance shows a clear trend of linear retreat or advance over years, though seasonal variations appear on the linear trend. The rate of shoreline retreat or advance is defined from the gradient of the straight line of linear trend. Figure 6 shows the result of the shoreline change rate for the studied area. The beach of Mamaia south is experiencing a rapid shoreline retreat of - 2m/year, but the other zone (Mamaia north to Navodari) does not exhibit significant shoreline changes, it remains as stable with some low accretion a proximity of jetty of media harbour with (+1.79m/year) and about (+0.53m/year) at Navodari, and some low retreat at Mamaia north with (a maximum -0.42m/year).The retreat of the shoreline at these zones is probably caused by decrease of the wave dissipation function of the jetty.

Scale	Human causes/factors	evolutions
Very long term (time scale: centurie	- human-induced climate change - major coastal structure - structural coastal (non) management - major reclamations and closures - major river regulation	- (quasi-)linear trends - fluctuations (from (quasi) cyclic to - trend changes (reversal, asymptotic
Long term (time scale: decades to centuries)	- river regulation - coastal structures - natural resource extraction (subsidence) - coastal (non)management - reclamations and closures	- trend changes (reversal, asymptotic - (quasi-)linear trends - fluctuations (from (quasi-) cyclic to
Middle term (time scale: years to decades)	- surf zone structures - shore nourishments - shore protection	- trend changes (reversal, asymptotic -fluctuations (from (quasi-) cyclic to
Short term (time scale: hours to years)	- surf zone structures - shore nourishments - shore protection	- trend changes (reversal, asymptotic -fluctuations (from (quasi-) cyclic to

Table 2: Some typical human-induced causes and factors and associated evolutions for shore and shoreline variability; see text for further explanation (based upon and adapted from (Stolk, 1989 and Stive et al., 1990)

4 Future evaluation of the Mamaia shore

We present in this work, the evolution of the shoreline of Mamaia north to Navodari using a mathematical model. It is based on the discretization by finite elements of the equations controlling the propagation of the swell, the littoral current, transport transversely and the line of coast. This model is applicable under the following conditions:

- Morphological assumptions: the littoral must be uniform, i.e. that the bathymetric lines are almost parallel and that the relative transverse profile is in balance.
- Sedimentological assumptions: granulometry must be uniform and of the same dimension
- Hydrodynamic assumptions: the hydrodynamic factors such as the swell characterized by its period T, its significant height Hs and its direction of propagation compared to the transverse profile of the coast, must be constant.

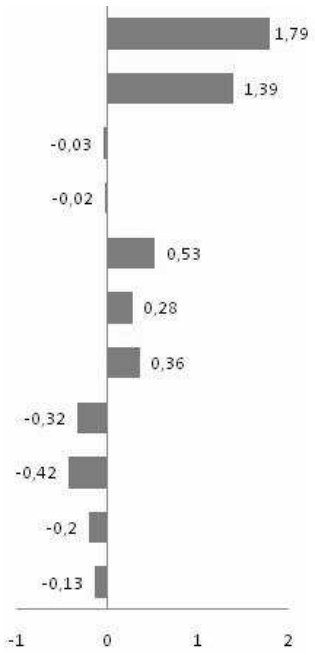


Figure 6: Shoreline variation

5 Intervention requirements

Coastal protection and flood defence techniques can be described in relation to the development of what are termed "hard and soft" engineering techniques. The hard engineering techniques involve the construction of solid structures designed to fix the position of the coastline, while soft techniques focus on the dynamic nature of the coastline and seek to work with the natural processes, accepting that its position will change over time.(see figure 8)

Starting from the 1980s, after the failure of many "hard" measures for protecting sandy beaches (like groins, breakwaters, jetties or sea-walls) became evident from both the technical and economic point of view, the interest of using "soft" remedial measures like beach nourishment has significantly increased. Once the beach erosion phenomenon has been defined, by identifying the causes (through a study on morphological processes) and the local interests (evaluating the different aspects related to safety, recreation, environment, and economy), the careful selection of the protective measures to be adopted is of primary interest. Nowadays beach nourishment, if necessary integrated with some "supporting structures", is an increasingly important option.

During a beach nourishment project, large volumes of beach-quality sand, called beach fill, are added from outside sources to restore an eroding beach. Or, a beach is constructed where only a small beach, or no beach, existed. Ultimately, beach nourishment widens a beach and advances the shoreline seaward.

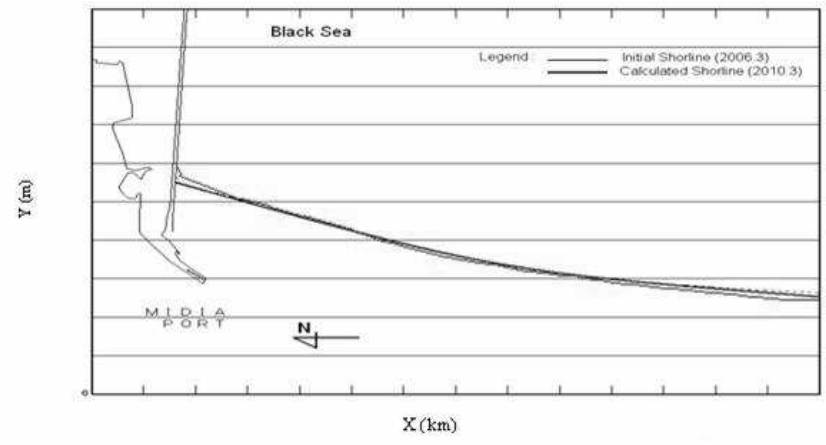


Figure 7: Expected evolution of the shoreline for 2010

From a morphological point of view there is not much preference where the sand is placed in the beach profile (provided it is between the breakerline and dune-foot or swash-line). The consequences of this sand placing:

- The gradual slope of the nourished beach causes waves to break in shallow water as they begin to feel bottom.
- As water rushes up the beach, wave energy dissipates.
- Water running back down the beach redistributes sediment, which is deposited in deeper water or moved along the shore.
- These sediments often create an offshore bar that causes waves to break farther offshore, again dissipating wave energy, and thus protecting people and property behind the beach.

To ensure that a nourished beach continues to provide protection and mitigate the effects of hurricanes and coastal storms, the project must be supplemented with additional quantities of sand, called periodic renourishment, as needed.

From the geometrical point of view, it can be shown that three types of nourished profiles are possible (Figure 9), depending on the volumes added and on whether the nourishment sand is coarser or finer than that originally present on the beach. These profiles are termed "intersecting," "nonintersecting," and "submerged," respectively (Dean, 1998). It can be shown that an intersecting profile requires the added sand to be coarser than the native sand, although this condition does not guarantee intersecting profiles, since the intersection may be at a depth in excess of the depth of closure. Nonintersecting or submerged profiles always occur if the sediment is of the same diameter or finer than the native sand.

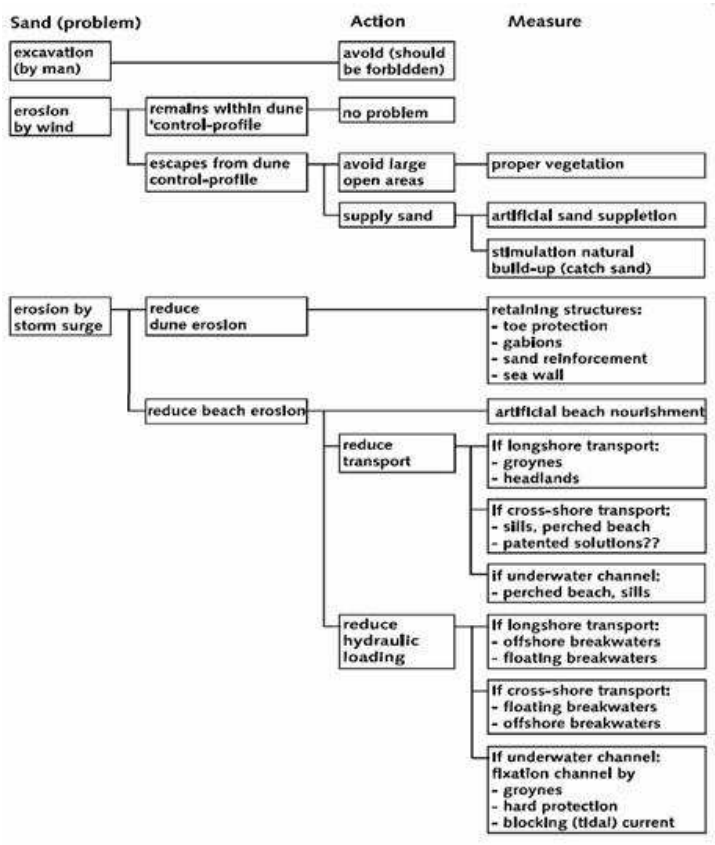


Figure 8: Schematic presentation of various shore protection measures (Krystian et al 2003)

During construction of a beach nourishment project, sand is placed so that natural coastal processes can reshape the nourished beach into the desired configuration as intended by coastal engineers The dry beach may seem overbuilt during construction, since sand is often placed on the shore at fairly steep slopes. After construction, it is normal for the newly nourished beach to readjust and change substantially within the first few months. Engineers expect modest waves to move and spread the sediment so that the nourished beach can begin assuming a more natural form. This sediment will continue to move offshore, so that larger waves are prevented from reaching the shore, and along the shore. This movement of sediment, while decreasing the width of the nourished beach somewhat, is not erosion; rather, it indicates that the project is performing as designed. Beach nourishment is a favoured means of beach protection for resort and high amenity beaches because it promotes amenity and natural character in the form of a wide sandy beach and, unlike some other structural measures,

generally it does not have adverse effects on adjacent areas of the coastline

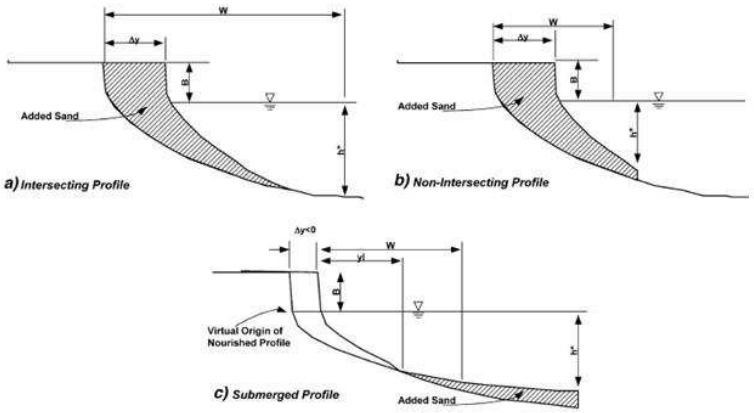


Figure 9: Intersecting, non-intersecting, and submerged profiles (Dean, 1998).

The rehabilitation of beaches by the artificial nourishment of sand is accompanied, on occasion, by auxiliary works, such as groins, detached breakwaters which are emergent or submerged. The aim of these works is to provide the new beach with a greater degree of frontal or lateral equilibrium. The new dimensions of the layout and the profile of the fill sometimes exceed the limits of the natural borders which provided stability to the original beach. To avoid the rapid deterioration of the newly nourished beach, certain works need to be constructed to give a complementary support to the sand, whether it be laterally, to avoid erosion caused by longshore current or frontally, to reduce the loss of sediment which causes the deformation of the profile. Groins are the structures used most frequently to provide lateral support for the beach fill. Breakwaters and Submerged Geotextile tubes placed parallel to the shoreline are the ones most used to contain the toe of the profile of a nourished beach. If the crest of the structures is close to the surface of the water or emerges, it can cooperate in the control of the energy of the waves which arrive at the beach and, thus, affect the sedimentary dynamics

In such realistic global situations, it may be more cost-effective to partly protect and/or to confine the artificial beaches with fixed structures which require less frequent nourishment. Partly-confined and/or sheltered beach types exist in the natural environment. These beach types can be emulated by man, namely:

- Reef beaches, sheltered by fixed or sacrificial offshore reefs which absorb some energy from the shorewards propagating waves, or

- Perched beaches, cross-shore confined by a low-crested or submerged artificial berm in the nearshore zone, or

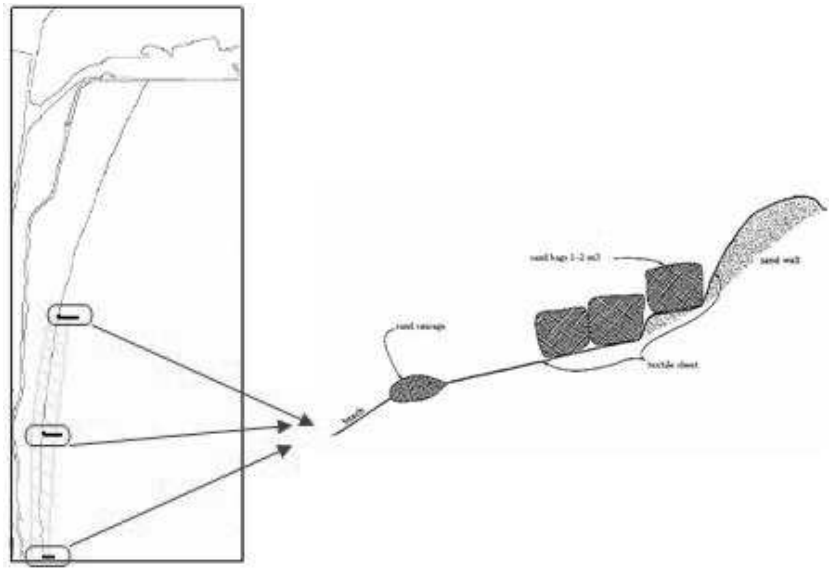


Figure 10: Shoreline protection with sand bags (variant 1)

- Pocket and tombolo beaches, of which the plan-form is governed by detached and/or shore-connected fixed breakwater structures, or

- The use of large sand-filled bags. Bags with a volume of m3 and 1 m3 and manufactured from a variety of textiles will be used. The bags will be placed in 2 different configurations (figure 10) against the sand wall, or

- The use of large sand-filled bags and sheets of textile to form an artificial sand bar. Low sand dune/bar will be constructed parallel to and in front of main sand wall. This sand bar will be covered with a textile sheet, the edges of which were pinned down with sand bags (figure11), or

- Geotextile T-groins will be constructed from sand bags (figure12). The central row of bags in the head of the T will be partially wrapped in a sheet of textile. Construction will be carried out in the surf zone in depth of about 1m.

- Geotextile bags, 'sausages' or tubes filled with sand are now gaining wide acceptance in coastal protection. They have been used as nearshore breakwaters (placed parallel to shore). Nearshore, low geotextile breakwaters

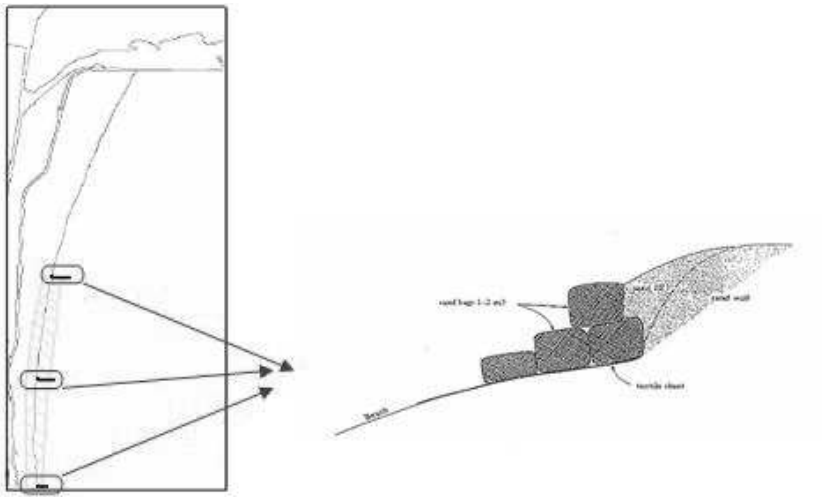


Figure 11: Shoreline protection with sand bags (variant 2)

are designed to a height sufficient enough to eliminate storm waves from reaching the shoreline but allow smaller waves to penetrate (figure13).

The design and placement of the geotextile breakwaters takes into account the height of the incident waves, depth, tidal range and site conditions.

Once the littoral process was evaluated, a tube cross section was designed and geosynthetic materials were defined in terms of their mechanical properties. The following considerations were mandatory:

1. Stresses on geosynthetics are very sensitive to the slurry pumping pressure when the tubes are filled. This pressure governs the criteria design for defining the estimated force of the required geosynthetics, working under load conditions.

2. Slurry pumping pressure does not have a significant influence on the final sectional area of the tube.

3. The apparent opening size of the geotextile is conditioned by sediment diameter D50.

4. Inlets separations are defined in terms also of D50. The larger the sediment diameter D50, the closer the inlets are.

5. The ultimate strength of required geosynthetics must consider, reduction factors for installation damage, chemical and biological degradation, treachery creep, and seam strength.

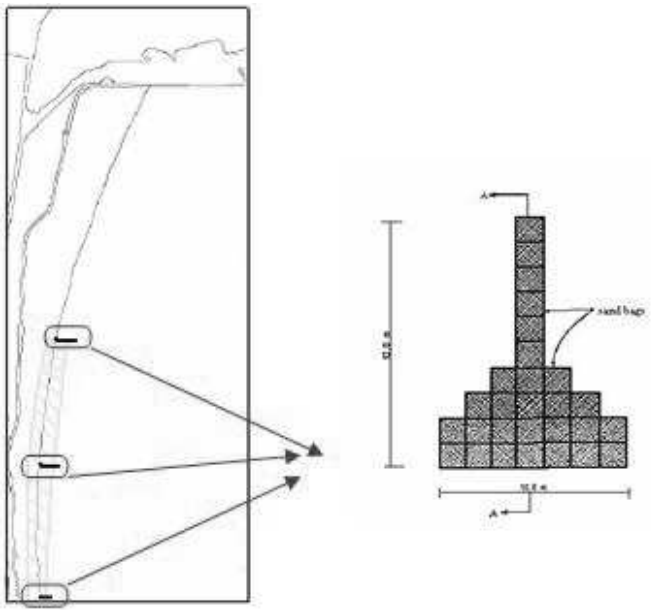


Figure 12: Shoreline protection with sand bag groin (variant 3)

$$T_{ult} = T_{work}(RF_{id}RF_dRF_cRF_{ss}) \tag{1}$$

Where T_{ult} is the ultimate strength of the required geosynthetic, T_{work} is the calculated tensile force under load conditions, and RF_{id}, RF_d, RF_c and RF_{ss} are the reduction factors for installation damage, chemical and biological degradation, creep, and seam strength, respectively.

6 Conclusion

The Romanian coast has experienced significant erosion and shoreline recession in historical times, but there is no evidence that rates of coastal recession and/or frequency of flooding have accelerated in recent decades as a result of sea level rise, increased storminess, or any other factor. The beaches of the Northern coast of Mamaia have been in a permanent erosion process that has dramatically increased in the past 10 years. Changes in the littoral dynamics, mainly due to human action, have generated a coastline regression rate, estimated at 1m per year and more. Tp cope ith these problems, protection works of the breakwater type have mainly used in the southern of Mamaia coast. The constructions of hard coastal defences (groins, detached breakwaters, sea-walls) are not always the optimal solution for the prevention from the beach erosion. They may even displace the coastal erosion problem from one area to another. A soft solution, of

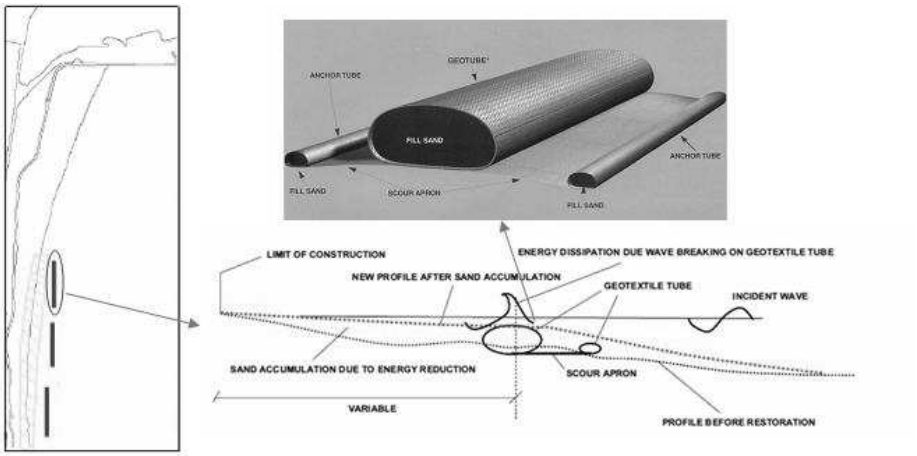


Figure 13: Schematic section of wave energy reduction, geotube (variant 4)

the beach nourishment type is recommended if and when economically feasible; otherwise the retreat or the abandonment of the coast is considered acceptable

References

[1] Bruun, P., 1962. Sea level rise as a cause of beacherosion. Proceedings ASCE. Journal of Waterway, Port and Coastal Engineering Division, 117-30.

[2] Chertic, E. et al.,1992. Studiul dinamic al caracteristicilor meteorologice pentru furtunile din bazinul vestic al Marii Negre in scopul determinarii campului vantului. Posibilitati de modelare si prognoza, Studii de Hidraulica, XXXIII, Minist. Mediului, ICIM, Bucuresti, pp.77- 103.

[3] Dean, R.G., Chen, R., Browder, A.E., 1997. Full scale monitoring study of a submerged breakwater, Palm Beach, Florida, USA. Coastal Engineering 29, 291-315.

[4] Dolan, R., Hayden, B.P., Heywood, J., 1978. Analysis of coastal erosion and storm surge hazards. Coast. Eng. 2, 41-53.

[5] Giosan L., Bokuniewicz H., Panin N., Postolache I. 1997. Longshore sediment transport pattern along Romanian Danube Delta coast. GeoEcoMarina 2, 11-23.

[6] JICA.,2006. studiul privind protectia i reabilitarea litoralului sudic al romniei la marea neagra

[7] Krystian,P. et al.,2003. Integrated approach and future needs in coastal. engineering: general remarks International Conference on Estuaries and Coasts, Hangzhou, China

[8] Nicholas, J. et al.,2001. Predicting coastal evolution using a sediment budget approach: a case study from southern England. Journal of Ocean and Coastal Management 44

[9] Panin, N.,1996. Danube Delta: Genesis, evolution and sedimentology. GeoEcoMarina 1, 11-34.

[10] Panin, N., 1996. Impact of global changes on geoenvironmental and coastal zone state of the Black Sea. GeoEcoMarina, 1, 1-11.

[11] Pethick, J., 1984. An introduction to coastal geomorphology. London: Edward Arnold (Pub.), 260pp

[12] Short, A.D., Hesp, P., 1982. Wave, beach and dune interaction in southeastern Australia. Mar.Geol. 48, 259- 284.

[13] Stive, M.J.F., Roelvink, J.A., De Vriend, H.J., 1990. Large-scale coastal evolution concept. The Dutch Coast. Proc. 22nd Int. Conf. Coast. Eng., vol. 9. ASCE, New York, pp. 1962- 1974.

[14] Stolk, A., 1989. "Sand system coast- a morphological characterisation-coastal defence after 1990" (in Dutch). Report GEOPRO 1989.02, p. 97.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Theoretical and Practical Methods Regarding the Absorbitors of Oscillations and the Multi-model Automatic Regulation of Systems

[1]Mircea Lupu, [2]Olivia Florea, [3]Ciprian Lupu
[1,2]Transilvania University of Brasov, Faculty of Mathematics and Informatics, Brasov, Romania
[3] "Politehnica" University of Bucharest, Faculty of Automatics and Computer Science, Bucharest, Romania
E-mails: [1]m.lupu@unitbv.ro, [2]olivia.florea@unitbv.ro, [3]cip@indinf.pub.ro

Abstract

It was modelled a hydraulic - system, regarding the dissipation of some discontinuous oscillations for obtaining the asymptotic stability. The amortization system was make by using the sensors with blocks of calculus and electronic control for the mathematical system at the input and output.

In the first part of this paper it's made the study regarding the dissipation of some discontinuous oscillations for obtaining the asymptotic stability. It is modeled a hydraulic - pneumatic system for the oscillations of rolling (level) absorption, like a response for the fluid and dry amortization.
The amortization system is a pumping servo-mechanism of the alternative fluid in two tanks: using the sensors with blocks of calculus and electronic control for the mathematical system at the input and output.
In the second part of this paper is presented a multi-model automatic regulation structure. This model is more advantageous compared to the classically. It is presenting the theoretical method close-loop, identification, R-S-T control algorithm and the adaptive control for a nonlinear dynamical system in the case of the level control for the fluid sloop for a tank to the other.

Keywords: R-S-T algorithm, closed-loop identification, adaptive control.

1 The Tantal's basin

We'll start to present the utility of the hydraulic automatic regulation with a first legendary example: "The Tantal's basin" from the Antique Greece. This

ISSN-1584-5990

was the first hydraulic system for the automatic regulation of the flow process based on auto-oscillations. The filling and discharge of the basin has a practical importance by these periodical effects. The legend: The monk Tantal has violated the rules and he was punished, tied by his legs near the basin, so the maximum level overtakes under his chin. In the maximum momentum, when he wants drink some water, the basin starts to discharge; so he make many bows.

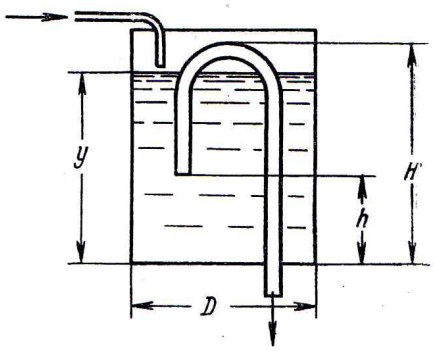


Figure 1: The Tantal's basin

The entire process has two phases:
1. In the basin the flow rate q is constant
2. When the fluid reaches the level H, in the basin it starts the discharge through the second pipe with the flow rate $Q > q$. When it reaches the level h both phases are repeating.

2 The amortization of the ships

It is known that in their movement, the maritime or aviation ships and the dynamical systems from mechanics are perturbed (by the waves, wind or mechanical parameters). In this way are forced to make the rolling oscillations represented by the angle of rotation $\theta(t)$ side by a fixed mark. The attenuation

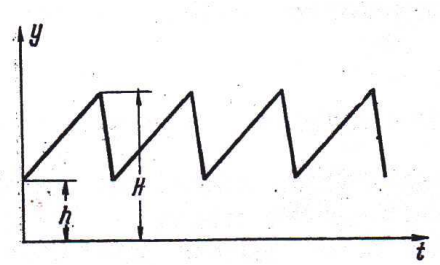


Figure 2: The auto-oscillations graphic of the fluid level in the Tantal's basin

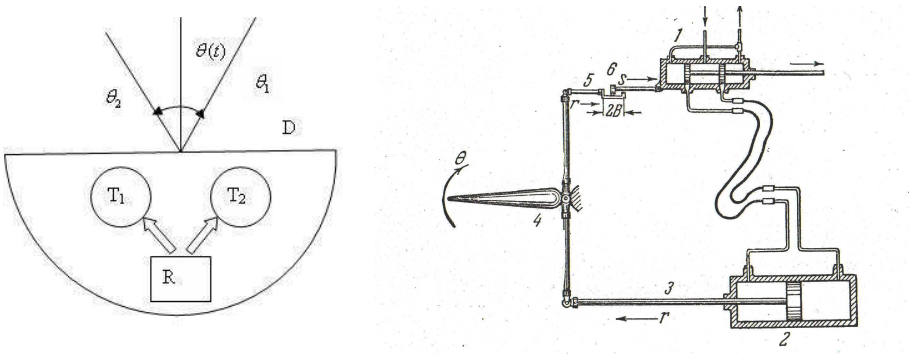


Figure 3: a)A pump connected to two tanks; b) The scheme of the hydraulic regulator

(the dissipation) of these oscillations to obtain again the stabile regime of movement it is made by using some pumps leaded servo-mechanics and electronic - to transfer the fluid by contra balance in two tanks mounted symmetrically of the longitudinal and vertical axis. In this mode of discharge and filling of the tanks in contra balance with the amplitude it is obtain a spell effect by the friction and the variation of the fluid mass (left - right), involving the asymptotic stability with intermittences. In the phase plane the spiral trajectories are wavy with the change of the movement sense tend to the stable asymptotic focus [1].

It is thought that R is the hydraulic regulator (fig. 1b)) is the schematic figure and in the figure (fig. 1a) R is lied by the two tanks T_1, T_2 connected by the pipes acting by the pumps P.
Simplifying the model, the equation of the system with the free degree $\theta(t)$ will be:

$$J\ddot{\theta} = -M\dot{\theta} - Kr \tag{1}$$

Where J is the inertia momentum of the ship, relating the vertical axis which goes through the load centre, $\theta(t)$ is the rolling angle side by the $\theta = 0$; $M\dot{\theta}$ is the linear momentum of the amortization by the friction of the viscous fluid; r is the displacement of the mechanism arm owing to the rolling. This is proportional of the θ angle, so that $Kr = N\theta$. This is considering to be the reaction momentum necessary that the ship come back to the normal course. The rolling is signalized by the θ angle in the gyroscopically quadrant mounted on the ship. So, through the variation of this angle θ, the server (1) is involved by the pump (ex: to the right) and is opening the input of the fluid in the half right of the servo-motors (2). This produce the movement r of the bar (3) proportional with the angle θ, which will be signalize by the rolling indicator (4), [2].
Because of the momentum obtained $N\theta$ like a response, the ship comes back to the normal course, making an amortized oscillation. For accelerating this amortization the hydraulic absorber must delimitate the rolling angle $\theta(t)$. In

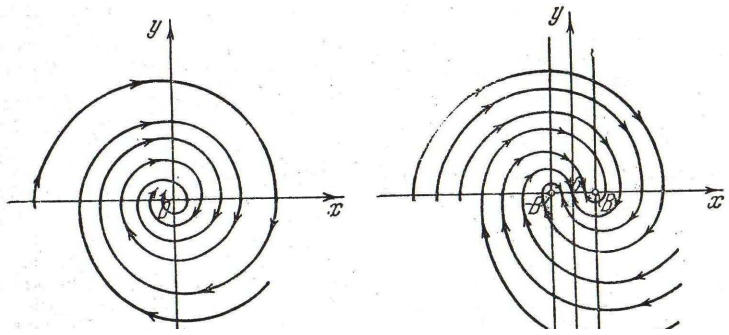


Figure 4: The trajectory in the phases plane: a) the homogeneous case; b) the non-homogeneous case

this way, this system is adding up an "inverse link" (5). The cylinder of the server acquires beside the rolling rotation to the right a double pressure. The server's valves are closing and the acting of the server is stopped. The reasoning is repeat acting to the left with the discharge of the right tank and the filling of the left one. So, we'll have a bigger amortization because of the momentum $N\theta$ and after a finite number of oscillations witch are amortizing by decreasing the amplitude, the ship is stabilized to the null solution [3].
The link is with "free bearing clearance" and for coulisse (6) an the $2B$ distance and the body' s server oscillating on the space $\pm B$, this mean $r \pm B \approx \theta$ or $r = \theta \pm B$ with the + sign for $\dot{\theta} > 0$ and $-$ for $\dot{\theta} < 0$.

$$\ddot{\theta} + 2r\dot{\theta} + k^2\theta = \pm k^2 B \tag{2}$$

In the phases plane the study of the movement is make thus: we are considering the movement from the right to the left with $\theta \in (-B, B)$. Noting $\theta = x_1 + B$:

$$\ddot{x}_1 + 2r\dot{x}_1 + k^2 x_1 = 0 \tag{3}$$

For $n^2 < k^2$ we have a low resistance and the solution is amortized linearly side by $x \equiv 0$ Having a displacing to the right face to the origin with B. The decrement of the oscillation is $\delta = e^{-n\pi/k_1}, k_1 = \sqrt{k^2 - n^2}$. If the displacement is from the left to the right $\theta = x_2 - B$, we obtain for x_2:

$$\ddot{x}_2 + 2n\dot{x}_2 + k^2 x_2 = 0 \tag{4}$$

The amortization law is the same like (3), but the oscillation will be displaced to the left with B.
So, the trajectory in the phases space will be obtain traced at the first time the logarithmic spiral in the (x, y) plane, $\dot{x} = y$ with the solution for the case (fig. 2a)

$$x = e^{-nt}\left(C_1 \cos k_1 t + C_2 \sin k_2 t\right) \tag{5}$$

And then, $\theta = x \pm B$ (fig. 2b) for the non-homogeneous equation (2); graphically it was made a cut on the $x'Ox$ axis and we are displacing the upper figure to

the right with B and the figure from the under half plane to the left with B. the spirals are merged by continuity and will be undulate to the origin thus: considering the initial position at $t = 0$, $\dot{x}_0^i > 0$ then with recurrence to the left:

$$x_0^{i+1} = (x_0^i - B)\delta - B = x_0^i\delta - B(1+\delta) \tag{6}$$

And the point will go to the left side of the origin if:
$x_0^i\delta - B(1+\delta) > 0$ or $x_0^i > B\left(1+\frac{1}{j}\right) > 2B$.

The (6) formulae is conditioned by $0 < x_0^i \leq B$. If $B < x_0^i \leq B\left(1+\frac{1}{j}\right)$ the trajectory falls to the origin through the inferior side without cross in the interval $(-B, B)$, see (fig. 2b).

The presented previous solution has need of the exact knowledge of the process parameters (the tanks, pumps, pipes, etc.) a difficult fact of realize. The practical control of the level in the ballast tanks, can have as solution an adaptive multi-model system [5]. The functionality of such system implies solving the next problems: the choice of the better model, the identification in the close loop of the adaptive model, the recalculation of the regulation algorithm. This article details some of the essential aspects of these problems.

3 Choice of the model

The model-error at the k instant is defined as the difference between the output y_i of the model M_i and the output y of the plant [4], [5], [6]:

$$\varepsilon_i(k) = y(k) - y_i(k) \tag{7}$$

The performance criterion which is used as the selection rule is defined below:

$$J_i(k) = \alpha\varepsilon_i^2(k) + \beta\sum_{j=1}^{k} e^{-\lambda(k-j)}e_i^2(j) \tag{8}$$

where $\alpha > 0$ and $\beta > 0$ are the weighting factors on the instantaneous measures and the long term accuracy; $\lambda > 0$ is the forgetting factor.

The choosing of the $\alpha, \beta$and λ parameters depends of the plant:

- $\alpha = 1$ and $\beta = 0 \rightarrow$ for the fast systems (good performances with respect to parameters changes, sensitive to disturbance);
- $alpha = 0$ and $\lambda = 0 \rightarrow$ or the slow systems (bad performances with respect to parameters changes, good performances with respect to disturbance).

4 Closed-loop recursive identification

A closed-loop adaptive method (filtered closed loop error- FCLOE identification) for the adjustable predictor is considered. This method computes the

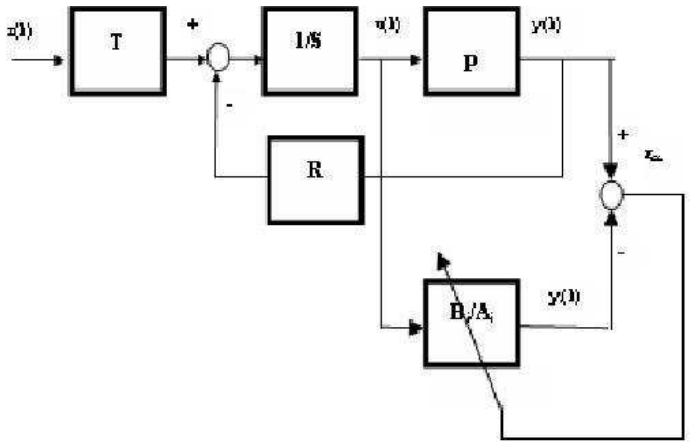


Figure 5: Close loop identification technique

parameters of the model in order to minimize the closed loop output prediction error ε_{CL} using the filtered data u and y. A FCLOE identification scheme is presented in Fig. 3:

The basic idea is to substitute (by filtering of u and y) the prediction error ε_{LS} with closed-loop output error ε_{CL}. The filter depends of the control algorithm. The FCLOE - algorithm in least squares recursive form is the following:

$$\theta(k+1) = \theta(k) + F(k)\phi_f(k)\varepsilon_{LS}(k+1) \tag{9}$$

$$F(k+1) = F(k) - \frac{F(k)\phi_f(k)\phi_f(k)^T F(k)}{1+\phi_f(k)^T F(k)\phi_f(k)}, F(0) = \alpha I, \alpha > 0 \tag{10}$$

$$\varepsilon_{CL}(k+1) = \frac{y(k+1) - \theta^T(k)\phi_f(k)}{1+\phi_f(k)^T F(k)\phi_f(k)} \tag{11}$$

where, $\theta(k)$ is the parameter vector; $\phi_f(k)$ is the filtered observation vector; $F(k)$ is the gain adaptation matrix; ε_{CL} is the closed-loop prediction error.

5 Model based control (re)design

For the M_i model we design a controller C_i that satisfies the desired nominal performances. The RST polynomial algorithm with two degrees of liberty, for C_i controller is proposed (see Fig. 4):

In this case the input $u(k)$ is:

$$u(k) = \frac{T(q^{-1})}{S(q^{-1})}r(k) - \frac{R(q^{-1})}{S(q^{-1})}y(k) \tag{12}$$

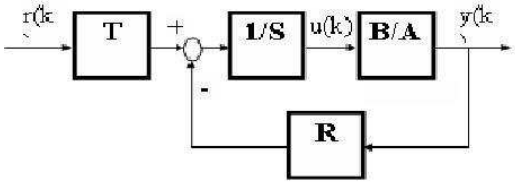


Figure 6: RST control algorithm

The disturbance rejection is ensured by the $R(q^{-1}), S(q^{-1})$ polynomials, obtained solving the equation:

$$P_C(q^{-1}) = A(q^{-1})S(q^{-1}) + B(q^{-1})R(q^{-1}) \tag{13}$$

where,

- pair $(A(q^{-1}), B(q^{-1}))$ represents the plant model;
- $P_C(q^{-1})$ is the closed-loop characteristic polynomial.

The reference tracking performance is ensured by the choice of the $T(q^{-1})$ polynomial. For each model $((A_i, B_i)$ a C_i control algorithm $(R_i, S_i, T_i$ polynomials) will be computed respectively.

The adaptive pole placement method is used for achieved performances in closed-loop.

There are two possibilities for the adaptive design approach:

5.1 Disturbance rejection adaptive algorithm:

1. Re-identification of the model M_{k+1} using the relation (9), where the filtered data is $\Phi_f(k) = \frac{S}{P_C}\Phi(k)$.

$$M_{k+1} = \frac{B_{k+1}(q^{-1})}{A_{k+1}(q^{-1})} \tag{14}$$

2. Evaluation of the pair $R_{k+1}(q^{-1}), S_{k+1}(q^{-1})$from equation:

$$P_C(q^{-1}) = A_{k+1}(q^{-1})S(q^{-1}) + B_{k+1}(q^{-1})R(q^{-1}) \tag{15}$$

3. Computation of the input $u(k+1)$:

$$M_{k+1} = \frac{B_{k+1}\left(q^{-1}\right)}{A_{k+1}\left(q^{-1}\right)} \tag{16}$$

5.2 Reference tracking adaptive algorithm:

1. Identification of the modelM_{k+1}:

$$M_{k+1} = \frac{B_{k+1}(q^{-1})}{A_{k+1}(q^{-1})} \tag{17}$$

2. Computation of $P_{Ck+1}(q^{-1})$ using the equation:

$$P_{Ck+1}(q^{-1}) = A_{k+1}(q^{-1})S(q^{-1}) + B_{k+1}(q^{-1})R(q^{-1}) \tag{18}$$

3. Computation $T_{k+1}(q^{-1})$ with relation:

$$T_{k+1}(q^{-1}) = \frac{P_{k+1}(1)}{B_{k+1}(1)} P_{C\ k+1}(q^{-1}) \tag{19}$$

4. Computation of the input $u(k+1)$

$$u(k+1) = \frac{T_{k+1}(q^{-1})}{S(q^{-1})} r(k) - \frac{R(q^{-1})}{S(q^{-1})} y(k) \tag{20}$$

The main experimental results from real time multi-models adaptive control system will be presented below.

6 Experimental results

We have evaluated the achieved performances of the adaptive multi-model control using an experimental installation as in Fig. 5. The main goal is to control in closed loop the level in Tank 1. There is a nonlinear relation between the level L and the flow F.

$$F = a\sqrt{2gL} \tag{21}$$

We consider three plant operating points P_1, P_2, P_3 on the nonlinear diagram $F = f(L)$ as in Fig. 6. The level values L_1, L_2, L_3 can be considered the set - points of the nominal level control system.

7 Conclusions

In this paper it was modeled a hydraulic - system, regarding the dissipation of some discontinuous oscillations for obtaining the asymptotic stability. The amortization system was make by using the sensors with blocks of calculus and electronic control for the mathematical system at the input and output.

An application of the multiple models adaptive control to a nonlinear model plant has been presented. A mechanism based on the performance model-error criterion for the choice of the best model in switching phase is considered. The multiple model adaptive control procedure proposed has the following advantages: a more precise model is chosen for the closed loop operating system, the R-S-T adaptive control ensures very good real time results for closed loop nonlinear system.

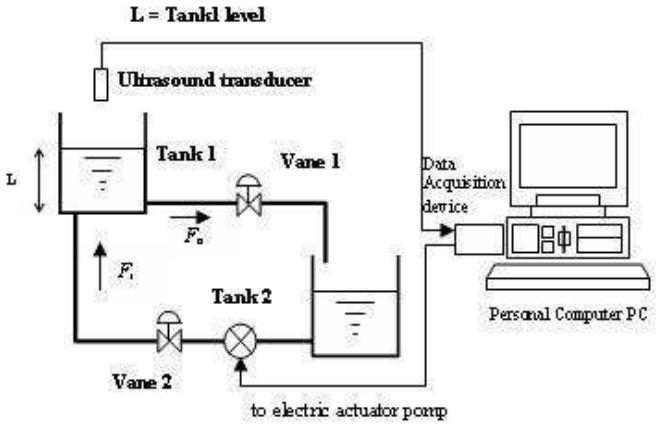


Figure 7: Experimental installation

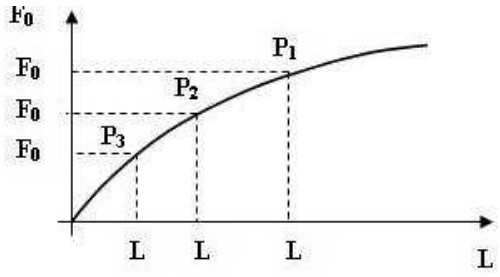


Figure 8: Plant operating points

8 Acknowledgement

This work was partially supported by CNCSIS IDEI Program of Romanian Research, Development and Integration National Plan II, Grant no. 1044/2007.

References

[1] Chiriacescu S., *Mechanical linear systems*, Ed. Acad. Romane, Bucuresti, 2007.

[2] Lupu M., Isaia F., *The mathematical modelling and the stability study of some speed regulators for nonlinear oscillating systems*, Analele Universitatii Bucuresti, Anul LV(2006), Nr. 2, pp. 203-212

[3] Florea O. , *Numerical and analytical methods for the dynamical system problems applied in the simulation of some hidraulic systems*, PhD thesis, Universitatea Transilvania din Brasov, 2009

[4] Lupu C., Popescu D., s.a *Sisteme de conducere a proceselor industriale.* Ed. Printech, Bucureşti, 2004

[5] Lupu C., Petrescu C., *Multi-models adaptive control*, UPB Bulletin, series C, 2006.

[6] Dumitrache I., *The engineering of automatic regulation*, Ed. Politehnica-Press, Bucuresti, 2005

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Frequential models for the precipitation evolution in Romania

Carmen Maftei and Alina Barbulescu
Ovidius University of Constanta, Romania
cmaftei@univ-ovidius.ro, alinadumitriu@yahoo.com

Abstract

The purpose of this article is to model the precipitation for a small Romanian watershed and to determine the IDF – curves.

Keywords: IDF - curves, Gumbel model

1 Introduction

Dimensioning the projects concerning hydraulic structures or water work projects implies to know the flood hydrograph associated with a return period (frequency). Since practically it is not economic to dimension the hydraulic works for the most intense precipitations, it is necessary to determine the optimum flow putting in balance the cost of an over sizing hydraulic work and the damage produced by a weak hydraulic structure. To solve this problem it is necessary to determine the maximum intensities of a rain having a frequency of a given event.

For small basins, the most used method to estimate the maximum annual discharge starting from the rainfall intensity is the "Rational Method". According to it the rainfall intensity (I) is considered for a duration that is at least equal to the time of concentration (T_c) of the basin. This means that for punctual rainstorms, a relationship between the intensity - duration - frequency (IDF curves) has to be established.

For a given return period (non-exceedable probability), a set of IDF curves represents the variation of the maximum annual rainfall intensity with the time interval length.

The purpose of this study is mainly to produce IDF - curves for Voinesti catchment, which is a part of the catchment Dambovitza (tributary of first order of Danube) in the Western of Sub Carpathians of Curvature.

The rainfall intensity data are deduced from the rainfall pluviograme originating from siphoning type recording rainfall gauges with the rim of the receiver at a level of 1.50 m above the ground, with an opening of 2.5 dm^2 and siphoning at an amount of 10 mm.

ISSN-1584-5990

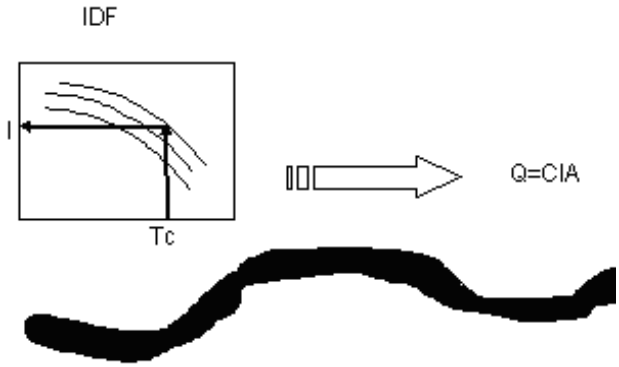


Figure 1: The schema of estimating the maximum annual discharge

The number of rainfall events recorded on Voinesti catchment was 132. The measument period was April - October. The maximum rainfall depths were searched for the durations of 5, 10, 15, 20 min.

2 Method

We propose to use the "frequency analysis" concept applied to the IDF problems.

The frequency analysis is a statistical method of prediction consisting in the study of the characteristics of a given process, in order to define the probabilities of their future appearance. The frequency analysis calls upon various statistical techniques and consists of a complex mechanism which is advisable to treat with much harshness.

The stages of a frequential analysis can be schematized in:

- Building the data series;
- The data series control;
- The choice of a frequential model;
- The adjustment control.

To understand the processes that intervene in the water cycle and to study their spatial and temporal variations, a database is essential. Building a database series is a process during which many errors could be made. On a first inspection, some of these may be identified and corrected at once, some are noted and marked, and others may remain undetected.

According to the errors' nature, various techniques and methods can be used. They are: "In situ", statistical investigations based on specific hypotheses and statistical test.

In our study, the control of the data series was made by statistical tests.

Gumbel's distribution is a special case of the Generalized Extreme value (GEV) distribution. It used in industrial applications and environmental sciences to model extreme values associated with flooding and rainfall. Gumbel's distribution has the form:

$$F(x) = \exp\left(-\exp\left(-\frac{x-a}{b}\right)\right), \tag{1}$$

where a and b are the model parameters.

Defining the reduced variable $u = \frac{x-a}{b}$, the relation (1) can be written:

$$F(x) = \exp(-\exp(-u) \tag{2}$$

and

$$u = -\ln(-\ln F(x)). \tag{3}$$

The advantage of reduced variable is that the quantile expression is linear.

In the case of an adjustment according to Gumbel's law, the quantile expression corresponds as graphical representation to a straight line. Consequently, the data points of the series that must be adjusted can be represented in a system of axes $x - u$; then it is possible to plot the straight line which fits the data and to deduce from it the parameters a and b which characterize the law.

So, the essential idea is to estimate the probability of un-surpassing $F(x_i)$ ascribable to every value x_i.

There are many formula to estimate the repartition function $\hat{F}(x)$ using the empirical distribution. They are based on the increasing ordering of the data series and the association of a rank r the each value.

The simulations made proved that for Gumbel's low the best results are given using the empirical frequency of Hazen:

$$f = \frac{r - 0.5}{n}, \tag{4}$$

where r is the rank and n is the sample volume.

In despite of the fact that it is an approximating method, the graphical adjustment has the big advantage to give a visual representation of the data and adjustment. It constitutes an essential aspect of the judgement on the adequation between the chosen law and the data, for any adjustment method used.

3 Results

3.1 The control of the series values

The data series are formed by the values of maximum intensity, at an interval equal to 5, 10, 15 and 20 min.

It can be seen that the maximum intensity values are as smaller when the intervals are longer.

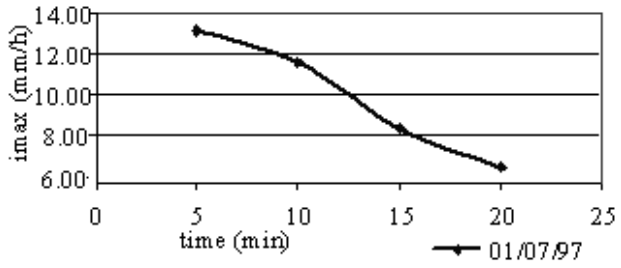


Figure 2: Example of a data series

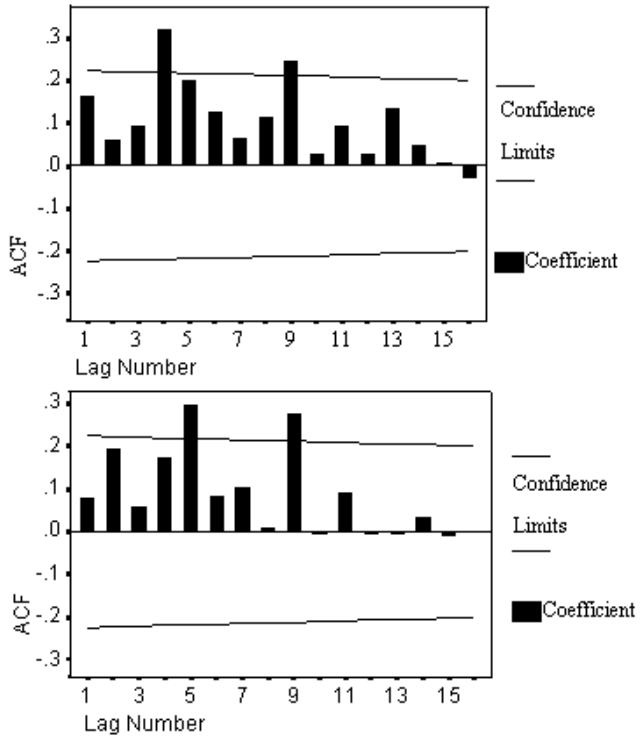


Figure 3: The ACF for the series at 5 and 10 minutes

Before using the data sets, it is essential to study the series tendency and the data correlation.

To test the data correlation, the autocorrelation function was used. The result of autocorrelation test for the rainfalls is presented in Figures 3 and 4. From the autocorrelation tests it results that the data are correlated. Indeed, the probabilities to reject the autocorrelation hypothesis are small (0.009-0.5) and there are few values of the autocorrelation function outside the confidence

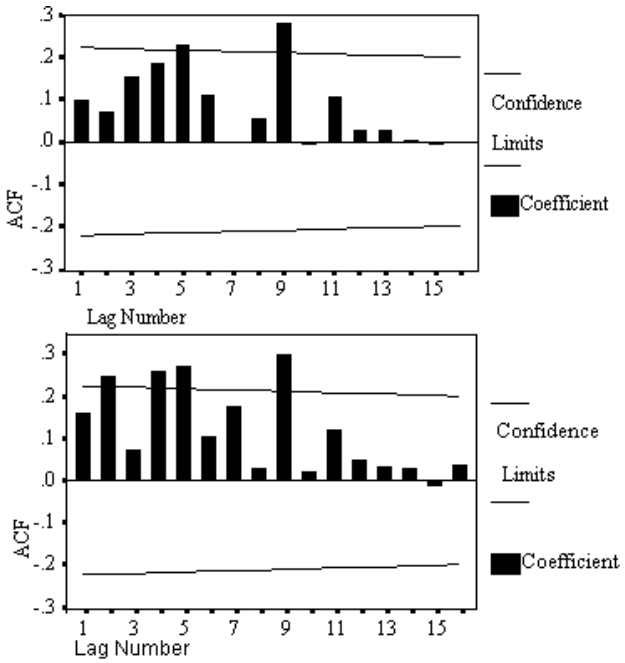


Figure 4: The ACF for the series at 15 and 20 minutes

intervals.

Spearman test was used for the tendency test of the series. The values of Spearman coefficients for the rainfall series were respectively: 0.385, 0.375, 0.421, 0.456, proving that the series are neither increasing, nor decreasing.

3.2 The frequential model

To determine the IDF curves of precipitation the following steps were made:

- Data preparation by:
  - Sorting the series values in the ascending order,
  - Allotting a rank order to each value;
- Calculation of the empirical frequency for each value (equation (4));
- Calculation of Gumbel's reduced variable u (equation (3));
- Plotting the couples (u_i, x_i) of the series to be fit;
- Fitting a linear relation of the type to the couples (u_i, x_i) and deducing the parameters a, b;
- Using the statistical model to estimate maximum rainfall intensity for a return period, T.

Gumble's diagram for the maximum rainfall intensities at 5, 10, 15, 20 minutes are represented in Figure 5.

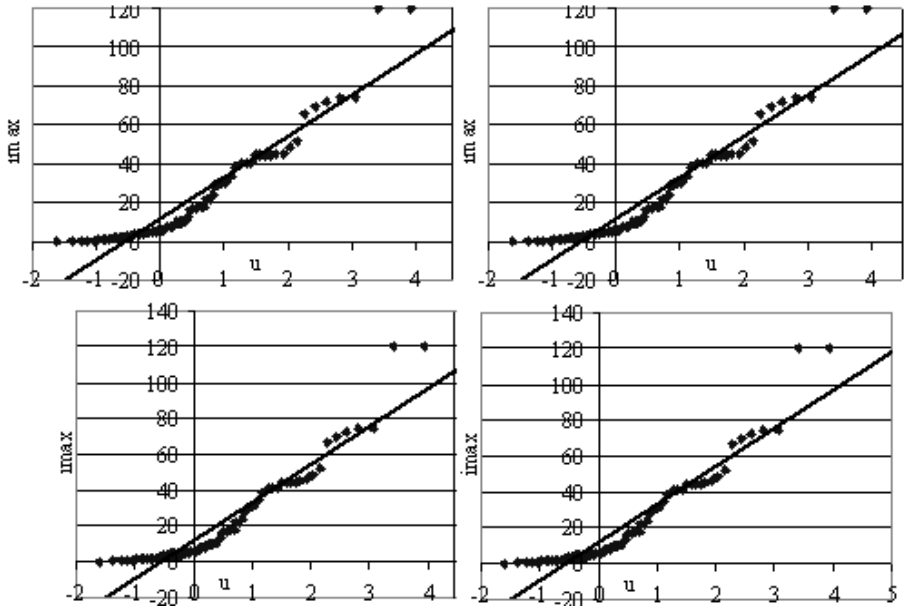


Figure 5: Gumbel's diagrams at 5, 10, 15 and 20 minutes

	a	b	R^2	Correlation coefficient	Standard deviation
5 min	11.542	21.278	0.9137	0.95	1.29
10 min	6.0311	14.385	0.8347	0.91	1.13
15 min	4.4234	9.9021	0.8486	0.92	1.05
20 min	4.1138	9.0491	0.8800	0.94	1.04

Figure 6: The modelling results

The numerical results for the parameters of Gumbel's distribution function corresponding to the different durations at the hydrometric station of Voinesti are presented in Figure 6, that contains:

- in the columns 2 and 3, the parameters a, b;
- in the column 4, the determination coefficient, which is big enough to say that the parameters are significant;
- in the column 5, the correlation coefficient;

- in the column 6, the standard deviation of the residuals resulted in the modelling.

3.3 The validation of model

To validate the choice of the Gumbel's distribution the residuals were tested.

- The results of normality tests - Jarques' Bera test and χ^2 - lead us to accept the hypothesis that the residuals are normal distributed, with the mean zero and the standard deviation around 0.385.
- Using the independence test (Durbin - Watson) and analysing the values of the autocorrelation function, we accept the hypothesis that the residuals are independent.
- The results of Bartlett test lead us to accept the hypothesis that the residuals are homoscedastic.

The results of the previous tests prove that Gumbel's model was well chosen.

Using the computed parameters, the IDF curves for different return periods can be determined, as they can be seen in Figure 7.

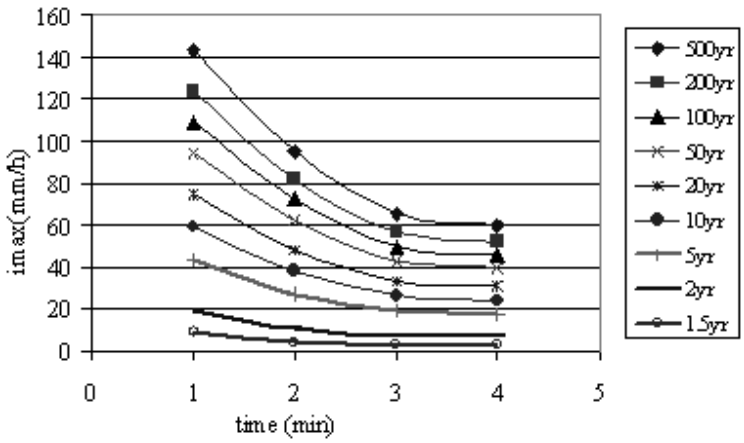


Figure 7: The IDF curves

4 Conclusions

In this article Gumbel' s frequential model on the maximum rainfall intensity registered by the pluviograph of the hydrometrical station Voinesti was applied, to obtain the IDF curves, for different calculus probabilities, necessary to dimension the hydrotechnic works.

Before the model application, the data concerning the maximum rainfall intensity calculated at 5, 10, 15 and 20 minutes were tested, in order to relieve the errors. The results of the tests were satisfactory.

To adjust the data, graphical methods were used. It was seen that the values of the maximum intensity can be fitted by straight lines, to which the determination coefficients were calculated. These being close to 1, the parameters of the adjustment lines, a, b, were determined.

The adjustment control were made by Jarques Bera and χ^2 tests. Since the tests gave satisfactory results, we conclude that Gumbel distribution can be successfully used in the estimation of the rainfall maximum intensity, with a certain response time, in the studied case.

In a future paper we shall compare the results obtained with this model with those obtained using other methods.

References

[1] Mohymont B., Demaree G. R., Faka D. N., Establishment of IDF-curves for precipitation in the tropical area of Central Africa - comparison of techniques and results, Natural Hazards and Earth System Sciences, 4, 2004, 375-387

[2] Maftei C., Modelisation spatialisee de l'ecoulement sur des petits basins versants, Ed. Ceremi, Iasi, 2004

[3] Meyan P., Musy A., Hydrologie frequentielle, Edition HGA, Bucharest, 1999

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Air Pollution Monitoring and Modeling near M. Kogalniceanu Civil Airport in 2008

[1]Octavian Thor Pleter, [2]Dana-Cristina Toncu,
[3]Gheorghita Toncu, [5]Virgil Stanciu
[1,2,3,4]Department of Aerospace Engineering,
Politehnica University of Bucharest
E-mail: cristinatoncu@canals.ro

Abstract

Major environmental problem worldwide, air pollution results also from transport activities, aerial inclusive. Air pollution monitoring and modeling seek to develop an accurate identification, measurement and evaluation of atmospheric state in order to facilitate forecast and decrease.

Choosing 2008 independently registered data in the aria of M. Kogalniceanu Civil Airport, MATHCAD Multiple Linear Regression analysis, MATLAB prediction and Air Quality Simulation Model based on diffusion equation were performed in order to forecast pollutants' evolution in 2009.

Keywords: air pollution; monitoring; modeling; air quality simulation model; diffusion equation.

1 Introduction

Air pollution is a major environmental issue worldwide. It is generated by important emission sources such as industrial plants, refineries, power plants, domestic heating and, more important, all kinds of transport. Hence, airports, such as M. Kogalniceanu Civil Airport, are considered to cause air pollution and, thus, discomfort. Noxious emissions influence both air quality and climate change, fundamentally changing the chemical content of the atmosphere. Nitrogen oxides (NO_x), CO and volatile organic compounds (VOCs) lead to photochemical smog and associated oxidants and facilitate the ozone production in the free troposphere and climate warming. Also major greenhouse gases (CO_2, CH_4, N_2O, halogenous compounds) are produced. Nitrogen oxide and sulfur oxide are also responsible for acid rain by atmospheric photochemical reaction, deteriorating ecosystem. Both direct carbonaceous aerosols emission and secondary aerosol precursors form simultaneously (NO_x, VOCs, SO_2, and NH_3), eventually affecting the climate.

ISSN-1584-5990

In the case of air pollutants (NO_x, CO, SO_2, and particulate matter), magnitudes and tempo-spatial distributions remain inaccurate. Despite sustained efforts for continuously monitoring, determined data barely presents patterns, being based on estimates and extrapolations. This occurs from two major causes: difficulty to predict the impact of emissions and pollutant distributions on air quality and thus, on human's health and local ecosystem viability; obstruction of accurate prediction for both tempo-spatial distributions and emission impact on large scales due to nonlinear atmospheric chemical processes of pollutant generation. Far from completion, the phenomenological description needs new tools and techniques to simple and exact pollution quantification and forecast.

With respect to air monitoring and modeling, previous efforts [1]–[9] include evaluation of several data acquisition systems and analysis instruments and methods applied mostly in meteorological studies. Additionally, air quality simulation models tend to reconsider time-dependent complex physical-chemical phenomena. Their results showed good agreement between simulations and experimental data.

This paperwork aims to develop an air pollution monitoring and modeling system relying on accurate data from a passive-mobile sampler collect and analysis complex, in anticipating a huge amount of information regarding the nature, sources, and distribution of pollutants. The resulting techniques and methods are to be further developed and validated for other areas.

2 Methodology

2.1 Measurements

The monitoring activity in M. Kogalniceanu Civil Airport was conducted in 2008 and focused on several important pollutants: sediment powder (particulate matter which deposit freely in time), nitrogen oxides, carbon monoxide and sulfur dioxide. Concentration levels were determined daily with both stationary and mobile stations, ranging from simpler approach (passive samplers) to more complex instrumental methods (air sampler, HPGC). Mobile pollution data measurement identified the distribution of narrow emission sources and correlated transport and reaction phenomenon with emanations, while stationary data collection simulated prompt response.

Sampling program was optimally designed for hazard prevention, assuming that errors in the forecast pollution came exclusively from inaccurate concentration record, since such uncertainties propagated in model, being induced through computational, process modeling and human factors.

2.2 Data analysis

The registered data obtained in the year 2008 near M. Kogalniceanu Civil Airport by independent sampling can be summarized in Figures 1-3. Their histogram, represented using MATLAB, is shown in Fig. 4, from which the unusual

distribution of determined pollutants can be noticed. Based on the collected data and shown pattern evolution, prediction graphs for 2009 were plotted in MATLAB.

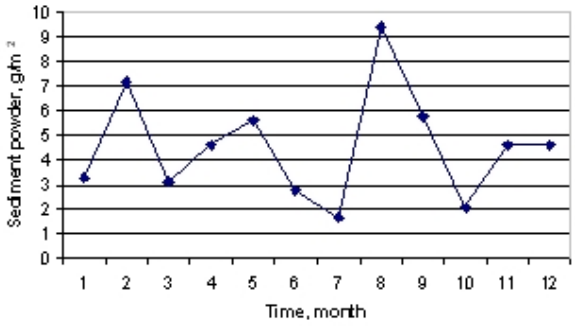


Figure 1: Average sediment powder for 2008

MATHCAD Multiple Linear Regression analysis tested both data quality evaluation and objective investigation data in airport complex terrain. It revealed how reasonable pollutants determination was, showing that stationary measured values were close to the dynamic ones. The resulting model helped to predict new stationary data for the case in which experiments couldn't be carried on from risk occurrence, as $R^2 \cong 1$(see Table 1).

3 Model description and simulation

Air Quality Simulation Models (AQSMs) aim to establish causal relations among emission levels and air quality, using numerical models which replicate transport and chemical transformation of pollutants in the atmosphere, either with the aid of prognostic modeling, based on the fundamental physiochemical principles governing air pollution, or through diagnostic modeling, related on statistical

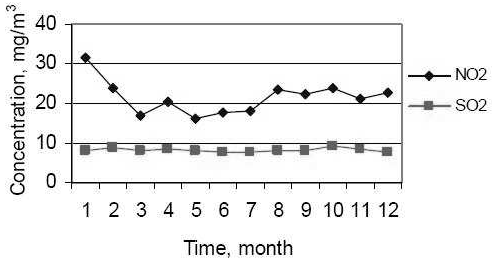


Figure 2: Average levels of NO_2 and SO_2 in 2008

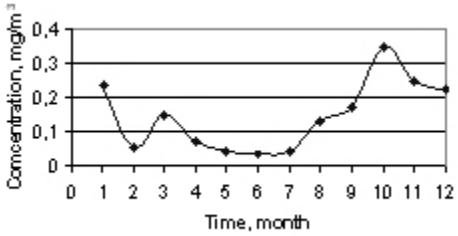


Figure 3: Monthly CO records for 2008

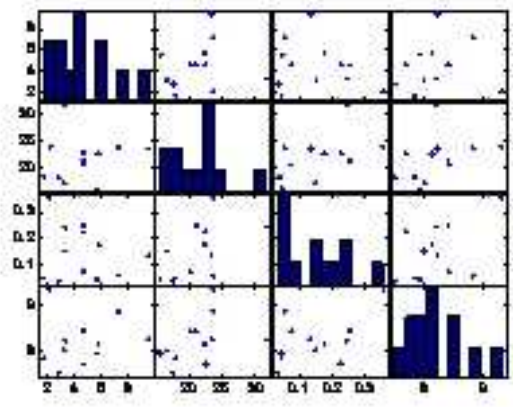


Figure 4: Histogram for chosen pollutants with gplot

description of observed data.

Air pollution involves three major phases: emission from different sources (both anthropogenic and natural); transport, mainly by wind fronts (advection); transformation: diffusion in the atmosphere, deposition to the surface (soil, water and vegetation), either dry (continuous during transport) or wet (only during rains), and chemical reactions in which secondary pollutants result. By consequence, it can be mathematically described using Euler approach relating the behavior of the species to a fixed coordinate system or Lagrange approach relating the concentrations changes to moving air.

The model used for M.Kogalniceanu Civil Airport based on both fundamental atmospheric transport and chemistry occurring, given by the diffusion equation applied to the atmospheric air:

$$\frac{\partial < c_i >}{\partial t} + \bar{u_i}\frac{\partial < c_i >}{\partial x_i} = \frac{\partial}{\partial x_j}(K_{jj}\frac{\partial < c_i >}{\partial x_j}) + R_i(\sum_{i=1}^{N} < c_i >) + S_i(x,t) \quad (1)$$

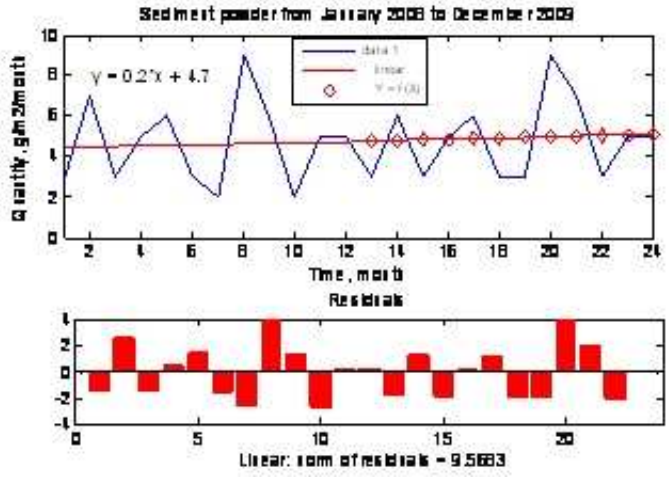


Figure 5: 2009 predicted values of sediment powder

where: $< c_i >$ – average pollutant concentration, mg/m^3;
$\bar{u_j}$ – mean wind speed, m^2/s;
R_i – rate of formation of i pollutant species, $mg/m^3 s$;
K_{jj} – diffusion coefficient,m^2/s; for gases $K = \frac{1}{3}\bar{\nu}\lambda$ (λ – average free road), and under normal conditions $Kappa \approx 10^{-5} m/s^2$;
S_i rate of emission of i pollutant species into the control volume, mg/m^3.
resulted from several approximations and negligence (gradient driven fluxes approximation of turbulent fluxes, neglect of molecular diffusion: D<<K, considering atmosphere as incompressible: the mean wind $\bar{u}$ and lateral average diffusivity K_{jj} are independent of y, and reaction rates approximation of ensemble average with ensemble averages):

$$D_A \frac{\partial^2 < c_A >}{\partial x^2} << \frac{\partial \bar{u_i} c_A}{\partial x_i} = -\frac{\partial}{\partial x_i}(K_A \frac{\partial < c_A >}{\partial x_j}) \tag{2}$$

$$\frac{\partial \bar{u}}{\partial x} = \frac{\partial \bar{v}}{\partial y} = \frac{\partial \bar{w}}{\partial z} \tag{3}$$

$$\frac{\sum_{i=1}^{N} c_i}{N} \approx \sum < c_i > \tag{4}$$

The 1-dimensional space diffusion model indicates that the rate of studied pollutant species change is proportional to the curvature of species density. Assumptions regard simultaneous dispersion and reaction of all individual pollutants and invariable dispersal abilities of individuals. The diffusion model assumes that the mean wind direction is along the x-axis, so the resulting rectangular coordinates were rotated according to the wind direction for each analysis interval. The chemical processes play an important role, but introduce nonlinearity in the model.

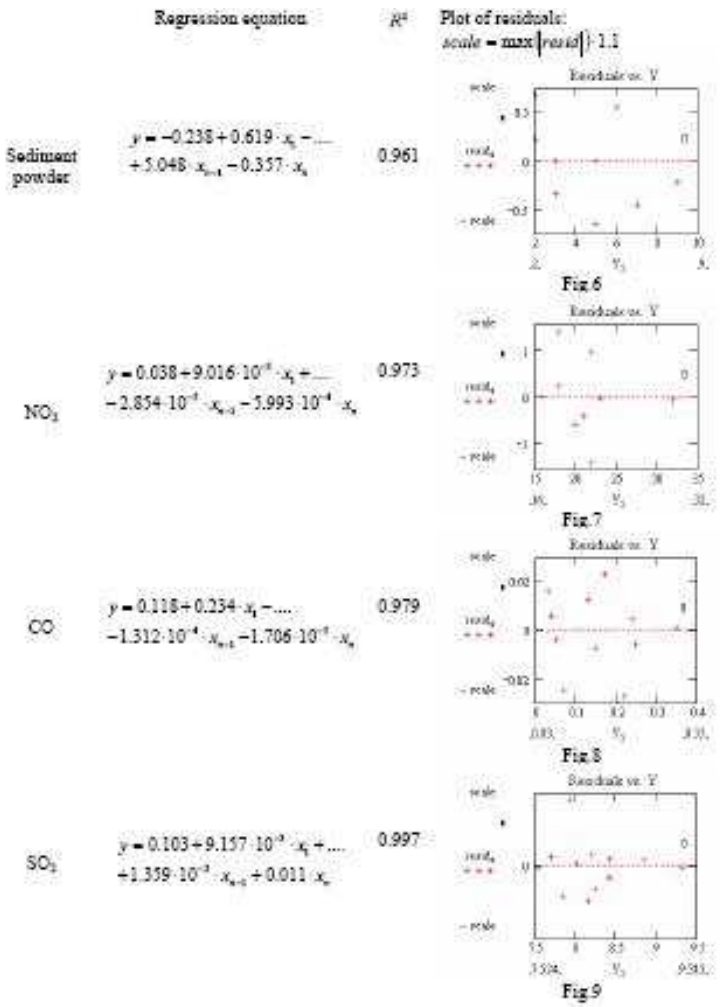


Figure 6: Multiple Linear Regression Analysis between stationary and dynamic data

The chemical reactions have the following form:

$$a_i A_i \xrightarrow{k} a_j Aj \quad R_s(c_i) = -\sum a_i c_i + \sum a_j c_{ij} \tag{5}$$

where a_i, a_j are the stoichiometric coefficients of the formation reaction of pollutants, and k the kinetic constant. If the chemical reaction is of 1st order, then reactive-diffusion is given by:

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} \pm kc \tag{6}$$

Initial and boundary conditions are added to the system of partial differential equations (1): $0 \leq x \leq x_{fin}$, $0 \leq t \leq t_{fin}$.

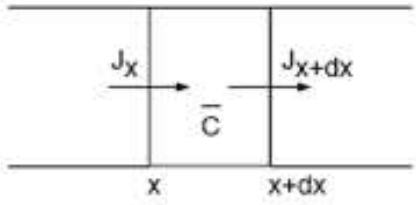


Figure 7: Volume element considered for calculus

The rate of pollutant accumulation into air volume elements is given by the following equation:

$$J_x - J_{x+dx} = \frac{\partial \bar{C}}{\partial x} dt \tag{7}$$

$\bar{C}$ is the average pollutant concentration in the volume element and cdx is the total amount of pollutant into the volume element.

In steady-state the solutions of the diffusion equation are linear when the diffusion coefficient is constant (mass flux J does not depend on time). K_{ij} varies with temperature and ionic strength, so differences in slope occur during heat flow. Also horizontal homogeneity was presumed.

For further simplification, advection equation term could be removed by assuming that the low of mass conservation is satisfied in the lower atmosphere level:

$$\frac{\partial C_A}{\partial t} = -\nabla N_A + R_{V,A} + S_{V,A} \tag{8}$$

where A - concentration of species A, ∇ - gradient operator, N_A - total molar flux of A relative to fixed coordinates, given by the sum of a convective and diffusive transport:

$$N_A = uC_A + J_A \tag{9}$$

where u is the wind vector (u, v, w) and J_A is the molar diffusive flux of species A. For a diluted gas in the atmosphere, the diffusive flux is approximated by Fick's law:

$$J_A = -D_A \nabla C_A \tag{10}$$

where D_A is the molecular diffusivity of the dilute species A in the air. In steady state:

$$\nabla uC_A = D_A \nabla^2 C_A + S_{V,A} \tag{11}$$

For sediment powder, mathematical formulation considered the case of suspended particles moving with speed v in gravitational field, and condition for impermeable volume frontier:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial z^2} - v\frac{\partial u}{\partial z} \tag{12}$$

with impermeability condition in plane $z = z_0$:

$$D\frac{\partial u}{\partial z} - vu = 0 \tag{13}$$

Computational method in MATLAB with adapted advection-diffusion code, pdepe function and Fourier analysis revealed both spatial and temporal accuracy of the model, as shown for CO in Figures 11-14. Simulation corresponds to both stationary collected data and phenomenological explanation of pollutant decrease in time for the case of discontinuous emission sources.

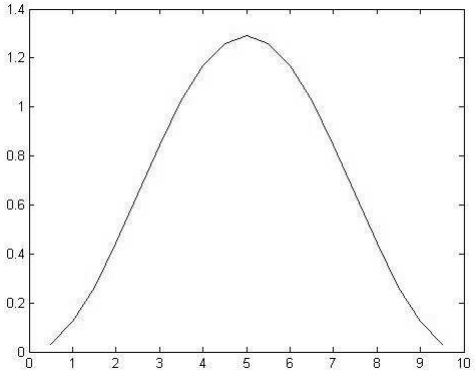


Figure 8: Advection-diffusion graph

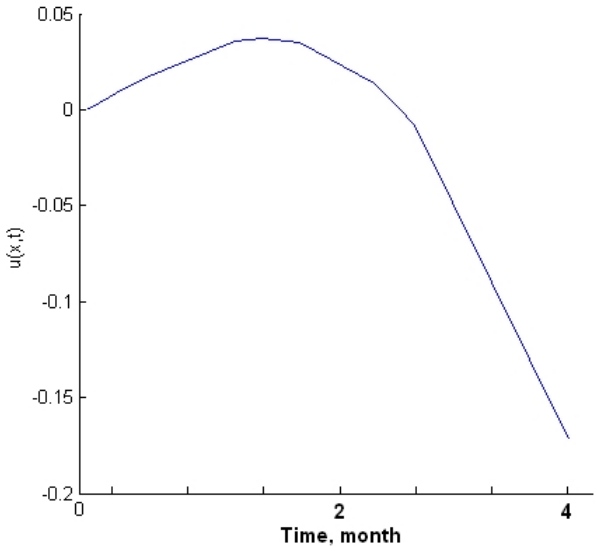


Figure 9: Time distribution with pdepe function

Factors of uncertainty in the computations are: the definition of the source of emissions (discontinuous and superposition of an infinite number of punctual sources), neglect of effects of temperature and pressure on diffusion, constant approximations for chemical rate and constants, diffusion coefficient, air properties, temperature, wind speed and direction.

4 Conclusion

Complex and advanced air pollution models are needed due to the nonlinear evolution of most pollutants and their reactions, as the configuration of the monitored area creates a departure from the Gaussian distribution.

Convection and diffusion are the two mechanisms by which pollutants moves through the open air, therefore pollution-dispersion modeling, despite the involved emissions uncertainty, is sufficient even for low-resolution two- and three-dimensional form, for studying pollutants' concentrations, giving good predictions.

The new mixed algorithm for pollution modeling is feasible to perform calculations for realistic computational domains and areas, providing high spatial, temporal and chemical resolution and the promise of pollution inventory.

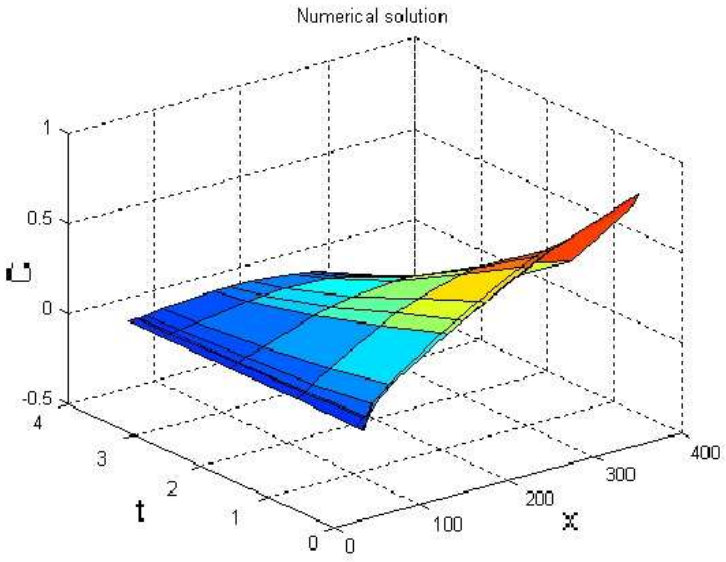


Figure 10: Numerical solution with pdepe function

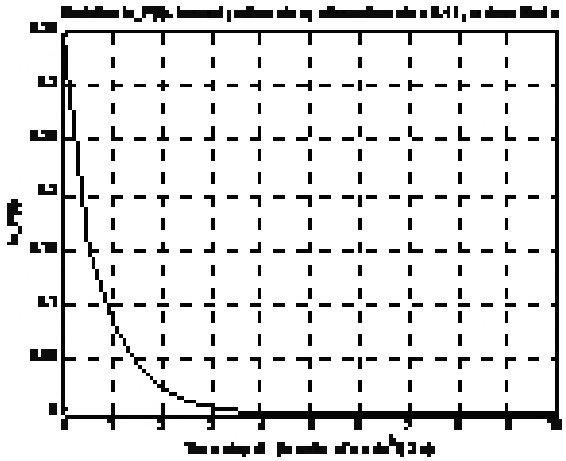


Figure 11: Transient diffusion with Fourier analysis

References

[1] O. Pummakarnchana et al., *Air pollution monitoring and GIS modeling: a new use of nanotechnology based solid state gas sensors*, Science and Technology of Advanced Materials 6(2005) 251-255.

[2] M. Lazar et al., *The Statistical System of Local Level Indicators and the*

Urban Audit Project, Buletinul UPG Ploiesti LVIII (2006) 53-58.

[3] Ronald van der A et al., *Air Pollution Monitoring with GOME and SCIAMACHY*, DRAGON Symposium, China 2004.

[4] Y.J. Jung et al., *Air Pollution Monitoring System Based on Geosensor Network*, IGARRS 2008 Boston.

[5] E. Petre, *A comparative Study of Some Air Pollution Software*, Buletinul UPG Ploiesti LIX (2007) 55-58.

[6] C. Miclaus et al., *Passive Samplers-Checking Method in the Air Quality Control Network*, EEMJ, 5 (2006) 1333-1340.

[7] M. Astitha et al., *Air pollution modelling in the Mediterranean Region: Analysis and forecasting of episodes*, Atmospheric Research (2008) doi: 10.1016/j.atmores.

[8] A.M. Abdulah et al., *The Application of Mm5-Cmaq for Modeling Urban Air Pollution In Selangor, Malaysia*, Computer modeling & applications, remote sensing, GIS (2008) 673-676.

[9] S.Ahmed et al., *Applicability of Air Pollution Modeling in a Cluster of Brickfields in Bangladesh*, Chemical Engineering Research Bulletin 12 (2008) 28-34.

[10] Medeiros, J. L., *Simulação Transiente de Dispersão Atmosférica de Poluentes Gasosos*, Anais do XIV C. Ibero Latino-Americano de M. Comp. em Eng., IPT, S. P., p. 1120-1129, 1993.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

One-and-a-half-dimensional Model of Cumulus Cloud with Two Cylinders. Research of Influence of Compensating Descending Flow on Development of Cloud

[1]Nikita Raba, [2]Elena Stankova,
[3]Natalya Ampilova
[1,2,3]Saint-Petersburg State University, Mathematics and Mechanics Faculty, Saint-Petersburg, Russia
E-mail: [1,2,3]no13@inbox.ru

Abstract

Modified 1,5-dimensional model of convective cloud with parameterized microphysics for liquid and solid phases is presented. A region of convective flow in the model is represented by two concentric cylinders. The inner cylinder corresponds to updraft flow region and the outer cylinder to surrounding downdraft flow region. The model shows more realistic state for the life cycle of the cloud (than original model with one cylinder) and reproduces all of the typical stages of its development (including the dissipation stage). Considerable effect of ratio of outer cylinder radius to inner cylinder radius on cloud parameters is shown.

Keywords: One-and-a-half-dimensional model; Convective cloud; Compensating descending flow.

1 Introduction

A mechanism of heat transfer and precipitation forming in convective clouds is interesting and important problem, because such dangerous phenomena as thunderstorm, hail, squall and tornado are connected with developing convective clouds. Numerical modelling is one of the most effective instruments of cloud investigation.

For scientific research oriented on detailed study of dynamical and microphysical processes in cloud it is necessary to use the models that fully reproduce natural processes (i.e. 2- or 3-dimensional models with detailed description of microphysical processes) [3, 5, 8]. However, for the purpose of efficiency of forecast (for example: in meteorological centers in airports) models with few

ISSN-1584-5990

computer resources are required. These models have to reproduce such parameters of the cloud (altitude of upper and lower boundaries of the cloud, maximum value of the water content, excess temperature, velocity of ascending flow, etc.) which may become an indication for the methods of forecast of dangerous convective effects. From this point of view time-dependent 1,5-dimensional models are the best [1, 2, 4, 6, 7, 8]. In this kind of models the cloud is generated in a cylindrical region. All of the parameters are averaged over cross section of the cylinder. The cylinder is surrounded by still atmosphere with constant parameters. However, this models can not adequately reproduce all of the life stages of the cloud (especially the dissipation stage) under some initial atmospheric conditions. In this case the cloud is actually stabilized after developing stage (both dynamical and microphysical parameters of the cloud are not changed). We believe that the reason of this effect is lack of compensating descending flow which suppresses the development of ascending flow.

Mechanism of compensating of ascending flow, however, is included in first 1,5-dimensional model of Asai and Kasahara [1] by means of additional outer cylinder. The inner cylinder corresponds to the ascending flow area (the cloud develops in this cylinder), the outer cylinder to the descending flow area. But this model includes very simple microphysics (without precipitation formation and solid phase).

We have developed a model of convective cloud that combines advantage of Asai and Kasahara model (i.e. takes into account compensating descending flow) and contains microphysics with precipitation formation and solid phase. Using this model we try to analyze the effect of compensating descending flow on convective cloud. Sufficiently interesting results are obtained.

2 Model description

The model represents 1,5-dimensional model of convective cloud with parametrization of microphysical processes for both liquid and solid phases.

In the model the region of convective flow is represented by two concentric cylinders (as in Asai and Kasahara model [1]). The inner cylinder (with constant radius a) corresponds to the updraft flow region (cloudy region) and the outer cylinder (with constant radius b) to the surrounding downdraft flow region (cloudless). The ratio of the area of cross section of inner cylinder to the area of cross section of outer ring-shaped cylinder is equal to

$$K_{ab} = a^2/(b^2 - a^2). \tag{1}$$

All of the equations are written in cylindrical coordinates. Parameters of cloud are represented as averaged over cross section of cylinder. Equations from [6] for the parameters have been modified because of effect of ascending flow. Moreover, pair equations for the parameters in outer cylinder have been added. Modifications in equations are marked with bold font. Differences between equations for parameters in outer cylinder and in inner cylinder are also marked. Interaction of compensating descending flow in outer cylinder with surrounding

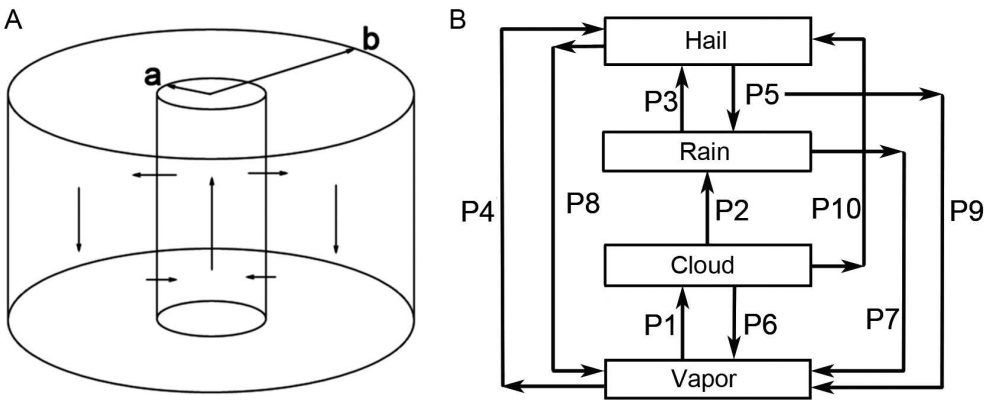


Figure 1: The schemes of (A) flow; (B) microphysical processes.

atmosphere are not considered because it is much less than interaction with ascending flow in inner cylinder.

The equations for the vertical velocity w are written as

$$\frac{\partial w_{in}}{\partial t} = -w_{in}\frac{\partial w_{in}}{\partial z} - \frac{2\alpha^2}{a}|w_{in} - \mathbf{w_{out}}|(w_{in} - \mathbf{w_{out}}) + \frac{2}{a}u_a(w_{in} - w_a) + \frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial w_{in}}{\partial z}) + g\frac{T_{v\,in} - \mathbf{T_{v\,mean}}}{\mathbf{T_{v\,mean}}} - g(Q_{c\,in} + Q_{r\,in} + Q_{i\,in}), \tag{2}$$

$$\mathbf{w_{out}} = -\mathbf{K_{ab}}\mathbf{w_{in}}, \tag{3}$$

where u is radial velocity, α^2 the coefficient for lateral eddy mixing, g the acceleration of gravity, ρ_a the air density, T_v the virtual temperature, $T_{v\,mean} = (a^2T_{v\,in} + (b^2a^2)T_{v\,out})/b^2$ the virtual temperature averaged over cross section of both cylinders, K_v the vertical eddy diffusion coefficient, Q_c, Q_r and Q_i are contents (mixing ratios) of cloud droplets, raindrops and hailstones (frozen droplets). Values at the lateral surface of the inner cylinder are denoted by subscript "a" (except ρ_a), values in the environment are marked by subscript "0", values in the inner cylinder (in the cloud) and in the outer cylinder are indicated by subscripts "in" and "out" respectively. u_a is determined by the equation of mass continuity under assumption of incompressibility which is given as

$$\frac{2}{a}u_a + \frac{1}{\rho_{a0}}\frac{\partial(\rho_{a0}w_{in})}{\partial z} = 0. \tag{4}$$

It is reasonable to assume that

$$A_a = \begin{cases} A_{out}, & if \quad u_a < 0, \\ A_{in}, & if \quad u_a \geq 0, \end{cases} \tag{5}$$

where instead of A can be placed any variable.

The equations for the temperature are written as

$$\begin{aligned}\frac{\partial T_{in}}{\partial t} &= -w_{in}(\frac{\partial T_{in}}{\partial z}+\Gamma_d)-\frac{2\alpha^2}{a}|w_{in}-\mathbf{w_{out}}|(T_{in}-\mathbf{T_{out}})\\ &+\frac{2}{a}u_a(T_{in}-T_a)+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial T_{in}}{\partial z})+F_{T\,in},\end{aligned} \tag{6}$$

where Γ_d is the dry adiabatic lapse rate, F_T is the quantity concerned with microphysics.

The equations for the contents of vapor, cloud droplets, raindrops and hailstones are given as

$$\begin{aligned}\frac{\partial T_{out}}{\partial t} &= -w_{out}(\frac{\partial T_{out}}{\partial z}+\Gamma_d)-\frac{2\alpha^2}{a}\mathbf{K_{ab}}|w_{in}-w_{out}|(T_{out}-T_{in})\\ &-\frac{2}{a}u_a\mathbf{K_{ab}}(T_{out}-T_a)+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial T_{out}}{\partial z})+F_{T\,out},\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial Q_{v\,in}}{\partial t} &= -w_{in}\frac{\partial Q_{v\,in}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-\mathbf{w_{out}}|(Q_{v\,in}-\mathbf{Q_{v\,out}})\\ &+\frac{2}{a}u_a(Q_{v\,in}-Q_{va})+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial Q_{v\,in}}{\partial z})+F_{v\,in},\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial Q_{c\,in}}{\partial t} &= -w_{in}\frac{\partial Q_{c\,in}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-\mathbf{w_{out}}|(Q_{c\,in}-\mathbf{Q_{c\,out}})\\ &+\frac{2}{a}u_a(Q_{c\,in}-Q_{ca})+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial Q_{c\,in}}{\partial z})+F_{c\,in},\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial Q_{r\,in}}{\partial t} &= -(w_{in}-V_{r\,in})\frac{\partial Q_{r\,in}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-\mathbf{w_{out}}|(Q_{r\,in}-\mathbf{Q_{r\,out}})\\ &+\frac{2}{a}u_a(Q_{r\,in}-Q_{ra})+\frac{Q_{r\,in}}{\rho_{a0}}\frac{\partial(\rho_{a0}V_{r\,in})}{\partial z}+F_{r\,in},\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\partial Q_{i\,in}}{\partial t} &= -(w_{in}-V_{i\,in})\frac{\partial Q_{i\,in}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-\mathbf{w_{out}}|(Q_{i\,in}-\mathbf{Q_{i\,out}})\\ &+\frac{2}{a}u_a(Q_{i\,in}-Q_{ia})+\frac{Q_{i\,in}}{\rho_{a0}}\frac{\partial(\rho_{a0}V_{i\,in})}{\partial z}+F_{i\,in},\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial Q_{v\,out}}{\partial t} &= -w_{out}\frac{\partial Q_{v\,out}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-w_{out}|\mathbf{K_{ab}}(Q_{v\,out}-Q_{v\,in})\\ &-\frac{2}{a}u_a\mathbf{K_{ab}}(Q_{v\,out}-Q_{va})+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial Q_{v\,out}}{\partial z})+F_{v\,out},\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial Q_{c\,out}}{\partial t} &= -w_{out}\frac{\partial Q_{c\,out}}{\partial z}-\frac{2\alpha^2}{a}|w_{in}-w_{out}|\mathbf{K_{ab}}(Q_{c\,out}-Q_{c\,in})\\ &-\frac{2}{a}u_a\mathbf{K_{ab}}(Q_{c\,out}-Q_{ca})+\frac{1}{\rho_{a0}}\frac{\partial}{\partial z}(K_v\frac{\partial Q_{c\,out}}{\partial z})+F_{c\,out},\end{aligned} \tag{13}$$

$$\frac{\partial Q_{r\,out}}{\partial t} = -(w_{out} - V_{r\,out})\frac{\partial Q_{r\,out}}{\partial z} - \frac{2\alpha^2}{a}|w_{in} - w_{out}|\mathbf{K_{ab}}(Q_{r\,out} - Q_{r\,in}) - \frac{2}{a}u_a\mathbf{K_{ab}}(Q_{r\,out} - Q_{ra}) + \frac{Q_{r\,out}}{\rho_{a0}}\frac{\partial(\rho_{a0}V_{r\,out})}{\partial z} + F_{r\,out}, \tag{14}$$

$$\frac{\partial Q_{i\,out}}{\partial t} = -(w_{out} - V_{i\,out})\frac{\partial Q_{i\,out}}{\partial z} - \frac{2\alpha^2}{a}|w_{in} - w_{out}|\mathbf{K_{ab}}(Q_{i\,out} - Q_{i\,in}) - \frac{2}{a}u_a\mathbf{K_{ab}}(Q_{i\,out} - Q_{ia}) + \frac{Q_{i\,out}}{\rho_{a0}}\frac{\partial(\rho_{a0}V_{i\,out})}{\partial z} + F_{i\,out}, \tag{15}$$

where F_v, F_c, F_r, F_i are the microphysical processes. V_r and V_i are the mean terminal falling velocities of raindrops and hailstones respectively (the equations for this velocities are given in [6]).

Microphysics in our model corresponds to the microphysics in [6]. Following processes are taken into consideration: condensation ($P1$), autoconversion ($P2_{auto}$) and collection ($P2_{coll}$), heterogeneous freezing of raindrops ($P3$), sublimation ($P4$), melting ($P5$), evaporation of cloud droplets ($P6$), evaporation of raindrops ($P7$), evaporation of hailstones ($P8$), evaporation of melting hailstones ($P9$), riming ($P10$). $P1, \ldots, P10$ are the rates of corresponding processes ($P2 = P2_{auto} + P2_{coll}$). Therefore the equations for F_T, F_v, F_c, F_r, F_i may be written as

$$F_T = \frac{L_v}{c_p}(P1 - P6 - P7 - P9) + \frac{L_s}{c_p}(P4 - P8) + \frac{L_f}{c_p}(P3 + P10 - P5), \tag{16}$$

$$F_v = -P1 + P6 + P7 + P8 - P4 + P9, \tag{17}$$

$$F_c = P1 - P2 - P6 - P10, \tag{18}$$

$$F_r = P2P3 + P5P7, \tag{19}$$

$$F_i = P3 + P4P5P8P9 + P10, \tag{20}$$

where c_p is the specific heat of air with constant pressure, L_v, L_s and L_f are the latent heats of evaporation, sublimation and fusion respectively.

3 Environmental, initial and boundary conditions

The height of the cylinder is 15 km. The temperature at the ground surface is 298K. The temperature laps rate is 9,8 K/km up to 2 km and is 6,3 K/km from 2

km to 10 km. The temperature is constant above 10 km. The relative humidity is 100% at the ground and decreases with lapse rate of 5%/km up to the top of cylinder. Initial contents of cloud droplets (Q_c), raindrops (Q_r), hailstones (Q_i) are equal to zero at all levels. Vertical (w) and radial (u_a) velocities and Q_c are assumed to be 0 at the top and at the bottom boundaries of the cylinder. Q_r and Q_i are 0 at top boundary. The initial disturbance of vertical velocity in the inner cylinder below 2 km is given as

$$w_{in} = \Delta w\, z\,(2 - z), \tag{21}$$

where Δw is taken as 1 m/sec. The coefficient for lateral eddy mixing is 0,1. The vertical eddy diffusion coefficient equals to 100 m^2/sec. The autoconversion threshold is $5 \cdot 10^{-4}$ kg/m^3. The autoconversion coefficient is 0,01 sec^{-1}.

4 Numerical procedures

The method of decomposition on physical processes is used for solving given system of equations. Only dynamical processes are taken into account at the first stage. Equations are numerical integrated using a finite difference method. Forward-upstream scheme is used. Vertical velocity is averaged over two grid points (in addition point below is taken if $w \geq 0$ or point above if $w < 0$). The final values with accounting microphysical processes are calculated on the second stage. A time step Δt of 1 sec and a height interval Δz of 200 m are used.

5 Results

A special framework is developed. It allows us to visualize the results of single experiment and series of experiments for pictorial showing the influence of some parameters to others. Following parameters are selected to study: the lifetime of the cloud, height of upper boundary of the cloud, maximum velocity of the ascending flow, maximum contents of cloud droplets, raindrops and hailstones. A research of dependences of selected parameters and development of cloud from radius of inner cylinder (a) and ratio of radius of outer cylinder to radius of inner cylinder (b/a) is implemented. Comparison with one cylinder model is made. The range of changes of a is from 0,6 km to 10 km, range of ratio b/a is 2–10 (i.e. range of K_{ab} is 1/3–1/399). Let us denote the model with one cylinder as M1C and our model with two cylinders as M2C. The same initial and boundaries conditions are used in both M1C and M2C.

In the experiment in M1C the cloud is stabilized (i.e. dissipation is not occurred, and form of the cloud is not changed in time) at about 30 minute with any radius. In M2C the cloud is dissipated. The greater ratio b/a, the later dissipation occurs (fig. 2f), the higher upper boundary of the cloud (fig. 2c), the lager maximum vertical velocity in the cloud (fig. 2b), the greater content of cloud droplets (fig. 2a). It is because with greater radius of outer cylinder

the compensating descending flow has the lower velocity and less prevents to development of the cloud. At the maximal values of ratio b/a (i.e. radius b of outer cylinder can be actually considered as infinity) these parameters in M2C tend to the corresponding parameters in M1C.

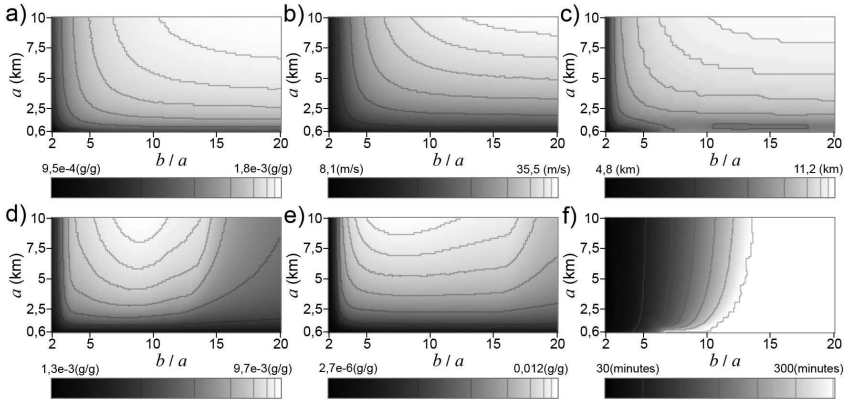


Figure 2: Dependences of a) maximum content of cloud droplets; b) maximum velocity of the ascending flow; c) height of upper boundary of the cloud; d) maximum content of raindrops; e) maximum content of hailstones; f) lifetime (dissipation time) of the cloud from radius a and ratio b/a. Note. The brighter point denotes the greater value of parameter. The scales with indication of range are placed under corresponding charts.

The above mentioned parameters increase with increasing of ratio b/a, but the contents of raindrops and hailstones reach their maximum in certain values of this ratio (from 7 to 10 depending on radius a) and exceed respective contents in M1C (fig. 2d and 2e).

All of the parameters (except lifetime of the cloud) increase with increasing radius a in both M1C and M2C. The stabilization of the cloud in M1C occurs a little earlier, and the lifetime of the cloud decreases with increasing radius a.

The presence of the outer cylinder essentially effects on the dynamics of development of the cloud. The time-height distribution of the content of cloud droplets is represented on fig. 3. One can see that M2C shows more realistic state for the life cycle of the cloud and reproduces all of the typical stages of its development, whereas the cloud in M1C stabilizes (i.e. M1C does not reproduce the dissipation stage).

We are going to implement the numerical experiments with real initial data and compare the results of modelling with the practical ones. This will allow us to adjust the model not only by ordinarily used parameters (the radius of cylinder, autoconversion threshold, coefficient of turbulence etc.) but also by tuning the ratio of outer cylinder radius to inner cylinder radius.

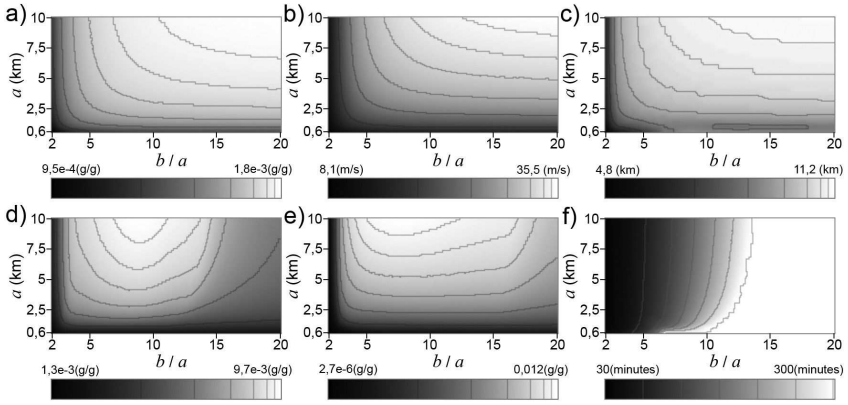


Figure 3: Content of cloud droplets. a) M1C with cylinder radius of 2 km; b) M2C, $a = 2$ km, $b = 10$ km ($K_{ab} = 1/24$); c) M2C, $a = 2$ km, $b = 14$ km ($K_{ab} = 1/49$); d) M2C, $a = 2$ km, $b = 20$ km ($K_{ab} = 1/99$). Note. On charts a), c) and d) the time scale is decreased after 1 hour point in order to show wider time interval.

6 Conclusions

1,5-dimensional time-dependent model of convective cloud with parameterized microphysics for liquid and solid phases is developed. The feature of the model is representation of convective clouds and its nearest surroundings as two concentric cylinders. The inner cylinder corresponds to updraft flow region (the cloud develops in this region) and the outer ring-shaped cylinder to surrounding downdraft flow region.

As results of numerical experiments have shown, the model with two cylinders reveals more realistic state for the life cycle of the cloud and reproduces all of the typical stages of its development (including the dissipation stage). Note that the behavior of the model when radius of the outer cylinder tends to infinity converges to that of the model with one cylinder.

The ratio of outer cylinder radius to inner cylinder radius can be used as a tuning parameter for adjustment of the model (for equalization of computation results to the results of natural experiment).

It is assumed to improve the microphysical part of the model by including size distribution of drops and crystals. It is also supposed to modify the dynamical part for taking into consideration the interaction of compensating descending flow in outer cylinder with surrounding atmosphere.

References

[1] Asai T., Kasahara A., *A Theoretical Study of the Compensating Downward Motions Associated with Cumulus Clouds*, Journal of the Atmospheric Sci-

ences, **24** (1967), 487-497.

[2] Dovgalyuk Yu.A., Veremey N.E., Sinkevich A.A., *Using a one-and-a-half-dimensional model for solving a fundamental and applied problems of cloud physics*, Saint-Petersburg, Gidrometeoizdat, 2007 (in Russian).

[3] Khain A., Pokrovsky A., Pinsky M., *Simulation of Effects of Atmospheric Aerosols on Deep Turbulent Convective Clouds Using a Spectral Microphysics Mixed-Phase Cumulus Cloud Model. Part I: Model Description and Possible Applications*, Journal of the Atmospheric Sciences, **61** (2004), 2963-2982.

[4] Ogura Y., Takahashi T., *The Development of Warm Rain in a Cumulus Model*, Journal of the Atmospheric Sciences, **30** (1972), 262-277.

[5] Seifert A., Baldauf M., Stephan K., Blahak U., Beheng K., *The Challenge of Convective-Scale Quantitative Precipitation Forecasting*, Proceedings of the 15th International Conference on Clouds and Precipitation, 2008.

[6] Shiino J., *A Numerical Study of Precipitation Development in Cumulus Clouds*, Papers in Meteorology and Geophysics, **29** (4) (1987), 157-194.

[7] Veremey N.E., Dovgalyuk Yu.A., Sinkevich A.A., *About forecast of development of convective clouds and associated dangerous phenomena*, Problems of cloud physics, Collection of selected articles, Saint-Petersburg, 2008, 104-116 (in Russian).

[8] Veremey N.E., Dovgalyuk Yu.A., Stankova E.N., *Numerical modelling of convective clouds developing in atmosphere with extraordinary situation (explosion, fire)*, Proceedings of the Russian Academy of Sciences, Physics of atmosphere and ocean, **23** (6) (2007), 792-806 (in Russian).

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

The Evolution of the Physical and Chemical Parameters of the Lakes in the Romanian Black Sea Littoral Area Which are Influenced by the Human Factors

Dacian Teodorescu
Romanian Water Authority from Dobroudja-Littoral Basins,
Constanta, Romania
E-mail: tdacian@gmail.com

Abstract

The present paper is a short synthesis of data results which are obtained through processing and interpretation of different parameters with direct monitoring methods on the lakes in the Romanian Black Sea Littoral from Dobroudja hydrographical space. Thus, the correlation between the results which are obtained from observations and measurements at the hydrometrical and meteorological stations with others data from literature papers don't lead us to very optimistical conclusions. The interdependence between physical factors and monitoring parameters such as: water level, water and air temperature, the potential evapo-perspiration process is influenced by the human factors which have a trend difficult to be predicted and a very fast evolution; the same influences modifying very much some biological and chemical characteristics of water of these lakes. Some considerations about climatic particularities which have an influence on the potential evapo-perspiration from continental hydrographical area in Dobruja. Case study: corelation between directly measurement outcomes which are obtained from different data evaporation stations with the calculated values from meteorological stations. For this type of analysis, it is propose to adapt and use the mathematical model from Cranfield Univesity. The model use measurement outcomes such as: maximum and minimum monthly many years values of air, relative humidity in percentages, shine times of the sun in hours and the wind speed at 2 meters high level in km/day. Through the simultaneously correlation of outcomes which are obtain in four different methods such as: Penman-Monteith, FAO, modified from Penman, Penman and Penman at surface open water. Le Turc method was used additional for comparison (even it is not so exactly). In different graphics are presented through comparison between the rainfalls quantity and outcomes obtain with Penman method at surface open water. Also it was makes correlation graphics between water and air temperature from hydrometrical and meteorological stations.

ISSN-1584-5990

Keywords: precipitation, catchwork, models

At the beginning the principal factors which participate to create bed foundation lake and water from it, was morphological and geological the first and climatic factors was the second.

The quantities of rainfalls which are registered on the Dobroudja area are very small.

Typical for the Dobroudja rainfalls is torrential regime, which are made that the extreme phenomenons to have major at last about variation regime of lakes especially which have therapeutical quality, about the erosion process and bed foundation mobility. It is also known that these big quantities of water roll big quantities of alluvial suspensions which are produced by specifically torrential rains with local character.

The levels regim are characterized by the minimum values which are registered in October, maximum values are registered in March-April period. Near we present the most important hydrometrical stations on the hydrographical network of Romanian Water Authority from Dobroudja-Black Sea Littoral area (Fig. 1).

About the multi - annual levels variation, we can notice that, all the Lakes in the Romanian Black Sea Littoral area have increase level trend in period which are present (except the Techirghiol Lake which have a decrease trend after 2000 year (Fig.2). Since the levels regime of many lakes in the Romanian Black Sea Littoral area are influenced by the human factors, it is very difficult to make a correlation between the modification of drainage basin which are supply water lakes and different level variations. In a year there are two stages of monthly medium levels such as: the increase stage - which start in september-octomber month and the decrease stage which start in april month. It is an exception Tabacarie Lake, who have the regime strict attached to evacuations from Siutghiol Lake and other against Black Sea which are control. It have two increase stages (between september-january and april-june) and two interpose decrease stages.

The absolutely gauge heigth and level amplitudes was produced in drainage basin as a result of the high floods from torrential rains. Such as: the different situations which are registred in September 1982, September 1999, February 2002, September 2004 and July 2005. These are proceed from very high water volumes of high floods after the level is increase.

Although the genesis conditions are relative the same, the evolution of lakes in the Romanian Black Sea Littoral area was different. The modiffications some time was irreversibile, these are not a result only of natural factors (the water rainfalls who have a very small quantity of salt substances have influence in dilution process, with a strongest effect at small surface lakes), in especially people factors influence. We can notice infact that the North half of Dobrouja lakes in the Romanian Black Sea Littoral area (until Midia Cape) are supply from surface flow and the South half of Dobrouja lakes in the Romanian Black Sea Littoral area are supply from subsurface sources. We can make a classifications such as:

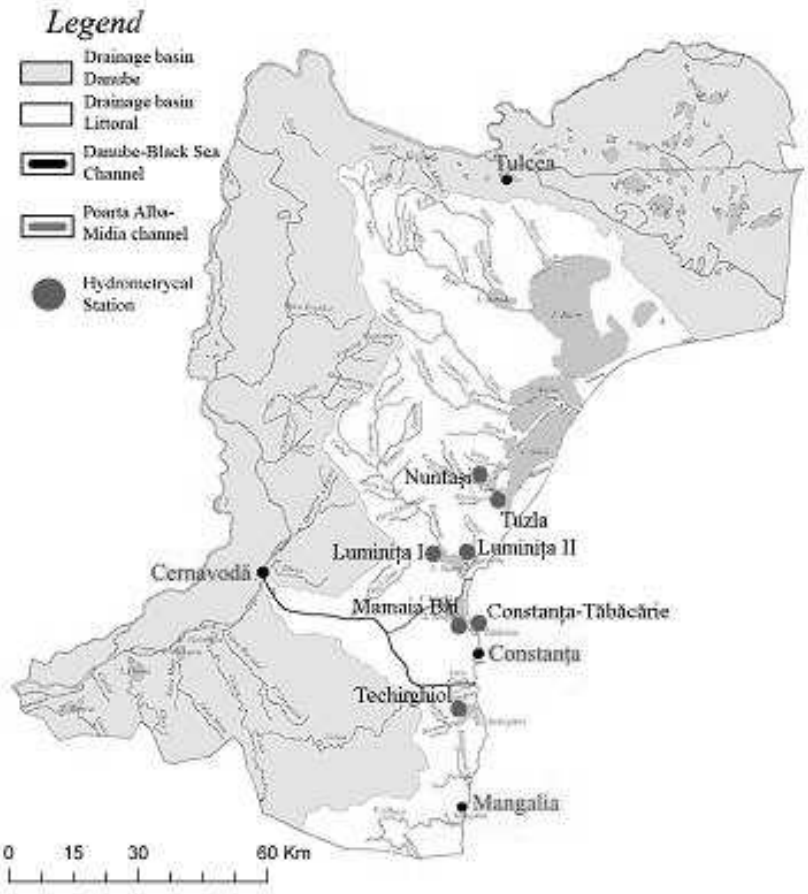


Figure 1: The hydrometry lake stations from hydrographical network of Romanian Water Authority from Dobrouda-Black Sea Littoral area

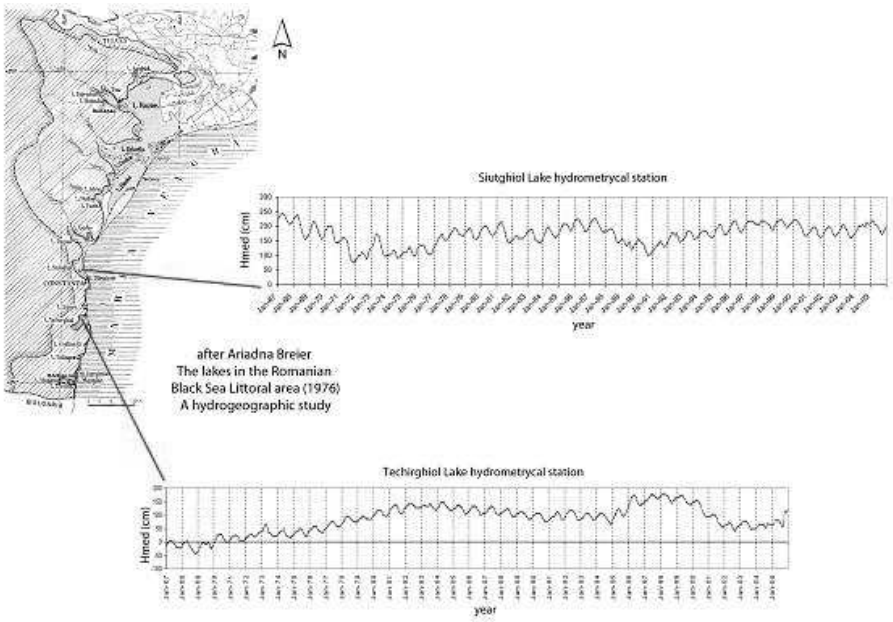


Figure 2: The monthly and year of medium level variation graphics at some lakes in the Romanian Black Sea Littoral area between 1967 - 2005

- Lakes without surface flow to the Black Sea (such as: Techirghiol Lake),
- Lakes which have a connection reversible with the Black Sea (such as: the Razim -Sinoe lakes complex),
- Lakes with surface flow to the Black Sea (such as: Tasaul Lake, Gargalac -Corbu Lake, Siutghiol Lake, Tabacarie Lake)

From the analysis of these data we can see that Le Turc method underestimate all values, FAO method make a overestimate (about 20 mm for both of them). For other methods the values which are getting through mathematical model are very near the measurement values (about 10 mm), some errors could be possible because all measurement intruments are very old (over 50 year of useing), even people factors influence which are near the stations primeters, and so on. Such as, Dobrouja climate specific feature have influenced the potential evapo-perspiration process (Fig.3).

From saturation deficit graphics we can see that all hydrographical area have a humidity deficit because evaporation value is increase and rainfall values are decrease. The maximum values are registred in july month with saturation deficit values about 110 mm through correlation between evaporation values from Constantza methorologycal station with rainfalls values from Mamaia Village and Techirghiol stations. We can notice that in all locations which are analysis it is an humidity excess it given by the rainfall quantity such as: for Mamaia Village evaporation station the real value about 25 mm and for Techirghiol evaporation station the real value about 20 mm (Fig.4). In time human fac-

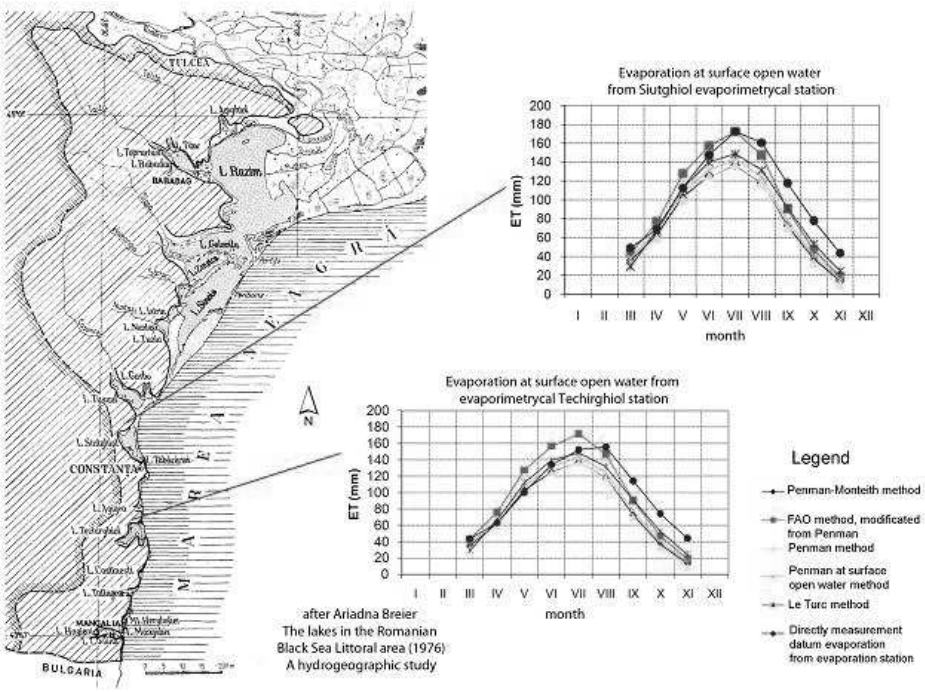


Figure 3: The monthly and many years of medium potential evapoperspiration variation graphics from evaporation stations at drainage basins of Siutghiol and Techirghiol Lakes between 1962-2005 period

tors influence gave dramatical modifications of water lakes chemical qualities evolution. Also these variation are produced in time and space. From present graphics we can see the quantity of salt is decrease through 700800 mg/dm3 in Siutghiol Lake to zero in present. And also we can see the salt values was about 7500095000 mg/dm3 at Techirghiol Lake in 1906 - 1907 period. Now this values is about 5600060000 mg/dm3 (Fig.5).

The groundwater which are subsurface flow in the South half of lakes in the Romanian Black Sea Littoral area are fresh, drinking and bicarbonate class.

The mineralization increase up 500 mg/l, the fixed residue value is between 500700 mg/l at Cismea catchwork, between 475550 mg/l at Caragea Dermen (drainage basin of Siutghiol Lake), between 600700 mg/l at Biruinta catchwork (in 161/2006 Order the limit value is between 500750 mg/l for ecological state and for good state to good. As we see the negative ions have the same equal sulphates values with clorine values at Caragea-Dermen i Cismea catchworks. In an other catchworks these values are increase even sulphates value is double (Fig.6).

Between 22-23 September 2005 in drainage basin of Costinesti Lake it was registered one flood with 162 cm maximum head and 200 m3/s return flow. At

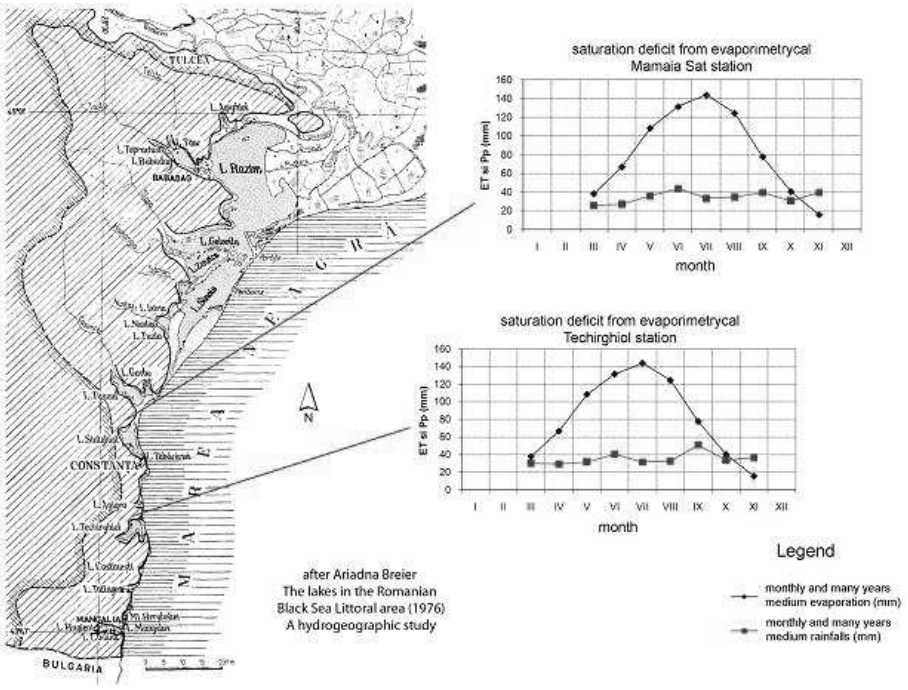


Figure 4: The monthly and multiannual medium saturation deficit variation graphics from evaporatin stations at drainage basins of Siutghiol and Techirghiol Lakes between 1962-2005

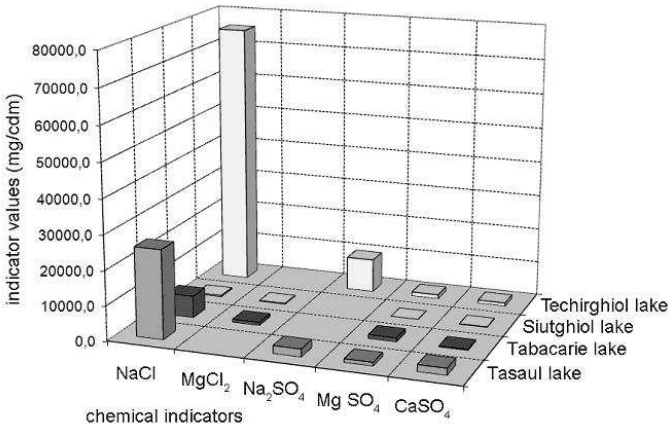


Figure 5: The water chemical parameters variation graphic of lakes in the Romanian Black Sea Littoral area between 1906 -1907 period (after the analysis certificate who was made by Romanian Geologycal Institute of Chemical Laboratory in mention period

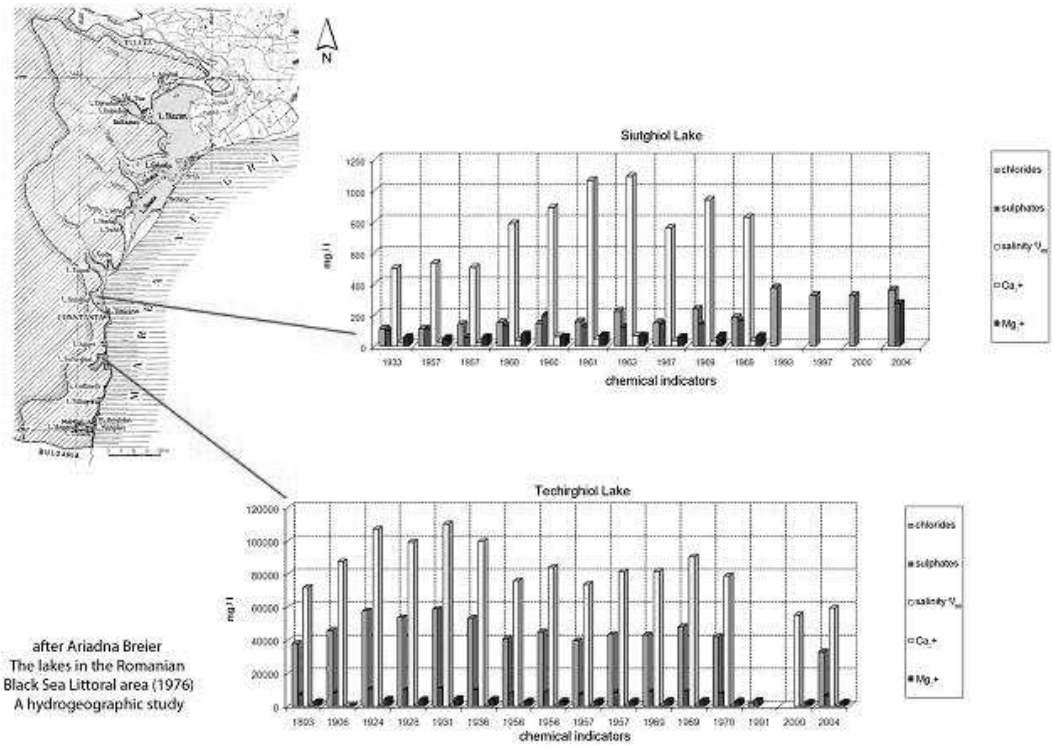


Figure 6: The water chemical indicators variation graphics of lakes in the Romanian Black Sea Littoral area between 1890 -2005

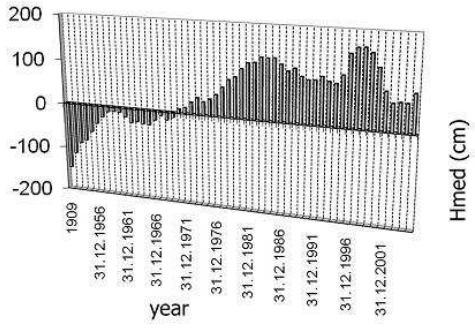


Figure 7: The monthly and many years medium levels variation graphic at Techirghiol Lake between 1952-2005

the Biruinta hydrometrycal station very near the upper watershed of drainage basin the rainfalls quatity was 222 mm/24 hours. This flow has a lot of damages and life loss, the water volume was broken on the seacoast line, which has separated lake on sea and it washing the sands over more hundreds meters lenght. In present the lake became what was long time ago a marine laguna.

Such as an accelerated process on people factors influence is Techirghiol Lake. Now these influences has an irreversible impact on water balance and to decrease salt value of this lake.

Started with 1909 the level was growing up by rainfalls, gauge heigth was about -150 cm, in 1952 gauge heigth was -115 cm. In 1954-1966 period years the level was growing up wit 71 cm; between 1967-1970 the levels are relative constant with a little decrease trend in 1968. After 1968 to 1984 it was an increase gauge heigth 132 cm, with decrease trend in 1974 to gauge heigth 29 cm (Fig.7).

The thermic regime of water lakes variation in a year is in function by the air temperature variation which is dependent, in fact we can see from the correlation graphic between medium monthly values of water and air temperature to Mamaia Bai hydrometrycal station on Siutghiol Lake and Techirghiol hydrometrycal station on Techirghiol Lake (Fig.8). The difference given by water rivers which is thermically influenced by the climatic factors through the water flow, the variation of water lakes is directly dependent by the air variation where the lakes have the bed fundation. The water thermic regime was calculated only for its surface, as it is well known that in fact the variation temperature is vertically.

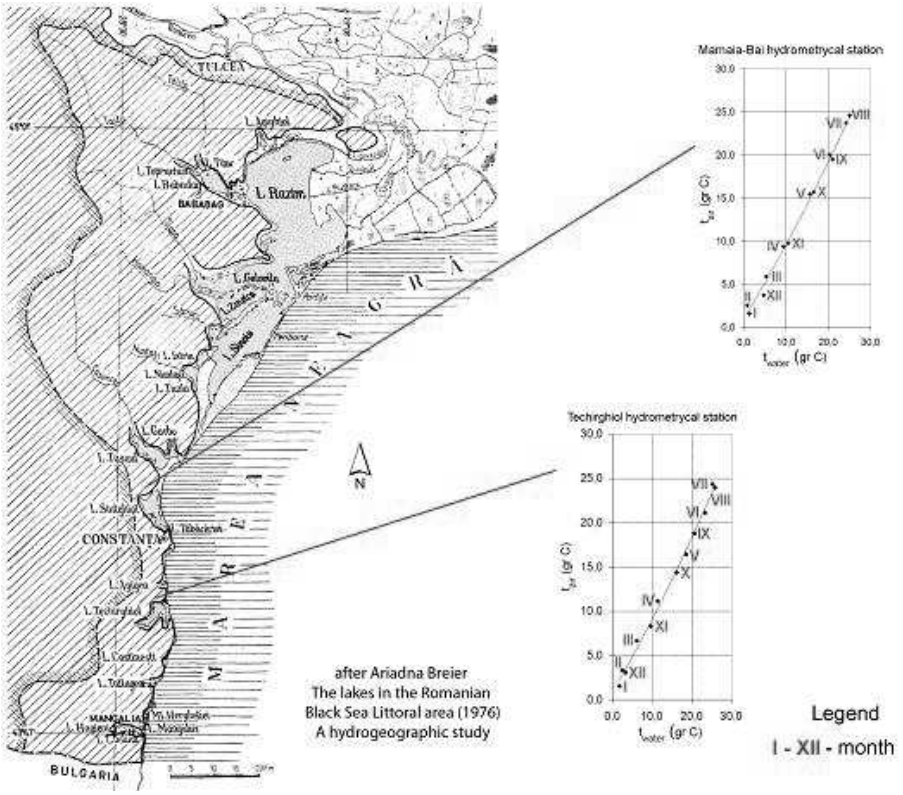


Figure 8: The correlation between water and air temperature at hydrometryal stations function by metheorologycal stations from drainage basins af Siutghiol and Techirghiol Lakes between 1989-2005

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Delayed Coking Modeling, Scheduling and Control

Dana-Cristina Toncu
Department of Aerospace Engineering,
Politehnica University of Bucharest
E-mail: cristinatoncu@canals.ro

Abstract

Modern Delayed Coking continues to have an important role in processing hydrocarbon residues, remaining a basic refinery operation. Therefore, modeling and scheduling this basic refinery operation has always been a challenge facing its complexity. Aiming to develop and apply a simple, flexible and accurate model for delayed coking process in order to improve product quality and profit, this paper seeks for a practical solution suitable for designing, planning, scheduling and, why not, optimization in dynamic, continuous coking unit system.

Keywords: delayed coking; modeling; scheduling; control.

1 Introduction

The delayed coking plant is mostly used for the conversion of residue not suitable for catalytic processes (with lower API gravity and higher carbon residue content) into distillates which may be further catalytically upgraded, with relatively low costs. Common feedstock presents large concentration of resins, asphaltenes, heteroatomic compounds (sulfur, nitrogen, oxygen, metals), which are considered catalyst poison and generally end up in coke. Delayed coking can process a wide variety of feedstocks, with considerable metals (especially nickel and vanadium), sulfur, resins, asphaltenes, but the typical feed is vacuum residue. Coking is a carbon rejection process in which hydrogen passes to the lighter products and metals concentrate in solid product, through a cycle of saturated and cyclo-paraffins and PNAs cracking and combining.

The overall reaction is favored by high temperatures and low pressures, low recycle ratios, and decreased cycle time for coking. Products obtained are gases, gasoline, gas oil and coke, which is the final product. Distillates feed the following plants:

- gases - catalytic cracking or gas desulphurization and sulfur recovery;
- gasoline - hydrotreatment;

ISSN-1584-5990

- light gas oil - hydrotreatment;
- heavy gas oil - catalytic cracking.

The main equipment is made of: heater; coke drum vessels; fractionation column; downstream vapor processing vessels.

The technological flow sheet consists of the following sections: feed and reaction; fractionation; gas compression; fast blowdown; water recovery.

The overall process of a 1 170 000 t/year plant takes 33.5 hours:

- feeding coke - 16 h;
- steam strip - 1.5 h;
- water cooling - 5 ... 5.5 h;
- drainage - 2 h;
- hydraulic de-coking - 5 h;
- drum closure - 2 h;
- steam heating for pressure test - 1 h;
- pressure test - 0.5 h.

There is also a general concern regarding ecologization of the plant, enclosing the blowdown system and treating the water from the hydraulic de-coking, which is extremely toxic and requires special methods.

In the last years, literature [1]–[11] has done some pertinent observations:

- Coke appears to form through a intermediate mesophase.
- The blending feed entering the drum is an isotropic phase.
- During reaction, strongly aromatic compounds become insoluble in the overall paraffinic phase, and so a second liquid phase.
- Under favorable conditions, spheres coalesce into larger regions before the reactions that transform the mesophase into coke occur.
- Coke forming tendency of the residue is an important variable in determining the optimum operating conditions.
- Coke quality depends on mesophase coalescence.
- Average size and quality of coke depends on the mesophase.
- A highly reactive mesophase reacts quickly to form shot coke (hard, dense material perturbing coking operations and reducing coke's value);
- When mesophase coalescences, higher value needle coke forms.

The main aim of any industrial process is to reach profit maximization through cost minimization and it can be achieved mainly with a proper model for design, exploiting, planning, scheduling and optimization.

2 Problem formulation

2.1 Formulation for Modeling

In the last years, there was a trend in developing [1]–[11] yield models and methods to predict coke quality.

The algorithm sets the requirement function as follows:

$$\min F \leq p \tag{1}$$

where: p is the required characteristic.
and fixes the parameters exhibitting bounded uncertainty:

$$a_{1,i} \leq a_{k,i} \leq a_{n,i} \tag{2}$$

Additional constant increases model accuracy:

$$\min F \leq p \pm c \tag{3}$$

2.2 Formulation for Scheduling

Given:

- production and storage capacity;
- equipment and their characteristics;
- production, storage and selling policy;
- production requirements;
- time considered;

determine:

- optimal task sequence;
- processed material at eact time;
- each task processing time;

in order to optimize at least one performance criterion: maximization of profit.

Qualitative transfer function between settled variables are then defined as followes:

$$minOperatingCost = InputCost + ProductionCost \tag{4}$$

where $Input\ Cost = Feedstock\ Cost + Inventory\ (stock)\ Cost$.

Material balance equation for the whole unit is compulsory for model accuracy:

$$M_{input} = M_{output} \tag{5}$$

where M states for material (or mass).
Literature reported some surces of uncertainty [5]:

- task processing time;
- product demand;
- product and feed price.

2.3 Modeling Methods

There are several approaches regarding modeling delayed coking:

- kinetic model, based on the chemical reaction and laboratory experiments;
- stochastic model (Monte Carlo method), constructing intermediar pseudocomponents and relying on analytical chemistry, cumulative probability density functions, phases equilibria and mass balance;
- empirical model, correlating coke yield with feed's physico-chemical properties, operating conditions, studies on pilot plants and commercial units;
- quality model, relating size of mesophase and coke quality.

The following model assumptions were drafted:

- order-driven operation parameters;
- deterministic model parameters;
- continuous/discontinuous process;
- single or multi-batch (of integer number) process;
- resouces, equipment or products (quality and yield) constraints.

Mathematical model has the following limitation objectives:

- allocation constraints;
- timing constraints;
- sequencing constraints.

3 Suggested Solution

The best approach to modeling a delayed coking process is the empirical one, based on actual operations.

A simple model was chosen, which included key parameters (physical and/or actions) and production constraints (relations between key parameters and quality requirements). These are encoded in model as components.

Modeling followed several steps:

- establishing correlations: settling terms and variable equation;
- finding correction;
- adjustment;
- ploting results;
- determining products' value (establishing and analysing products);
- expressing performance criterion: minimizing costs, maximizing profit (substracting costs from product value);
- objective function verification.

Chosen model variables are:

- feed and its characteristics;
- operating conditions: temperature, pressure, heavy gas oil recycle;

The feed property p is expressed as a function of unit capacity and input streams' characteristics:

$$p_F = f(Q_F, \sum p_{S,i}) \tag{6}$$

Each product capacity is given by a function of feed capacity and properties, on one hand, and operating variables, on the other hand:

$$Q_S = f(Q_{F_i}, p_{F_i}, \sum V_n) \tag{7}$$

where V_n are the operating variables.

Output stream characteristics (or product characteristics) are set as functions of feed properties and operating variables:

$$p_{S,o} = f(p_F, V_n) \tag{8}$$

The last two equations define the process model for delayed coking plant.

The system equation is completed with mass balance and energy balance equations.

The total mass balance may not be satisfied, due to material losses and unmodeled streams on one hand, and to the static or dynamic approach, on the other hand.

In the case of a dynamic system, a differential equation describes the process:

$$\frac{\partial x}{\partial t} = t_{x,c} \tag{9}$$

with initial conditions $x(0)$ and x_0 , where c is the control function of control parameters:

$$c = f(cp) \tag{10}$$

It results that:

$$\frac{\partial x}{\partial t} = f(x, g(cp, x)) \tag{11}$$

All variables are considered to be measurable.

Plant performance is a function of state and control:

$$J = \int_0^{t_f} f(x, c)dt \tag{12}$$

Constaints are expressed as inequalities:

$$f(x) \leq x_{set} \tag{13}$$

$$f(c) \leq c_{set} \tag{14}$$

3.1 Experimental

Due to the crude oil's price, feed was a blend containing different residua obtained in the refinery, slops and slurry. Laboratory analysis for feed and products was carried out.

3.2 Results and Discussions

The chosen unit to model is operated under conditions shown in Fig.1 in order to produce coke having the characteristics given in Fig.2.

As known from literature, a simple correlation was found between coke yield and Conradson Carbon residue of feedstock, revealed by Figs.4-5.

The overall mass balance for the plant is presented in Fig.3.

2-nd order polynomial regression coefficient obtained was 0.858, which is closed to optimum value 1. It resulted a different equation from the one reported in literature:

Coking the blend residue (A), the following products result: gases with 1-4 carbon atoms in molecule $(C_1 - C_4)$ (A_1), gas oil with final boiling point around

Delayed Coking operating conditions		
Parameter	Fractionation column	Coking drum
Pressure, bar	1,21 ... 1,25	1,96 ...2,14
Temperature, C	492,41 ... 493,43	490 ...495
Recycle rate	10% of feed (14379 kg/h)	
Inner drum vapour rate, m/s	0,0725 ... 0,082	

150-180 Celsius (A_2), two sorts of gas oil, one with initial boiling point at 182 Celsius (A_3), the other one at 222 Celsius (A_4) and coke (A_5), according to the following scheme:

$$Y = -703.949 \cdot CCR^2 + 51.688 \cdot CCR - 0.928 \tag{15}$$

Characteristic	Value
Humidity, %m/m	7
Ash, %m/m	1
Sulfur, %m/m	4,5 ... 5,0
Volatile matter, %m/m	13
Superior calrific power, kJ/kg	34300

$$A \rightarrow \nu_{1,1}A_1 + \nu_{1,2}A_2 + \nu_{1,3}A_3 + \nu_{1,4}A_4 + \nu_{1,5}A_5 \tag{16}$$

where $\nu_{i,j}$ are the product yields.

Considering that both sorts of gas oil suffer further thermal cracking into gases (A_1), gasoil (A_2) and coke (A_5) after a similar scheme:

$$A_3 \rightarrow \nu_{2,1}A_1 + \nu_{2,2}A_2 + \nu_{2,5}A_5 \tag{17}$$

$$A_4 \rightarrow \nu_{3,1}A_1 + \nu_{3,2}A_2 + \nu_{3,5}A_5 \tag{18}$$

and that $\nu_{i,j}$ are product yields of total conversion of feed, characteristic to tha reaction (hence mass conversion coefficient):

$$\nu_{1,1} + \nu_{1,2} + \nu_{1,3} + \nu_{1,4} + \nu_{1,5} = 1 \tag{19}$$

$$\nu_{2,1} + \nu_{2,2} + \nu_{2,5} = 1 \tag{20}$$

Mass balance for delayed coking commercial plant [m3/h]			
Input		Output	
Residue	2514	Gases with hydrogen sulfide	250
Light fraction hydrotreatment	27	Gasoline	243
Catalytic cracking gas oil (slurry)	178	Gas oil 1	684
Slops	50	Gas oil 2	943
Total	2769	Petroleum coke	586
		Fraction from column	20
		Technological loss	43
		Total	2769

Figure 1: Unit mass balance

$$\nu_{3,1} + \nu_{3,2} + \nu_{3,5} = 1 \tag{21}$$

knowing gas oil yields, gas, gasoline and coke yields can be calculated with the following serie of formulae:

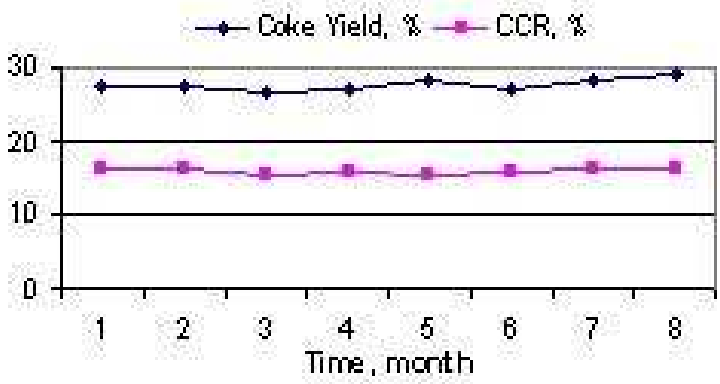
$$gA_1 = \nu_{1,1} + \nu_{1,3} \cdot \nu_{2,1} + \nu_{1,4} \cdot \nu_{3,1} - \nu_{2,1} \cdot g \cdot A_3 - \nu_{3,1} \cdot g \cdot A_4 \tag{22}$$

$$gA_2 = \nu_{1,2} + \nu_{1,3} \cdot \nu_{2,2} + \nu_{1,4} \cdot \nu_{3,2} - \nu_{2,2} \cdot g \cdot A_3 - \nu_{3,2} \cdot g \cdot A_4 \tag{23}$$

$$gA_5 = \nu_{1,5} + \alpha \cdot \nu_{1,3} \cdot \nu_{2,5} + \beta \cdot \nu_{1,4} \cdot \nu_{3,5} - \nu_{2,5} \cdot g \cdot A_3 - \nu_{3,5} \cdot g \cdot A_4 \tag{24}$$

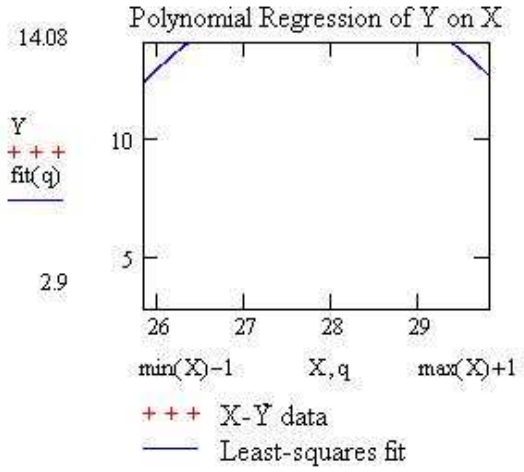
where gA_i represents the products yield, in mass, relating to feedstock A, and α, β are coefficients.

All coefficients are determined according to the particular feedstock processed or to be processed. For example, for several Russian residua, in the case of anly one gas oil sort obtained, $\nu_{1,1} = 0.07$, $\nu_{1,2} = 0.07$, $\nu_{1,3} = 0.8$, $\nu_{1,4} = 0.05$, $\nu_{2,1} = 0.22$, $\nu_{2,2} = 0.25$, $\nu_{2,4} = 0.52$, for A_1 - gas, A_2 - gasoline, A_3 - gas oil, A_4 - coke. Experiments shown that coefficients remain constants to process parameters. The hypothesis according to which they are not changed by the feedstock's nature [12] could not be validated as different blends were used, therefore coefficients had different values.



4 Conclusion

Experience proved that delayed coking can be modeled in a simple logic approach, with satisfactory accuracy for further control, schedule and optimization. Far for being exhaustive, research can be further focused on finding a flexible model for new processing challenge - residua.



References

[1] M.Joly, L.F.L.Moro, and J.M.Pinto, *Planning and Scheduling for Petroleum Refineries Using Mathematical Programming*, Braz. J. Chem. Eng. 19 (2002).

[2] O.O. Bello, B.T. Ademodi, S.R.A. Macaulay, G.K. Latinwo, *Effects of operating conditions on compositional characteristics and reaction kinetics of liquid derived by delayed coking of nigerian petroleum residue*, Braz. J. Chem. Eng. 20 (2002).

[3] J. M. Pinto and L. F. Maro, *A Planning Model for Petroleum Refineries*, Braz. J. Chem. Eng. 19 (2000).

[4] K. Rajagopal, *Modelling Pyrolysis and Carbonization of Petroleum Distillation Residues in Delayed Coking Operations*, Proceedings FOCAPO 2003 509-512.

[5] Xiaoxia Lin et al., *A new robust optimization approach for scheduling under uncertainty: I. Bounded uncertainty*, Computers and Chemical Engineering 28 (2004) 1069-1085.

[6] Zhenya Jia, Marianthi Ierapetritou, *Efficient short-term scheduling of refinery operations based on continuous time formulation*, Computers and Chemical Engineering 28 (2004) 1001-1019.

[7] Donald E. Shobrys, Douglas C. White, *Planning, scheduling and control systems: why cannot they work together*, Computers and Chemical Engineering 26 (2002) 149-160.

[8] Christodoulos A. Floudos, Xiaoxia Lin, *Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review*, Computers and Chemical Engineering 28 (2004) 2109-2129.

[9] Carlos A. Mendez, Jaime Cerda, *Dynamic scheduling in multiproduct batch plants*, Computers and Chemical Engineering 27 (2003) 1247-1259.

[10] Giulia Bozzano, Mario Dente, *A Mechanicistic Approach to Delayed Coking Modelling*, Proceeding of European Symposium on Computer Aided Process Engineering 529-535.

[11] Tom B. Bechtel, *A Simplified Numerical Approach to Feedback Controller Design*, C. J. Ch. E. 82 (2004) 1319-1325.

[12] A. G. Sardanasvili, A. I. Lvova, *Oil and Gas Processing, Exercises and problems*, 2nd Edition, Ed. Tehnica (1985) 145-152.

SECTION B

MECHANICS

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Equilibrium Points in the Rein's Model for Semi-averaged Planar Elliptic Restricted Three-body Problem

Mihai Bărbosu
SUNY Brockport, Dept. of Mathematics, Brockport, NY, USA
mbarbosu@brockport.edu

Tiberiu Oproiu
Astronomical Observatory, Cluj-Napoca, Romania

Abstract

The paper deals with the "simplified semi-averaged" scheme for the planar elliptic restricted three-body problem, which admits a prime integral analogous to the Jacobian integral from the corresponding circular problem. The abscissa of the $\overline{L}_1$ collinear double point (between primaries) was studied as function of the finite masses ratio $m = m_2/m_1$ for $m \in [0.1, 1.0]$ and for eccentricity $e \in [0.0, 0.1]$. The abscissa of $\overline{L}_1$ does not essentially depend on the orbital eccentricity.

Keywords: three-body problem, mass, averaging method

As it is [7], the planar elliptic three-body problem consists of studying the motion of a body P under the gravitational action of two bodies P_1, P_2, with the following restrictions:

(i) the body P has infinitesimal mass; therefore it does not influence the motion of the bodies P_1 and P_2 which have finite masses m_1 and m_2, respectively;

(ii) the motions of P_1 and P_2 are given as solution of a two-body problem; they describe elliptic orbits having a common focus in the common mass centre O, the eccentricity e, the semimajor axes a_1 and a_2, and the mean motion n;

(iii) the motion of P takes place in the orbital plane of P_1 and P_2. In order to study the motion of P, one chooses a uniformly rotating frame $O\xi\eta$. This frame rotates around O with the constant angular velocity n. The coordinates of the bodies on this frame are $P_1(\xi_1, \eta_1)$, $P_2(\xi_2, \eta_2)$ and $P(\xi, \eta)$.

In the rotating frame, the bodies P_1 and P_2 describe closed curves around the "mean" points $\overline{P}_1(-a_1, 0)$ and $\overline{P}_2(a_2, 0)$, respectively. The positions of P_i, $i = 1, 2$, are given in this frame by:

$$\begin{aligned} \xi_i &= (-1)^i a_i(1-e^2)'cos(v-nt)/(1+e\cos v), \\ \eta_i &= (-1)^i a_i(1-e^2)\sin(v-nt)/(1+e\cos v),\ i=1,2, \end{aligned} \tag{1}$$

ISSN-1584-5990

where v stands for the true anomaly.

The planar motion of the infinitesimal mass with respect to the 0-frame is described by the differential system:

$$\begin{aligned} \ddot{\xi} - 2n\dot{\eta} &= U_\xi, \\ \ddot{\eta} + 2n\dot{\xi} &= U_\eta, \end{aligned} \tag{2}$$

where subscripts mark the corresponding partial derivatives, and:

$$U = (n^2/2)(\xi^2 + \eta^2) + k^2 m_1/r_1 + k^2 m_2/r_2. \tag{3}$$

In the above formulae, k is Gauss' constant, while r_1 and r_2 are given by the expressions:

$$r_i^2 = (\xi - \xi_i)^2 + (\eta - \eta_i)^2, \quad i = 1, 2. \tag{4}$$

Since the coordinates (ξ_i, η_i), $i = 1, 2$, are expressed by eqs. (1), the force function of the problem becomes an explicit function of t, namely $U = U(\xi, \eta; t)$. For this reason, the differential system (1) does not admit a first integral analogous to Jacobi's integral which appears in the case of the circular restricted three-body problem.

Several authors have proposed simplified schemes for the elliptic restricted three-body problem, by using the averaging method. In this paper we shall use Rein's "semi-averaging" scheme [6], which is valid in the following hypotheses:

(i) As it is known, one can express the tim-dependence of the coordinates (ξ_i, η_i), $i = 1, 2$, given by eqs. (1), by means of infinite power series of the eccentricity:

$$\begin{aligned} \xi_i &= (-1)^i a_i(1 + e(-e/2) - (1 + 3e^2/8)\cos(nt) + \\ &\quad + (e/2)\cos(2nt) + (3e^2/8)\cos(3nt) + \dots) + \dots), \\ \eta_i &= (-1)^i a_i e((2 - 3e^2/8)\sin(nt) + (e/4)\sin(2nt) + \\ &\quad + (7e^2/24)\sin(3nt) + \dots), \ i = 1, 2. \end{aligned} \tag{5}$$

One can obtain finite expressions (depending on t) for the coordinates of P_1 and P_2 in the case of e small enough to neglect e^j, $j \geq 2$; denoting these coordinates by $\overline{\xi}_i, \overline{\eta}_i$, we have:

$$\overline{\xi}_i = (-1)^i a_i(1 - e\cos(nt)), \ \overline{\eta}_i = 2(-1)^i a_i e\sin(nt), \ i = 1, 2. \tag{6}$$

These equations represent ellipses having the centres in the points $\overline{P}_1$ and $\overline{P}_2$, respectively, and the major axes parallel to $O\eta$-axis. The length of the semiaxes A_i and B_i, $i = 1, 2$, of these ellipses can be determined by using the formulae:

$$A_i = 2a_i e, \quad B_i = a_i e, \quad i = 1, 2. \tag{7}$$

(ii) Instead of the force function U, one considers its time-averaged value $\overline{U}$, given by:

$$\overline{U} = (n^2/2)(\overline{\xi}^2 + \overline{\eta}^2) + (1/T)\int_{t_0}^{t_0+T} (k^2 m_1/r_1 + k^2 m_2/r_2)dt, \tag{8}$$

with $T = 2\pi/n$, and r_1, r_2 given by (4) in which ξ_i, η_i were replaced by $\overline{\xi}_i, \overline{\eta}_i$. In order to emphasize that the motion of P is governed by the force function $\overline{U}$, we denoted the frame $O\xi\eta$ by $O\overline{\xi}\overline{\eta}$.

The averaged function $\overline{U}$ can also be written in the form (see [6]):

$$\overline{U} = (n^2/2)(\overline{\xi}^2 + \overline{\varepsilon}^2) + k^2 m_1 M_1/\nu_1 + k^2 m_2 M_2/\nu_2, \tag{9}$$

where $\nu_i^2 = \lambda_i^2 - \mu_i^2$ (λ_i, μ_i = elliptic coordinates in the system of ellipses having the same foci as the ellipses of P_1 and P_2, respectively), and:

$$M_i = (2/\pi)K(\chi_i), \quad i = 1, 2, \tag{10}$$

while $K(\chi_i)$ represent the elliptic integrals of first kind:

$$K(\chi_i) = \int_0^{\pi/2} (1 - \chi_i^2 \sin^2 x)^{-1/2} dx, \quad i = 1, 2, \tag{11}$$

the modules χ_i, $i = 1, 2$, being determined from:

$$\chi_i = (4a_i^2 e^2 - \mu_i^2)/(\lambda_i^2 - \mu_i^2), \quad i = 1, 2. \tag{12}$$

The coordinates (λ_i, μ_i), $i = 1, 2$, can be determined from the Cartesian coordinates $(\overline{\xi}, \overline{\eta})$ by means of the following formulae [1,2,6]:

$$\begin{aligned} \lambda_i^2 &= (f_i + (f_i^2 - g_i)^{1/2})/2, \\ \mu_i^2 &= (f_i - (f_i^2 - g_i)^{1/2}/2, \ i = 1, 2, \end{aligned} \tag{13}$$

where:

$$\begin{aligned} f_i &= (\overline{\xi} - (-1)^i a_i)^2 + 3a_i^2 e^2, \\ g_i &= 12 a_i^2 e^2 \overline{\eta}^2, \ i = 1, 2. \end{aligned} \tag{14}$$

The differential equations of the relative motion of the body P in Rein's semi-averaged scheme have the form:

$$\begin{aligned} \ddot{\overline{\xi}} - 2n\dot{\overline{\eta}} &= \overline{U}_{\overline{\xi}}, \\ \ddot{\overline{\eta}} + 2n\dot{\overline{\xi}} &= \overline{U}_{\overline{\eta}}. \end{aligned} \tag{15}$$

Eqs. (15) admit a prime integral analogous to the Jacobian integral from the circular restricted three-body problem [1]:

$$\dot{\overline{\xi}}^2 + \dot{\overline{\eta}}^2 = 2(\overline{U} + h), \tag{16}$$

h being the integration constant.

The zero relative velocity curves are given by the equation:

$$\overline{U}(\overline{\xi}, \overline{\eta}) + h = 0. \tag{17}$$

It is easier to study these curves in the $\overline{P}_1xy$-frame, which is obtained by translating the $O\overline{\xi}\overline{\eta}$-frame along the $O\overline{\xi}$-axis, such that O coincides with $\overline{P}_1$. The relations between the old and the new coordinates will be:

$$x = \overline{\xi} + a_1, \quad y = \overline{\eta}, \tag{18}$$

where:

$$a_1 = am_2/(m_1 + m_2), \quad a_2 = am_1/(m_1 + m_2), \tag{19}$$

with $a =$ distance between $\overline{P}_1$ and $\overline{P}_2$.

Introducing the normalization with respect to a:

$$\overline{q} = q/a, \quad q \in \{x, y, r_i, \overline{\lambda}_i, \overline{\mu}_i, \overline{\nu}_i\}, \quad i = 1, 2, \tag{20}$$

the zero relative velocity curves will be given by [6]:

$$\overline{U}(\overline{x}, \overline{y}) + h = (k^2/a)((m_1 + m_2)\overline{r}_1^2/2 - m_2\overline{x} + m_2^2(2(m_1 + m_2)) +$$

$$+m_1M_1/\overline{\nu}_1 + m_2M_2/\overline{\nu}^2) + h = 0, \tag{21}$$

where:

$$\overline{r}_1^2 = \overline{x}^2 + \overline{y}^2, \quad \overline{r}_2^2 = (\overline{x} - 1)^2 + \overline{y}^2. \tag{22}$$

The double points of the zero relative velocity curves are given by the real roots of the algebraic system $\overline{U}_{\overline{x}} = 0$, $\overline{U}_{\overline{y}} = 0$, that is:

$$\begin{array}{l}(m_1 + m_2)\overline{x} - m_2 + m_1(M_1/\overline{\nu}_1)_{\overline{x}} + m_2(M_2/\overline{\nu}_2)_{\overline{x}} = 0,\\ (m_1 + m_2)\overline{y} + m_1(M_1/\overline{\nu}_1)_{\overline{y}} + m_2(M_2/\overline{\nu}_2)_{\overline{y}} = 0.\end{array} \tag{23}$$

As to the circular restricted three-body problem, the system (23) has five roots represented by three collinear points $\overline{L}_i(\overline{x}_i, 0)$, $i = 1, 2, 3$, and two quasi-equilateral points $\overline{L}_i(\overline{x}_i, \overline{y}_i)$, $i = 4, 5$.

One can find numerical algorithms to determine the coordinates of these points in [3,4,5,6].

The slope in the collinear double point $\overline{L}_1$ (situated between $\overline{P}_1(0, 0)$ and $\overline{P}_2(1, 0)$) was determined from the expression [7]:

$$\tan^2 \varphi_{\overline{L}_1} = -(\overline{U}_{\overline{x}\overline{x}}/\overline{U}_{\overline{y}\overline{y}})_{\overline{x}=\overline{x}_1, \overline{y}=0}, \tag{24}$$

for different values of the eccentricity and of the mass ratio $m = m_2/m_1$ [4].

In this note we studied the variation of $\overline{L}_1$ collinear double point's abscissa as function of m and e, The results are given in Table 1.

Figure 1 plots $\overline{x}_{L1}$ versus m for two values of the eccentricity; $e = 0$ and $e = 0.1$. Examining Table 1 and Figure 1, we can conclude that:

(i) the values for $\overline{x}_1$ do not depend sensibly on m; for small values of the mass ratio m the differences are more important;

(ii) from Table 1 we can see that $\overline{x}_{L_1}(m, e) \to 0.5$ as $m \to 1$.

Table 1: The values of $x_{\overline{L1}}$ vs. m and e

m	$e = 0.00$	$e = 0.02$	$e = 0.04$	$e = 0.06$	$e = 0.08$	$e = 1.00$
0.10	0.717512	0.718013	0.719454	0.721662	0.724343	0.727096
0.20	0.658556	0.658885	0.659850	0.661384	0.663375	0.665667
0.30	0.620866	0.621105	0.621807	0.622940	0.624446	0.626243
0.40	0.592947	0.598125	0.593650	0.594502	0.595649	0.597043
0.50	0.570751	0.570884	0.571276	0.571916	0.572782	0.573845
0.60	0.552343	0.552439	0.552726	0.553194	0.553832	0.554618
0.70	0.536634	0.536701	0.536889	0.537226	0.537670	0.538221
0.80	0.522950	0.522992	0.523116	0.523320	0.523597	0.523942
0.90	0.510844	0.510863	0.510922	0.511018	0.511149	0.511312
1.00	0.5	0.5	0.5	0.5	0.5	0.5

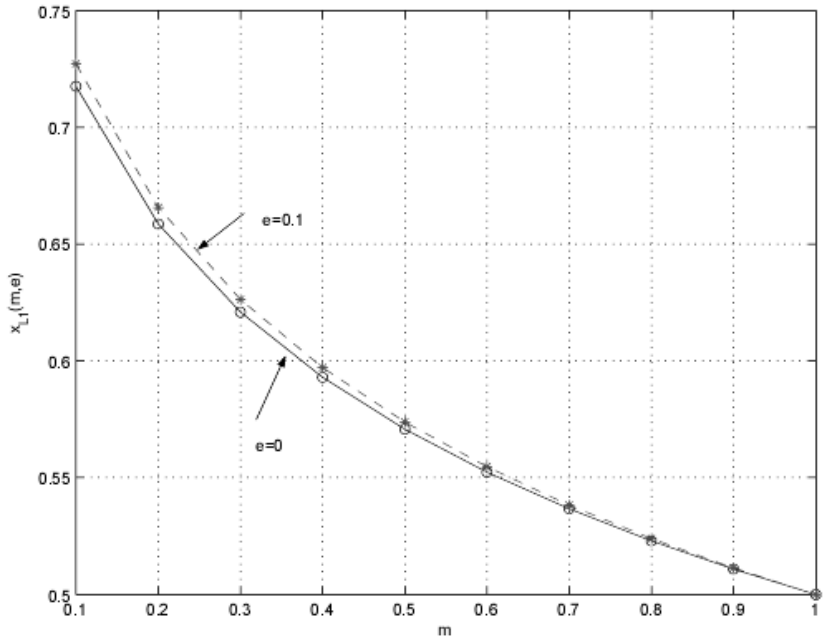


Figure 1: Collinear libration point $\overline{L}_{L1}$

References

[1] Pál, Á., *A Mathematical Model of the Elliptical Restricted Three-Body Problem*, Babeş-Bolyai Univ., Fac. Math. Res. Sem., **4**(1982), No. 3, 114-123.

[2] Pál, Á., *Application of the Model of the Elliptical Restricted Three-Body Problem to the Study of the Binary Stars*, Babeş-Bolyai Univ., Fac. Math. Res. Sem., **5**(1983), No. 4, 9.

[3] Pál, Á., Oproiu, T., *On the Determination of the Collinear Double Points in the Averaged Plane Elliptic Restricted Three-Body Problem*, Babeş-Bolyai Univ., Fac. Math. Phys. Res. Sem., **10**(1988), No. 10, 135-147.

[4] Pál, Á., Oproiu, T., *Determination of the Slopes of "Zero Relative Velocity" Curves from the Elliptic Restricted Three-Body Problem with an Averaging Method*, Romanian Astron. J., **1**(1991), 97-102.

[5] Pál, Á., Oproiu, T., Macaria, R., *Soluţii de echilibru în problema restrânsă eliptică a celor trei corpuri*, Communication held at the Scientific Session dedicated to the centennary of the birth of Professor Constantin Pârvulescu, Bucharest, 8-9 November 1990.

[6] Rein, N. F., *O kachestvennykh kharakteristikakh dvizheniya v uproschchennoj poluosrednennoj ellipticheskikh ogranichennoj probleme trekh tel*, Trudy GAIS, **14**(1940), 127-152.

[7] Szebehely, V., *Theory of Orbits. The Restricted Problem of Three Bodies*, Academic Press, New York, London, 1967.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Results of the Application of d'Alembert's Principle for Rigid Bodies in Rotation Motion

[1]Mihail Boiangiu, [2]Aurel Alecu
[1,2]Department of Mechanics, "Politehnica" University of Bucharest
Splaiul Independentei, no.313, sector 6, Bucharest, Romania
E-mails: [1]mboiangiu@gmail.com, [2]aurel_alecu@yahoo.com

Abstract

In this paper the authors approached few problems of application of the d'Alembert's principle. One problem is the position of central axis of the d'Alembert's fictitious forces system for a rigid body in rotation motion. Other problem approached is the calculus of the centrifugal moments for plane plates. A calculus formula based on the coordinates of centers of mass is obtained. Also in this paper is studied the position of the support of the resultant vector of d'Alembert's fictitious forces system, for plane bars and plates having uniform rotation motion.

Keywords: D'Alembert's principles, rotation, d'Alembert's fictitious forces, central axis, centrifugal moments, plane plates, plane bars.

1 Introduction

In more problems, where appear bodies having rotation motion when we apply the d'Alembert's principle, it is good to know the position of the central axis of the d'Alembert's fictitious forces. The determination of the central axis is some time difficult.

For bars and plates having uniform rotation motion, the solving of the problem of determination of the reactions is simple if we know the position of the support of resultant vector of d'Alembert's fictitious forces system.

Also, in problems where appear plates and bars having rotation motion, when we apply the theorem of the angular momentum, we need the centrifugal moments. The determination of the centrifugal moments can be some time laborious. In literature, the values of the centrifugal moments are calculated by integration.

ISSN-1584-5990

2 Central axis of the d'Alembert's fictitious forces system

Let us consider a rigid body having rotation motion (figure 1a). The axis of rotation is defined by two distinct fixed point (smooth spherical joints) $A\,(0,0,h_1)$ and $B\,(0,0,h_2)$. The body is acted by the given forces (including the force of gravity) $\vec{F}_i\,(i=1,...,n)$. We consider a rigid Cartesian reference system $O_1x_1y_1z_1$ so that O_1z_1 is rotation axis, and a movable Cartesian reference system $Oxyz$ so that Oz is the rotation axis, $O\equiv O_1$ and the center of mass of the rigid body is situated on the Ox axis, its coordinates being $C\,(\xi,0,0)$. We denote by R_x, R_y, R_z, M_x, M_y, M_z the projections of the resultant vector $\vec{R}$ and respectively the resultant moment $\vec{M}_o$ of given forces $\vec{F}_i$ (with respect to the point O) on the axes of movable reference system $Oxyz$. We replace also the spherical joints from A and B by the reactions $\vec{R}'\left(R'_x,R'_y,R'_z\right)$, respectively $\vec{R}''\left(R''_x,R''_y,R''_z\right)$ (figure 1b).

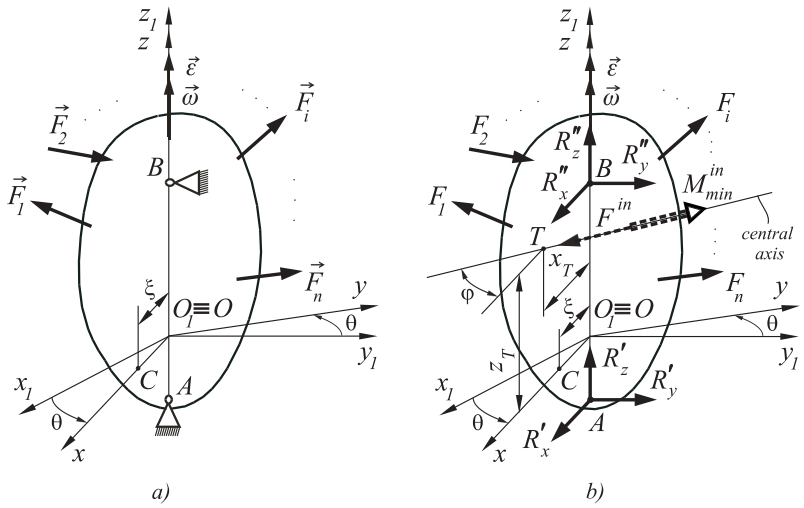


Figure 1: Central axis of the d'Alembert's fictitious forces system.

The moments with respect to the point O of the reaction $\vec{R}'$ and $\vec{R}''$ are respectively $\vec{M}_o\left(\vec{R}'\right)=-h_1R'_y\,\vec{i}+h_1R'_x\,\vec{j}$ and $\vec{M}_o\left(\vec{R}''\right)=-h_2R''_y\,\vec{i}+h_2R''_x\,\vec{j}$.

The resultant vector and the resultant moment, with respect to the point O, of the d'Alembert's fictitious forces are [1]:

$$\begin{cases} \vec{F}^{in}=-m\vec{a}_c \\ \vec{M}_o^{in}=-\frac{d\vec{K}_o}{dt} \end{cases} \tag{1}$$

where:

m is the mass of body;

$\vec{a}_c$ is the acceleration of the center of mass;
$\vec{K}_o$ is the angular momentum.

Keeping account of the position vector of the center of mass, $\vec{r}_c = \xi\vec{i}$, the angular velocity $\vec{\omega} = \omega\vec{k}$, the angular acceleration $\vec{\varepsilon} = \varepsilon\vec{k} = \dot{\omega}\vec{k}$ and the expression of the angular momentum with respect to the point O, $\vec{K}_o = -J_{xz}\omega\vec{i} - J_{yz}\omega\vec{j} + J_z\omega\vec{k}$, where J_{xz}, J_{yz} are the centrifugal moments and J_z is the moment of inertia with respect to the rotation axis, the resultant vector and the resultant moment of the d'Alembert's fictitious forces become:

$$\vec{F}^{in} = m\xi\omega^2\vec{i} - m\xi\varepsilon\vec{j} \tag{2}$$

$$\vec{M}_o^{in} = \left(J_{xz}\varepsilon - J_{yz}\omega^2\right)\vec{i} + \left(J_{yz}\varepsilon + J_{xz}\omega^2\right)\vec{j} - J_z\varepsilon\vec{k}. \tag{3}$$

We apply the d'Alembert's principle and we obtain the equations system:

$$\begin{aligned}
0 &= R_x + R'_x + R''_x + m\omega^2\xi \\
0 &= R_y + R'_y + R''_y - m\varepsilon\xi \\
0 &= R_z + R'_z + R''_z \\
0 &= M_x - h_1R'_y - h_2R''_y + J_{xz}\varepsilon - J_{yz}\omega^2 \\
0 &= M_y + h_1R'_x + h_2R''_x + J_{yz}\varepsilon + J_{xz}\omega^2 \\
0 &= M_z - J_z\varepsilon
\end{aligned} \tag{4}$$

Because $\vec{F}^{in}\cdot\vec{M}_o^{in} = -mJ_{yz}\xi\left(\varepsilon^2+\omega^4\right) \neq 0$, the d'Alembert's fictitious forces system is equivalent with the resultant force $\vec{F}^{in}$ acting along the central axis of the system and a couple acting in a plane perpendicular to the central axis, whose moment has the value

$$M_{\min}^{in} = \frac{\vec{F}^{in}\cdot\vec{M}_o^{in}}{\left|\vec{F}^{in}\right|} = -J_{yz}\sqrt{\varepsilon^2+\omega^4} \tag{5}$$

or

$$\vec{M}_{\min}^{in} = \frac{\vec{F}^{in}\cdot\vec{M}_o^{in}}{\left|\vec{F}^{in}\right|}\cdot\frac{\vec{F}^{in}}{\left|\vec{F}^{in}\right|} = -J_{yz}\omega^2\vec{i} + J_{yz}\varepsilon\vec{j}. \tag{6}$$

The equations of the central axis of the d'Alembert's fictitious forces

$$\frac{M_{Ox}^{in} - yF_z^{in} + zF_y^{in}}{F_x^{in}} = \frac{M_{Oy}^{in} - zF_x^{in} + xF_z^{in}}{F_y^{in}} = \frac{M_{Oz}^{in} - xF_y^{in} + yF_x^{in}}{F_z^{in}} \tag{7}$$

become

$$\frac{J_{xz}\varepsilon - J_{yz}\omega^2 - zm\xi\varepsilon}{m\xi\omega^2} = \frac{J_{yz}\varepsilon + J_{xz}\omega^2 - zm\xi\omega^2}{-m\xi\varepsilon} = \frac{-J_z\varepsilon + xm\xi\varepsilon + ym\xi\omega^2}{0} \tag{8}$$

It results for the central axis the equations:

$$\begin{cases} \varepsilon x + \omega^2 y = \frac{J_z \varepsilon}{m\xi} \\ z = \frac{J_{yz}}{m\xi} \end{cases} . \tag{9}$$

It is easily to see that the central axis of the d'Alembert's fictitious forces system pricks the plane Oxz (defined by the center of gravity and the rotation axis) in a point, denoted T, with the coordinates $x_T = \frac{J_z}{m\xi}$, $y_T = 0$, $z_T = \frac{J_{xz}}{m\xi}$.

This point is known in Mechanics as center of percussions. The center of percussions is a point, in the plane defined by the center of gravity and the rotation axis, where if we apply a percussion, the percussions in joints are equal with zero.

We can analogously defined the point T to be the point where if we apply a force $\vec{F} = m\vec{a}_c$ and a couple whose moment is $\vec{M}_T = J_{yz}\omega^2\vec{i} - J_{yz}\varepsilon\vec{j}$, the supplementary reactions in joints are equal with zero (the dynamic reactions are equal with the static reactions).

The demonstration is simple. If we denote the position vector of the point T $\vec{r}_T = \frac{J_z}{m\xi}\vec{i} + \frac{J_{xz}}{m\xi}\vec{j}$, it follows for the moment of the force $\vec{F}$ with respect to the point O: $\vec{M}_o\left(\vec{F}\right) = \vec{r}_T \times \vec{F} = -J_{xz}\varepsilon\vec{i} - J_{xz}\omega^2\vec{j} + J_z\varepsilon\vec{k}$. We replace $\vec{F}$, $\vec{M}_o\left(\vec{F}\right)$, $\vec{M}_T$ in the equations system (4) and we obtain:

$$\begin{aligned} 0 &= R_x + R'_x + R''_x \\ 0 &= R_y + R'_y + R''_y \\ 0 &= R_z + R'_z + R''_z \\ 0 &= M_x - h_1 R'_y - h_2 R''_y \\ 0 &= M_y + h_1 R'_x + h_2 R''_x \\ 0 &= M_z \end{aligned} \tag{10}$$

Because the properties of the point T in the impulsive motions (collisions) are similar to the properties in the dynamics of the rigid body, we suggest to change the name of the point T from center of percussions in, for example, neutral point.

3 Centrifugal moments for a plane plate having rotation motion

Let us now to consider a homogeneous plane plate (figure 2a). This plate rotates around of a vertical axis which is identical with a straight leg. The axis is contained in the plane of the plate. We know also $B\,(0, 0, h_2)$ and $A\,(0, 0, h_1)$.

We report the plate to a Cartesian reference system (figure 2a) so that the center of mass C to be situated on the Ox axis and the Oz axis to be rotation axis.

Let us apply now the d'Alembert's principle. For a some point P, with mass m_P, the d'Alembert fictitious force is (figure 2b):

$$\vec{F}_P = -m_P\vec{a}_P = m_P\omega^2 x_P\vec{i} - m_P\varepsilon x_P\vec{j} = F^{in\nu}\vec{i} + F^{in\tau}\vec{j}. \tag{11}$$

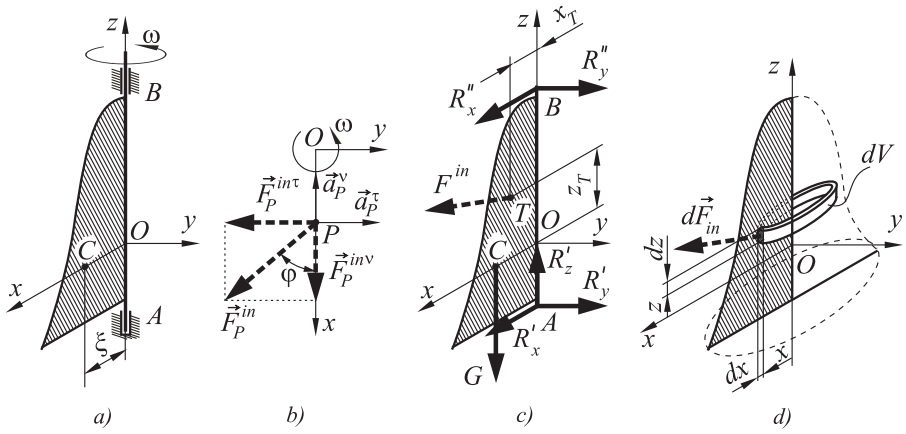


Figure 2: Plane plate having rotation motion.

The angle between the d'Alembert's fictitious force and a line parallel to the Ox axis is given by the relation:

$$tg\varphi = \frac{\left|\vec{F}^{in\tau}\right|}{\left|\vec{F}^{in\nu}\right|} = \frac{\varepsilon}{\omega^2}. \tag{12}$$

This value not depends of the position of the point P. It follows that the system of d'Alembert's fictitious forces is a parallel forces system and can be replaced by the resultant vector (figure 2c) applied in the point T (because only this point of the plate is situated on the central axis). So, the point T will be the center of parallel forces and its position can be calculated with the relations [2]:

$$x_T = \frac{\int_{(D)} x dF^{in}}{\int_{(D)} dF^{in}}; \qquad z_T = \frac{\int_{(D)} z dF^{in}}{\int_{(D)} dF^{in}}. \tag{13}$$

For determination of the coordinate z_T we isolate an element of little infinite area $dA = dxdz$ (figure 2d). The elementary d'Alembert's fictitious force $d\vec{F}^{in}$ is written:

$$d\vec{F}^{in} = -\vec{a}\,dm = \omega^2 x dm\vec{i} - \varepsilon x dm\vec{j} = \omega^2 x\rho dxdz\vec{i} - \varepsilon x\rho dxdz\vec{j} \tag{14}$$

where ρ represents the surface density.

The absolute value of the elementary d'Alembert's fictitious force is:

$$\left|d\vec{F}^{in}\right| = \rho\,|x|\,dxdz\sqrt{\varepsilon^2 + \omega^4}. \tag{15}$$

Because the plate is all situated on the same side of the rotation axis, all the d'Alembert's fictitious forces have the same sign. It follows for z_T:

$$z_T = \frac{\int_{(D)} z dF^{in}}{\int_{(D)} dF^{in}} = \frac{\iint_{(D)} \rho\sqrt{\varepsilon^2+\omega^4} z x dx dz}{\iint_{(D)} \rho\sqrt{\varepsilon^2+\omega^4} x dx dz} = \frac{\iint_{(D)} z x dx dz}{\iint_{(D)} x dx dz} = \\ = \frac{\iint_{(D)} z\, 2\pi\, x dx dz}{\iint_{(D)} 2\pi\, x dx dz} = \frac{\iint_{(D)} z\, dV}{\iint_{(D)} dV} = z_{rc} \quad (16)$$

where z_{rc} is the center of mass of the rotation body generated by plate by rotation around the Oz axis (figure 2d).

Comparing the relations obtained for z_T ($z_T = \frac{J_{xz}}{m\xi}$ and $z_T = z_{rc}$) the followed formula results for the centrifugal moment:

$$J_{xz} = m\xi\, z_{rc}, \quad (17)$$

or

$$J_{xz} = \rho A \xi\, z_{rc}, \quad (18)$$

where A is the aria of the plate.

In conclusion, the centrifugal moment is equal with the product between the mass of the plate, the coordinate of center of mass of the rotation body, generated by plate, and the coordinate of center of mass of the plate on the axis perpendicular on the rotation axis.

Keeping account of the second Guldin's law, $V_{Oz} = 2\pi\xi A$ (where V_{Oz} is the volume of the rotation body generated by plate by rotation around the Oz axis), the relation (18) becomes:

$$J_{xz} = \frac{\rho z_{rc} V_{Oz}}{2\pi}. \quad (19)$$

The geometric centrifugal moment will be:

$$I_{xz} = A\xi\, z_{rc}, \quad (20)$$

or

$$I_{xz} = \frac{z_{rc} V_{Oz}}{2\pi}. \quad (21)$$

Let us consider a plate as in figure 3a. With the axes as in figure, when we rotate the plate about the Ozaxis we obtain $J_{xz} = \rho A \xi\, z_{rc}$ and when we rotate the plate about the Ox axis we obtain $J_{xz} = \rho A \zeta\, x_{rc}$. It follows:

$$\xi\, z_{rc} = \zeta\, x_{rc}, \quad (22)$$

or

$$\frac{\xi}{\zeta} = \frac{x_{rc}}{z_{rc}}. \quad (23)$$

From the relation $\frac{\xi}{\zeta} = \frac{x_{rc}}{z_{rc}}$ and the figure 3a it results that the triangles OAC and $C_{rx}OC_{rz}$ are similar.

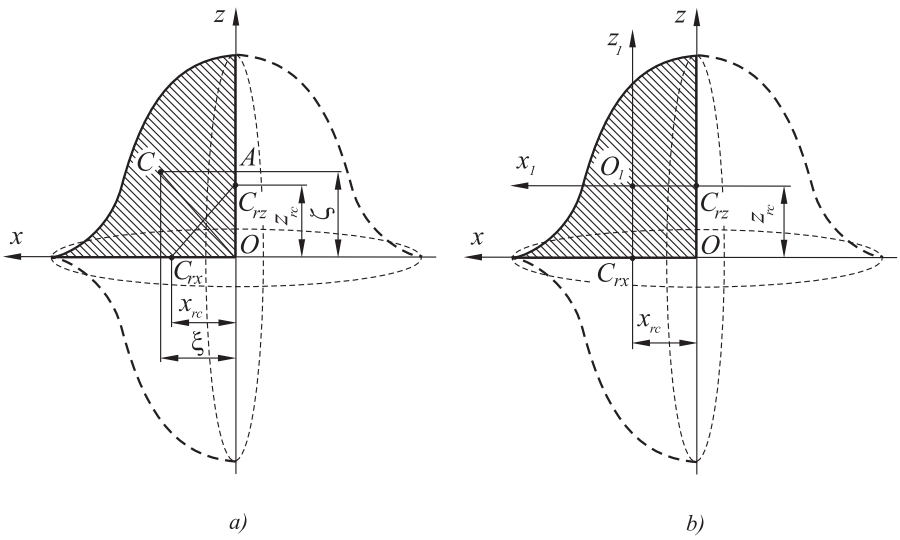


Figure 3: Plane plate rotated around two orthogonal axes.

From the relation $J_{xz} = m\xi\, z_{rc}$ it results that, if $z_{rc} = 0$, then $J_{xz} = 0$. It follows that the axis $C_{rz}x_1$ which passes by the center of mass of the rotation body, generated by plate, is a principal axis of inertia for plate (figure 3b). Also, from the relation $J_{xz} = \rho A\zeta\, x_{rc}$ it results that the axis $C_{rx}z_1$ is a principal axis of inertia for plate.

In the situations $\xi = 0$ or $\zeta = 0$, the formulas $J_{xz} = m\xi\, z_{rc}$ respectively $J_{xz} = m\zeta\, x_{rc}$ don't lead to $J_{xz} = 0$, because in these cases $\vec{F}^{in} = 0$ and the central axis does not exist.

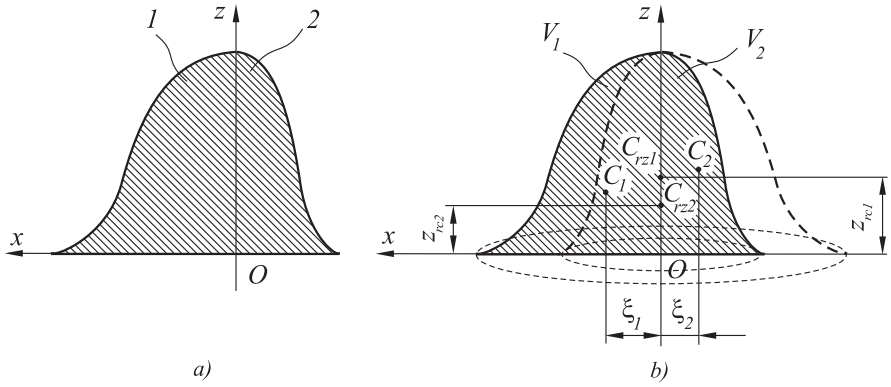


Figure 4: Plane plate cut by an rotation axis.

The relations (17), (18), (19), (20), (21) are available if the plate is situated completely on the same part of the rotation axis. If the rotation axis cuts the

plate, we split the plate in two parts with areas A_1, A_2 and coordinates of centers of mass ξ_1 , respectively ξ_2 (figure 4). These two parts generate by rotation two rotation bodies with the volumes V_{1Oz} and respectively V_{2Oz}. In this case the centrifugal moment is:

$$J_{xz} = J_{1xz} + J_{2xz} = \rho A_1 \xi_1 z_{rc1} + \rho A_2 \xi_2 z_{rc2} = \frac{\rho \left(2\pi A_1 \xi_1 z_{rc1} + 2\pi A_2 \xi_2 z_{rc2}\right)}{2\pi} =$$

$$= \frac{\rho}{2\pi} \left[2\pi A_1 \xi_1 z_{rc1} - 2\pi A_2 \left(-\xi_2\right) z_{rc2}\right] = \frac{\rho \left(z_{rc1} V_{1Oz} - z_{rc2} V_{2Oz}\right)}{2\pi}. \quad (24)$$

The geometric centrifugal moment will be:

$$I_{xz} = A_1 \xi_1 z_{rc1} + A_2 \xi_2 z_{rc2} = \frac{\left(z_{rc1} V_{1Oz} - z_{rc2} V_{2Oz}\right)}{2\pi}. \quad (25)$$

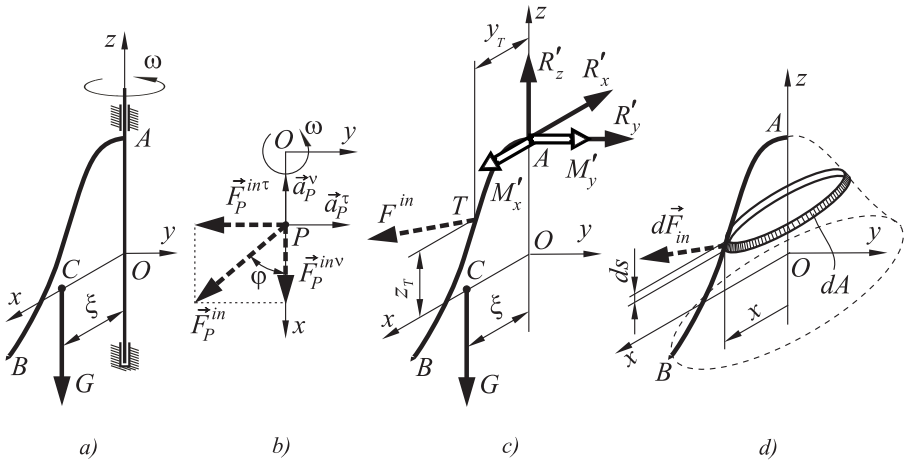


Figure 5: Plane bar having rotation motion.

For a plane bar, with mass m, having rotation motion around an axis from its plane, we can write for the centrifugal moment a relation similar to the relation written for plates (figure 5):

$$J_{xz} = m\xi z_{rc}, \quad (26)$$

or

$$J_{xz} = \rho l \xi z_{rc}, \quad (27)$$

where:

l is the length of the bar;
ρ is the linear density;
ξ is the coordinate of the center of mass of the bar on the perpendicular axis on the rotation axis;
z_{rc} is the coordinate of the center of mass of the rotation

surface generated by bar by rotation.

Keeping account of the first Guldin's law, $A_{Oz} = 2\pi\xi l$ (where $A_{Oz}$is the area of the rotation surface generated by bar by rotation around the Oz axis), the relation (27) becomes:

$$J_{xz} = \frac{\rho z_{rc} A_{Oz}}{2\pi}. \tag{28}$$

4 Plane bars and plates having uniform rotation motion

In case of plane bars and plates having uniform rotation motion ($\varepsilon = 0$) the support of the resultant vector of d'Alembert's fictitious forces system (central axis) crosses the rotation axis. Therefore, when we apply the d'Alembert's principle, it is necessary to know only the coordinate z_T.

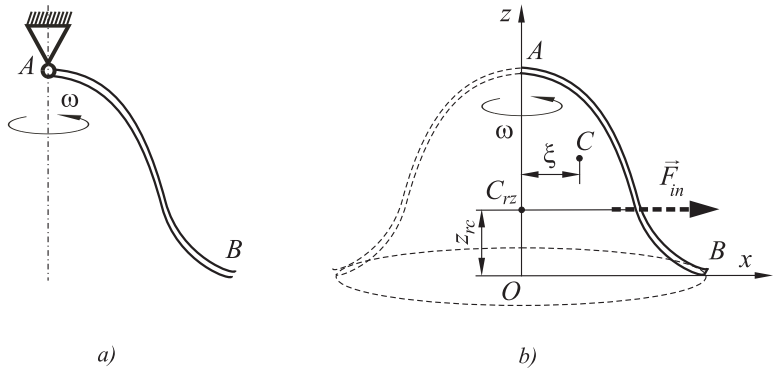


Figure 6: Plane bar having uniform rotation motion.

Let us first time to consider a homogeneous plane curve bar, articulated at a head A. This bar rotates with a constant angular velocity ω, around of a vertical axis which passes through the head A and is contained in the plane of the curve (figure 6).

The system of d'Alembert's fictitious forces (parallel forces) reduces at a resultant on the central axis. In accordance with the relation (16) the support of the resultant vector crosses the rotation axis in the center of mass of the rotation surface generated by bar by rotation.

We consider now a homogeneous plane plate (figure 7). This plate rotates with a constant angular velocity ω, around of a vertical axis which is identical with a straight leg. The axis is contained in the plane of the plate.

The system of d'Alembert's fictitious forces (parallel forces) reduces at a resultant on the central axis. In accordance with the relation (16) the support of the resultant vector crosses the rotation axis in the center of mass of the rotation body generated by plate by rotation.

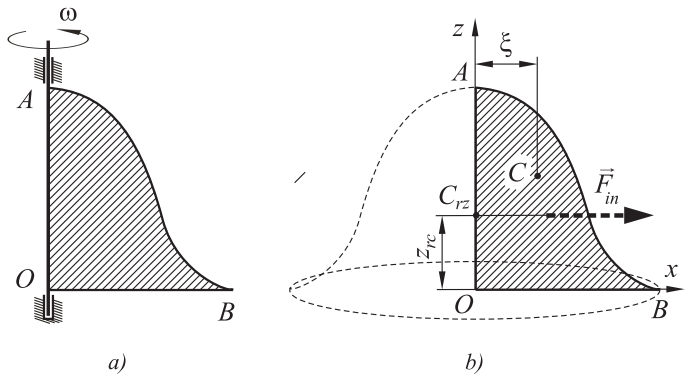


Figure 7: Plane plate having uniform rotation motion.

5 Conclusion

In this paper the authors proposed alternative methods for calculus of position of the central axis and the centrifugal moments, based on the coordinates of centers of mass. Keeping account that in literature is easily to find the positions of the centre of mass for more bodies, the method proposed here is accessible because it replaces the calculus with integrals which can be some time laborious.

References

[1] R. Voinea, D. Voiculescu, V. Ceausu, *Mecanica*, Editura Didactica si Pedagogica, Bucuresti, 1975.

[2] M. Radoi, E. Deciu, *Mecanica*, Editura Didactica si Pedagogica, Bucuresti, 1993.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Homogenization of human cortical bone. Numerical approach. Homogenization and mechanical behavior of human cortical bone. A numerical approach

M. Racila
Department of Applied Mathematics, University of Craiova, Romania
E-mail: mracila@yahoo.com

J.M. Crolet
ISIFC, France, E-mail: jmcrolet@univ-fcomte.fr

Abstract

The purpose of this study is the elaboration of a mathematical model of human cortical bone which allows studying the mechanical behaviour of this heterogeneous and complex structure knowing the properties of its basics components (collagen, hydroxyapatite (Hap) and bony fluid) and its architectural configuration. At long term it could be helpful to understand the bone remodelling which obviously needs the knowledge of the mechanical information transfer (what information receives a cortical bony cell when the bone is solicited?). The model that we propose could answer at some questions concerning the bony remodelling.

From a mathematically point of view, asymptotic homogenization method [1] in a piezoelectric framework is used and the finite element method is developed for solving the cells problems.

The developed computational methods have been packed into a software (called SiNuPrOs) made in Matlab; it is in free access on the website http://isifc.univ-fcomte.fr/SINUPROS/accueil.htm.

Keywords: human cortical bone, homogenization, piezoelectricity, multi-scale method, multi-physic model

1 Introduction

We present here a mathematical model of the human cortical bone which allows studying the mechanical behavior of this heterogeneous and complex structure knowing the properties of its basics components (collagen, hydroxyapatite (Hap) and bony fluid) and its architectural configuration. At long term it could be

ISSN-1584-5990

helpful to understand the bone remodeling which obviously needs the knowledge of the mechanical information transfer.

Many models that have been developed for cortical bone oversimplify much of the architectural and physical complexity. Our model proposes a more complete approach: it is multi scale because it contains five structural levels and multi physic because it takes into account simultaneously structure (with various properties: elasticity, piezoelectricity, porous medium), fluid and mineralization process modelization.

The main idea of the model consists in regarding the architecture of the cortical bone as a multi levels structure. One uses the asymptotic homogenization method (AHM) [1] in a piezoelectric framework (since collagen is piezoelectric) and a finite element method is developed for solving the local problems obtained by homogenization. This method of homogenization which one proposes for our model is a method on five hierarchical levels, by using asymptotic developments. This one was employed successfully for the study of several problems concerning the multiple scales. The method of the asymptotic developments initially makes it possible to compute the homogenized properties by knowing the properties of the basic components of the composite material considered with its microscopic geometry, and in the second time it gives access to local information, at the microscopic level. This second utility could be very important in the remodeling process since it is already known that the bone remodeling is first started by some cells (called osteocites) which "respond" to different micro stresses existing at inferior levels of the cortical architecture.

One solves the so-called local problems in the particular case of heterogeneous piezoelectric structure with monoclinic components, thus we will have to solve local microscopic problems of elastic or piezoelectric type, function of the elastic and piezoelectric tensor's forms. Solving these microscopic boundary value problems defined on the reference microscopic cell, one obtains the periodic corrector basis functions which are involved in the formulae for the effective tensors of elasticity, dielectricity and piezoelectricity coupling.

The above method is applied for each level of the cortical architecture in order to obtain the physical properties at lamellar, osteonal and cortical level.

The modelization of collagen as a piezoelectric medium has needed the development of a new behavior law allowing a better simulation of the effect of a medium considered as evolving during the mineralization process

In a first step, one describes the architectural model which one uses, then one develops the mathematical tools necessary to describe how this structure, at a given level, will behave, provided that one knows the physical properties of the basic components. Finally, some numerical simulations are presented.

The main interest of SiNuPrOs deals with the possibility to study, at each level of the cortical architecture, either the elastic properties or the piezoelectric effects or both of them.

The computational methods have been packed into software made on a Matlab platform, allowing a large number of predictive simulations [3] corresponding to various different configurations.

2 SiNuPrOs: a model of cortical bone

We shall present the SiNuPrOs model from two points of view: firstly, one presents the architecture of cortical bone that we considered in our modelization, then, one presents the mathematical tools which are used to calculate and simulate the mechanical properties respectively the mechanical behavior of human cortical bone.Thus, the description of SiNuPrOs contents:

a) the definition of the five encased structural scales

b) the data, for each one of these scales, of what constitutes the structure, the spaces occupied by the fluid and the behavior laws

c) the passage from a behavior law existing at a given scale to the behavior law existing at the scale which is immediately above him

The mathematical formalism is based on the mathematical theory of homogenization and the finite element method.

2.1 The bony architecture

Since many years cortical bone has been considered as a hierarchical structural composite where several organizational levels can be identified. We present here succinctly these embedded structures used in our model and the associated parameters: the Haversian system, the osteon, the lamella, the collagen fiber and the mineral structure.

The Haversian system is made of osteons and interstitial system (IS) which is also made of old surmineralized osteons. The interface between the osteons and the interstitial system is known as the cement line. The porous portion of cortical bone is made of a system of channels: the Haversian channels which run longitudinally through the center of the osteons and the Volkmann's channels, randomly orientated in a plane perpendicular to the long axis of the Haversian channels. At this level of the architecture we considered the following parameters: the osteonal diameter, the distance between two osteons, the haversian porosity, the Volkmann's porosity, the distance between two Volkmann's channels and the thickness of the cement line.

The osteon considered as a tube is not homogeneous: it is constituted by concentric lamellae encased one in the others; the characteristics are the lamellar thickness and the distance between two lamellae. These lamellae are crossed by small channels (canaliculae) that leave from the Haversian channel; one takes into account this volume of the canaliculae both in the current osteon and in the interstitial system.

The lamella is not homogeneous and it has to be considered as a composite material too. The fibers are made of collagen that one assimilates to cylinders. In each lamella, they are parallel between them and they have a specific orientation in respect with the longitudinal direction: this angle is one of the various parameters of the model SiNuPrOs. The matrix is a relatively complex medium formed of crystals and fluid being able to circulate between them, which allows us to consider cortical bone as a multiscale porous medium. Thus the parameters that it is necessary to introduce at this level are the diameter of collagen

fiber and the distance between two such fibers.

The collagen fiber is not a homogeneous medium and its architecture is also complex. It is made of microfibrils and a mineralization process is possible at the microfibrillar level.

Concerning *the mineral structure*, nowadays there is no modelization of the Hap crystal elaboration because it is generated through a very complex process. Since it is already known that the spatial disposition of Hap crystals isn't regular and that they are not geometrically isolated but rather joined one to another and arranged by blocks, we have suggested in a first approach [2] the use of an entity called *EVMC* (*Elementary Volume of Mineral Content*). Such an EVMC is made of several Hap crystals and it contains only Hap crystals and fluid. A Hap crystal is made of a solid part and of a "fluid" part called "linked water" which can be rather considered as a gel. No fluid motion is possible in this linked water. In [4] a model of these EVMC is developed with a regular alternation and explicit relationships are obtained for the Young's moduli in the three directions. The four parameters associated to these EVMC are the degrees of mineralization and the percentage of linked water, these parameters being respectively defined for a current osteon and for the pieces of IS's osteons and the coefficient of nanoscopic anisotropy.

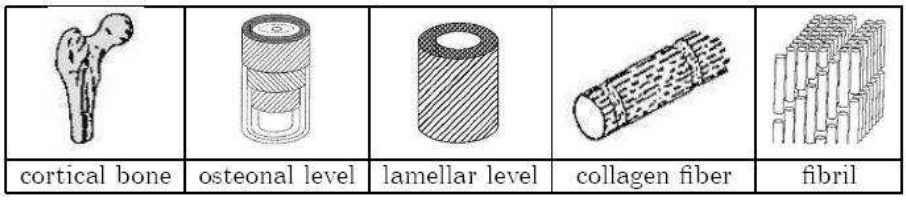


Figure 1: Cortical architecture

Following this architectural organization (Figure 1), the multi scale aspect of the human cortical bone is modeled in our model by using 18 structural parameters in a specific application of the mathematical theory of homogenization and 10 other physical parameters are necessary for the multi physic aspect (the Hap and collagen densities, the Young's moduli of Hap and collagen, the Poisson's ratios of Hap and collagen, the piezoelectric and dielectric tensors of collagen and the dielectric tensors for EVMC and fluid). Globally there are 28 parameters in the SiNuPrOs model, making it one of the most complex models of cortical bone. Indeed, this model takes into account almost all the entities occurring in the description of the architecture of human cortical bone.

2.2 Homogenization and mechanical behavior

The method which one proposes for our model is a method of homogenization on five hierarchical levels, by using asymptotic developments. This one was employed successfully for the study of several problems concerning the multiple scales. The method of the asymptotic developments initially makes it possible

to compute the homogenized properties knowing the properties of the basic components of the composite material considered with its microscopic geometry, and in the second time it gives access to local information, at the microscopic level, which is very important when simulate the bony remodeling.

Two points need to be investigated. Firstly, collagen being a piezoelectric medium, it is necessary to use homogenization method written in an adapted framework. Secondly, it is necessary to introduce an evolving behavior law to take into account the fact that the mineralization process changes locally the physical properties. Since collagen is taken into consideration at a very fine scale, it is natural to introduce its piezoelectric properties.So, let us consider the general case of a composite periodic structure Ω where the two components have piezoelectric and/or elastic properties; these two media verify simultaneously the two following equations:

$$\begin{cases} -\dfrac{\partial}{\partial x_j}[C^{\varepsilon}_{ijkl}\ e_{kl}(\mathbf{u}^{\varepsilon}) + \mu \cdot g^{\varepsilon}_{kij}\dfrac{\partial \varphi^{\varepsilon}}{\partial x_k}] = b_i \\ \dfrac{\partial}{\partial x_j}[\mu \cdot g^{\varepsilon}_{jkl}\ e_{kl}(\mathbf{u}^{\varepsilon}) - \mu \cdot \epsilon^{\varepsilon}_{jk}\dfrac{\partial \varphi^{\varepsilon}}{\partial x_k}] = 0 \end{cases}, \quad 1 \leq i,j,k,l \leq 3$$

with the Dirichlet conditions:

$$\begin{cases} \mathbf{u}^{\varepsilon} = \mathbf{0} & \text{on} \quad \partial\Omega \\ \varphi^{\varepsilon} = 0 & \text{on} \quad \partial\Omega \end{cases}$$

where:

- ε is the size of each microstructure (the scale parameter)
- $e_{kh}(\mathbf{u}^{\varepsilon}) = \dfrac{1}{2}(\dfrac{\partial u^{\varepsilon}_k}{\partial x_h} + \dfrac{\partial u^{\varepsilon}_h}{\partial x_k})$ is the strain tensor associated to the displacement field u^{ε}
- b is the density of forces in Ω
- φ^{ε} is the electrical potential
- C_{ijkl}, g_{ijk} and ϵ_{ij} are respectively the elastic, piezoelectric and dielectric tensors
- $\mu = \frac{(md_{th}-md)_+}{md_{th}}$ where md is the mineralization degree and md_{th} is the threshold value of the mineralization (which is the value of the mineralization from which the piezoelectric effect vanishes)

Recalling the heterogeneous character of the material, all the material parameters as well as the mechanical and electric fields depend on the scale parameter ε. Above and throughout the paper we use the Einstein summation convention.

The above equations can be homogenized using the well-known approach based on the asymptotic expansions [1]:

$$\begin{aligned} u^{\varepsilon}(x,y) &= u^0(x,y)+\varepsilon u^1(x,y)+\varepsilon^2 u^2(x,y)+..... \\ \varphi^{\varepsilon}(x,y) &= \varphi^0(x,y)+\varepsilon\varphi^1(x,y)+\varepsilon^2\varphi^2(x,y)+.... \end{aligned}$$

(where $x$and y are the "slow" and the "fast" variables, respectively; $y=\frac{x}{\varepsilon}$) and one obtains for the unit cells the following "strong" problems:

$$\begin{cases} -\frac{\partial}{\partial y_j}\left[C^{\varepsilon}_{ijkl}\cdot\frac{\partial\chi_l^{mn}}{\partial y_k}+\mu\cdot g^{\varepsilon}_{kij}\cdot\frac{\partial\Psi^{mn}}{\partial y_k}\right]=\frac{\partial}{\partial y_j}C^{\varepsilon}_{ijmn} \\ -\frac{\partial}{\partial y_j}\left[\mu\cdot g^{\varepsilon}_{jkl}\cdot\frac{\partial\chi_l^{mn}}{\partial y_k}-\mu\cdot\epsilon^{\varepsilon}_{jk}\cdot\frac{\partial\Psi^{mn}}{\partial y_k}\right]=\frac{\partial}{\partial y_j}g^{\varepsilon}_{jmn} \end{cases} \tag{2.1}$$

and

$$\begin{cases} -\frac{\partial}{\partial y_j}\left[\mu\cdot g^{\varepsilon}_{kij}\cdot\frac{\partial R^m}{\partial y_k}+C^{\varepsilon}_{ijkl}\cdot\frac{\partial\Phi_l^m}{\partial y_k}\right]=\frac{\partial}{\partial y_j}g^{\varepsilon}_{mij} \\ -\frac{\partial}{\partial y_j}\left[\mu\cdot\epsilon^{\varepsilon}_{jk}\cdot\frac{\partial R^m}{\partial y_k}-\mu\cdot g^{\varepsilon}_{jkl}\cdot\frac{\partial\Phi_l^m}{\partial y_k}\right]=\frac{\partial}{\partial y_j}\epsilon^{\varepsilon}_{jm} \end{cases} \tag{2.2}$$

for $1\le j,k\le 2$ and $1\le m,n,l\le 3$, which will be "weakened" and then solved by FEM method in order to compute the corrector basis functions $\chi_l^{mn},\Psi^{mn},R^m$ and Φ_l^m, for $1\le m,n,l\le 3$.The reader interested in details on the formal approach of the asymptotic method is referred to [5].

Using these so-called local problems in the particular case of heterogeneous piezoelectric structure with monoclinic components, we will have to solve local microscopic problems of elastic or piezoelectric type, function of the elastic and piezoelectric tensor's forms. Thus, the next cases will appear:

- If $(m,n)=(1,1);(2,2);(3,3);(1,2)$ the unit cell problems (2.1) becomes elastical problems:

$$\begin{cases} -\dfrac{\partial}{\partial y_j}[C^{\varepsilon}_{ijkl}\dfrac{\partial\chi_l^{mn}}{\partial y_k}]=\dfrac{\partial}{\partial y_j}C^{\varepsilon\varepsilon}_{ijmn} \\ \chi_l^{mn}\ \text{is a}\ U-\text{periodic function} \end{cases} \tag{2.3}$$

U being the unit cell used for the respective level.

- If $(m,n)=(1,3);(2,3)$ one must solve the piezoelectric type unit cell problems:

$$\begin{cases} -\dfrac{\partial}{\partial y_j}[C^{\varepsilon}_{3jk3}\dfrac{\partial\chi_3^{m3}}{\partial y_k}+g_{k3j}\dfrac{\partial\Psi^{m3}}{\partial y_k}]=\dfrac{\partial}{\partial y_j}C_{3jm3} \\ -\dfrac{\partial}{\partial y_j}[g_{jk3}\dfrac{\partial\chi_3^{m3}}{\partial y_k}-\epsilon_{jk}\dfrac{\partial\Psi^{m3}}{\partial y_k}]=\dfrac{\partial}{\partial y_j}g_{jm3} \\ \chi_3^{m3}\ \text{and}\ \Psi^{m3}\ \text{are}\ U\text{-periodic functions} \end{cases} \tag{2.4}$$

For the unit cell problems (2.2) we distinguish two other cases:

- For $m = 1, 2$ one has piezoelectric problems on the unit cell:

$$\begin{cases} -\dfrac{\partial}{\partial y_j}[C^{\varepsilon}_{3jk3}\dfrac{\partial \Phi^m_3}{\partial y_k} + g^{\varepsilon}_{k3j}\dfrac{\partial R^m}{\partial y_k}] = \dfrac{\partial}{\partial y_j} g^{\varepsilon}_{m3j} \\ -\dfrac{\partial}{\partial y_j}[-g^{\varepsilon}_{jk3}\dfrac{\partial \Phi^m_3}{\partial y_k} + \epsilon^{\varepsilon}_{kj}\dfrac{\partial R^m}{\partial y_k}] = \dfrac{\partial}{\partial y_j}\epsilon^{\varepsilon}_{jm} \\ R^m \text{ and } \Phi^m_3 \text{being } U\text{-periodic functions} \end{cases} \tag{2.5}$$

- whereas for $m = 3$ we have elastical problems to solve:

$$\begin{cases} -\dfrac{\partial}{\partial y_j}(C_{ijkl}\dfrac{\partial \Phi^3_l}{\partial y_k}) = \dfrac{\partial}{\partial y_j} g_{3ij} \\ \Phi^3_l \ U - \text{periodic } (1 \le i, j, l, k, \le 2) \end{cases} \tag{2.6}$$

Solving these microscopic boundary value problems defined on the reference microscopic cell U, one obtains the U-periodic corrector basis functions which are involved in the formulae for the effective tensors of elasticity, dielectricity and piezoelectricity coupling. For reasons of tensor's symmetry, we will have to determine 18 corrector functions solving the problems (2.3)-(2.6) by Finite Element Method using different periods (Figure 2) according to each architectural level:

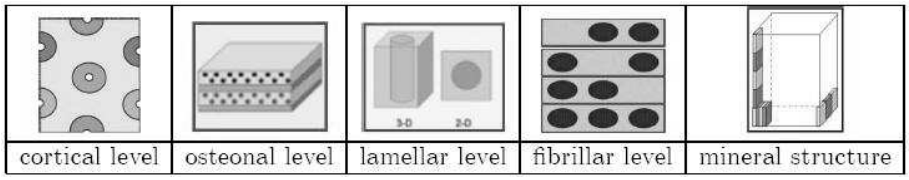


Figure 2: Various periods used

With the corrector basis functions thus obtained one can find the homogenized tensors. Finally, the homogenized problem which results is:

$$\begin{cases} C^h_{ijkl}\dfrac{\partial^2 u_k}{\partial x_j x_l} + g^h_{kij}\dfrac{\partial^2 \varphi}{\partial x_j \partial x_k} + b_i &= 0 \\ g^h_{jmn}\dfrac{\partial^2 u_m}{\partial x_j x_n} - \epsilon^h_{jm}\dfrac{\partial^2 \varphi}{\partial x_j x_m} &= 0 \end{cases}$$

and the homogenized elastic properties are calculated by the following relations:

$$\begin{aligned} C^h_{ijkl} &= < C_{ijmn}(y)e_{mny}(\chi^{kl}(y)) + g_{mij}(y)\frac{\partial \Psi^{kl}(y)}{\partial y_m} + C_{ijkl}(y) > \\ g^h_{kij} &= < g_{kij}(y) + g_{mij}(y)\frac{\partial R^k(y)}{\partial y_m} + C_{ijmn}(y)\frac{\partial \Phi^k_m(y)}{\partial y_n} > \end{aligned}$$

where$\langle\rangle$ represents the average on U of a function F which means: $< F >= \frac{1}{|U|}\int_U F dy$

The homogenized dielectric coefficients are computed similarly. Their expressions are:

$$\epsilon^h_{jm} = < \epsilon_{jm}(y) + \epsilon_{jk}(y)\frac{\partial R^m(y)}{\partial y_k} - g_{jkl}(y)e_{kly}(\Phi^m(y)) >$$

The above method is applied for each level of the cortical architecture in order to obtain the physical properties at lamellar, osteonal and cortical level. The phase of computation of homogenized coefficients is hierarchic and this allows an ascendant determination of physical properties. The homogenization of macroscopic level is made using a period containing four types of osteons, interstitial system and fluid in haversian channels. These types are defined from the investigations made by Ascenzi et al. [6] on orientations of collagen fibers in consecutive lamellae.

2.3 Microscopic fields

Once the homogenization phase is finished, one can solve the problem at the macroscopic level. Moreover, it is possible to obtain information at the microscopic level and that it is what we will present in what follows.

We consider now that the composite structure is subjected to a given loading. By the technique of homogenization, here before exposed, one can substitute an equivalent homogeneous structure to it, and at a given loading one can associate, in each point of the homogeneous structure the strain, stress and electric fields, etc..

The localization process is the step which enables us to associate at a deformations macroscopic field, the field of micro stresses reigning in the cell (the period) which is in the vicinity of this item x.

Let us return to the initial relations (the behaviors laws):

$$\left\{ \begin{array}{lcl} \sigma_{ij} & = & C_{ijkl}e_{kl} + g_{kij}\frac{\partial \varphi}{\partial x_k} \\ D_i & = & g_{ikl}e_{kl} - \epsilon_{ik}\frac{\partial \varphi}{\partial x_k} \end{array} \right.$$

One uses the asymptotic development of the solution $(\mathbf{u}^\varepsilon, \varphi^\varepsilon)$ as already presented in 2.2 and one introduces it into the precedent behavior laws.

Thus, one obtains, by taking into account the new operator of derivation,

the stress tensor:

$$\begin{array}{rc} \sigma^\varepsilon_{ij} = & C^\varepsilon_{ijkl} \cdot \left[e_{klx}(\mathbf{u}^\varepsilon) + \frac{1}{\varepsilon} \cdot e_{kly}(\mathbf{u}^\varepsilon)\right] + g^\varepsilon_{kij} \cdot \frac{\partial \varphi^\varepsilon}{\partial x_k} \\ & C^\varepsilon_{ijkl} \cdot [e_{klx}(\mathbf{u}^0(x,y) + \varepsilon \mathbf{u}^1(x,y) + \varepsilon^2 \mathbf{u}^2(x,y)) \\ & + \frac{1}{\varepsilon} \cdot e_{kly}(\mathbf{u}^0(x,y) + \varepsilon \mathbf{u}^1(x,y) + \varepsilon^2 \mathbf{u}^2(x,y))] \\ & + g^\varepsilon_{kij} \cdot \left[\begin{array}{c} \frac{\partial}{\partial x_k}(\varphi^0(x,y) + \varepsilon\varphi^1(x,y) + \varepsilon^2\varphi^2(x,y)) + \\ + \frac{1}{\varepsilon} \cdot \frac{\partial}{\partial y_k}(\varphi^0(x,y) + \varepsilon\varphi^1(x,y) + \varepsilon^2\varphi^2(x,y)) \end{array} \right] \end{array}$$

$$\begin{aligned} = \quad & \left\{ C^{\varepsilon}_{ijkl} \cdot e_{klx}(\mathbf{u}^0) + C^{\varepsilon}_{ijkl} \cdot e_{kly}(\mathbf{u}^1) + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^0}{\partial x_k} + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^1}{\partial y_k} \right\} \\ & + \varepsilon \cdot \left\{ C^{\varepsilon}_{ijkl} \cdot e_{klx}(\mathbf{u}^1) + C^{\varepsilon}_{ijkl} \cdot e_{kly}(\mathbf{u}^2) + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^1}{\partial x_k} + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^2}{\partial y_k} \right\} \\ & + \varepsilon^2 \cdot \{ \ldots\ldots\ldots\ldots\ldots\ldots \} \end{aligned}$$

Using the asymptotic development of the stress tensor $\sigma^{\varepsilon}_{ij}$, we'll have:

$$\sigma^{\varepsilon}_{ij}(u^{\varepsilon}) = \sigma^0_{ij}(u^{\varepsilon}) + \varepsilon \cdot \sigma^1_{ij}(u^{\varepsilon}) + \varepsilon^2 \cdot \sigma^2_{ij}(u^{\varepsilon}) + \ldots\ldots$$

Following the powers of ε one obtains:

$$\sigma^0_{ij}(u^{\varepsilon}) = \left\{ C^{\varepsilon}_{ijkl} \cdot e_{klx}(u^0) + C^{\varepsilon}_{ijkl} \cdot e_{kly}(u^1) + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^0}{\partial x_k} + g^{\varepsilon}_{kij} \cdot \frac{\partial \varphi^1}{\partial y_k} \right\}$$

Using the calculations made at the homogenization phase, the form of u^1 and that of φ^1, one will obtain in an indirect way the analytical relations of the micro stresses for a given macroscopic loading $\left(\mathbf{u}^0, \varphi^0\right)$:

$$\begin{aligned} \sigma^0_{ij}(x,y) = \quad & e_{mnx}(\mathbf{u}^0) \cdot \left[C^{\varepsilon}_{ijmn}(y) + C^{\varepsilon}_{ijkl}(y) \cdot e_{kly}(\chi^{mn}(y)) + g^{\varepsilon}_{kij}(y) \cdot \tfrac{\partial \Psi^{mn}(y)}{\partial y_k} \right] \\ & + \tfrac{\partial\, \varphi^0}{\partial x_m}(x) \cdot \left[C^{\varepsilon}_{ijkl}(y) \cdot e_{kly}(\Phi^m(y)) + g^{\varepsilon}_{mij}(y) + g^{\varepsilon}_{kij}(y) \cdot \tfrac{\partial R^m(y)}{\partial y_k} \right], \; i,j = 1,2,3 \end{aligned}$$

If one notes by

$$K_{ijmn} = C^{\varepsilon}_{ijmn}(y) + C^{\varepsilon}_{ijkl}(y) \cdot e_{kly}(\chi^{mn}(y)) + g^{\varepsilon}_{kij}(y) \cdot \frac{\partial \Psi^{mn}(y)}{\partial y_k}$$

and

$$G_{mij} = g^{\varepsilon}_{mij}(y) + C^{\varepsilon}_{ijkl}(y) \cdot e_{kly}(\Phi^m(y)) + g^{\varepsilon}_{kij}(y) \cdot \frac{\partial R^m(y)}{\partial y_k}$$

then the preceding relations can be written as follows:

$$\sigma^0_{ij}(x,y) = K_{ijmn} \cdot e_{mnx}(\mathbf{u}^0) + G_{mij} \cdot \frac{\partial\, \varphi^0}{\partial x_m}(x), i,j = 1,2,3$$

Thus, the knowledge of the micro stresses $\sigma^0_{ij}(x,y)$ for a fixed point x of the section of the unit cell requires the evaluation of the coefficients $K_{ijmn}(y_1,y_2)$ in this point.

The $K_{ijmn}(y_1,y_2)$ are called *the elementary micro stresses*. After the resolution of the cells problems, these quantities are calculated once for all, for a given configuration of the composite. The micro stresses field corresponding to a field of macroscopic strains (or of macroscopic stresses) is obtained like linear combination of these last quantities and the elementary micro stresses. One can thus test the effect of the various macroscopic loadings.

3 Results

Many results can be obtained with such a model. In this paper, we present only the major results which are possible to get on the elastic properties at the macroscopic scale, i.e. the scale of the cortical bone for a Reference Configuration [3]. More precisely, we are interested in the variations of the components C_{11} and C_{33}.

It is of course possible to do different studies for different orientations of collagen fibers. On the following pictures, we compare the effects of the main parameters on the coefficient C_{11} (Figure 3) and on the coefficient C_{33} (Figure 4) in the framework of 3 different coupling of orientations of the collagen fibers in consecutive lamellae [6]. Three different kinds of architecture are considered:

- architectures $0/0$: all the collagen fibers are parallel and oriented according the z axis

- architectures $0/90$: there is an alternation in the orientation of fibers between two consecutive lamellae, namely all fibers are parallel and oriented according to $z-$axis in a lamella (orientation of $0°$) and they are all parallel and oriented according to the x- axis in the following lamella (orientation of $90°$)

- architectures $45/-45$: there is an alternation in the orientation of fibers between two consecutive lamellae, namely all fibers are parallel and oriented in such a way that the angle between their direction and the z-axis is $45°$ in a lamella (orientation of $45°$ and $-45°$ in the following lamella (orientation of $-45°$)

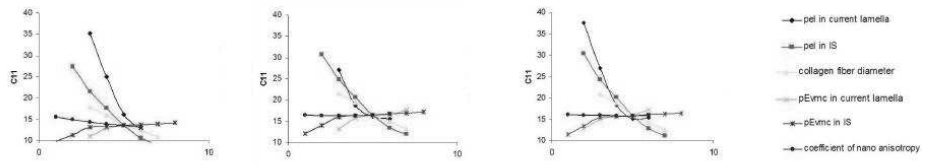


Figure 3: Effects on the coefficient C_{11} of the main parameters

For the C_{11} component we have the following variations:

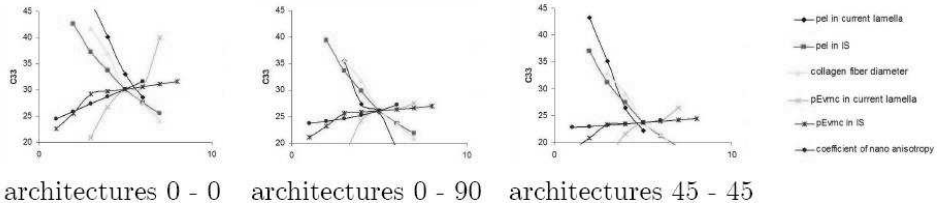
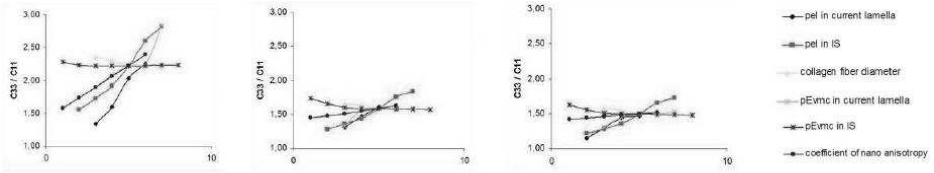


Figure 4: Effects on the coefficient C_{33} of the main parameters

One can note that, generally, the variations of the coefficients have the same effects, the representative curves having the same geometrical shapes. In the

particular case of the coefficients C_{11}, architectures (0 - 90) and (45 - -45) for which the alternation of orientation corresponds to an angle of 90 degrees give similar results. Indeed, the mean values obtained for the configuration of reference are 16.43 GPa for architectures (0 - 90) and 15.74 GPa for architectures (45 - -45). In addition, if the collagen fibers all are vertically oriented, there is a global decreasing of all the values from approximately 2 GPa, the mean value of C_{11} obtained for architectures (0 - 0) being of 14.00 GPa.

For the C_{33} coefficients, a similar analysis can be done but the variations are more important: the mean values obtained for the configuration of references being respectively of 23.50, 26,08 and 30.63 for architectures (45 - -45), (0 - 90) and (0 - 0). As previously, a comparative analysis (Figure 5) can be done between the three kinds of architectures.



architectures 0 - 0 architectures 0 - 90 architectures 45 - 45

Figure 5: Effects on the anisotropy ratio C_{33}/C_{11} of the main parameters

Figures 3, 4 and 5 present results from this study; they are in general in good agreement with the experimental literature [7]. We present here, as an example, the values that Katz [8] gives for the elastic coefficients obtained from experimental results on a human tibia. It is always possible to find architectural configurations corresponding to these results; Table 1 points out two different configurations giving similar physical properties with SiNuPrOs:

GPa	C_{11}	C_{13}	C_{33}	C_{44}	C_{66}
Katz [8]	11.60	6.10	22.50	4.91	2.41
SiNuPrOs configuration 1	11.19	5.50	21.78	4.99	3.95
variations (%)	4	10	3	2	39
SiNuPrOs configuration 2	11.66	6.02	22.47	5.54	3.93
variations (%)	0.53	1.33	0.13	11.37	38.68

Table 1. Elastic properties of cortical bone in a human tibia

4 Conclusion

The aim of this work was to describe a numerical model allowing the calculation of the mechanical properties of the multi-scale composite structure of the human cortical bone. The model which was realized is very complex and practically all the possible configurations can be tested, the model allowing a direct intervention on the various data dependent on the architecture of the bone. Many

parametric studies can be made and thus the effect of the microstructure on the mechanical behavior of cortical bone could be investigated with this model; also, the homogenization method allows the access at the microscopic fields where the cells are living and we hope that in a next future we will be able to understand and simulate the bone remodeling.

The basic components (collagen, Hap crystals and "bony fluid") are taken into account with their largest properties: the collagen is considered as a piezoelectric medium and dielectric properties can be used to model the presence of ions in the bony fluid. A pseudo periodicity in the architecture is assumed and the theory of homogenization is the mathematical tool used to link all these architectural levels in order to get the bony properties at each architectural level.

No mineralization process has been introduced because the lack of information in the literature but several important parameters are present such as the percentage of EVMC or the percentage of linked water. Relationships between all these coefficients will be introduced according to the future models of mineralization process.

The problem of validation is important. There are papers in the literature published on the mechanical properties of cortical bone but it is quite difficult to use them because they generally present mean values obtained from many experimental data without any precision on the values of the local properties or on the architectural characteristics.

The informatic program SiNuPrOs is in free access on the Internet at: http://isifc.univ-fcomte.fr/SINUPROS/accueil.htm.

References

[1] Panasenko G., *Multiscale Modelling for Structures and Composites*, Springer, Dordrecht, Netherlands, 2005

[2] Crolet J.M., Racila M., Mahraoui R., Meunier A., *New numerical concept for hydroxyapatite in human cortical bone*, Computer Methods in Biomechanics and Biomedical Engineering, Vol. 8 (2), pp. 139-143, 2005

[3] Racila M., Crolet J. M., *Human cortical bone: the SiNuPrOs model. Part I - Description and macroscopic results*, Comput. Meth. in Biomec. and Biomed. Eng., vol. 11, pp. 169 – 187, 2008

[4] Racila M., Crolet J. M., *Nano and macro structure of cortical bone: numerical investigations*, Mechanics of Advanced Materials and Structures, Vol. 14, Issue 8, pp. 655–663, 2007

[5] Racila M., Ph.D Thesis: *Mathematical modeling of multiscale transfer of mechanical signals in human cortical bone. Theoretical aspects and computational methods*, University of Franche-Comté, 2005

[6] Ascenzi A., Benvenuti A., *Orientation of collagen fibers at the boundary between two successive osteonic lamellae and its mechanical interpretation,* J. of Biomechanics, no. 19, pp. 349-361, 1986

[7] Rho J-Y et al., *Mechanical properties and the hierarchical structure of bone*, Medical Engineering & Physics, vol. 20, pp. 92-102, 1998

[8] Katz J.L., *Mechanics of hard tissue, in Biomechanics, Principle and Applications,* D.R. Peterson and J.D. Brozino (Eds), CRC Press, 2008

This study was supported by the Romanian Research Project PN-II-RU-RP-2008 no. 1 from 3/11/2008

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On the Cauchy Problem of Navier-Stokes flow

Silviu Sburlan
Mircea cel Bătrân Naval Academy, Constanţa, Romania
E-mail: ssburlan@yahoo.com

Abstract

Consider the Cauchy semilinear problem of the Navier-Stokes flow of incompressible fluids - one of the *Millenium Prize Problems* (see [1]). By standard arguments we can formulate the problem as an abstract equation and prove the existence and the uniqueness of the strong solution. The proof is constructive and it is based on the Fourier method developed in the energetical space of the Stokes operator (on the complete sequence of the eigenvectors of the duality map). Some open problems are also appended.

Keywords: Navier - Stokes equations, abstract Fourier method, solution.

1 Introduction

Consider the Navier-Stokes system for incompressible fluids filling all of $\mathbb{R}^N$, ($N = 2$ *or* 3):

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = \nu \Delta u - \nabla p + f, \tag{1}$$

$$\nabla \cdot u = 0, \ in \ \mathbb{R}^N, \ t \geq 0, \tag{2}$$

$$u(x, 0) = u^0(x), \ x \in \mathbb{R}^N, \tag{3}$$

where $\nu \geq 0$ (*dynamical viscozity*), f (*body forces*) and u^0 (*initial velocity*) are given. These equations are to be solved for an unknown *velocity vector* $u : \mathbb{R}^N \times [0, \infty) \mapsto \mathbb{R}^N$ and the *presure* $p : \mathbb{R}^N \times [0, \infty) \mapsto \mathbb{R}$. The *Euler equations* can be obtained for $\nu = 0$ in (1)-(3).

For physically reasonable solutions, we ask that u does not grow large as $|x| \to \infty$. Hence we will restrict attention to data f and u^0 that satisfy:

$$|\partial_x^\alpha \partial_t^n f(x, t)| \leq C_{\alpha n K}(1 + |x| + t)^{-K}, \ x \in \mathbb{R}^N, t \geq 0, \tag{4}$$

and

$$|\partial_x^\alpha u^0(x)| \leq C_{\alpha K}(1 + |x|)^{-K}, \tag{5}$$

ISSN-1584-5990

for any α, n and K. A solution $[u,p]$ is physically reasonable only if it is enough smooth and *it has bounded energy*, i.e.,

$$\int_{\mathbb{R}^N} |u(x,t)|^2 dx < C, \forall t \geq 0. \tag{6}$$

Let u^0 be any smooth, divergence-free vector field (i.e., $\nabla \cdot u = 0$) satifying (5). We have to prove that either there exists a smooth solution $[u,p] \in C^\infty(\mathbb{R}^N \times [0,\infty))$ that satisfy (1),(2),(3) and (6), or there exist no such solutions. Remark that these problems are not solved yet for $\nu > 0$, $f = 0$ and $N = 3$ (see [1]).

Let $\phi : \mathbb{R}^N \times [0,\infty) \mapsto \mathbb{R}^N$ be a compactly supported vector field. Then, multiplying (1) and (2) by ϕ, a formal integration by parts yields:

$$\iint_{\mathbb{R}^N \times \mathbb{R}_+} u \cdot \frac{\partial \phi}{\partial t} dxdt - b(u,u,\phi) =$$

$$= \iint_{\mathbb{R}^N \times \mathbb{R}_+} \nabla u \cdot \nabla \phi dxdt + \iint_{\mathbb{R}^N \times \mathbb{R}_+} (f - \nabla p) \cdot \phi dxdt, \tag{7}$$

where

$$b(u,u,\phi) := \sum_{ij} \iint_{\mathbb{R}^N \times \mathbb{R}_+} u_i u_j \frac{\partial \phi_i}{\partial x_j} dxdt$$

and

$$\iint_{\mathbb{R}^N \times \mathbb{R}_+} u \cdot \nabla_x \phi dxdt = 0. \tag{8}$$

A solution of (7)-(8) is called a *weak solution* of Navier-Stokes system. Observe that these equations make sense for all $u \in L^2$ and $p \in L^1$.

2 Abstract Fourier Method

Define *the space of incompressible fluids*:

$$C^\infty_{0,\sigma} := \{y \in (C^\infty_0(\mathbb{R}^N))^N;\ \nabla \cdot u = 0\},$$

and let X be its completion with respect to the norm $\|\cdot\|_2$. Then X is a Hilbert space with the scalar product:

$$(y,w) := \int_{\mathbb{R}^N} y \cdot w = \sum_{i=1}^N \int_{\mathbb{R}^N} y_i w_i dx.$$

Let E be its subspace:

$$E := \{y \in X;\ y \in (W^{1,2}(\mathbb{R}^N)^N\}.$$

Since $W^{1,2}(\mathbb{R}^N) = W_0^{1,2}(\mathbb{R}^N)$ (see [2]), E can be viewed as the completion of $C_{0,\sigma}^{\infty}$ in the norm of $W_0^{1,2}$, namely:

$$E := \overline{C_{0,\sigma}^{\infty}}^{\|\cdot\|_{1,2}} = \{y \in (W_0^{1,2}(\mathbb{R}^N))^N;\ \nabla \cdot y = 0\}.$$

Remark that the scalar product in X is in fact the duality pairing between E^* and E.

Consider *the Stokes operator* $A \in L(E, E^*)$ defined by:

$$(Ay, w) := \sum_{i=1}^{N} \int_{\mathbb{R}^N} \nabla y_i \cdot \nabla w_i dx,\ \forall y, w \in E,$$

and define the three-linear form:

$$b(y, z, w) := \sum_{i,j=1}^{N} \int_{\mathbb{R}^N} y_i D_i z_j w_j dx,$$

that determines the nonlinear operator $C : E \mapsto E^*$ by:

$$C(y, w) := b(y, y, w),\ \ \forall y, w \in E.$$

Then the equation (7) can be written as:

$$\frac{d}{dt}(u, v) = (f - Au - C(u), v),\ \forall v \in E, \tag{9}$$

and we obtain the following weak formulation of Navier-Stokes system:
Given $f \in L^2(0, T; E^*)$ *and* $u^0 \in X$ *find* $u \in L^2(0, T; E)$ *such that* $u_t \in L^2(0, T; E^*)$ *and:*

$$\begin{cases} \frac{du}{dt} + \nu Au + C(u) = f \ \ on\ (0, T), \\ u(0) = u^0. \end{cases} \tag{10}$$

Here $u_t := \frac{du}{dt}$ and $T > 0$ is any large real number. The weak solution that is enough smooth, i.e, $u \in C((0, T); E) \cap C^1((0, T); X)$, is called *the strong solution* of Navier-Stokes system.

As $A : E \mapsto E^*$ is symmetric $(Ay, w) = (y, Aw)$ and strongly monotone $(Ay, y) \geq \|y\|^2$ and E is its energetic space, there exist $[e_n, \lambda_n] \in E \times [0, \infty)$ solutions of the eigenvalue problem:

$$(Ae_n, v) = \lambda_n (e_n, v),\ \forall v \in E,$$

with $(e_i, e_j) = \delta_{ij}$ and $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \to \infty$. Then $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis in E, $(\sqrt{\lambda_n} e_n)_{n\in\mathbb{N}}$ is an orthonormal basis in X and $(\lambda_n e_n)_{n\in\mathbb{N}}$ is an orthonormal basis in E^*. Moreover, A is continuous and can be viewed as the duality mapping $J : E \mapsto E^*$ (see [4]), that is

$$\langle Ju, v\rangle = (u, v)_E := (Au, v), \quad \|u\|_E := (u, u)_E^{1/2} = \|u\|_{1,2}.$$

Hence, $u \in L^2(0,T;E) \Rightarrow Au \in L^2(0,T;E^*)$ and we have the implications (see [8] p.281):

$$C(u) \in L^1(0,T;E^*) \Rightarrow f - \nu Au - C(u) \in L^1(0,T;E^*) \Rightarrow u_t \in L^1(0,T;E^*).$$

Therefore u is $(a.e. =)$ continuous from $[0,T)$ to E^*.

We try to find the weak solution of the Cauchy problem (10) as the Fourier series in E:

$$u(x,t) := \sum_{n=0}^{\infty} b_n(t)e_n(x) = \lim_{n\to\infty} \sum_{k=0}^{n} b_k(t)e_k(x). \tag{11}$$

To do this denote by $D_i v := \frac{\partial v}{\partial x_i}$ and describe in detail the convective term:

$$(v\cdot\nabla)v := (v_1D_1 + \cdots + v_ND_N)v = (v_1D_1v_k + \cdots + v_ND_Nv_k)_{1\le k\le N}.$$

Denoting $e_n := (e_n^m)1 \le m \le N$ and $v_n(x,t) := \sum_{p=0}^{n} b_p(t)e_p(x)$ for all $n \in \mathbb{N}$, the k^{th}-component, $(1 \le k \le N)$, of the vector $(v_n\cdot\nabla)v_n$ will be:

$$v_n^1D_1v_n^k + \cdots + v_n^ND_Nv_n^k = \sum_{m=1}^{N}\left(\sum_{p=0}^{n} b_pe_p^m \cdot \sum_{p=0}^{n} b_pD_me_p^k\right).$$

Then $(u\cdot\nabla)u = \lim(v_n\cdot\nabla)v_n$ by Mertens theorem and:

$$((v_n\cdot\nabla)v_n(\cdot,t),e_j) := \sum_{k=1}^{N}\int_{\mathbb{R}^N}(v_n^1D_1v_n^k + \cdots + v_n^ND_Nv_n^k)\cdot e_j^k(x)dx =$$

$$= \sum_{p=1}^{n}\sum_{q=1}^{p} b_{p-q}(t)b_q(t)\left[\int_{\mathbb{R}^N}\sum_{k=1}^{N}\left(\sum_{m=1}^{N} e_{p-q}^m(x)D_me_q^k(x)\right)\cdot e_j^k(x)dx\right] =:$$

$$=: c_j(b_1(t),\dots b_n(t)),\ \ j \ge 1.$$

Taking (11) and $v := e_j$, $j \ge 1$, we formally obtain that $b_j(t)$ must satisfy the scalar Cauchy problem:

$$\begin{cases} b_j' + \nu\lambda_jb_j + c_j(b_1,\dots,b_n) = f_j(t), \\ b_j(0) = u_j^0,\ 1 \le j \le n, \end{cases} \tag{12}$$

where u_j^0 and $f_j(t)$ are the Fourier coefficients in E of u^0 and $f(\cdot,t)$, respectively:

$$u_j^0 := (u^0,e_j)_E, \quad f_j(t) := (f(\cdot,t),e_j)_E, \quad \forall j \in \mathbb{N}.$$

Let $n \in \mathbb{N}$ be any fixed number. The homogeneuos linear system $b_j' + \nu\lambda_jb_j = 0$, $1 \le j \le n$, has the solution $b_j(t) := k_je^{-\nu\lambda_jt}$ with $k_j \in \mathbb{R}$, $1 \le j \le n$. By variation of constants $k_j := h_j(t)$ we deduce that h_j are solutions of the following system

$$\begin{cases} h_j' = e^{\nu\lambda_jt}f_j(t) - \lambda_je^{\nu\lambda_jt}c_j(e^{-\nu\lambda_1t}h_1(t),\dots,e^{-\nu\lambda_nt}h_n(t)), \\ h_j(0) = u_j^0,\ 1 \le j \le n \end{cases},$$

with c_j some sums of quadratic terms in $h_1, \dots, h_n$. Consider the vectors $h := (h_j)_{1\le j\le n}$ and

$$F(t,h) := (e^{\nu\lambda_j t} f_j(t) - \lambda_j e^{\nu\lambda_j t} c_j(e^{-\nu\lambda_1 t}h_1(t), \dots, e^{-\nu\lambda_n t}h_n(t)))_{1\le j\le n}.$$

Then, in vectorial form, the above system can be expressed as the Cauchy problem

$$\begin{cases} h' = F(t,h), \\ h(0) = u^0, \end{cases} \tag{13}$$

with F continuous in $[t,h]$ and of class C^1 in h. Hence there exists a unique solution $h = h(t)$ and $h \in C^1([0,\tau])$, with $0 < \tau \le T$, by Picard theorem.

Consequently, the Cauchy problem (12) has a unique solution, that determine the coefficients of the Fourier series (11) and so $u(x,t)$. To prove that this is the weak solution of the Navier-Stokes system (10) we have to prove that:

(a) The series $\sum\limits_{n=1}^{\infty} b_n(t)e_n(x)$ converges uniformly to $u \in C([0,T];E)$;

(b) The series $\sum\limits_{n=1}^{\infty} b_n'(t)e_n(x)$ converges uniformly to $u_t \in C([0,T];E^*)$;

(c) This $u(x,t)$ is the unique weak solution of the Cauchy problem (10).

3 Existence, Uniqueness and Smoothness

The answers to the above three problems, (a),(b) and (c), depend on the convergence of the numerical series:

$$\sum \lambda_n^{-\frac{N+2}{2}}.$$

Suppose that this series is convergent in $\mathbb{R}$ and remark that

$$M_1 := \sup\{|h_n(t)|; n \in \mathbb{N},\ t \in [0,T]\} < +\infty,$$

$$M_2 := \sup\{|h_n'(t)|; n \in \mathbb{N},\ t \in [0,T]\} < +\infty.$$

Then

$$|b_n(t)|^2 = |e^{-\nu\lambda_n t}h_n(t)| = e^{-2\nu\lambda_n t}|h_n(t)|^2$$

and thus for any $\delta \in (0,T)$ there exists $n_\delta \in \mathbb{N}$ such that:

$$e^{-2\nu\lambda_n t} < e^{-2\nu\lambda_n \delta} < \frac{1}{\lambda_n^{\frac{N+2}{2}}},\ \forall n \ge n_\delta,\ t \in [\delta, T].$$

Hence

$$|b_n(t)|^2 \le M_1^2 \frac{1}{\lambda_n^{\frac{N+2}{2}}},$$

that is, the function series $\sum |b_n(t)|^2$ is convergent for $t \in [\delta,T]$, with $\delta > 0$ arbitrarily choosen in $(0,T)$. Thus the series $\sum b_n(t)e_n$ is uniformly convergent to $u \in C([0,T];E)$ because

$$\|\sum b_n(t)e_n\|_E^2 = \sum |b_n|^2 < +\infty.$$

Further, $b_n'(t) = -\nu\lambda_n e^{-\nu\lambda_n t}h_n(t) + e^{-\nu\lambda_n t}h_n(t)h_n'(t)$ and thus:

$|b_n'(t)|^2 \leq 2\nu^2\lambda_n^2 e^{-2\nu\lambda_n t}|h_n(t)|^2 + 2e^{-2\nu\lambda_n t}|h_n'(t)|^2 \leq 2(\nu^2\lambda_n^2 M_1^2 + M_2^2)e^{-2\nu\lambda_n t}$.

Then

$$\left|\frac{b_n'(t)}{\lambda_n}\right|^2 \leq 2\nu^2 M_1^2 e^{-2\nu\lambda_n t} + 2M_2^2\frac{e^{-2\nu\lambda_n t}}{\lambda_n^2}$$

and, similarly, we can conclude that:

$$\left|\frac{b_n'(t)}{\lambda_n}\right|^2 \leq 2\nu^2 M_1^2\frac{1}{\lambda_n^{\frac{N+2}{2}}} + 2M_2^2\frac{1}{\lambda_n^{\frac{N+2}{2}}},\ \forall n \geq n_\delta,\ t \in [\delta, T].$$

Consequently, the series $\sum b_n'(t)e_n$ converges absolutely to $\tilde{u} \in C((0,T);E^*)$ because

$$\|\sum b_n'(t)e_n\|_{E^*}^2 = \left\|\sum\frac{b_n'(t)}{\lambda_n}\lambda_n e_n\right\|_{E^*} = \sum\left|\frac{b_n'(t)}{\lambda_n}\right|^2 < +\infty.$$

We will show that $\tilde{u} = \frac{du}{dt}$ as a distribution on $(0,T)$. Indeed:

$$\int_0^1\left(\sum_{n=1}^m b_n'(t)e_n\right)\phi(t)dt = \left[\left(\sum_{n=1}^m b_n(t)e_n\right)\phi(t)\right]_0^T - \int_0^T\left(\sum_{n=1}^m b_n(t)e_n\right)\phi'(t)dt =$$

$$= -\int_0^T\left(\sum_{n=1}^m b_n(t)e_n\right)\phi'(t)dt,\ \ \forall\phi \in C_0^\infty(0,T),\ m \in \mathbb{N},$$

that is $u' = \tilde{u}$ as a distribution from $(0,T)$ into E^*.

Let us show that $u := \sum b_n(t)e_n$ is a weak solution of the Cauchy problem (10):
Indeed, for $t = 0$ we have:

$$u(x,0) = \sum b_n(0)e_n(x) = \sum u_n^0 e_n(x) = u^0(x).$$

On the other hand, remark that

$$\sum[b_j' + \nu\lambda_j b_j + \lambda_j c_j(b_1,\ldots,b_n)]e_j = \sum_{j=1}^n f_j e_j$$

and thus, multiplying in X by $e_k,\ 1 \leq k \leq n$, we deduce:

$$(\sum_{j=1}^n b_j'e_j, e_k) + (\sum_{j=1}^n \nu\lambda_j b_j e_j, e_k) + (\sum_{j=1}^n \lambda_j c_j e_j, e_k) = (\sum_{j=1}^n f_j e_j, e_k).$$

Since $Ae_j = Je_j = \lambda_j e_j$ and $\{e_k;\ k \in \mathbb{N}\}$ is an orthonormal basis in E, letting $n \to \infty$ we deduce that:

$$(\frac{du}{dt}, w) + \nu(Au, w) + ((u\cdot\nabla)u, w) = (f, w),\ \ \forall w \in E,$$

that is, u is the weak solution of the problem (10). Remark that $u \in C([0,T];E) \cap C^1((0,T);X)$, that is, u is the strong solution of the Cauchy problem.

To show the uniqueness of this strong solution, let u_1 and u_2 be two such solutions, i.e.,

$$u_1(x,t) := \sum p_n(t)e_n(x) = \lim v_{1n}(x,t),$$

$$u_2(x,t) := \sum q_n(t)e_n(x) = \lim v_{2n}(x,t),$$

where

$$v_{1n}(x,t) := \sum_{j=1}^{n} p_j(t)e_j(x), \quad v_{2n}(x,t) := \sum_{j=1}^{n} q_j(t)e_j(x).$$

Thus, for any $\varepsilon > 0$ there exist $n'_\varepsilon, n''_\varepsilon \in \mathbb{N}$ such that:

$$|(\frac{d}{dt}v_{1n} - \nu\Delta v_{1n} - (v_{1n}\cdot\nabla)v_{1n}, v) - (f,v)| < \varepsilon,\ \forall v \in E,\ n \geq n'_\varepsilon,$$

and

$$|(\frac{d}{dt}v_{2n} - \nu\Delta v_{2n} - (v_{2n}\cdot\nabla)v_{2n}, v) - (f,v)| < \varepsilon,\ \forall v \in E,\ n \geq n''_\varepsilon.$$

Denote by $n_\varepsilon := \max\{n'_\varepsilon, n''_\varepsilon\}$ and take $v := e_j,\ 1 \leq j \leq n$. Then we obtain:

$$k_j := |p'_j(t) + \nu\lambda_j p_j(t) + \lambda_j c_j(p_1,\ldots,p_n) - f_j(t)| < \varepsilon,\ \forall n \geq n_\varepsilon,\ t \in (0,T)$$

and

$$r_j := |q'_j(t) + \nu\lambda_j q_j(t) + \lambda_j c_j(q_1,\ldots,q_n) - f_j(t)| < \varepsilon,\ \forall n \geq n_\varepsilon,\ t \in (0,T).$$

Let $p := (p_1,\ldots,p_n)$ and $q := (q_1,\ldots,q_n)$ be, respectively, the solutions of the following problems:

$$\begin{cases} b' = F_1(t,b) \\ b(0) = u^0 \end{cases}, \quad \begin{cases} b' = F_2(t,b) \\ b(0) = u^0 \end{cases},$$

where $F_1(t,b) := (-\lambda_j b_j - c_j(b) + f_j(t) + k_j(t))_{1\leq j\leq n}$ and $F_2(t,b) := (-\lambda_j b_j - c_j(b) + f_j(t) + r_j(t))_{1\leq j\leq n})$. As F_1 and F_2 are continuous in t and of class C^1 in b, the solutions depend continuously on the second term, that is:

$$|F_1 - F_2| := \max_{1\leq j\leq n} |F_{1j} - F_{2j}| = \max_{1\leq j\leq n} |k_j(t) - r_j(t)| \leq$$

$$\leq \max_{1\leq j\leq n} \{|k_j| + |r_j|\} < 2\varepsilon,\ \forall t \in (0,T).$$

It follows that:

$$|p(t) - q(t)| := \max_{1\leq j\leq n} |p_j(t) - q_j(t)| < g(\varepsilon),\ \forall t \in (0,T)$$

with $\lim_{\varepsilon\to 0} g(\varepsilon) = 0$, that is $p_j(t) = q_j(t)$ and so $u_1 = u_2$.

We can conclude:

Theorem: *If the series* $\sum \lambda_n^{-\frac{N+2}{2}}$ *is convergent, then there exists an unique (weak) strong solution of the Cauchy problem of Navier-Stokes system.*

Along to same line we mention that this series is convergent in the case of a bounded domain in $\mathbb{R}^N$. The proof is based on the Weyl's law:

$$N(\lambda) = \frac{\mu(\Omega)\omega_N}{(2\pi)^N}\lambda^N + R(\lambda),$$

where $N(\lambda) := card\{j \in \mathbb{N}; \sqrt{\lambda_j} \leq \lambda\}$, ω_N is the volume of the unit ball in $\mathbb{R}^N$ and $R(\lambda) = \mathcal{O}(\lambda^{N-1})$, which depends on the measure of the domain $\mu(\Omega)$ in $\mathbb{R}^N$ (see [3] for details). In the case of the unbounded domain, particularly when the domain is all of $\mathbb{R}^N$, this convergence is still unproved.

References

[1] Fefferman L.Ch.: Existence and Smoothness of Navier-Stokes Equation, in *Millenium Prize Problems* (J.Carlson, A.Jaffe, A.Wiles eds.), pp.57-67, AMS-Providence, NY 2006.

[2] Nečas J.: *Les méthodes directes en théorie des équations elliptiques*, Academia, Prague, 1967.

[3] Sburlan C. On the Solvability of Navier-Stokes Equations, Bull.Transilvania Univ.Braşov, vol.13(48), New Ser.,2006,pp.321-329.

[4] Sburlan S.: The Fourier Method for Abstract Equations, An.Şt.Univ. Ovidius Constanţa, Ser.Mat.,3,1,1995,pp.186-193.

[5] Sburlan S.: *Topological and Functional Methods for Partial Differential Equations*, Surv.Ser.Math., Analysis 1, Ovidius.Univ.Constanţa, 1995.

[6] Sburlan S., Moroşanu G.: *Monotonicity Methods for Partial Differential Equations*, MB-11/PAMM, TU-Bp, Budapest,1999.

[7] Sohr H.: *The Navier-Stokes Equations*, Birkhäuser Verlag, Basel, 2001.

[8] Temam R.: *Navier-Stokes Equations*, North-Holland Publ.Comp., Amsterdam, 1977.

SECTION C

DYNAMICAL SYSTEMS

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Dynamical Systems with Memory Effects. Applications in Plasma Physics

D. Constantinescu
Department of Applied Mathematics,
University of Craiova, Romania, dconsta@yahoo.com

Abstract

Phenomena involving some memory effects can be easily identified in physics, chemistry, biology, ecology, social sciences. Their transcription in mathematical formalism is a real challenge. In the domain of dynamical systems there are at least three types of models with memory. For the short time memory we can speak about systems with delay and about the systems with hysteresis and the long term memory can be expressed using the fractional dynamical systems.

In this paper we present some theoretical results concerning the systems with memory effects (fractional systems and systems with hysteresis) and we focus on the use of these models for explaining some experimental observations on the fusion of plasma in tokamaks.

Two applications are proposed: the fractional diffusion equation is considered in order to fit some experimental data related to the heat transport and a dynamical system with discontinuous hysteresis of delayed relay type is used to reproduce experimental data on saw-tooth crash of the central temperature in ASDEX Upgrade tokamak

Using these mathematical models we point out that the memory effects are important parts of a realistic description of phenomena of the real world.

Key words dynamical systems, memory effects, fusion plasma physics

1 Introduction

In many applications it is assumed that the evolution of a system is governed by its evolution law (expressed by ordinary or partial differential equations that involve the state and the rate of change of this state) and by its state at the initial moment of the study (expressed by the initial conditions at a given moment t_0). This is the principle of causality which means that the future states of the system are independent of the past states and are determined only by the evolution law and the state of the system at the moment t_0. The corresponding Cauchy problem, the prototype of the autonomous systems, is

$$x'(t) = f\left(x(t)\right), \qquad x\left(t_0\right) = x_0$$

ISSN-1584-5990

Sometimes this principle is only a first approximation of the true situation and more realistic models must include some of the past states of the systems. These models involves the memory effects.

The simplest type of past dependence is through the state variable and not the derivative of the state variable, the so-called retarded differential equations or retarded difference equations of systems with delay [1]. For a given time delay τ, the systems with delay are described by

$$x'(t) = f\left(x\left(t\right), x\left(t-\tau\right)\right), \qquad x\left(t\right) = \varphi\left(t\right), \quad t \in [t_0 - \tau, t_0]$$

These systems are important in several branches of engineering, in physics [2], economics [3], optimal production decision, in the study of complex systems [4] as well as in many biological research topics [5] but they are also very interesting from the mathematical point of view. The study of the action of some time-delayed controller for the stabilization of the unstable periodic orbits embedded in chaotic attractors [6], the evaluation of the largest Liapunov exponent [7], the study of some bifurcations [8], [9] in systems with delays are only a few directions where interesting results were obtained.

An other type of past dependence is the hysteresis effect. The term "hysteresis" means "to lag behind" and originates from ancient Greeks. The hysteresis is defined as a "rate independent memory effect". A system which incorporates the hysteresis effect is described by

$$x'\left(t\right) = f\left(\left(Hx\right)\left(t\right), x\left(t\right)\right), \qquad x\left(t_0\right) = x_0.$$

The hysteresis operator H is specific for each system, but its main characteristic is that $\left(Hx\right)\left(t\right)$ is determined by the value of the input x at the moment t_0 and by some of the values of x in the time interval (t_0, t). There is no dependence on the derivative of x. In many cases the hysteresis output is determined by the extremum points of the input, while the speed of the input variation between the extremum points has no influence on the output. In this type of systems only the "recent memory" is used, but the memory time interval is not fixed as in the case of the systems with delay.

Although Volterra's studies on rate dependent memory phenomena date back to the beginning of the XX-th century, the history of hysteresis (i.e. rate independent effects) is quite shorter. As far as we know it was only in 1966 that a rigorous, functional approach was used for the description of several hysteresis phenomena [10] . In the period 1970-1980 M. A. Krasnosel'ski, A. V. Pokrovski and co-workers proposed a mathematical formulation of some physical models in terms of hysteresis operators and they realized a systematic analysis of the mathematical properties of these operators. In 1989 an important monograph written by the Russian group was translated into English [11]. Many other important monograph were published from that moment [12], [13].

Hysteresis effects were first related with the studies of the ferromagnetism (since 1887, when Lord Rayleigh proposed a model of ferromagnetic hysteresis) but now they are encounted in different area of science: electromagnetism, mechanics, elasto-plasticity, superconductivity, optics, nuclear physics, economics.

In the last time the hysteresis memory effects were pointed out in fusion plasma physics, which is a very important research topic because its aim is the production of energy by the thermonuclear controlled fusion. Clear hysteresis-like effects have been observed and described in the gyrotron oscillators [14]. The study of the L/H transition [15] or of the neoclassical tearing modes [16] can be also related with the hysteresis effect. Some models based on the hysteresis effect and the catastrophe theory were proposed for the study of the Edge Localized Modes [17] and of the fast magnetic reconnection [18] in tokamaks. These new results show the importance of the memory effects in the nuclear fusion.

An other type of memory effect is involved in the fractional dynamical systems, which are based on the fractional calculus (i.e. calculus of integral and derivatives of any arbitrary real order). The theory of the fractional dynamical systems has gained considerable importance during the past three decades, due mainly to its applications in diverse and wide spread field of science and engineering [20], [21], [22]. It is widely accepted that all the definitions proposed for fractional derivatives are "non-local" in the sense all the values of the function $x = x(t)$ in the interval $[t_0, t]$ are used for the computation of $(D^{\alpha}x)(t)$ when $\alpha \notin N$, contrary to the usual derivatives which are local operators. For example the Caputo fractional derivative (which was introduced in 1969 [23]) is defined by $({}_{t_0}D_C^{\alpha}x)(t) = \frac{1}{\Gamma(n-\alpha)}\int_{t_0}^{t}\frac{x^{(n)}(\tau)}{(t-\tau)^{n+1-\alpha}}d\tau$ for $\alpha \in [n-1, n)$. We interpret this "non-locality" as a memory effect in the sense that the evolution of a system

$$({}_{t_0}D_C^{\alpha}x)(t) = f(x), \qquad x(t_0) = x_0$$

is influenced not only by the initial condition $x(t_0) = x_0$, but also by all the values of x in the interval $[t_0, t]$ and this is due to the definition of the fractional derivative.

The three type of systems we presented (delayed, with hysteresis and fractional) are different approaches of the use of memory effects: the delayed systems consider the influence of the past using a constant, prescribed, memory-time interval, for the hysteresis effect the length of the memory-time interval is not constant and the fractional systems take into account the whole evolution of the system from the initial moment of the study to present.

In what follows we will use the memory effect systems in order to explain some phenomena which were experimentally observed in tokamaks (toroidal devices used for the production of energy through the thermonuclear controlled fusion).

The paper is organized as follows: in Section 2 a hysteresis model is used for the description of sawtooth crash of the central temperature in ASDEX-Upgrade tokamak, in Section 3 a fractional model is used in order to explain some feature of the anomalous diffusion observed in many tokamaks, a summary and conclusions are given in Section 4.

2 Hysteresis and sawtooth oscillations

The control of the plasma's temperature is crucial for the success of experiments in tokamaks. Many phenomena which were experimentally observed are not yet entirely explained. One of them is the sawtooth oscillation of the plasma temperature in the center of the ASDEX-Upgrade tokamak: during the stable ramp phase heating raises the temperature (the rise time of the is about 0.007 s) and at the collapse the associated thermal energy is released in the outer part of the plasma in the form of a heat pulse (the crash time about 0.00005 s). In order to reproduce some experimental data a model based on hysteresis effect of delayed relay (thermostat) type can be used [24].

The model is based on the one-dimensional diffusion equation:

$$\frac{\partial T}{\partial t}(r,t) = D(t)\frac{\partial^2 T}{\partial r^2}(r,t) + S(r) \tag{1}$$

The temporal coordinate is t, the spatial coordinate is $r \in [0,a]$ (representing the radial coordinate in the toroidal device with the poloidal radius a) , $S : [0,a] \to R$ is the source term, $D : [0,\infty) \to R$ is the diffusion coefficient and $T : [0,a] \times [0,\infty) \to R$ describes the temperature of plasma. This diffusion coefficient incorporates the hysteresis effect. It satisfies the equation

$$D'(t) = \frac{H\left(\frac{\partial T}{\partial r}(r_0,t)\right) - D(t)}{\tau} \tag{2}$$

The values of the parameters in the equation (2) are $r_0 = 0.15$ (specific to configuration of the ASDEX-Upgrade tokamak) and $\tau = 0.00004$ (specific for a particular experiment, #20975). The source term is $S(r) = S_0 \cdot (1 - r/0.15)$.

The hysteresis operator of delayed relay type [12], acting on $C([0,\infty))$, depends on the parameters $D_{\min}, D_{\max}, \beta, k$ and on the initial value $\eta_0 \in \{D_{\min}, D_{\max}\}$.The value $(Hx)(0)$ is defined by:

$$(Hx)(0) = \begin{cases} D_{\min}, & x(0) \le \beta k \\ \eta_0, & k < x(0) < k \\ D_{\max}, & x(0) \ge k \end{cases}. \tag{3}$$

For $t > 0$ one considers $X_t = \{\tau \in (0,t],\ \ x(\tau) \in \{D_{\min}, D_{\max}\}\}$ and the value $(Hx)(t)$ is defined by

$$(Hx)(t) = \begin{cases} (Hx)(0), & X_t = \varnothing \\ D_{\min}, & x(\max X_t) = D_{\min} \\ D_{\max}, & x(\max X_t) = D_{\max} \end{cases}. \tag{4}$$

The initial conditions of the system are

$$\begin{aligned} \frac{\partial T}{\partial t}(0,t) &= 0 \\ T(r,0) &= 2.18 \\ T(0.15,t) &= 2.18 \end{aligned} \tag{5}$$

and the values of the parameters are $D_{\min} = 1$, $D_{\max} = 2521$, $k = 3.6$. The initial conditions (5) represent the Neumann boundary condition in the plasma interior, the experimental temperature before the sawtooth crash, respectively the Dirichlet condition at the position of the $q = 1$ surface.

Proposition 1 *For any time $P > 0$ the problem (1), (2), (5) has an unique solution $T \in H^1\left(0, P; L^2[0, 0.15]\right) \cap L^\infty\left(0, P; H_0^1\left([0, 0.15]\right)\right)$.*

For the proof, which is very technical, the method presented in [12], pp 334-339, was adapted.

In order to solve numerically the equation (1) by taking into account the relations (2), (3) and (4) we used a grid method, namely the Crank-Nicolson implicit algorithm. For the discretization of the hysteresis operator a equidistant temporal grid $0 = t_0 < t_1 < ... < t_n < ...$was considered and the following expression was obtained from (4):

$$(Hx)(t_{n+1}) = \begin{cases} (Hx)(t_n), & \beta k < x(t_{n+1}) < k \\ D_{\min}, & x(t_{n+1}) \leq \beta k \\ D_{\max}, & x(t_{n+1}) \geq k \end{cases}.$$

The temporal step was $h_t = 0.000001$ and the spatial step was $h_r = 0.003$.

The system has two other parameters: the strength S_0 of the source term and the hysteresis parameter β. By changing these two parameters one can try to reproduce the experimental full cycle of the sawtooth crash in ASDEX Upgrade.

It was observed that the model reproduces correctly the two time scales of the sawtooth crash in ASDEX Upgrade tokamak (the slow rise time around 7 ms and the rapid crash time around 50 μs) for $\beta = 0.5$ and $S_0 = 240$.

A systematic study of the dependence of the sawtooth's full cycle upon the parameters values was accomplished. It was observed that it is possible to regulate the width and the height of sawtooth peaks by changing the values of the parameters and . For example larger values S_0 shorten the sawtooth period, but larger values of β prolong it.

3 A fractional diffusion equation and application in plasma physics

The transport of a scalar field in one dimension is, according to the standard diffusive paradigm,

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(D \cdot \frac{\partial T}{\partial x}\right) - \frac{\partial (V \cdot T)}{\partial x} + S. \tag{6}$$

In this equation $D = D\left(x, t, T, \frac{\partial T}{\partial x}\right)$ is the diffusion coefficient, $V = V(x, t)$ is the velocity pinch and $S = S(x, t)$ is the source term. The term $\frac{\partial}{\partial x}\left(D \cdot \frac{\partial T}{\partial x}\right)$ is the diffusive transport and the term $\frac{\partial (V \cdot T)}{\partial x}$ is the convective transport.

This equation may be used for the description of the heat transport in tokamaks. The goal of the transport modelling is to find V and D based on theory, numerics or experimental evidence. The "V-D" paradigm is a local description that assumes a well-defined transport scale and that widely separated regions do not interact. There are evidences that indicate that this assumption might be too restrictive: non-diffusive scaling, the non-Gaussian statistics, fast propagation phenomena, and non-local transport. An important current problem in fusion research is to develop models capable to describe these non-diffusive transport phenomena and many fractional diffusion equations are good candidates. For example It was shown [25] that the equation

$$\frac{\partial T}{\partial t}=\frac{\partial}{\partial x}\left(A\frac{\partial T}{\partial x}+B\cdot{}_{a}^{R}D_{x}^{\alpha}T+C\cdot{}_{x}^{R}D_{b}^{\alpha}T\right)+S\left(x,t\right)$$

is a model of the heat transport which fit the experimental data obtained in the Joint European Torus (JET) tokamak. In this equation A, B, C are specific coefficients and ${}_{a}^{R}D_{x}^{\alpha}T$ respectively ${}_{x}^{R}D_{b}^{\alpha}T$ represents the left, respectively the right spatial derivatives of T is the Riemmann sense. This equation has not memory effect because the time derivative $\partial T/\partial t$ is a local operator.

In order to study the memory effect in the diffusion equation we will consider the model

$$\frac{\partial_{C}^{\alpha}T}{\partial_{C}t^{\alpha}}-\frac{\partial^{2}T}{\partial x^{2}}=0.$$

For $\alpha\in(0,1)$ this equation is known as the fractional diffusion homogenous equation The Caputo time-derivative is defined by

$$\frac{\partial_{C}^{\alpha}T}{\partial_{C}t^{\alpha}}\left(x,t\right)=\frac{1}{\Gamma\left(1-\alpha\right)}\int_{0}^{t}\frac{\left(\partial T/\partial t\right)\left(x,\tau\right)}{\left(t-\tau\right)^{2-\alpha}}d\tau.$$

It is appropriate for the study of linear differential equations because it is compatible with the Laplace transform.

The aim of our study is to point out some traveling-wave solutions for the fractional diffusion equation.

The traveling-wave solution has the form $T\left(x,t\right)=T\left(x+ct\right)=T\left(\zeta\right)$. In terms of ζ, the homogenous fractional diffusion becomes

$$T''\left(\zeta\right)-c\cdot\left(D_{C}^{\alpha}T\right)\left(\zeta\right)=0 \tag{7}$$

Proposition 2 *The general solution of the equation (7) is*

$$T_{\alpha}\left(\zeta\right)=K_{1}+K_{2}\cdot\zeta\cdot\sum_{k=0}^{\infty}\frac{\left(c\cdot\zeta^{2-\alpha}\right)^{k}}{\Gamma\left(2+\left(2-\alpha\right)\cdot k\right)} \tag{8}$$

where K_1 and K_2 are arbitrary constants and $\zeta\in R$

Proof: For solving the linear differential equation (7) the Laplace transform will be used. From the operational equation associated with (7), one obtains

$$\left(LT\right)\left(p\right)=T\left(0\right)\cdot\frac{1}{p}+T'\left(0\right)\cdot\frac{1}{p^{\alpha}\left(p^{1-\alpha}-c\right)}.$$

In order to find the original $T(\zeta)$ we use the formula $L\left(t^{\beta-1}E_{\gamma,\beta}\left(\lambda t^{\gamma}\right)\right)(p) = \frac{p^{\gamma-\beta}}{p^{\gamma}-\lambda}$ (see [22], p. 50) for $\gamma = 2-\sigma$, $\beta = 2$ and $\lambda = c$. In this formula $E_{\gamma,\beta}$ represents the generalized Mittag-Lefler function, namely the analytic function $E_{\gamma,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\gamma\cdot k+\beta)}$. The equality (8) is obvious because $T(0)$ and $T'(0)$ are arbitrary numbers.

Proposition 3 *The traveling wave solution of the classical diffusion equation with the initial conditions $T(0) = a$, $T'(0) = b$ is punctually approximated by the traveling wave solution of the fractional diffusion equation with the same initial conditions.*

Proof For $\alpha \to 1$, $\alpha < 1$ we have $\lim_{\alpha\to 1}\left(a + b\cdot\zeta\cdot\sum_{k=0}^{\infty}\frac{\left(c\cdot\zeta^{2-\alpha}\right)^k}{\Gamma(2+(2-\alpha)\cdot k)}\right) = a + \frac{b}{c}\cdot\sum_{k=0}^{\infty}\frac{(c\cdot\zeta)^{k+1}}{\Gamma(2+k)} = a + \frac{b}{c}\left(e^{c\zeta}-1\right)$, i.e. $\lim_{\alpha\to 1} T_{\alpha}(\zeta) = T_1(\zeta)$.

The previous equation is the simplest case of the fractional-time model of the heat transport in tokamaks, namely $\frac{\partial_C^{\alpha}T}{\partial t^{\alpha}} = \frac{\partial}{\partial x}\left(D\cdot\frac{\partial T}{\partial x}\right) - \frac{\partial(V\cdot T)}{\partial x} + S$.This equation will be studied in future works for various velocity pinchs and sources.

4 Conclusions

Some basic information about the systems that involve memory effect was presented and two applications were proposed in order to reproduce some experimental data obtained in experiments in tokamaks (toroidal devices where the energy is obtained through the thermonuclear controlled fusion). Using these mathematical models we point out that the memory effects are important parts of a realistic description of phenomena of the real world.

References

[1] G. Stepan, Retarded dynamical systems, stability and characteristic functions, Longman Scientific and Technical, England, 1989

[2] I. D. Albu, M. Neamtu, M. Opris, Dissipative mechanical systems with delay,arXiv:math.DS/0412396v1

[3] G. Mircea, M. Neamtu, D. Orpis, Bifurcatii Hopf pentru sisteme cu argument intarziat si aplicatii, Editura MIRTON, Timisoara, 2004

[4] H. R. Thieme, X.-Q. Zhao, Asymptotic speeds of spread and traveling waves for integral equations and delayed reaction-diffusion models, J. Differential Equations, 195 (2003), 430-470.

[5] E. Liz, M. Pinto, V. Tkachenko and S. Trofimchuk, A global stability criterion for a family of delayed population models, Quart. Appl. Math., 63 (2005) 56-70.

[6] M. Chen, D. Zhou, Y. Shang, A simple time-delayed method to control chaotic systems, Chaos, Solitons and Fractals 22 (2004), 1117-1125

[7] A.Stefanski, A. Dabrowski, T. Kapitaniak, Evaluation of the largest Lyapunov exponent in dynamical systems with time delay, Chaos, Solitons and Fractals 23 (2005) 1651–1659

[8] X. Liao, G. Chen, Local stability, Hopf and resonant codimension–two bifurcation in a harmonic oscillator with two time delays, International Journal of Bifurcation and Chaos, Vol. 11, No. 8(2001), 2105–2121.

[9] Anca-Veronica Ion, An example of Bautin type bifurcation in a delay differential equations, J. Math. Anal. Appl. 329 (2007) 777–789

[10] R. Bouc, Solution periodique de e-equation de la ferroresonance avec hysteresis, C. R. Acad. Sci. Paris, Serie A 263 (1966), 197-199

[11] M. A. Krasnosel'ski, A. V. Pokrovski, Systems with hysteresis, Springer, Berlin,1989, Russian edition: Nauka, Moscow, 1983

[12] A. Visintin, Differential Models of Hysteresis, Springer-Verlag Berlin Heidelberg , 1994

[13] I.D. Mayergoyz, Mathematical Models of Hysteresis and Their Applications, Elsevier, 2003

[14] O. Dumbrajs et al, Hysteresis-like effects in gyrotron oscillators, Physics of Plasmas 10, no.3 (2003), 1183-1186

[15] S. Toda, et al, Double hysteresis in L/H transition and compound dithers, Plasma Phys. Control. Fusion 38, (1996)1337 .

[16] H. Zohm, et al, Neoclassical MHD in ASDEX Upgrade and COMPASS-D, Plasma Phys. Control. Fusion 39, (1997) B237 .

[17] S.-I. Itoh et al, Physics of collapses: probabilistic occurrence of ELMs and crashes, Plasma Phys. Control. Fusion 40, (1998), 737

[18] P.A. Cassaket al, Catastrophe model for fast magnetic reconnection onset, Phys. Rev. Letters 95, (2005), 235002

[19] W. Deng, J. Lu, Generating multi-directional multi-scroll chaotic attractors via a fractional differential hysteresis system, Physics Letters A 369 (2007) 438–443

[20] E. Hilfer (Ed), Applications of fractional calculus in physics, World Sci. Publishing, New York, 2000

[21] G. M. Zaslavsky, Chaos in fractional dynamics, vol 511of Lect. Notes in Physics, Oxford Univ. Press, 2005

[22] A.A. Kilbas, H.M. Srivastava, J. J. Trujillo, Theory and applications of fractional differential equations, Elsevier, Amsterdam, 2006

[23] M. Caputo, Elasticita e Dissipatione, Zanichelle, Bologna, 1969

[24] O. Dumbrajs, V. Igochine, H.Zohm and ASDEX-Upgrade team, Hysteresis in sawtooth crash in ASDEX Upgrade tokamak, private communication, presented at the International workshop "Anomalous transport in plasma fusion", Craiova, October 6-8, 2008

[25] D. del-Castillo-Negrete, Fractional diffusion models of non-local transport Phys. Plasmas 13, 082308 (2006)

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On exponential stability of linear skew-evolution semiflows in Banach spaces

Nicolae Lupa and Ioan-Lucian Popa
Department of Mathematics
Faculty of Mathematics and Computer Science
West University of Timişoara, Romania
E-mail: nlupa@math.uvt.ro,popa@math.uvt.ro *

Abstract

In this paper we consider an exponential stability concept for linear skew-evolution semiflows in Banach spaces and we deduce the versions of some well-known theorems due to Datko, Krein-Daletckij and Rolewicz.

Keywords evolution equations, exponential stability.

1 Notions and preliminary results

Let X be a Banach space, Θ a locally compact metric space and $\mathcal{B}(X)$ the Banach algebra of all linear bounded operators acting on X. Let Δ be the set defined by

$$\Delta=\{(t,s)\in\mathbb{R}_+^2 : t\geq s\}.$$

Definition 1.1 A family of operators $\mathcal{U}=\{U(t,s)\}_{t\geq s\geq 0}\subset\mathcal{B}(X)$ is called *evolution family* on X if it satisfies the following two properties:

$E1)$ $U(t,t)=I$(the identity on X), for all $t\geq 0$

$E2)$ $U(t,s)U(s,t_0)=U(t,t_0)$, for all $t\geq s\geq t_0\geq 0$.

An evolution family is called *strongly continuous* if for every $x\in X$ the function

$$\Delta\ni(t,s)\longmapsto\parallel U(t,s)x\parallel\in\mathbb{R}_+$$

is continuous.

If there exist the constants $M\geq 1$ and $\omega>0$ such that

$$\parallel U(t,s)\parallel\leq Me^{\omega(t-s)},\ \text{for all } t\geq s\geq 0$$

then the evolution family is said to be with *exponential growth.*

*The work was supported by the Exploratory Research Grant PN II ID 1080 / 2009

ISSN-1584-5990

We consider the following non-autonomous Cauchy problem:

$$(CP) \quad \begin{cases} \dot{u}(t) = A(t)u(t),\ t \geq s \geq 0 \\ u(s) = x \end{cases}$$

where $A(t) : D(A(t)) \subset X \longrightarrow X$ is in general an unbounded linear operator on the Banach space X, for every $t \geq s$ and $x \in X$.

Definition 1.2 We said that $A(\cdot)$ *generate an evolution family on* X if the Cauchy problem (CP) is "well-posed", in the sense that there exists a strongly continuous evolution family on X with exponential growth solving (CP), i.e., the solution of (CP) is given by $u(t) = U(t,s)u(s)$, for $t \geq s$(see, e.g. [6]).

Definition 1.3 A continuous mapping $\varphi : \Delta \times \Theta \longrightarrow \Theta$ is called *evolution semiflow* on Θ if it satisfies the following properties:

$s1)$ $\varphi(t,t,\theta) = \theta$, for all $t \geq 0$

$s2)$ $\varphi(t,t_0,\theta) = \varphi\left(t,s,\varphi(s,t_0,\theta)\right)$, for all $t \geq s \geq t_0 \geq 0$ and all $\theta \in \Theta$.

We will study the existence and uniqueness of the solution for the perturbation Cauchy problem:

$$(A,B) \quad \begin{cases} \dot{u}(t) & = [A(t) + B(\varphi(t,t_0,\theta))]u(t),\ t \geq t_0 \\ u(t_0) & = x \end{cases}$$

where $A(\cdot)$ generate an evolution family on the Banach space X , φ is an evolution semiflow on the compact metric space Θ and $B : \Theta \longrightarrow \mathcal{B}(X)$ is a strongly continuous mapping, $t_0 \geq 0$ and $x \in X$.

As in [2] it can be proved the following result:

Lemma 1.4 *If* $B : \Theta \longrightarrow \mathcal{B}(X)$ *is strongly continuous and* $u : \mathbb{R}_+ \longrightarrow X$ *is a continuous function, then for each* $\theta \in \Theta$ *and each* $t_0 \geq 0$ *the mapping* $[t_0,\infty) \ni t \longmapsto B(\varphi(t,t_0,\theta))u(t) \in X$ *is continuous, where* φ *is an evolution semiflow on the compact metric space* Θ.

According with previous Lemma we can consider the following integral equation:

$$u(t) = U(t,t_0)x + \int_{t_0}^{t} U(t,\tau)B(\varphi(\tau,t_0,\theta))u(\tau)\,d\tau, \tag{1}$$

for all $t \geq t_0 \geq 0$ and for all $\theta \in \Theta$.

Definition 1.5 A solution of equation (1) is called *mild solution* of the Cauchy problem (A,B).

Theorem 1.6 *If* $A(\cdot)$ *generate an evolution family on* X *and* $B : \Theta \longrightarrow \mathcal{B}(X)$ *is strongly continuous then the Cauchy problem (A,B) has an unique mild solution* $\Phi(t,t_0,\theta)x$ *given by*

$$\Phi(t,t_0,\theta)x = U(t,t_0)x + \int_{t_0}^{t} U(t,\tau)B(\varphi(\tau,t_0,\theta))\Phi(\tau,t_0,\theta)x\,d\tau \tag{2}$$

for all $t \geq t_0$.

Moreover, the mapping $C : \Delta \times X \times \Theta \longrightarrow X \times \Theta$ *given by*

$$C(t, t_0, x, \theta) = (\Phi(t, t_0, \theta)x, \varphi(t, t_0, \theta))$$

satisfies the following properties:

1) $\Phi(t, t_0, \theta) \in \mathcal{B}(X)$, for all $(t, t_0) \in \Delta$ and $\theta \in \Theta$
2) $\Phi(t_0, t_0, \theta) = I$, for all $(t_0, \theta) \in \mathbb{R}_+ \times \Theta$
3) $\Phi(t, t_0, \theta) = \Phi(t, s, \varphi(s, t_0, \theta))\Phi(s, t_0, \theta)$, for all $t \geq s \geq t_0 \geq 0$ and $\theta \in \Theta$
4) *The function*

$$[t_0, \infty) \ni t \longmapsto \| \Phi(t, t_0, \theta)x \| \in \mathbb{R}_+$$

is uniformly continuous on $\theta \in \Theta$, *for all* $t_0 \geq 0$ *and* $x \in X$.

Proof For $t_0 \geq 0$, $\theta \in \Theta$ and $x \in X$ we consider the sequence:

$$\Phi_0(t; t_0, \theta)x = U(t, t_0)x$$

$$\Phi_1(t; t_0, \theta)x = U(t, t_0)x + \int_{t_0}^{t} U(t, \tau)B\left(\varphi(\tau, t_0, \theta)\right)\Phi_0(\tau; t_0, \theta)x\, d\tau$$

$$\vdots$$

$$\Phi_n(t; t_0, \theta)x = U(t, t_0)x + \int_{t_0}^{t} U(t, \tau)B\left(\varphi(\tau, t_0, \theta)\right)\Phi_{n-1}(\tau; t_0, \theta)x\, d\tau$$

Using Lemma 1.4 it follows inductively that the function

$$[t_0, \infty) : t \longmapsto \Phi_n(t; t_0, \theta)x \in X$$

is continuous for all $n \in \mathbb{N}$, and hence the sequence $\{\Phi_n(t; t_0, \theta)x\}$ is well defined. For $t \geq t_0$ we have:

$$\begin{aligned} \| \Phi_1(t; t_0, \theta)x - \Phi_0(t; t_0, \theta)x \| &\leq \int_{t_0}^{t} \| U(t, \tau)B\left(\varphi(\tau, t_0, \theta)\right)U(\tau, t_0)x \|\, d\tau \\ &\leq M^2 \sup_{\theta \in \Theta} \| B(\theta) \| (t - t_0)e^{\omega(t-t_0)} \| x \| \\ &= M^2 L(t - t_0)e^{\omega(t-t_0)} \| x \| \end{aligned}$$

and by induction we obtain:

$$\| \Phi_n(t; t_0, \theta)x - \Phi_{n-1}(t; t_0, \theta)x \| \leq M^{n+1}L^n \frac{(t - t_0)^n}{n!} e^{\omega(t-t_0)} \| x \|$$

By the above inequality we deduce that

$$\| \Phi_{n+p}(t; t_0, \theta)x - \Phi_n(t; t_0, \theta)x \| \leq M \sum_{k=n+1}^{n+p} \frac{(ML)^k}{k!}(t - t_0)^k e^{\omega(t-t_0)} \| x \|$$

Hence, the sequence $\{\Phi_n(t;t_0,\theta)x\}$ is uniform convergence on any compact interval that contain t and t_0.
Denoting by

$$\Phi(t,t_0,\theta)x = \lim_{n\to\infty} \Phi_n(t;t_0,\theta)x$$

we obtain that

$$\Phi(t,t_0,\theta)x = U(t,t_0)x + \int_{t_0}^{t} U(t,\tau)B\left(\varphi(\tau,t_0,\theta)\right)\Phi(\tau,t_0,\theta)x\,d\tau$$

Let $\Psi(t,t_0,\theta)x$ be another solution of integral equation (2).
Then we obtain:

$$\Phi(t,t_0,\theta)x - \Psi(t,t_0,\theta)x = \int_{t_0}^{t} U(t,\tau)B(\varphi(\tau,t_0,\theta))[\Phi(\tau,t_0,\theta)x - \Psi(\tau,t_0,\theta)x]\,d\tau.$$

Hence

$$\| \Phi(t,t_0,\theta)x - \Psi(t,t_0,\theta)x \| \leq \int_{t_0}^{t} Me^{\omega(t-\tau)}L \| \Phi(\tau,t_0,\theta)x - \Psi(\tau,t_0,\theta)x \|\, d\tau \tag{3}$$

where $L = \sup_{\theta\in\Theta} \| B(\theta) \|$.
We consider now $f(t) = e^{-\omega t} \| \Phi(t,t_0,\theta)x - \Psi(t,t_0,\theta)x \|$. Then from (3) we have that

$$f(t) \leq ML\int_{t_0}^{t} f(\tau)d\tau.$$

Using Gronwall's Lemma we obtain that $f(t) = 0$ and hence $\Phi(t,t_0,\theta)x = \Psi(t,t_0,\theta)x$.
It is obvious that $\Phi(t_0,t_0,\theta) = I$ and $\Phi(t,t_0,\theta) \in \mathcal{B}(X)$ for all $(t,t_0) \in \Delta$ and $\theta \in \Theta$.
To prove properties 3) we shall use the Gronwall's Lemma and the following relation:

$$\Phi(t,t_0,\theta)x - \Phi(t,s,\varphi(s,t_0,\theta)\Phi(s,t_0,\theta)x =$$

$$= \int_{s}^{t} U(t,\tau)B(\varphi(\tau,t_0,\theta))\left[\Phi(\tau,t_0,\theta)x - \Phi(\tau,s,\varphi(s,t_0,\theta))\Phi(s,t_0,\theta)x\right]d\tau$$

It remains to prove 4). Indeed, for $t \geq t_0$ and $h > 0$ we have

$$\| \Phi(t+h,t_0,\theta)x - \Phi(t,t_0,\theta)x \| \leq \| U(t+h,t_0)x - U(t,t_0)x \| +$$

$$+\int_{t_0}^{t} \| U(t+h,\tau) - U(t,\tau) \| \cdot \| B(\varphi(\tau,t_0,\theta)) \| \cdot \| \Phi(\tau,t_0,\theta)x \| \, d\tau$$

$$+\int_{t}^{t+h} \| U(t+h,\tau)B(\varphi(\tau,t_0,\theta))\Phi(\tau,t_0,\theta)x \| \, d\tau$$

From Lebesque's dominated convergence theorem we obtain that $\Phi(\cdot,t_0,\theta)x$ is continuous on the right uniformly on Θ. In the same way we can prove the continuity on the left.

Definition 1.7 A mapping $\Phi : \Delta \times \Theta \longrightarrow \mathcal{B}(X)$ is called *evolution cocycle* over an evolution semiflow φ if:

(c_1) $\Phi(t,t,\theta) = I$,(the identity on X), for all $(t,\theta) \in \mathbf{R}_+ \times \Theta$

(c_2) $\Phi(t,s,\varphi(s,t_0,\theta))\Phi(s,t_0,\theta) = \Phi(t,t_0,\theta)$, for all $t \geq s \geq t_0 \geq 0,\ \theta \in \Theta$.

Definition 1.8 The mapping $C : \Delta \times X \times \Theta \to X \times \Theta$ defined by the relation

$$C(t,s,x,\theta) = (\Phi(t,s,\theta)x, \varphi(t,s,\theta)), \tag{4}$$

where Φ is an evolution cocycle over an evolution semiflow φ, is called *linear skew-evolution semiflow* on $X \times \Theta$.

Definition 1.9 A linear skew-evolution semiflow C is said to be :

(i) *uniformly exponentially stable* if there exist the constants $N \geq 1$ and $\nu > 0$ such that

$$\|\Phi(t,t_0,\theta)x\| \leq Ne^{-\nu(t-s)} \|\Phi(s,t_0,\theta)x\|, \tag{5}$$

for all $(t,s),(s,t_0) \in \Delta$ and all $(x,\theta) \in X \times \Theta$;

(ii) *exponentially stable* if there exist the constants $N \geq 1$ and $\nu > 0$ such that for each $x \in X$ there exists $t_0 = t_0(x) \geq 0$ with

$$\|\Phi(t,t_0,\theta)x\| \leq Ne^{-\nu(t-s)} \|\Phi(s,t_0,\theta)x\|, \tag{6}$$

for all $t \geq s \geq t_0$ and all $\theta \in \Theta$.

Remark 1.10 If a linear skew-evolution semiflow on $X \times \Theta$ is uniformly exponentially stable then it is exponentially stable.

The following example shows that the converse is not valid.

Example 1.11 For $X = \mathbb{R}^2$ with the euclidian norm and $\varphi : \Delta \times \Theta \longrightarrow \Theta$ be an arbitrary evolution semiflow on a locally compact metric space Θ, the mapping $C : \Delta \times X \times \Theta \longrightarrow X \times \Theta$, $C(t,t_0,x,\theta) = (\Phi(t,t_0,\theta)x, \varphi(t,t_0,\theta))$, with

$$\Phi(t,t_0,\theta)(x_1,x_2) = (\xi_1,\xi_2)$$

where

$$\xi_1 = e^{-(t-t_0)} \cos t\,(x_1 \cos t_0 + x_2 \sin t_0) + e^{t-t_0} \sin t\,(x_1 \sin t_0 - x_2 \cos t_0)$$

$$\xi_2 = e^{-(t-t_0)} \sin t\,(x_1 \cos t_0 + x_2 \sin t_0) - e^{t-t_0} \cos t\,(x_1 \sin t_0 - x_2 \cos t_0)$$

for all $(t,t_0,\theta) \in \Delta \times \Theta$, is an exponentially stable linear skew-evolution semiflow on $X \times \Theta$, which is not uniformly exponentially stable (for proof see [4]).

2 The main results

Proposition 2.1 A linear skew-evolution semiflow $C = (\Phi, \varphi)$ is exponentially stable if and only if there are two constants $N \geq 1$ and $\nu > 0$ such that for each $x \in X$ there exists $t_0 = t_0(x) \geq 0$ with

$$\| \Phi(t+h, t_0, \theta)x \| \leq Ne^{-\nu h} \| \Phi(t, t_0, \theta)x \| \tag{7}$$

for all $h \geq 0$, $t \geq t_0$ and for all $\theta \in \Theta$.

Proof *Necessity* is obvious.
Sufficiency rezults from relation :

$$\Phi(t, t_0, \theta)x = \Phi(s + t - s, t_0, \theta)x$$

Proposition 2.2 *Let $C = (\Phi, \varphi)$ be a linear skew-evolution semiflow on $X \times \Theta$ with exponential growth such that for all $x \in X$, $\theta \in \Theta$ and for all $t \geq 0$ the function*

$$\mathbb{R}_+ \ni \tau \longmapsto \| \Phi(t+\tau, t, \theta)x \| \in \mathbb{R}_+$$

is continuous. Then the following statements are equivalent:

(i) *C is exponentially stable*;

(ii) *There exist $T > 0$ and $c \in (0, 1)$ such that for each $x \in X$ there exists $t_0 = t_0(x) \geq 0$ and for each $\theta \in \Theta$ exists $\eta = \eta(x, \theta) \in (0, T]$ with*

$$\| \Phi(t+\eta, t_0, \theta)x \| \leq c \| \Phi(t, t_0, \theta)x \|, \ for\ all\ t \geq t_0 \tag{8}$$

(iii) *There exist $T > 0$ and $c \in (0, 1)$ such that for each $x \in X$ there exists $t_0 = t_0(x) \geq 0$ and for each $\theta \in \Theta$ and $t \geq t_0$ exists $\eta = \eta(x, \theta, t) \in (0, T]$ with*

$$\| \Phi(t+\eta, t_0, \theta)x \| \leq c \| \Phi(t, t_0, \theta)x \| \tag{9}$$

Proof (i) $\Longrightarrow$ (ii). Let $T > 0$ such that $Ne^{-\nu T} < 1$ and let $c = Ne^{-\nu T}$. For $x \in X$, $t_0 = t_0(x)$ and $\theta \in \Theta$, we consider $\eta = T$. Then from Proposition 2.1 we obtain the conclusion.
(ii) $\Longrightarrow$ (iii) is obvious.
(iii)$\Longrightarrow$ (i). Let $x \in X$. From hypothesis, there exists $t_0 = t_0(x)$ such that for each $t \geq t_0$ and each $\theta \in \Theta$ exists $\eta_1 \in (0, T]$ with

$$\| \Phi(t+\eta_1, t_0, \theta)x \| \leq c \| \Phi(t, t_0, \theta)x \| .$$

If we get $t \geq t_0$ and $t_1 = t + \eta_1$ then there exists $\eta_2 \in (0, T]$ with

$$\| \Phi(t+\eta_1+\eta_2, t_0, \theta)x \| \leq c^2 \| \Phi(t, t_0, \theta)x \| .$$

Inductively, we obtain that there exists a non-decreasing sequence $s_n = \eta_1 + \eta_2 + \cdots + \eta_n$, with $s_n \leq nT$ and

$$\| \Phi(t+s_n, t_0, \theta)x \| \leq c^n \| \Phi(t, t_0, \theta)x \| \tag{10}$$

for all $n \in \mathbb{N}^*$.

Case 1. If $\lim\limits_{n\to\infty} s_n = \infty$ then for $h > 0$ exists $n \in \mathbb{N}^*$ such that $s_n \leq h < s_{n+1} \leq (n+1)T$.
The following relations holds:

$$\begin{aligned} \| \Phi(t+h, t_0, \theta)x \| &= \| \Phi(t+h, t+s_n, \varphi(t+s_n, t_0, \theta))\Phi(t+s_n, t_0, \theta)x \| \\ &\leq Me^{\omega(h-s_n)} \| \Phi(t+s_n, t_0, \theta)x \| \\ &\leq Me^{\omega T}c^n \| \Phi(t, t_0, \theta)x \| \\ &= Me^{\omega T}e^{\nu T}e^{-\nu h} \| \Phi(t, t_0, \theta)x \| \end{aligned}$$

where $c = e^{-\nu T}$.
From Proposition 2.1 we deduce that C is exponentially stable.

Case 2. If $\lim\limits_{n\to\infty} s_n = \gamma \in \mathbb{R}$ then we obtain that

$$\Phi(t+\gamma, t_0, \theta)x = 0.$$

Moreover

$$\Phi(t+h, t_0, \theta)x = 0, \tag{11}$$

for all $h \geq \gamma$.
Let $h \in [0, \gamma)$, there exists $n \in \mathbb{N}^*$ such that $s_n \leq h < s_{n+1}$.
In a similar way as in Case 1 it result that there exist the contants $N \geq 1$ and $\nu > 0$ such that

$$\| \Phi(t+h, t_0, \theta)x \| \leq Ne^{-\nu h} \| \Phi(t, t_0, \theta)x \| \tag{12}$$

for all $t \geq t_0$, $h \geq 0$ and for all $\theta \in \Theta$. In conclusion, we obtain that C is exponentially stable.

Theorem 2.3 *Let $C = (\Phi, \varphi)$ be a linear skew-evolution semiflow on $X \times \Theta$ with exponential growth such that for all $x \in X$, $\theta \in \Theta$ and for all $t \geq 0$ the function*

$$\mathbb{R}_+ \ni \tau \longmapsto \| \Phi(t+\tau, t)x \| \in \mathbb{R}_+$$

is continuous. Then C is exponentially stable if and only if there exists a non-decreasing function $R : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, with $R(ab) \leq R(a)R(b)$ and $R(t) > 0$, for all $t > 0$ and exists a positive constant $K > 0$ such that for each $x \in X$ exists $t_0 \geq 0$ with

$$\int_0^\infty R(\| \Phi(t+\tau, t_0, \theta)x \|)\, d\tau \leq KR(\| \Phi(t, t_0, \theta)x \|) \tag{13}$$

for all $t \geq t_0$ and $\theta \in \Theta$.

Proof *Necessity.* It is a simple verification for $R(t) = t$, for all $t \geq 0$.
Sufficiency. We suppose that C is not exponentially stable. Then from Proposition 2.2 it follows that for every $T > 0$ and for every $c \in (0, 1)$ there exists $x_0 \in X$ such that for all $t \geq 0$ exists $\theta_0 \in \Theta$ and $\tau \geq t$ with

$$\| \Phi(\tau+\eta, t, \theta_0)x_0 \| > c \| \Phi(\tau, t, \theta_0)x_0 \| \tag{14}$$

for all $\eta \in (0,T]$. Let $K>0$ and let R a function like before. Then for $c \in (0,1)$ and $T > KR(\frac{1}{c})$, it follows that there exists $x_0 \in X$ such that for all $t \geq 0$ exists $\theta_0 \in \Theta$ and $\tau \geq t$ such that the relation (14) is true.
Hence, we obtain :

$$\begin{aligned} R\left(\frac{1}{c}\right)\int_0^\infty R(\parallel \Phi(\tau+\theta,t,\theta_0)x_0 \parallel)\, d\theta &\geq \int_0^\infty R\left(\frac{1}{c}\parallel \Phi(\tau+\theta,t,\theta_0)x_0 \parallel\right) d\theta \\ &\geq \int_0^T R\left(\frac{1}{c}\parallel \Phi(\tau+\theta,t,\theta_0)x_0 \parallel\right) d\theta \\ &\geq TR(\parallel \Phi(\tau,t,\theta_0)x_0 \parallel) \end{aligned}$$

which implies that

$$\int_0^\infty R(\parallel \Phi(\tau+\theta,t,\theta_0)x_0 \parallel)\, d\theta > KR(\parallel \Phi(\tau,t,\theta_0)x_0 \parallel$$

which contradicts the hypothesis.

Remark The previous theorem still remain valid if we get a non-decreasing function $\overline{R}:\mathbb{R}_+ \longrightarrow \mathbb{R}_+$, with $\overline{R}(ab) \geq \overline{R}(a)\overline{R}(b)$ and $\overline{R}(t)>0$, for all $t>0$.

Corollary Let $C=(\Phi,\varphi)$ be a linear skew-evolution semiflow on $X \times \Theta$ with exponential growth such that for all $x \in X$, $\theta \in \Theta$ and for all $t \geq 0$ the function

$$\mathbb{R}_+ \ni \tau \longmapsto \parallel \Phi(t+\tau,t,\theta)x \parallel \in \mathbb{R}_+$$

is continuous.Then C is exponentially stable if and only if there exist two positive constants $K>0$ and $p>0$ such that for each $x \in X$ exists $t_0 \geq 0$ with

$$\left(\int_0^\infty (\parallel \Phi(t+\tau,t_0,\theta)x \parallel)^p\, d\tau\right)^{1/p} \leq K \parallel \Phi(t,t_0,\theta)x \parallel \tag{15}$$

for all $t \geq t_0$ and $\theta \in \Theta$.

Theorem *Let $C=(\Phi,\varphi)$ be a linear skew-evolution semiflow on $X \times \Theta$ with exponential growth such that for all $x \in X$, $\theta \in \Theta$ and for all $t \geq 0$ the function*

$$\mathbb{R}_+ \ni \tau \longmapsto \parallel \Phi(t+\tau,t,\theta)x \parallel \in \mathbb{R}_+$$

is continuous. Then C is exponentially stable if and only if there exist two positive constants $K,\alpha>0$ such that for each $x \in X$ exists $t_0 \geq 0$ with

$$\sup_{t>s}\frac{1}{t-s}\int_s^t e^{\alpha(\tau-s)} \parallel \Phi(\tau,t_0,\theta)x \parallel\, d\tau \leq K \parallel \Phi(s,t_0,\theta)x \parallel \tag{16}$$

for all $s \geq t_0$ and $\theta \in \Theta$.

Proof *Necessity.* For $\alpha = \frac{\nu}{2}$ we obtain:

$$\begin{aligned}\frac{1}{t-s}\int_s^t e^{\alpha(\tau-s)} \parallel \Phi(\tau,t_0,\theta)x \parallel d\tau &\leq \frac{1}{t-s}\int_s^t e^{-\frac{\nu}{2}(\tau-s)}\, d\tau \parallel \Phi(s,t_0,\theta)x \parallel \\ &= \frac{1}{t-s}\int_0^{t-s} e^{-\frac{\nu}{2}u}\, du \parallel \Phi(s,t_0,\theta)x \parallel \\ &= \frac{2}{\nu}\frac{1-e^{-\frac{\nu}{2}(t-s)}}{t-s} \parallel \Phi(s,t_0,\theta)x \parallel \\ &\leq K \parallel \Phi(s,t_0,\theta)x \parallel .\end{aligned}$$

Sufficiency. We suppose that C is not exponentially stable. We consider the constants $K, \alpha > 0$ and $c \in (0,1)$.
Since $\lim\limits_{T\to\infty} \frac{e^{\alpha T}-1}{\alpha T} = \infty$, we obtain that there exists $\delta > 0$ such that $\frac{e^{\alpha T}-1}{\alpha T} > \frac{K}{c}$, for every $T > \delta$.
By Proposition 2.2 for $T > \delta$ it follows that there exists $x_0 \in X$ such that for any $t_0 \geq 0$ exists $\theta \in \Theta$ and $s \geq t_0$ with

$$\parallel \Phi(s+\eta,t_0,\theta)x_0 \parallel > c \parallel \Phi(s,t_0,\theta)x_0 \parallel$$

for all $\eta \in (0,T]$.
It follows that

$$\begin{aligned}\frac{1}{T}\int_s^{s+T} e^{\alpha(\tau-s)} \parallel \Phi(\tau,t_0,\theta)x_0 \parallel d\tau &\geq \frac{c}{T}\int_0^T e^{\alpha\tau}\, d\tau \parallel \Phi(s,t_0,\theta)x_0 \parallel \\ &> K \parallel \Phi(s,t_0,\theta)x_0 \parallel\end{aligned}$$

which contradicts the relation (16). Hence, C is exponentially stable.

References

[1] Buşe C., *Real Integrability Conditions for the Nonuniform Exponential Stability of Evolution Families on Banach Spaces*, International Series of Numerical Mathematics, Vol. 157, 31-42

[2] Chow S.-N., Leiva H., *Dynamical Spectrum for Time Dependent Linear Systems in Banach Spaces*, Japan J. Indust. Appl. Math., 11(1994), 379-415

[3] Daleckij J. L., Krein M. G., *Stability of Differential Equations in Banach Space*, Amer. Math. Soc., Providence, RI, 1974

[4] Lupa N., Popa I.L., *On asymptotic behaviour of linear differential equations in Banach spaces*, Proceding of the International Symposium-Research and Education in Innovation Era, 2nd Edition, Arad(2008)

[5] Megan M., Stoica C., Buliga L., *On asymptotic behaviors for linear skew-evolution semiflows in Banach spaces*, Carpathian J. Math., 23(2007), No. 1-2, 117-125

[6] Nickel G., *On evolution semigroups and wellposedness of nonautonomous Cauchy problems*, Ph. D. thesis, Tübingen, 1996

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Herman Ring Classification on Function $h(z)\,\Pi_i\{exp(g_i(z))[(a_i - z)/(1 - \bar{a}_i z)]\}$

David C. Ni
Direxion Technology, Taipei, Taiwan, R.O.C.
E-mail: davidcni@yahoo.com

Chou Hsin Chin
National Chiao Tung University
Department of Electrophysics, Hsin Chu, Taiwan, R.O.C.
E-mail: jchchin@cc.nctu.edu.tw

August 3, 2009

Abstract

Herman Rings represent a class of fractals with hierarchical structure in meromorphic dynamical systems. In this paper, we classify the complex function, $h(z)\prod_i\{exp(g_i(z))[(a_i - z)/(1 - \bar{a}_i z)]\}$, which is constructed based on Relativity and interaction energy. We list some interested features of this function set, such as limited number of partial Herman Ring domains, skip-symmetry, and symmetry broken.

Keywords: Herman Ring, dimension, fractal, skip-symmetry, symmetry broken

1 Introduction

Herman Rings represent a class of fractals with hierarchical structure in meromorphic dynamical systems. Due to the richness of mathematical elaboration and computing sophistication, we have observed limited efforts on simple rational functions [1-4].

In our previous efforts, we explored complex function: $h(z)\prod_4\{exp(g_i(z))[a_i - z)/(1 - \bar{a}_i z)]\}$ and revealed some interested characteristics of this function [5]. We also extended our observations to the applications in antenna areas [6, 7] and studied the mathematical foundation [8]. In this paper, we further classify this set of function in details, and provide some interesting observations on the function set.

ISSN-1584-5990

2 Function Construction and Class Formation

We define the function set, $f = z^q \prod_i C_i$, which may have the form, $f = z^q C_1 C_2 C_3$, where z is a complex variable, q is an integer, and C_i has following form :

$$C_i = exp(g_i(z))[(a_i - z)/(1 - \bar{a}_i z)] \tag{1}$$

Here $\bar{a}_i$ is the complex conjugate of complex number a_i. We propose this form based on the following form known in the theory of special relativity by A. Einstein:

$$u'_x = (u_x - v)/(1 - vu_x/c^2) \tag{2}$$

The z^q term represents the interaction of C_i. The term $exp(g_i(z))$ in equation (1) represents the phase, where $g_i(z)$ is a complex function. A given domain can be a domain of complex numbers, $z = x+yi$, with $(x^2+y^2)^{1/2} \leq R$. Here, R is a real number. We also use a simpler domain, $(x^2+y^2)^{1/2} = R$, to observe directly a mapping or a transformation from the original domains defined aforemetioned to a set of $x + yi$ after function iteration. There are two sets of domains will be discussed in this paper. Firstly, the remained domain, which is a converging subset of an original domain after function iteration, and secondly, the mapped domain of an original domain after function iteration. The function f will go through iteration as:

$$f^n(z) = f \circ f^{n-1} \tag{3}$$

Here, n is a positive integer indicating the order of function iteration. We selected $q = -1$ in the z^q term representing potential energy. The selection of q value indicates different physical implications, which will be discussed in the future. In this paper, we will focus on the following functions:

- $Z^{-1}C$ - first order
- $Z^{-1}C_1C_2$ - second order
- $Z^{-1}C_1C_2C_3$ - third order
- $Z^{-1}C_1C_2C_3C_4$ - fourth order with extension to higher orders

3 Computation and Transformation

In order to obtain the remained domains and mapped domains, we have to determine the criteria of convergence, divergence, and oscillation of function values after a given number of iterations. We adopted function values of several designated iterations, such as 10, 100, and 1000 times of iteration for this purpose. We define the parametric space or normalized space including the following parameters.

- z
- a - the normalized energy item
- $exp(g_i(z))$ - the normalized phase item
- Iteration

The selection of these parameters can transform the fractal domains into chaotic domains and vice versa.

4 Classification

4.1 Hierarchical Herman Rings of Original Domains

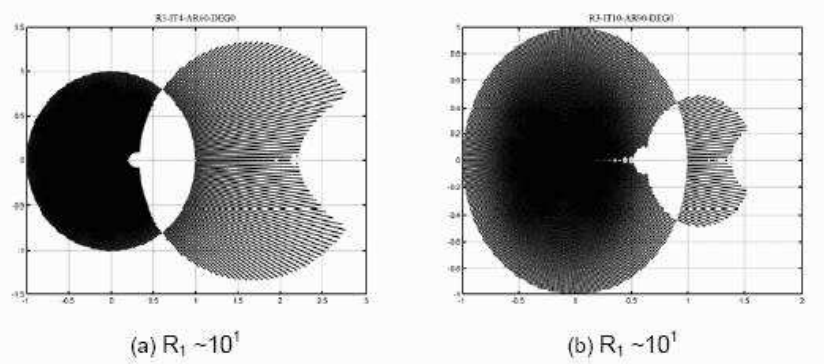


Figure 1: $Z^{-1}C$ domains

We will classify the function set based on the remained domains created by $Z^{-1}C$, $Z^{-1}C_1C_2$, $Z^{-1}C_1C_2C_3$, $Z^{-1}C_1C_2C_3C_4$ and some higher order functions. Figure 1 shows the $Z^{-1}C$ domains. Figure 1 (a) and (b) show a domain $z = x + yi$, with $(x^2 + y^2)^{1/2} \leq R_1$. This remained domain forms a a set of 2-level structure of partial Herman-Ring fractal domains based on three different parameter: $\{a\}$, R_1 and iteration numbers. We have not observed other remained domains of different R_1 values, where $(x^2 + y^2)^{1/2} \leq R_1$. The scale of these 2-level domains is in the range of $R_1 \sim 10^1$.

Figure 2 shows the $Z^{-1}C_1C_2$ domains. Figure 2 (a), (b) and (c) show a domain $z = x + yi$, with $(x^2 + y^2)^{1/2} \leq R_2$. The remained domain forms a set of 3-level structure of partial Herman-Ring fractal domains based on three different sets of parameter : {a}, R_2, and iteration number. Figure 2(a) shows several disconnected domains at scale of $R_2 \sim 10^{-2}$. Figure 2(b) shows several disconnected domains at scale of $R_2 \sim 10^1$. The domains in Figure 2(a) locate around $z = 0$ point at Figure 2(b). Figure 2(c) shows several disconnected domains at scale of $R_2 \sim 10^3$. Figure 2(b) is actually part of leftmost domain on the Figure 2(c).

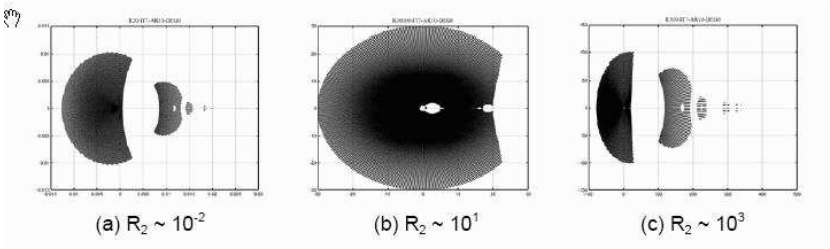


Figure 2: $Z^{-1}C_1C_2$ domains

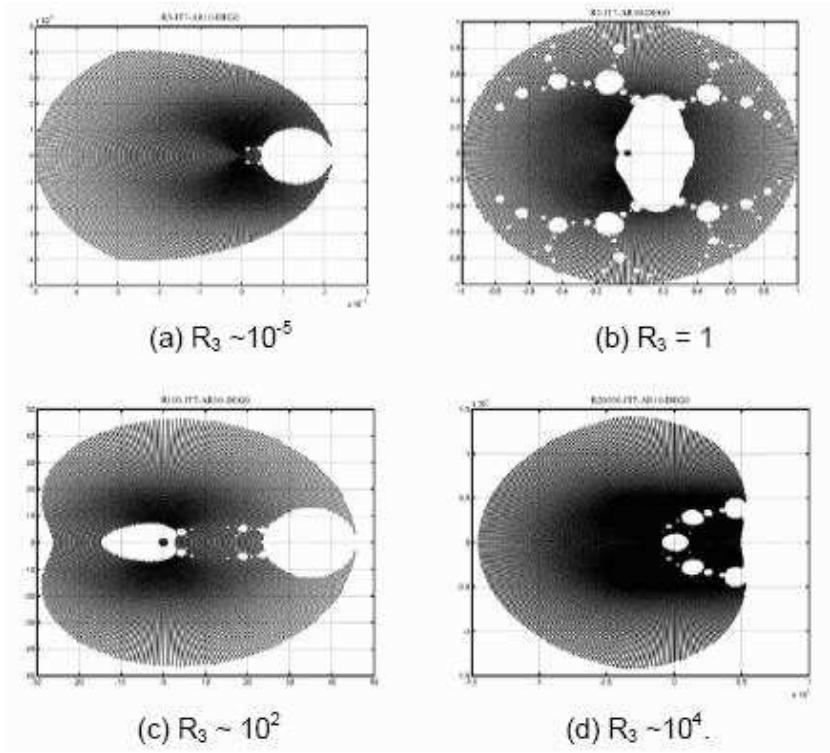


Figure 3: $Z^{-1}C_1C_2C_3$ domains

Figure 3 shows the $Z^{-1}C_1C_2C_3$ domains. Figure 3 (a), (b), (c) and (d) show a domain $z = x + yi$, with $(x^2 + y^2)^{1/2} \leq R_3$. The remained domain forms a set of 4-level structure of partial Herman-Ring fractal domains based on three different sets of parameter R_3, $\{a\}$ and iteration number. Figure 3(a) shows a partial Herman-Ring domain at scale of $R_3 \sim 10^{-5}$. Figure 3(b) shows domain a partial Herman-Ring domain at scale of $R_3 = 1$, the unit circle. The domain in Figure 3(a) can be seen locating around $z = 0$ point at Figure 3(b). Figure 3(c) shows a partial Herman-Ring domain at scale of $R_3 \sim 10^2$. Figure 3(b) can be seen locating around $z = 0$ point at Figure 3(c). Figure 3(d) shows a partial Herman-Ring domain at scale of $R_3 \sim 10^4$. Figure 3(c) is locating around $z = 0$ point at Figure 3(d). We have explored whether or not that there exists structures beyond the found limit scales, namely, $R_3 \sim 10^{-5}$ and $R_3 \sim 10^4$ similarly to the $Z^{-1}C$ and $Z^{-1}C_1C_2$ cases. For the lower bound beyond $R_3 \sim 10^{-5}$, the result is straightforward since there is no hole existing for accommodating another level of Herman Rings in smaller scale. For the upper bound, we have elaborated up to 10^{100} range and could not find other levels. Based on symmetry of hierarchical structures, we conclude that there are only four levels existing. This effort was also applied to the domains of higher order functions.

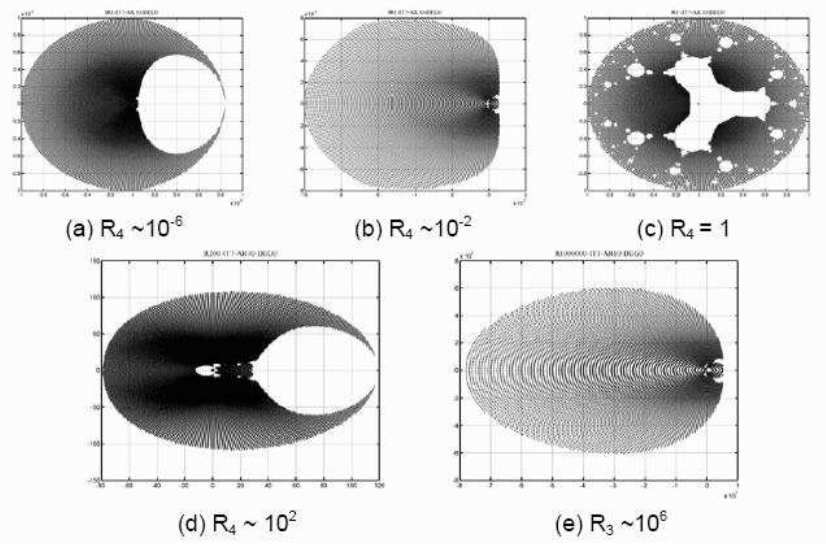


Figure 4: $Z^{-1}C_1C_2C_3C_4$ domains

Figure 4 shows the $Z^{-1}C_1C_2C_3C_4$ domains. Figure 4 (a), (b), (c), (d) and (e) show a domain $z = x + yi$, with $(x^2 + y^2)^{1/2} \leq R_4$. The remained domain forms a set of 5-level structure of partial Herman-Ring fractal domains based on three different sets of parameter R_4, $\{a\}$ and iteration number. Figure 4(a) shows a partial Herman-Ring domain at scale of $R_4 \sim 10^{-6}$. Figure 4(b) shows domain a partial Herman-Ring domain at scale of $R_4 \sim 10^{-2}$. Figure 4(c) shows

a partial Herman-Ring domain at scale of $R_4 = 1$ the unit circle. The domain in Figure 4(b) can be seen locating around $z = 0$ point at Figure 4(c). Figure 4(d) shows a partial Herman-Ring domain at scale of $R_4 \sim 10^2$. Figure 4(c) can be seen locating around $z = 0$ point at Figure 4(d). Figure 4(e) shows a partial Herman-Ring domain at scale of $R_4 \sim 10^6$. Figure 4(d) is locating around $z = 0$ point at Figure 4(e). The partial Herman-Ring domain at scale of $R_4 \sim 10^{-6}$ is topologically similar to that at $R_4 \sim 10^2$ and the domain at scale of $R_4 \sim 10^{-2}$ is topologically similar to that at $R_4 \sim 10^6$. We call this phenomenon as skip-level symmetry.

When we explore higher order than $f = Z^{-1}C_1C_2C_3C_4$, we found that the number of levels keeps the same as those of $Z^{-1}C_1C_2C_3C_4$, namely, only 5 levels are found at 11th order or 20th order for example. The skip-level symmetry is also observed in the higher-ordered domains. Figure 5 shows the unit-circle fractals at different orders.

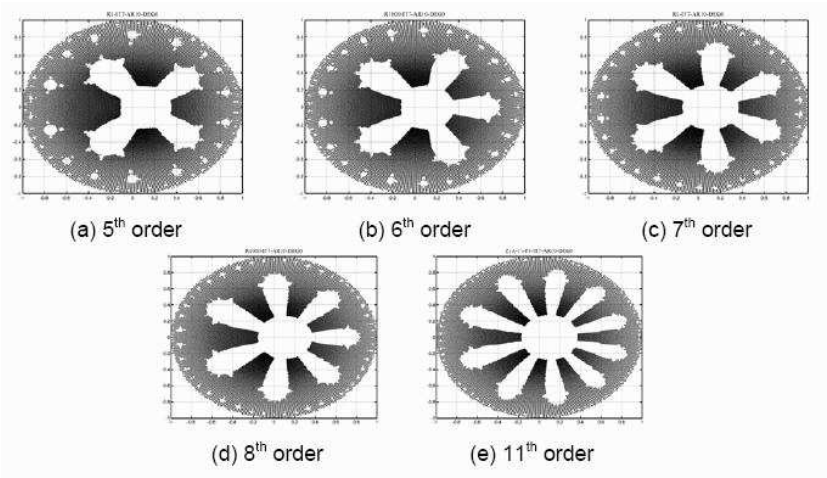


Figure 5: Unit circle domains at different functional order

The influence of the normalized phase item, $exp(g_i(z))$, can be shown in Figure 6. Figure 6(a) shows fractals with three branches of fractals propagating from $z = 0$ to the boundary of unit circle. When the phase factor included in the transformation, we observed symmetry broken as the rotation angles change from 0 degree to 60 degrees and 75 degrees. One of three branches grows and dominates. When the rotation angle reaches at 90 degrees, the fractal becomes chaotic [9].

Table 1 shows that the levels and corresponding scales of the domains subjected to different functional orders.

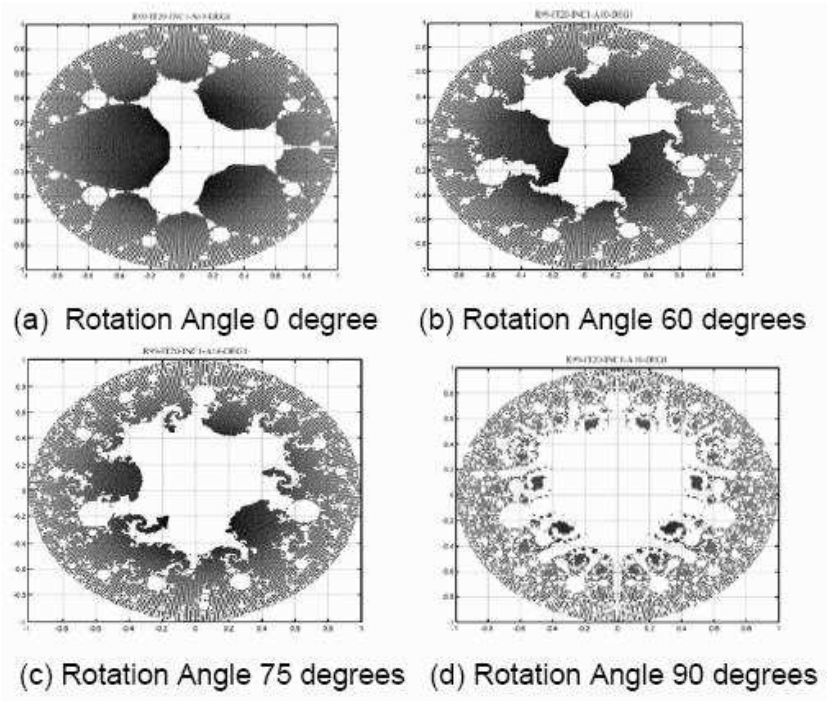


Figure 6: $Z^{-1}C_1C_2C_3C_4$ Unit-circlr domains under different rotation angles

$Z^{-1}\prod C_i$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
Levels	2	3	4	5	5	5	5	5
Scale 1	10^1	10^{-2}	10^{-5}	10^{-6}	10^{-7}	10^{-9}	10^{-10}	10^{-11}
Scale 2	10^1	10^1	1	10^{-2}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Scale 3		10^3	10^2	10^{-2}	1	1	1	1
Scale 4			10^4	10^2	10^3	10^3	10^3	10^4
Scale 5				10^6	10^7	10^9	10^9	10^{11}

Table 1: Functional Orders and Domain Scale of leveled structure

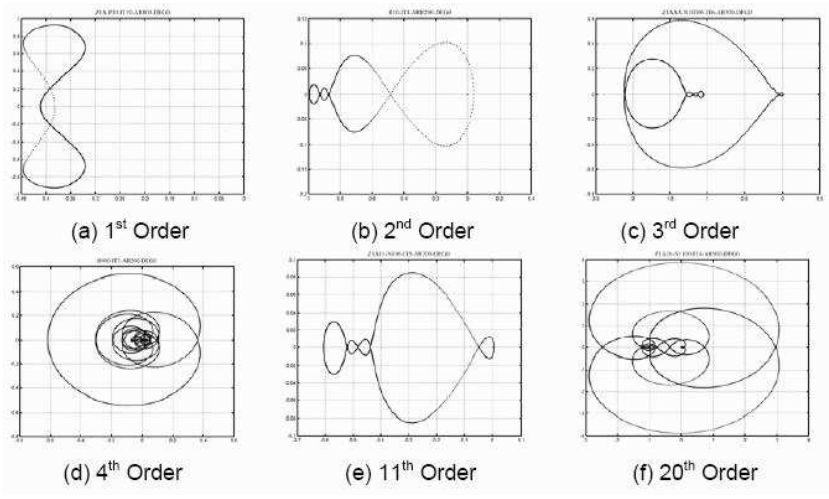


Figure 7: Mapped domains at different orders

4.2 Herman Rings of Mapped Domains

The mapped domains become very simple, such as a point or a portion of algebraic curves depending on they converge, diverge, or oscillate. To simplify the analysis, we adopt a set of points on a circle on complex plane as original domain and observe the transformed domains. Figure 7 shows some mapped domains at different orders.

We have observed that the closed curves as shown in the Figure 7 iterated to chaotic or limit point sets when the parameters are modified. For example, the small increments of parameter a will result the mapped curves rotating around an axis as in Figure 7(b). For the first and second order mapped domains, we have not observed curves similar to 7(d) or 7(f) in the parametric space we explored. Figures 7(d) and 7(f) may be seen as quasi-periodic orbits in the n-body problems [4].

5 Remark

In this paper, we explore both remained domains and mapped domains with the defined parametric space of function $f = h(z)\prod_i exp(gi(z))[(a_i - z)/(1 - \bar{a}_i z)]$. There are some important and interesting observations from our studies:

- There are at most five levels of partial Herman Ring domains at different z scales.
- For the orders, which are equal and higher than four orders, there exist five and only five level domains. In addition, skip-symmetry is observed for these high order domains.

- The unit circle domains preserve circular boundary, while other domains at other scales show partial fractal curves in conjunction with smooth algebraic curves.
- The normalized phase parameter can induce symmetry broken.

These observations are very interesting and may potentially be implacable for explanations of other mathematical fields as well as physical fields.

References

[1] Benoit B. Mandelbrot, The Fractal Geometry of Nature, New York, W.H. Freeman and Company, 1983.

[2] K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, New York, John Wiley & Sons, 1990.

[3] J. Milnor, Dynamics in One Complex Variable, Vieweg, 2000.

[4] A. Fathi and J. -C. Yoccoz (Eds.), Dynamical Systems: Michael Herman Memorial Volume, Cambridge University Press, Feb. 2006.

[5] David C. Ni and C. H. Chin, $Z^{-1}C_1C_2C_3C_4$ System and Application, TIENCS workshop, Singapore, August 1-5, 2006.

[6] David C. L. Ni and C. H. Chin, A Novel Approach for Designing Fractal Antennas, Proc. of International Conference on Wireless Information Networks and Systems, pp. 157-160, (Eds. S.O. Mohammand et al.), Barcelona, Spain, July 2007.

[7] David C. Ni and C. H. Chin, Broadband Fractal Antennas, Proc. of International Symposium on Antennas and Propagation, Taipei, Taiwan, pp. 68 (Abstract), October 27-30, 2008

[8] C. H. Chin,Hypertranscendental Geometry, Preprint NCTU-IMS-0801, Taiwan, R.O.C.

[9] David C. Ni and C. H. Chin, Symmetry Broken in Low Dimensional N-body Chaos, accepted for publication in the Chaos 2009 conference, Chania, Crete, Greece, June 1-5, 2009

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On the Construction of an Invariant Measure of a Symbolic Image of a Dynamical System

E.I. Petrenko
Math. and Mech. Faculty, St.Petersburg State University
Russia
Email: zhene@mail.ru

Abstract

For a dynamical system we suppose to construct an approximation to its invariant measure using the technique of symbolic image. Given a mapping f and a fixed covering of the phase space the symbolic image G is defined as an oriented graph, being the vertices correspond to the cells of the covering and edges mark the existence of nonempty intersections of the covering cells with their images. We construct an invariant measure (a stationary process) on the set of edges of G. It was shown in [13] that applying the subdivision process we can construct a sequence of invariant measures that converges to an invariant measure of the dynamical system f. We consider two computational techniques.

First way is based on the prime cycles enumeration, being the invariant measure on the graph is defined as a convex linear combination of prime cycles measures.

The second way of the construction of an invariant measure allows assigning value to all edges. We apply a linear programming technique based on a method of the sequential balance of the vertices measures.

The results of numerical experiments are given.

Keywords: dynamical system; symbolic image; entropy; stationary process; linear programming.

1 Introduction

This paper is dedicated to an elaboration of numerical methods of the construction of an invariant measure of a dynamical system. We use the notion of symbolic image that was introduced by G.S. Osipenko [8] and became one of main tools for the investigation of dynamical system by symbolic analysis methods. The advantage of such a method is that many problems (localization of periodic orbits and invariant sets, estimation of Lyapunov exponents, estimation of topological entropy) may be solved using well known algorithms for directed graphs.

ISSN-1584-5990

The algorithm of the construction of invariant measure using prime cycles was designed in [1]. In a prime cycle with l edges the value $1/l$ is assigned to every edge. A coefficient (weight) is designated to every prime cycle, being the sum of weights equals to one. The measure of the edge belonging to more than one cycle is defined as the sum of the measures which the edge has in every cycle. If an edge does not belong to any cycle, its measure is zero. The measure of a vertex is the sum of measures of outcoming (or incoming) edges. This method, while clear, has an evident disadvantage: the number of prime cycles may be very significant and the algorithm becomes time-consuming. An optimization may lead to cycles missing. Hence, the measure is not assigned to all edges of the graph.

The second method is aimed at the construction of an invariant measure, such that to assign a value to every edge. To solve the problem we apply a linear programming technique. It allows us to construct a stationary process on the graph (with a given accuracy), using a method of the sequential balance of the vertices measures. L.M. Bregman proved the convergence of the method in [2]. The entropy computed with regard to the measure is an estimation of the stationary process entropy. Numerical experiments show that this value less than the entropy of corresponding topological Markov chain.

The paper is organized as follows: in the next section the definitions of symbolic image and stationary process on a graph are given. Section 3 describes the algorithms of the construction of invariant measure. Finally, in sections 4 and 5 we give the data of numerical experiments and summarize our results.

2 Main definitions

Let ϕ be a discrete dynamical system generated by a homeomorphism f on a compact $M \in R^n$. Symbolic image of a dynamical system f [8] is an oriented graph G, constructed in accordance with a covering $\{M_i\}, i = 1, \ldots, k$ of M by closed sets, being vertices correspond to the covering cells and the existence of the edge (i, j) means that $f(M_i) \cap M_j \neq \emptyset$. The symbolic image is a finite approximation of the system f.

It depends on the covering and may be specified by the following parameters: d — diameter of the covering (the largest of diameters of M_i); q — upper bound of the symbolic image (the largest of diameters of $f(M_i)$); r — lower bound of the symbolic image, which is the minimum of the distances between $f(M_i)$ and M_j, if $f(M_i) \cap M_j = \emptyset$. Being a relationship between the parameters and an value ε is given, there is a correspondence between the ε-orbits of the system and paths on G [11]. The process of sequential subdivision of the set M results in obtaining sequence of symbolic images.

Let for a Markov chain on $G = (V, E)$ a vector $\mathbf{p}$ and a matrix P be defined, i.e a measure μ is assigned to vertices and edges, such that $p_I = \mu(I), I \in V$ and $\sum\limits_I P_{IJ} = 1$.

Definition 1 *[7] The Markov chain is said to be stationary if the equality holds:*

$$\mathbf{p}P = \mathbf{p} \tag{1}$$

The stationarity means that for any vertex the sum of measures of incoming edges equals the sum of measures of outcoming ones.

It should be noted that a stationary process on $G = (V, E)$ may be defined as a stationary process on the set $A = E$ of edges of G, being its support is in the set of the admissible paths on G. Hence a stationary Markov chain on a graph G is a stationary process on G, and the stationarity means the invariance of μ with regard to the shift map σ.

3 Construction of an invariant measure μ

3.1 Prime cycles approach

Let $p = (e_1, e_2, \ldots, e_k)$ be a prime cycle on symbolic images. One may construct the following invariant measure on p

$$\mu_p(e) = \frac{1}{k}, \quad e \in p. \tag{2}$$

The measure of the edge belonging to more than one cycle is defined as the sum of the measures which the edge has in every cycle. If an edge does not belong to any cycle, its measure is zero. The measure of a vertex is the sum of measures of outcoming (or incoming) edges.

We propose approximating algorithm with complexity of $O(m \times n)$ where m — the number of nodes and n — the number of edges.

Let $G = (V, E)$ be an oriented graph. We refer to $e = (i \to j)$ as the edge from i to j.

Definition 2 *Node c is said to be **ancestor** of a node p if and only if p is the parent of c or if the parent node of c is predecessor of p.*

The algorithm uses BFS [4] graph traversal algorithm for each strongly connected component of a symbolic image. We construct search tree having edge $p \to c$ only if BFS search has found node c in the list of the nodes accessible from p. It is only necessary to store a reference to the parent node for every node.

A node of the tree is represented by a tuple of (node, pointer to the parent node in the BFS search tree). We denote the operation of taking reference to a tuple as ref, being *null* is empty pointer. We create a queue q of such pairs.

3.1.1 Algorithm of the enumeration of prime cycles

1. Select a node i.
2. Add a new pair $(i, null)$ to the queue q.

3. While q is not empty pop a pair (n, p) from it.
4. Skip this pair if n was visited by the algorithm. Return to the step 3.
5. For every edge $n \to m$ do
   (a) If m was visited and m is ancestor of n then a cycle is found. Form a list of the nodes of the cycle checking parent references.
   (b) If m is not ancestor of n add the pair $(m, ref(n, p))$ to q.

   Return to the step 3.

We use a bit flag for every node to check weather it was visited by the algorithm. Every node is visited no more than once.

Note, that this algorithm may skip some prime cycles. To workaround this issue we repeat this algorithm for every node in the graph that has not yet been found as a part a prime cycle. We limit search depth to increase performance.

3.1.2 Construction of the invariant measure using prime cycles

Definition 3 *[5] An **iterator** is an object that allows a programmer to traverse through all elements of a collection, regardless of its specific implementation.*

We use an iterator of prime cycles to put aside the method of finding prime cycles. The algorithm processes cycles as the iterator finds them. i.e. the measure is assigned to the next found cycle.

- At the beginning we set a measure for all edges to be 0. The algorithm sets the weights of the edges using a function $g : \mathbb{N}_0 \to \mathbb{R}$.
- Take a cycle l from the iterator. Do the following:
  - compute $z = \frac{g(|l|)}{|l|}$;
  - for all edges e from the cycle l set $\mu(e) := \mu(e) + z$.
- Norm the measure μ. Compute $w := \sum_{e \in E} \mu(e)$. For each edge e set $\mu(e) := \frac{\mu(e)}{w}$.

For g we tried several functions:

1. $g(l) = 1$. It is the case of equal weights.
2. $g(l) = l$. We assign greater weight to greater cycle.
3. $g(l) = l^2$. As in previous case.

3.2 Linear programming approach

Assign probabilities to all edges of the graph G arbitrary. Denote by $P = \{p_{ij}\}, i, j = 1, \ldots, m$, the matrix formed by these values. Our goal is to transform P in such a way to obtain a stationary process on G.

This problem may be formulated as a part of the following linear programming task.

Maximize the function $\sum\limits_{i,j} x_{ij} \ln \frac{p_{ij}}{x_{ij}}$ on conditions

$$\begin{array}{l} \sum\limits_{j=1}^{m} x_{ij} = a_i, \quad \sum\limits_{i=1}^{m} x_{ij} = b_j, \quad x_{ij} \geq 0; \\ \sum\limits_{i=1}^{m} a_i = \sum\limits_{j=1}^{m} b_j; \quad a_i, b_j > 0; \quad p_{ij} \geq 0, \quad \sum\limits_{i,j} x_{ij} = 1. \end{array} \tag{3}$$

Our problem may be considered as a particular case when $a_i = b_i, i = 1, \ldots, n$.

A method of approximative solution of (3) based on the successive balance of rows and columns of P was proposed by G.V.Sheleihovsky, its convergence was proved by L.M.Bregman [2].

The formula of transformation for x_{ij} has the form

$$x_{ij}^{l+1} = x_{ij}^{l} \sqrt{\frac{in(i)}{out(i)}},$$

for i^{th} row and

$$x_{ki}^{l+1} = x_{ki}^{l} \sqrt{\frac{out(i)}{in(i)}}$$

for i^{th} column, being $in(i)$ and $out(i)$ are sums of elements in column i and row i respectively. It should be noted that diagonal elements are not changed. The convergence of the algorithm was proved in [13].

Algorithm Let $i \in V$ and

$$beg(i) = \{e \in E, e = (i, j), j \in V\},$$

$$end(i) = \{e \in E, e = (j, i), j \in V\}.$$

- Assign measures to all edges of G. As the normalization step may be fulfilled at the end of operating period, we assume $\mu(e) = 1$, $\forall e \in E$.
- For each vertex i calculate its balans

$$q(i) = |\sum_{e \in beg(i)} \mu(e) - \sum_{e \in end(i)} \mu(e)|.$$

Construct the queue Q of the vertices of G, being a vertex i with the maximum $q(i)$ has the maximum priority. So, we assign the greatest priority to the most unbalanced vertex.

- In the cycle: select the next vertex i from Q.
  - If $q(i) < \varepsilon$, then complete the processing of i and go out from the cycle. (In view of the structure of Q such an inequality holds for all remaining elements.)
  - Else calculate
    - $out(i) = \sum_{e \in beg(i)} \mu(e)$
    - $in(i) = \sum_{e \in end(i)} \mu(e)$
    - $\forall e \in end(i)$ set $\mu(e) := \mu(e)\sqrt{\frac{out(i)}{in(i)}}$.
    - $\forall e \in beg(i)$ set $\mu(e) := \mu(e)\sqrt{\frac{in(i)}{out(i)}}$.
    - If some of values $out(i), in(i), \sqrt{\frac{out(i)}{in(i)}}$ is too large, we fulfill the normalization.
- Fulfill the normalization. The algorithm is completed.

To provide the efficiency of the second algorithm we have to save both forth and back directions of the edges, which results in the representation of the graph with using two hash-tables. Priority queue has been implemented using Fibonacci trees [4].

Example 1 *Construct an invariant measure of a symbolic image for Henon map [6].*

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{pmatrix} 1 - ax^2 + by \\ x \end{pmatrix}, \tag{4}$$

for $a = 1.4$, $b = 0.3$. *We consider area* $D = [-10, 10] \times [-10, 10]$ *and use linear and point methods [15] to construct a symbolic image. The initial partition consists from 9 cells. On each step every cell is subdivided in 4 cells (on 2 for each of two axis). We performed 10 steps of construction of symbolic image. The last step was performed with point method [15]. Cell size is about* 0.0065×0.0065. *We received about 5 000 nodes in symbolic image. The chain recurrent set is shown on pic. 1.*

Example 2 *We computed an invariant measure for Ikeda map [9].*

$$\begin{array}{rcl} \tau(x,y) & = & C_1 - \frac{C_3}{1+x^2+y^2} \\ \begin{pmatrix} x \\ y \end{pmatrix} & \rightarrow & \begin{pmatrix} d - C_2(x\cos\tau(x,y) - y\sin\tau(x,y)) \\ C_2(x\sin\tau(x,y) + y\cos\tau(x,y)) \end{pmatrix}, \end{array} \tag{5}$$

where $d = 2$, $C_1 = 0.4$, $C_2 = 0.9$, $C_3 = 6$.

We consider area $D = [-10, 10] \times [-10, 10]$ *and use linear and point methods [15] to construct a symbolic image. The initial partition consists from 9 cells. On each step every cell is subdivided in 4 cells (on 2 for each of two axis).*

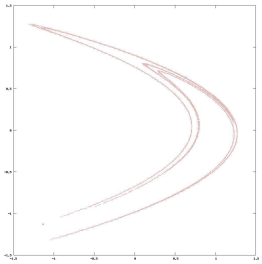
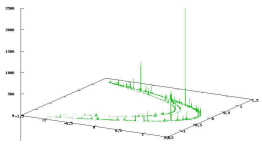
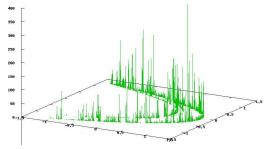


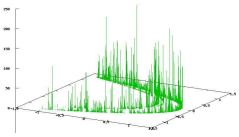
Figure 1: Henon chain-recurrent set



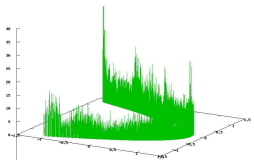
Method 1. $g(x) = 1$



Method 1. $g(x) = x$



Method 1. $g(x) = x^2$



Method 2

Figure 2: Henon map. Invariant measure density.

We performed 10 steps of construction of symbolic image. The last step was performed with point method [15]. Cell size was about 0.0065×0.0065. *We received about 104 799 nodes in symbolic image. The chain recurrent set is shown on pic. 3.*

We visualize a density of the invariant measure with segments drawn on axis Z on pic. 4.

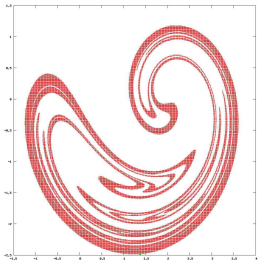
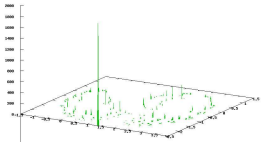
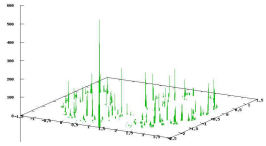


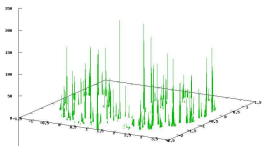
Figure 3: Ikeda map chain-recurrent set



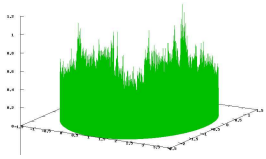
Method 1. $g(x) = 1$



Method 1. $g(x) = x$



Method 1. $g(x) = x^2$



Method 2

Figure 4: Ikeda map. The density of the invariant measure.

4 Conlusion

In this paper two approaches to the construction of an invariant measure of a symbolic image are described. Constructing a sequence of symbolic images and a sequence of unvariant measures we obtain an approximation to an invariant measure of an initial dynamical system.

Acknowledgements

Authors are greatly indebted to J.V.Romanovsky, N.B.Ampilova and G.S. Osipenko for suggesting the problem and for many stimulating conversations.

References

[1] Ampilova N. B., *On the construction of an invariant measure of a symbolic image*, Int.Congress Nonlinear dynamical analysis-2007, June 4-8 2007, (St.Petersburg, Russia) 359.

[2] Bregman L.M., *The proof of convergence of the Sheleihovsky method for the problem with transport constraints (in Russian)*, Journal of computational mathematics and mathematical physics, **7(1)** (1967), pp. 147–156.

[3] Bregman L.M., *Relaxation method for obtaining of the common point of convex sets and its application to the solving of convex programming tasks*, Journal of computational mathematics and mathematical physics, **7(3)** (1967), pp. 620–631.

[4] Cormen T., Leiserson C., Rivest R., *Introduction to algorithms* (in Russian), M., 2001.

[5] Gamma E., Helm R., Johnson R., Vlissides J. M., *Design Patterns: Elements of Reusable Object-Oriented Software (Addison-Wesley Professional Computing Series*, Addison-Wesley Professional, 1994. P. 416.

[6] Henon M., *A two-dimensional mapping with a strange attractor*, Comm. Math.Phys., **50** (1976), pp. 69–77.

[7] Lind D., Marcus B., *An introduction to symbolic dynamics and coding*, New York, 1995.

[8] Osipenko G.S., *On symbolic image of a dynamical system*(in Russian) In Boundary problems, Perm, 1983, pp. 101–105.

[9] Osipenko G., *Numerical Explorations of the Ikeda mapping dynamics*, E-Journal Differential equations and control processes, http://www.neva.ru/journal 2, 2004.

[10] Osipenko G.S., Ampilova N.B., *Introduction to symbolic analysis of dynamical systems (in Russian)*, SPbGU, 2005.

[11] Osipenko G.S., *Dynamical Systems, Graphs, and Algorithms*, Lect.Notes in Math., 1889, Springer, 2007.

[12] Osipenko G.S., Krupin A.V., Bezruchko A.A., Petrenko E.I., Kapitanov A.Y., *On a construction of invariant measures of dynamical systems*, E-Journal Differential equations and control processes, http://www.neva.ru/journal 4, 2007.

[13] Osipenko G.S., *On the problem of approximations of measures of dynamical systems*, E-Journal Differential equations and control processes, http://www.neva.ru/journal 2, 2008.

[14] Petersen, K., *Ergodic Theory*, Cambridge Univ.Press., Cambridge, 1989.

[15] Petrenko, E.I., *Design and implementation of the algorithms of construction of a symbolic image*, E-Journal Differential equations and control processes, http://www.neva.ru/journal 3, 2006.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Evolutionary Differential Based on Poincaré Section

Ali Sanayei
Department of Electrical Engineering,
Sahand University of Technology, Tabriz, Iran
sanayeiali8@gmail.com

Abstract

Most mathematicians believe that the primary target in differential equations is to study the transformation of the median universe and nothing except transformation is constant.[1] Has this target practically resulted? Is classical differential equations (raised by Newton and Leibniz) fully efficient to recognize and review the natural systems?

"Differential" means comparative transformation, comparative transformation means motion, and motion means evolution. Evolution means creation of information to get more adaptable. Classical differential in not defined in discrete, whereas the creation of information occurs in discrete. Whatever is placed in nature's essence is uncertainty. Uncertainty is an attribute of information[2,3], not matter and energy.

For reviewing the natural systems in this paper, evolutionary differential based on Poincaré section has been replaced instead of classical differential and we have stated some practical examples with their results.

Keywords: System, holism, self-organization, information, interaction, evolution, differential, iterated maps, fractal dimension, uncertainty, Poincaré section, A.S. method, A.S. diagram.

1 Introduction

In this paper, after expressing the theoretical and philosophical foundations of the system theory, we will explain the differences of natural systems and man-made systems. After that, we will expand dynamics of systems based on system theory, evolution, and recursive equations. Then, we will define Poincaré section technique and its differences with Poincaré map. Finally, we will review the dynamic of a famous system based on the evolutionary differential and a new method.

ISSN-1584-5990

2 Theoretical and philosophical foundations of systems

Developments of science in the last decades have caused lots of doubts about mechanical paradigm. The new viewpoints and theories, such as "Holomovement" theory raised by David Bohm and Karl Pribram, "Non-equilibrium thermodynamics" raised by Ilya Prigogine,"Cybernetics" raised by Norbert Wiener and "Chaos" theory, perceived the Newtonian viewpoint which is based on reductionism is not able to analyze the nature. Also the famous paper that was published by Einstein, Podolsky and Rosen in 1935 [4], shook the pillars of quantum mechanics.

Reductionism viewpoint claims for the cognition of any systems (natural or unnatural), you can reduce them to their components. You can recognize the properties of a system with the sum of the properties of its components. From the aspect of reductionism, "system" is defined as: a set of some components that they work together to achieve a goal (or goals).[5] Now, I clear the deficiency of that viewpoint, with an example. Suppose if sodium(Na) which is a violently reactive soft metal and chlorine(Cl), which is a poisonous gas, will be combined together, the result will be the kitchen salt(NaCl). This material isn't poisonous and it is quite eatable. It means NaCl has some new properties, that you cannot find them only in Na or Cl themselves. How can reductionism viewpoint explain the new properties of NaCl? They have been created! Now, assume that the human body is a system. We can say, it consists from some subsystems, such as head, body, bones, muscles, stomach, blood, nerves, cells, etc. Is reductionism viewpoint able to explain a book summarization procedure based on the sum of the components properties? If it claims that it can do that, and if one of the components will be disappeared or missed, some errors will be created in the summarization process. Whereas we have seen disable people without legs, that they can summarize a book. Because the goal of a system (in our case: summarizing) exists in the entire of the system (similar to a field).

The logic that Newtonian viewpoint has sat on it, is very simple. It was explained by Descartes that: to understand any complex phenomenon, you need to take it apart, i.e. reduce it to its individual components. If these are still complex, you need to take further steps for your analysis, and look at their components.[6,7] In fact, the foundation of Newtonian ontology is matter, and you cannot see other things, such as mind, life, evolution and organization.

It is true that we can use "superposition principle" to review some of the properties of unnatural systems. For example, the weight of a satellite is approximately the sum of its components wight. But the reductionism viewpoint is not able to characterize its goal(s) yet. Therefore, we need to choose a new viewpoint. I name it "system theory based on Holism". Based on that, the "system" is defined as: a whole that is *created* by interacting of its components. It means the whole is much more than its components. Therefore, the information is a new dimension that it manifests in matter and energy. In other words, information has a non physical essence that it exists in the whole of configuration

of the systems. Thus, you cannot determine clear places for it.

At the last explanation in this section, we can say the natural systems are *self-organized*. It means, they have functional structures which appear and maintain spontaneously. Therefore, they are robust and flexible, and will adapt their organization to any changes in the environmental changes and learn new tricks to cope with unforeseen problems.[8,9]

3 Dynamical systems based on recursive differential and evolution

The inputs of a natural system are *fuzzy*. An important question is that: how are its outputs stabilized? We can answer to that question based on global interactions of natural system components and self-organizing. A natural system according to self-organizing can distinguish stable and instable situations and choose the stable situation for adapting itself against environmental changes. Nevertheless, in stable situations, the system is dynamic (*far from the equilibrium*). Therefore, one of the principal principles of the system theory is "selective retention" [10]. Thus, the instable poles of a system in transformation function are very important to choose stabilized goal(s). I name this process "Evolution", and it is the answer of the above-mentioned question.

Now, it's time to define evolutionary differential. According to our previous discussions, the nature is moving and changing every time, and "differential" means comparative transformation. On the other hand, "evolution" is an essential for the nature. Therefore, our definition should contain evolution concept.

Classical differential is defined as follow:
Suppose that $f(t)$ ('t' indicates time) is a function with respect to t. The differential of $f(t)$ is defined:

$$\frac{df(t)}{dt} := \lim_{\triangle t \to 0} \frac{f(t+\triangle t) - f(t)}{\triangle t} \tag{1}$$

Eq.(1) tells us: first,

Theorem [14] **:** if $f(t)$ will not be continues in the point $t = t_0$, then $f(t)$ is not derivative in that point.

Second, the transformation of a function relative to a very small temporal interval (i.e. $\Delta t \to 0$). Third, its geometric conception in one point is the tangent in that point, and forth, you can replace a curve by a lot of segments (*going to limit)*[11]. In other words, if a curve will be one-dimensional, you can cover it with some of one-dimensional lines.

But, I have some critiques about that definition. First, suppose, we intend to study the annual comparative change of the animal's population in an ecosystem. It means: $\triangle t = 1$ year. Can Eq.(1) review this natural system? It says $\triangle t$ should move toward zero. Second, why should $f(t)$ be continues? Whereas creation of information and sudden changes occur in discrete points (bifurcation points), and third, fractal dimension tells us, we cannot always cover a line(for

example) with a lot of very short lines. For instance, take straight line and remove its middle third. Remove the middle third of each of the two smaller segments remaining after that, and so on ad infinitum. In the limit, you get a fractal called a *Cantor set*[13].

We can compute that measure by noting that the Mth stage of construction. The length of the line segments remaining is given by:

$$length = 1 - 1(\frac{1}{3}) - 2(\frac{1}{3})^2 - 2^2(\frac{1}{3})^3 - ... - 2^{M-1}(\frac{1}{3})^M \tag{2}$$

In the limit, the amount left is:

$$length = 1 - \frac{1}{3}\sum_{i=0}^{\infty}(\frac{2}{3})^i \tag{3}$$

Eq.(3) is a geometric series whose value is 3. Therefore, we see that the measure (length) of the Cantor set is 0 . Let us now calculate the "box-counting" dimension of this set. The box-counting dimension D_b of a geometric object is determined by the following: Construct "boxes" of side length R to cover the space occupied by the geometric object under construction. For a one-dimensional set, the boxes are actually line segments of length R. We then count the *minimum* number of boxes, $N(R)$, needed to contain all the points of the geometric objects. As we let the size of each box to get smaller, we will need a larger number of the smaller boxes to cover all the points of the object. The box-counting dimension D_b is defined to be the number that satisfies:

$$N(R) = \lim_{R\to 0} kR^{-D_b} \tag{4}$$

where k is a proportionality constant. In practice, we find D_b by taking the logarithm of both sides Eq.(4) (before taking the limit) to find:

$$D_b = \lim_{R\to 0}[-\frac{\log N(R)}{\log R} + \frac{\log k}{\log R}] \tag{5}$$

As R becomes very small, the last term in Eq.(5) goes to 0, and we may define:

$$D_b = -\lim_{R\to 0}\frac{\log N(R)}{\log R} \tag{6}$$

For the Cantor set construction, at the Mth stage of construction, we need a minimum of 2^M boxes with $R = (\frac{1}{3})^M$. If we use those values in Eq.(6), we find(if "$R \to 0$" then "$M \to \infty$"):

$$D_b = -\lim_{M\to\infty}\frac{\log 2^M}{\log(\frac{1}{3})^M} = \frac{\log 2}{\log 3} = 0.63... \tag{7}$$

It is a surprising result. Based on "going to limit" idea, we can always cover a one-dimensional line, with lots of one-dimensional lines segment. But in

Cantor set, we cannot cover the lines with lots of one-dimensional lines segment, because its dimension is nearly 0.63, not 1. It has a fractal dimension. In fact, non integer dimensions indicate the interaction of components of a system (similar to the effects of the sea waves on the beach).

Now, "Evolutionary Differential" (d_e) could be defined as follow:
Suppose $f(t)$ ('t' indicates time) is a function with respect to t, then:

$$\frac{d_e f(t)}{d_e t} := \frac{f(t+\triangle t) - f(t)}{\triangle t} \tag{8}$$

where $\triangle t$ is a real number and unequal 0.
As you see, we deleted the limitation of $\triangle t \to 0$. With that changing, classical differential is changed into a recursive equation (we will expand this just a later). It has a memory and the system can evolve in uncertainty environment.

4 Reviewing a natural system based on evolutionary differential

In this section, we review the Logistic differential equation. It is a mathematical model of biological population growth. The Logistic differential equation is as follow:

$$\dot{x}(t) = \lambda x(t)[1 - x(t)] \tag{9}$$

where λ is a positive constant.
If we solve Eq.(9) with an initial condition $x(t=0) = x_0$, its response could be as follow:

$$x(t) = \frac{x_0 e^{\lambda t}}{x_0 e^{\lambda t} - x_0 + 1} = \frac{x_0}{x_0 - e^{-\lambda t}(x_0 - 1)} \tag{10}$$

For instance, we intend to study the annual growth of an animal species population in an ecosystem. Therefore, we need to know when or how does the population growth go to annihilation?, when or how does it increase?, when or how is it periodical?, what are the bifurcation points?, when or how does it go into the chaotic behavior?, and etc. Can Eq.(10) answer those questions?
Now, we replace the evolutionary differential(Eq.(8)), instead of $\dot{x}(t)$ in Eq.(9). Therefore:

$$\frac{x(t+\triangle t) - x(t)}{\triangle t} = \lambda x(t)[1-x(t)] \Rightarrow x(t+\triangle t) - x(t) = \triangle t.\lambda x(t)[1-x(t)] \tag{11}$$

As we explained, the choosing $\triangle t$ value depends on our selection based on our case. We are not able to do that based on the classical differential. Now, we choose $\triangle t$=1 and replace it in Eq.(11). The result could be as follow:

$$x(t+1) - x(t) = \lambda x(t)[1 - x(t)] \Rightarrow x(t+1) = x(t)[1 + \lambda - \lambda x(t)] \tag{12}$$

If we define: $y(t) = \frac{\lambda x(t)}{1+\lambda}, \xi = 1 + \lambda$, and replace them in Eq.(12), after some algebraic manipulation we obtain:

$$y(t+1) = \xi y(t)[1 - y(t)] \tag{13}$$

The actual calculation runs as follow: start with some value of y_0, compute y_1, then y_2, and so on. It means:

$$y(1) = \xi y(0)[1 - y(0)], y(2) = \xi y(1)[1 - y(1)], ... \tag{14}$$

We call this a sequence of iterations. Therefore, we rewrite Eq.(14) by following to *indicate* that each term is created by the previous term. Thus, we use subscript n:

$$y_{n+1} = \xi y_n(1 - y_n) \equiv \ell_\xi(y) \tag{15}$$

Eq.(15) is called: "Logistic map", and ξ is called: control parameter, because it can control the system. The function $\ell_\xi(y)$ is sometimes called an *iterated map function*, since it maps one value of y, say y_0, in the range $0 \leq y \leq 1$ (basin of attraction) into another value of y, which we call y_1, in the some range if ξ is in the range $1 \leq \xi \leq 4$. A y value, call it $\tilde{y}$, which gives

$$\tilde{y}_\xi = \ell_\xi(\tilde{y}_\xi) \tag{16}$$

is called a *fixed point* of the iterated map. (The subscript ξ indicates $\tilde{y}$ depends on the value of ξ.) Based on Eq.(16), the logistic map has two fixed points:

$$\tilde{y}_\xi = 0 \tag{17}$$

$$\tilde{y}_\xi = 1 - \frac{1}{\xi} \tag{18}$$

If $\xi < 1$,then $\tilde{y}_\xi = 0$ is the only fixed point in the range of y that is of interested for our biological model. Therefore, we can say: the population dies out ($y \to 0$) as n increases. See fig.1 .

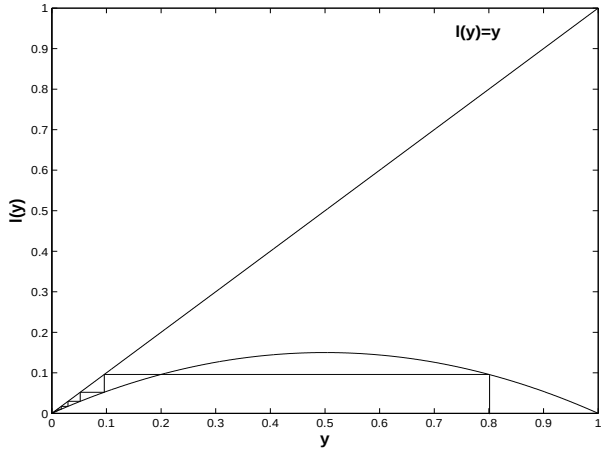


Figure 1: A graphic representation of the Logistic map , $\xi = 0.6$, $y_0 = 0.8$.

Now, we choose a value for ξ that is greater than 1 (for example: ξ=1.5) and we assume that the trajectory evolves from $y = 0.1$. Now heads for the fixed point $\tilde{y}_\xi = 1 - \frac{1}{\xi} = \frac{1}{3}$. In fact, if a trajectory evolves in the range $0 < y \leq 1$, it will be attracted to $y = \frac{1}{3}$ (an attracting or a stable fixed point). When $\xi > 1$, the system tells us: given any initial number y_0 lying between 0 and 1, the population fraction eventually approaches the attracting fixed point $\tilde{y}_\xi = 1 - \frac{1}{\xi}$. In fact, for $\xi > 1$, $\tilde{y}_\xi = 0$ has become a repelling fixed point (unstable), since, trajectories that start near $y = 0$, move away from that value. See fig.2.

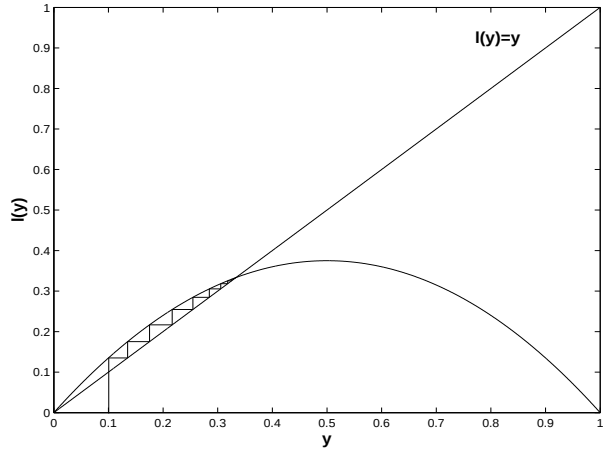


Figure 2: A graphic representation of the Logistic map , $\xi = 1.5$, $y_0 = 0.1$.

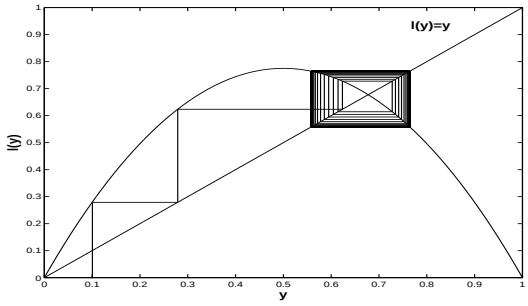


Figure 3: A graphic representation of the Logistic map , $\xi = 3.1$, $y_0 = 0.1$.

Now, If ξ is just greater than 3, we will see that the trajectory doesn't settle to an attractor value. It means, it oscillates between $y = 0.558...$ and $y = 0.764...$. In biological aspect, the population fraction is high in one year, low the next, then high again, then low again, and so on. Since the population

returns to the same value every 2 years, we call this, period-2 behavior. This changing into period-2 behavior is called: *period-doubling bifurcation*. See fig.3.

As we explained, "bifurcation" describes any sudden changes in the system behavior. If we increase ξ in this manner, another period-doubling bifurcation occurs. In fact, the attractor consists of four points. Further increase in ξ, lead to period-8, period-16, and so on.

For ξ just greater than 3.5699... , the trajectory values never seems to repeat. The behavior is *chaotic*. See fig.4 .

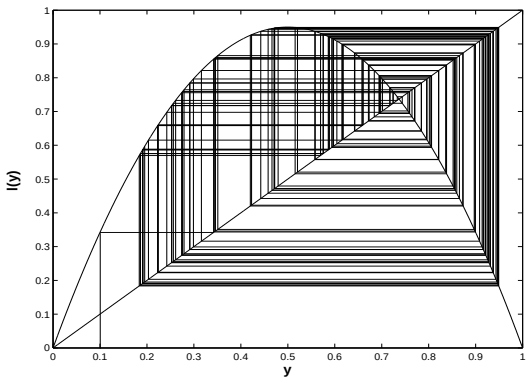


Figure 4: A graphic representation of the Logistic map , $\xi = 3.8$, $y_0 = 0.1$.

We can summarize all the information about the system behavior, and all of figures(1 to 4) in one diagram. It is called the bifurcation diagram. See fig.5 . One of the most important properties of bifurcation diagram is to show

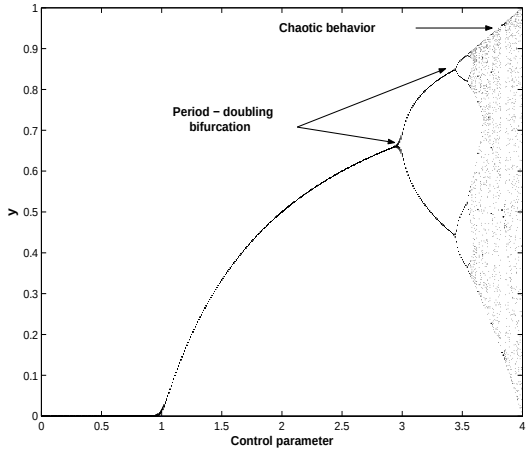


Figure 5: Bifurcation diagram for the Logistic map function .

the communication of chaos and fractal geometry based on the *stretching and folding* in the chaotic mood.

As you see, we could answer all of the questions about Eq.(9) .

5 Poincaré section and Poincaré map

A Poincaré section (named after French mathematician Henri Poincaré) is merely a geometric picture of trajectory's activity at a cross section of the attractor. The cross section is a slice or section though the attractor, transverse (not parallel) to the "flow" or bundle of trajectories. In fact, the Poincaré section is a surface (line or plane) through phase space, transverse to the trajectories.[12]

The main purpose of the Poincaré section is recognizing of the dynamic system in different viewpoints. In crossing of the Poincaré section with a trajectory, there exists some dots. Every dots has a characterized address. They indicate "event" and the system behavior. Based on the dots, we can recognize the real behavior of a system. In other words, the Poincaré section is a transformation from determinism environment into the uncertainty environment that it doesn't disconnect the interaction of the system components.

Poincaré map is a rule, function, graph or model that tells where the trajectory will next cross the Poincaré section, given the location of the previous crossing at the selected Poincaré section. In fact, It is an iterated map. Formally, let $\{\alpha_1, \alpha_2, \alpha_3, ...\}$ be the set of intersections of a trajectory with the surface of Poincaré section. The Poincaré map is a relation $\mathcal{P}$ such that $\mathcal{P}(\alpha_n) = \alpha_{n+1}$.

As we explained, if a Poincaré plane intersects the trajectory, some points will exist. Sometimes we are able to find a recursive equation (Poincaré map). If we find a recursive equation, all of the information about the system will be known. But usually, we aren't able to find the Poincaré map. In this case, we should recognize the system behavior based on geometry of the points. You may ask: how are we able to recognize the system behavior based on some *points*? In reply to the mentioned question, we introduce a method in the next section.

6 A.S. method

The "Ali Sanayei's (A.S.)" method contains two stages:

Stage 1: Changing classical differential into the evolutionary differential (Eq.(8)). Then by using the previous processes (e.g. Eqs.(11-14)), we will own one (or some) iterated map(s).

Stage 2: Plot (one of) the iterated map in order versus the number of iterations. I name it "A.S. diagram".

In section 6, we reviewed four cases about the Logistic map. Now, we review again those cases based on the A.S. method. You can compare the results with the previous results. In Eqs.(11-15), we obtained the Logistic map function (Eq.(15)). Therefore, stage one has been done. For stage two, we consider four subsections:

1) Let $\xi = 0.6 < 1, y_0 = 0.8$. You can see the A.S diagram for this case in fig.6 .

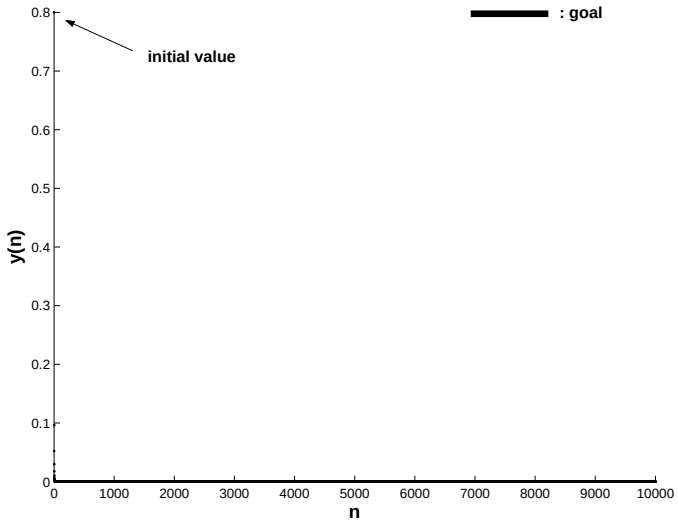


Figure 6: A.S. diagram for 10000 iterations, $\xi = 0.6, y_0 = 0.8$.

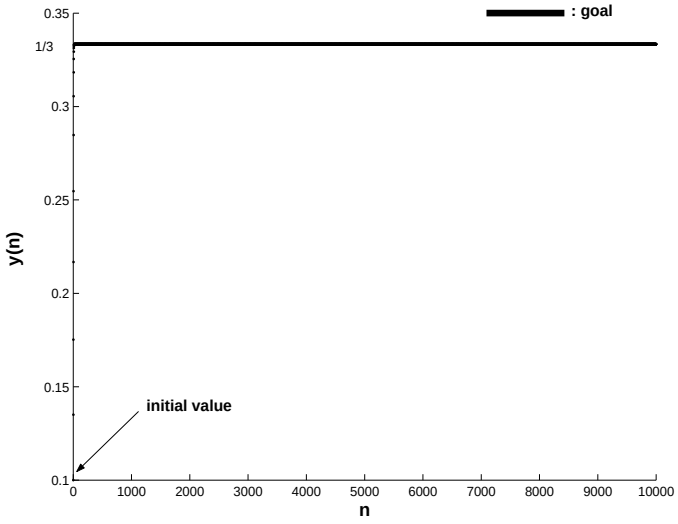


Figure 7: A.S. diagram for 10000 iterations, $\xi = 1.5, y_0 = 0.1$.

As you see, we have only one fixed point and its value is 0 . Therefore, the

trajectory goes toward 0. It means the population dies out as n increases.

2) Let $\xi = 1.5, y_0 = 0.1$. See the A.S. diagram in fig.7 . As you see in fig.7, the fixed point has been changed into $1 - \frac{1}{\xi} = \frac{1}{3}$. Therefore, it attracts the trajectory toward itself.

3) Let $\xi = 3.1, y_0 = 0.1$. See the A.S. diagram in fig.8 . As you see the trajectory oscillates between $y = 0.558...$ and $y = 0.764...$. Therefore, the system behavior is periodic.

4) Let $\xi = 3.8, y_0 = 0.1$. See the A.S. diagram in fig.9 . It shows that we have gone to the chaotic mood. Fig.9 indicates a chaotic mood, not a randomness behavior. You can test it by *divergence of nearby the trajectory* using the Lyapunov exponent approach.

As another example, consider the following model:

$$\begin{cases} \dot{x}(t) = -x^2(t) - x(t) + 0.4y(t) + \mu \\ \dot{y}(t) = x(t) - y(t) \end{cases} \tag{19}$$

where μ is a constant.

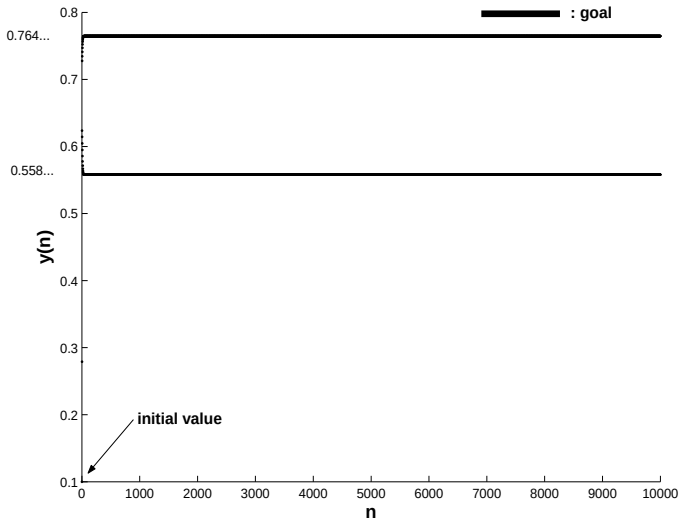


Figure 8: A.S. diagram for 10000 iterations, $\xi = 3.1, y_0 = 0.1$.

Stage 1: We change Eqs.(19) into two iterated maps based on the evolutionary differential (Eq.(8) and $\triangle t = 1$). After some algebraic manipulation we find out that:

$$\begin{cases} x_{n+1} = -x_n^2 + 0.4y_n + \mu \\ y_{n+1} = x_n \end{cases} \tag{20}$$

where μ is a control parameter.

Eqs.(20) is the *Hénon map*.

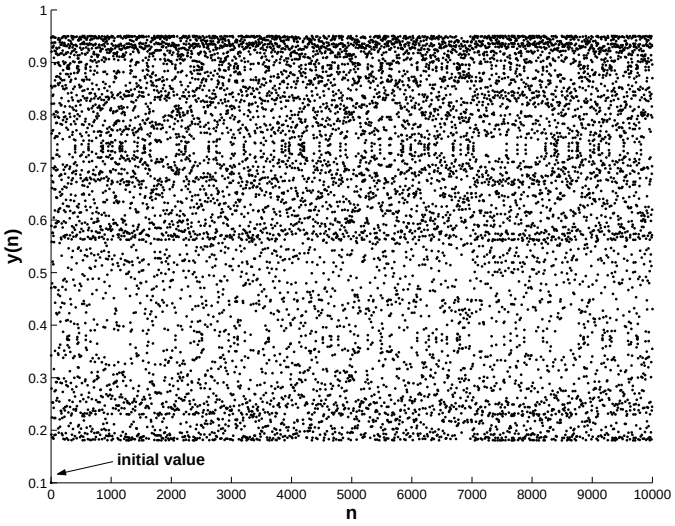


Figure 9: A.S. diagram for 10000 iterations, $\xi = 3.8, y_0 = 0.1$.

Stage 2: For instance, we plot the A.S. diagram for $\mu = 0.98$. The result is fig.10 . As you see, we have three period-doubling bifurcation points.

7 Conclusion

Based on the sections 1-6, we can conclude that "Evolutionary Differential" (Eq.(8)), is able to review the real dynamics of a system in uncertainty environment. Also, if we combine that with Poincaré section, we can obtain an iterated map that it has a memory and control parameter (or some control parameters). Therefore, we are able to recognize that the system dynamics with various dimensions. If we obtain the iterative maps, we can have all the information about the system behavior, otherwise, we are able to recognize and analyze the system behavior based on geometry of the points that appear on the Poincaré section. By A.S. method, we are able to recognize the system behaviors based on geometry of the points. It is very useful in the case that we are not able to find the Poincaré map. Also its simplicity is one of its advantages in comparison with the plotting x_{n+1} versus x_n . In fact, this method is based on the evolutionary differential and geometry of the points. Therefore, we can name it "Evolutionary differential based on Poincaré section".

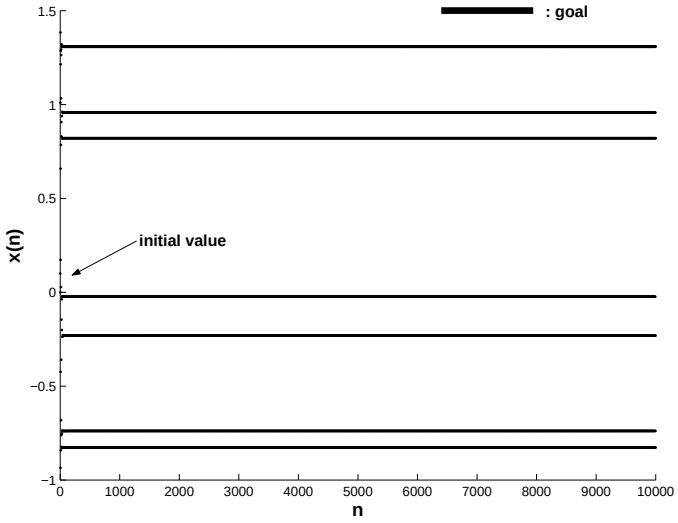


Figure 10: A.S. diagram for 10000 iterations, $\mu = 0.98, x_0 = y_0 = 0.1$.

8 Acknowledgment

I acknowledge Full Professor S.M.R. Hashemi Golpayegani for his recommendations and Miss M. Jalali for grammer edition in this paper.

References

[1] G.F.Simmons, Differential equations with applications and historical notes, McGraw-Hill Inc., 1972.

[2] L.A.Zadeh, Toward a generalized theory of uncertainty(GTU)- an outline, *Information Sciences* 172, Elsevier, pp.1-40, 2005.

[3] L.A.Zadeh, Generalized theory of uncertainty(GTU) - principle concepts and ideas, *Computational Statistics and Data Analysis* 51, Elsevier, pp.15-46, 2006.

[4] A.Einstein, B.Podolsky, N.Rosen, Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?, *Physical Review*, Vol 47, 1935.

[5] K.Ogata, Modern Control Engineering, Prentice-Hall, 1970.

[6] F.Hehlighen, P.Cillers, C.Grenshenson, Complexity and Philosophy, *Radcliffe publishing*, Oxford, 2007.

[7] J.D.Collier, Holism in the New Physics, *Descant* 79/80, pp.135-154, 1993.

[8] F.Heylighen, C.Grenshenson, The Meaning of Self-Organizing in Computing, *IEEE Inteligent Systems*, 18:4, pp.72-75, 2003.

[9] C.Grenshenson, F.Heylighen, When Can we Call a System Self-organizing?, *7th European Conference, ECAL*, 2003.

[10] F.Heylighen, Principles of Systems and Cybernetics: an evolutionary perspective, *World Science*, Singapore, pp.3-10, 1992.

[11] M.Baranger, Chaos, Complexity, and Entropy - A physics talk for non-physicists, MIT, Cambridge.

[12] G.P.Williams, Chaos Theory Tamed, Joseph Henry press, Washington,D.C., 1997.

[13] R.C.Hilborn, Chaos and Nonlinear Dynamics, second edition, Oxford university press, 2002.

[14] G.B. Thomas, R.L. Finney, Calculus and Analytic Geometry, seventh edition, Addison-Wesley, 1988 .

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Invariant Sets of Dynamical Systems — the Computation by Methods of Interval Arithmetic

Sergey Terentev
S.Petersburg State University
Faculty of Mathematics and Mechanics, Russia
E-mail: waterq@yandex.ru

Abstract

In computer simulation of dynamical systems the division of phase space into cells is widely applied. Such a cell is reasonable to interpret as an interval vector in the space of appropriate dimension and to use interval arithmetic. In the present work the algorithm of localization of invariant sets of dynamical systems using a library of interval calculation is described. A number of examples and comparisons with the techniques implemented by conventional(real) arithmetic is given.

Keywords: computer simulation, dynamical systems, numerical methods, symbolic image, interval arithmetic.

1 Introduction

One of well-known investigative techniques of dynamical systems is the construction of its symbolic image [1] — an approximation of the initial system. According to a selected covering of the phase space and a dynamical system an oriented graph is consructed, being the nodes correspond to the coverage cells and edges mean the existence of nonempty intersections of cells and their images. Symbolic image depends on the covering. Subsequent subdivision of the covering cells allows for more accurate approximation of the system phase portrait. Such a construction makes it possible to apply algorithms on graphs to obtain many essential characteristics of a dynamical system, viz. invariant sets and their isolating neighborhoods, Morse spectrum or entropy. In this work , linear and point techniques are dealt with [6]. According to the point technique,for each cell uniformly spread points are selected, and the union of cells containing images of points is assumed to be the cell image. According to the linear technique, the least n-dimensional axially-oriented parallelepiped containing the images of cell nodes serves as the image. It seems to be natural to consider a cell in the phase space as an interval vector in the space of the appropriate dimension, and use interval arithmetic to realize a cell image construction algorithm. In contrast to the conventional arithmetic, such an approach allows for

ISSN-1584-5990

partial elimination the problem of round-off errors, since we always deal with a number's neighborhood, not with a number itself. When interval arithmetic is applied to, the so-called problem of upward bias may arise. As shown in [7], if an interval function satisfies the Lipschitz condition, the computational result in approximate interval arithmetic tends to the computational result in exact interval arithmetic, being machine accuracy tends to infinity. The algorithm of localization of invariant sets of dynamical systems using libraries of interval computations has been designed and implemented.The results of numerical experiments are given. Comparison characteristics of the algorithm and the algorithms described in [6] are given as well.

2 Interval arithmetic

Definition 1 *By interval we mean the set expressed as:*

$$x = [\underline{x}, \overline{x}] = \{\tilde{x} \in R \mid \underline{x} \leq \tilde{x} \leq \overline{x}\}.$$

Symbol IR denotes the space of all intervals over R.

Definition 2 *Define the radius of interval x as follows:*

$$rad(x) = \frac{(\overline{x} - \underline{x})}{2}.$$

Definition 3 *Let $x_1, \ldots, x_m \in IR$. Then we set*

$$\breve{x} = (x_1, \ldots, x_m) = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_m, \overline{x}_m]$$

as interval vector and IR^m denotes the space of all interval vectors.

Definition 4 *The radius of interval vector is defined as follows:*

$$rad(\breve{x}) = (rad(x_1), \ldots, rad(x_m))$$

$$x_1, \ldots, x_m \in IR.$$

Let $a \subset R$ — be an nonempty limited set. Then

$$\square\, a = [\ inf(a)\ ,\ sup(a)\ ]$$

is the hull of a. Put it in other way, "$\square$" is the operation of construction of minimal interval containing the set a.

Definition 5 *The operation "$\square$" can be introduced for the set from R^m. Let $a \subset R^m$ — be a nonempty limited set. Then*

$$\square\, a = [\ inf(a_1)\ ,\ sup(a_1)\ ] \times \ldots \times [\ inf(a_m)\ ,\ sup(a_m)\ ],$$

where a_i — is the projection of a on i axis, $i = 1, \ldots, m$.

A set of function (Φ) (called elementary) which are real functions continuous on each closed interval of their range of definition is introduced. As a rule, initially, $\Phi = \{\ sqr,\ sqrt,\ ln,\ exp,\ sin,\ cos\ \}$ and it can be extended in action. The set of intervals are extended by adding the symbol ω (indeterminedness):

$$R^* = R \cup \{\omega\}\ \ IR^* = IR \cup \{\omega\}\,.$$

The number ω is assumed to be more than any real number. Any function $\varphi \in \Phi$ is extended on interval argument as follows:

$$\breve{\varphi} = \square\,\{\varphi(\tilde{x}) | \tilde{x} \in x\}\,, x \in IR$$

3 Symbolic image of a dynamical system

Let be a compact in R^q. Consider a discrete dynamical system, generated by a homeomorphism $f : M \to M$ $C = \{M_1, \ldots, M_n\}$ — a covering of M with closed sets. The sets $\{M_1, \ldots, M_n\}$are called cells. The oriented graph G is constructed in accordance with M and f as follows: vertex i corresponds to the cell M_i and there is the edge $i \to j$ if $f(M_i) \cap M_j \neq \emptyset$. Such a graph is called the symbolic image of f relating to C [1]. Let us define subcovering $C_i = \{M_j \mid M_j \cap f(M_i) \neq \emptyset\}$ and set $R_i = \bigcup_{j\ \in\ c_i} M_j$.

4 Algorithm of localization of invariant sets of dynamical systems

The main task of the symbolic image construction is the construction of a cell image. This procedure is implemented using the library of interval functions. Symbolic image gives more and more accurate approximation of the system's behavior at subsequent subdivision. That is why the algorithm implies the construction of a sequence of symbolic images. Note that we do not construct a graph according to the algorithm suggested, but work with a set containing cells that participate in the forming of the symbolic image.

Let $f : \breve{M} \to \breve{M}$ be a homeomorphism on compact $\breve{M} \subseteq IR^q$.Let $C_0 = \left\{\breve{M}_1, \ldots, \breve{M}_n\right\}$ be a covering of the compact $\breve{M}$ with cells $\breve{M}_i \subseteq IR^q; i = 1, \ldots, n$. Compute the set $\breve{R}_i$ for each cell $\breve{M}_i$ and construct the set-representation of G corresponding to the symbolic image G of f relating to C_0, to be more exact, $G_{C_0} = \bigcup_{i\in[1,n]} \breve{R}_i$. This set contains all the cells that correspond to the vertices of G. The next step is to apply subdivision to the elements of G_{C_0} and obtain a graph G_1 and its set-representation G_{C_1}. At k-step we consider the subdivision

$$C_k = \left\{\breve{M}_{j1}, \ldots, \breve{M}_{jm} | \breve{M}_{ji} \subset G_{k-1}\right\}$$

and construct the corresponding set-representation G_{C_k}.

4.1 Cell representation

Taking into account geometrical interpretation of interval vectors, it is reasonable to consider the cells of a phase space as interval vectors of an appropriate dimension. The image of an interval vector is an interval vector as well, thus we always deal with n-dimensional parallelepipeds in phase space of dimension n. There is a metric on the space of interval vectors, but in our problem it is more suitable to use the Hausdorf distance between sets. In the algorithm implementation the cells have the same size and integer system of coordinates, where a unit of distance represents cell size. This allows us to present a cell by the coordinate of its left upper corner and thereby to decrease the memory capacity required for data storage.

4.2 Time complexity

Main operations are the calculation of sets R_i, adding them to the set G_{C_k} and the computation of the value of function on the interval vector. The adding operation requires O(m) steps, where m is the cardinality of G_{C_k}. Calculation of R_i may be reduced to the calculation of the function f($\breve{M}_i$) and to the location of intersection of the image obtained with C_0. As the integer system of coordinates is used, the intersection operation is reduced to the comparison of integer numbers. Thus, the calculation of R_i takes O(m1 + m2), where O(m1) is the function calculation time, O(m2) is intersection location time. These operations are performed for each of n coverage cells. As a result, the algorithm operation time is estimated as O((m1 + m2)n) = O(n).

5 Implementation

The algorithm of localization of invariant sets of dynamical systems is implemented in C++ language with using a development tool kit cygwin [10]. Owing to implementation of "mixed computation" support [2], performance of the tool implemented has essentially increased. Vizualizing results are implemented with the aid of GNUPLOT technology [8]. Implementation of interval arithmetic from BOOST library [12] was applied. The tool has been developed for Windows-based platform, it being transportable into Unix-like systems at the source code level. Mixed computations had been implemented by Antlr technology [11] and a development tool kit MinGW [9].

5.1 User Interface

Interaction between the user and the application is implemented by means of command line. The user enters a command specifying a configuration file and receives both the text file with the computation results and GIF-file containing graphical representation of the computational result. The command format is : -I configuration-file-ot output.txt.-oi output.gif. The configuration file contains the system description in the following format:

Dimension = $d \in N$
Decomposition $\{n_1, \ldots, n_d | n_i \in N\}$
Bounds $\{[l_1, u_1], \ldots, [l_d, u_d] | l_i, U_i \in R\}$
Cells Covering = "file-path"
System variables $x_1 = v_1, \ldots, x_d = v_d, v_i \in N$
Approximations $a \in N$
User definitions
$d_1(x_1, \ldots, x_d) =$ expression
...
$d_k(x_1, \ldots, x_d) =$ expression
System
$f_1(x_1, \ldots, x_d) =$ expression
...
$f_d(x_1, \ldots, x_d) =$ expression

Dimension parameter describes the system dimension. Decomposition parameter defines the subdivision of current covering. Bounds parameter specifies the area in the phase space. User definitions parameter specifies user functions to be used in describing the system with a view to simplification of arithmetical functions. Cells covering parameter specifies file name containing the results of computation of covering cells in subsequent subdivision. If file name is not specified, C_0 is used as the covering. System variable specifies axis designation. Approximations parameter specifies the number of approximation. System parameter s pecifies a dynamical system. Symbol # points at the beginning of one-line comment. Symbol { points at the beginning of multi-line comment. Symbol } points at the end of multi-line comment.

5.2 Mixed computations

The term "mixed computations" has been introduced to specify the optimization technique for computer programs. Its principle consists in the following. Let P_g be the program that inputs any data from the data set I and outputs data from the data set O. Let P_l be the program that inputs any data from the data set $I_l, I_l \subset I$ and outputs data from the data set $O_l, O_l \subset O$. When doing this, the programs output similar results on similar input data. The P_l program structure being essentially simpler than the structure of P_g, the data processing rate with P_l program is higher than the same characteristic for P_g program. Assuming that data from I_l are always the input for P_g program, we can use P_l for their processing [2].

When creating REFAL-compiler routine ("super-compiler"), V.F. Turchin and S.A. Romanenko used a similar optimization technique for computer programs: Super-compiler receives the program source code and returns a new one, performing its task more quickly [3].

The present algorithm being implemented, a library containing the means for creating object approximation of the source system in the form of some class instance created dynamically and storing fundamental system properties

(dimension and phase space boundaries, etc.) and allowing for effective manipulations, e.g. efficient operation with the system function. Using such an approach, the dynamical system (DS) description is derived from the configuration file and is the input to the special analyzer [4]. The analyzer criticizes the description and builds a pars tree. Further, correlation between functions and variables included into the system description, and their definitions is made. Semantic pars tree is constructed and in accordance with it a special generator [4] generates a file containing the code of a class that is an abstract model of the source system. After that compiling and feeding the class into the addressing space of the tool run is performed.

6 Examples

6.1 Henon map

$x_1 = 1 - 1.4x^2 + y, \quad x_2 = 0.3y$

$$\breve{M} = [-1.5, 1.5] \times [-5, 5]$$

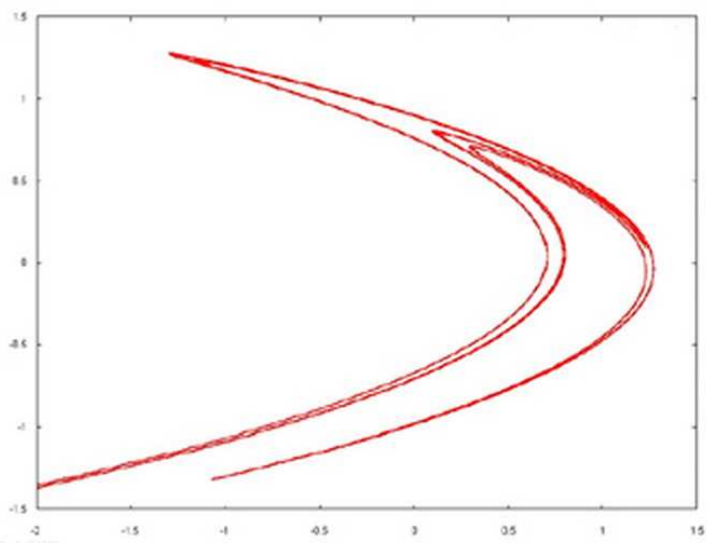


Figure 1: Attractor of Henon map

Technique	Time
Linear	24 047 ms
Point	40 656 ms
Interval	18 000 ms

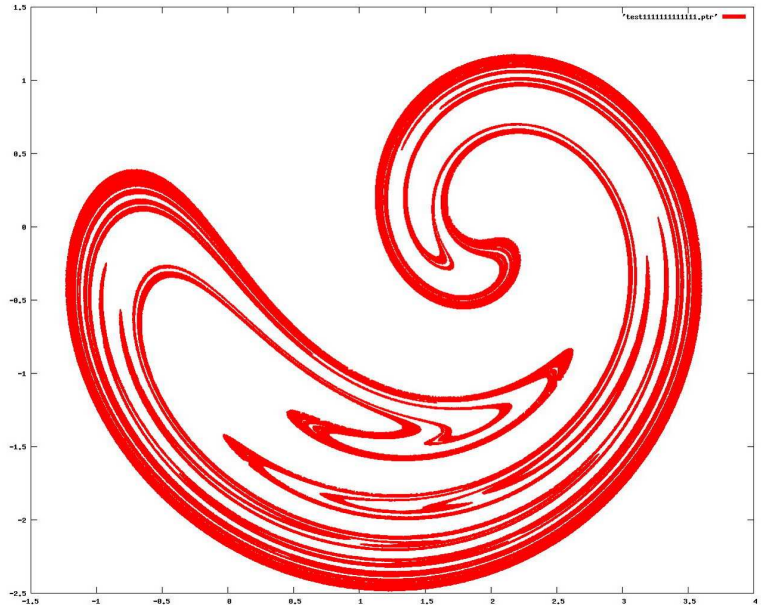


Figure 2: Invariant set for Ikeda map

6.2 Ikeda map

$x_1 = 2{-}0.9(xcos\tau(x,y){-}ysin\tau(x,y))$, $x_2 = 0.9(xsin\tau(x,y){+}ycos\tau(x,y))$, $\tau(x,y) = 0.4 - \frac{6}{1+x^2+y^2}$

$$\breve{M} = [-10,10] \times [-10,10]$$

Technique	Time
Linear	172 484 ms
Point	112 359 ms
Interval	34 000 ms

6.3 Delayed map

$x = y, \quad y = ay(1-x)$,,a=2.27

$$\breve{M} = [0,1] \times [0,1]$$

Technique	Time
Linear	17 297 ms
Point	28 406 ms
Interval	15 000 ms

References

[1] G. Osipenko, N. Ampilova, *Introduction into symbolic analysis of dynamic systems*, Saint-Petersburg State Univerity, 2005.

[2] Ershov A.P. Mixed Computations, *In the world of science*, http://ershov.iis.nsk.su/archive/eaindex.asp?did=2596, 14.02.1984. p.

[3] Romanenko S.A., Turchin V.F., *REFAL-COMPILER. TRANSACTIONS OF THE 2-ND ALL-UNION PROGRAMMING CONFERENCE*, Conference B, Novosibirsk, 1970, February, 3-6, http://www.refal.org/origins/RfcVkp2/index.htm

[4] A. Aho , R. Sethi, Jeffrey D. Ullman, *Compilers: Principles, Techniques, and Tools*,"Williams" Publishing house, 2003. -768 p.

[5] N. Ampilova, S. Terentyev, *The Application of Interval Arithmetic Methods to the Problem of the Symbolic Image Construction*, Electronic journal of differential equations and process control,http://www.neva.ru/journal, #4, 2006.

[6] E. Petrenko, Development and Implementation of Algorithms for Construction of the Symbolic Image, Electronic journal of differential equations and process control,http://www.neva.ru/journal, #3, 2006.

[7] Neumaier A., *Interval Methods for systems of equations*, Cambridge University Press 1990.

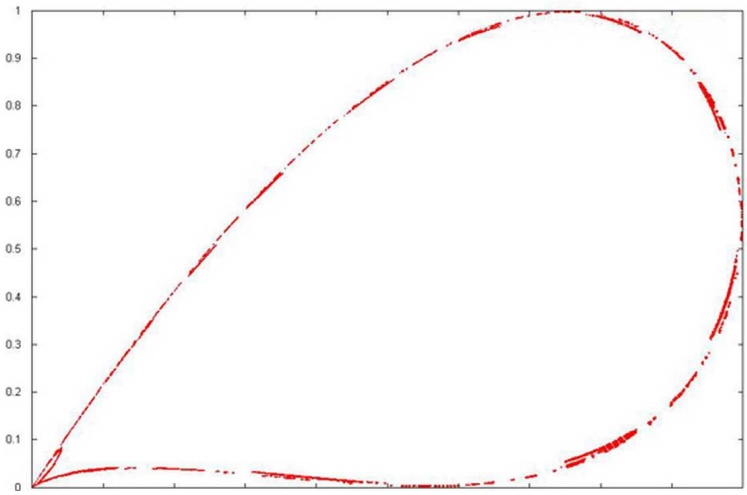


Figure 3: $a = 2.27$. Invariant set after 22 iterations

[8] *GNUPLOT* http://gnuplot.sourceforge.net/

[9] *MinGW* http://www.mingw.org/

[10] *Cygwin* http://www.cygwin.com/

[11] *ANTLR* http://www.antlr.org/

[12] *BOOST* http://www.boost.org/

SECTION D

NUMERICAL METHODS

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

A Necessary and Sufficient Condition for Characterization of Spline Functions

Adrian Branga
"Lucian Blaga" University of Sibiu,
Department of Mathematics,
str. Dr. I. Raţiu, nr. 5-7, 550012, Sibiu, Romania,
e-mail: adrian_branga@yahoo.com

Abstract

In the introduction of this paper is presented the definition of the spline functions as solutions of a variational problem and are shown some theorems regarding to the existence and uniqueness. The main result of this article consist in a necessary and sufficient condition for characterization of spline functions based on the properties of the spaces, operator and interpolatory set involved.

Keywords: spline functions, best approximation, characterization of the solution.

1 Introduction

Definition 1 *Let X_1 be a real linear space, $(X_2, \|.\|_2)$ a normed real linear space, $T : X_1 \to X_2$ an operator and $U \subseteq X_1$ a non-empty set. The problem of finding the elements $s \in U$ which satisfy*

$$\|T(s)\|_2 = \inf_{u \in U} \|T(u)\|_2, \tag{1}$$

is called the general spline interpolation problem, corresponding to the set U.

A solution of this problem, provided that exists, is named general spline interpolation element, corresponding to the set U.

The set U is called interpolatory set.

In the sequel we assume that X_1 is a real linear space, $(X_2, (.\,,.)_2, \|.\|_2)$ is a real Hilbert space, $T : X_1 \to X_2$ is a linear operator and $U \subseteq X_1$ is a non-empty set.

Theorem 1 (Existence Theorem) *If U is a convex set and $T(U)$ is a closed set, then the general spline interpolation problem (1) (corresponding to U) has at least a solution.*

ISSN-1584-5990

The proof is shown in the papers [2, 4].
For every element $s \in U$ we define the set

$$U(s) := U - s. \tag{2}$$

Lemma 1 *For every element $s \in U$ the set $U(s)$ is non-empty ($0_{X_1} \in U(s)$).*

The result follows directly from the relation (2).

Theorem 2 (Uniqueness Theorem) *If U is a convex set, $T(U)$ is a closed set and exists an element $s \in U$ solution of the general spline interpolation problem (1) (corresponding to U), such that $U(s)$ is linear subspace of X_1, then the following statements are true*

i) *For any elements $s_1, s_2 \in U$ solutions of the general spline interpolation problem (1) (corresponding to U) we have*

$$s_1 - s_2 \in Ker(T) \cap U(s); \tag{3}$$

ii) *The element $s \in U$ is the unique solution of the general spline interpolation problem (1) (corresponding to U) if and only if*

$$Ker(T) \cap U(s) = \{0_{X_1}\}. \tag{4}$$

For a proof see the papers [2, 3].
Further we define the set

$$V := T(U) \tag{5}$$

and for every element $s \in U$ we consider the set

$$V(s) := T(U(s)). \tag{6}$$

Lemma 2 *For every element $s \in U$ the following statements are true*

i) *$V(s)$ is non-empty set ($0_{X_2} \in V(s)$);*

ii) *$V = T(s) + V(s)$;*

iii) *If $U(s)$ is linear subspace of X_1, then $V(s)$ is linear subspace of X_2.*

A proof is presented in the paper [2].
In the sequel for every element $s \in U$ we define the set

$$W(s) := \{w \in T(X_1) \mid (w, \widetilde{v})_2 = 0,\ (\forall)\ \widetilde{v} \in V(s)\}. \tag{7}$$

Lemma 3 *For every element $s \in U$ the set $W(s)$ has the following properties*

i) *$W(s)$ is non-empty set ($0_{X_2} \in W(s)$);*

ii) *$W(s)$ is linear subspace of X_2.*

A proof is shown in the paper [2].

2 Main result

Theorem 3 *An element $s \in U$, such that $U(s)$ is linear subspace of X_1, is solution of the general spline interpolation problem (1) (corresponding to U) if and only if the following equality is true*

$$\{T(s)\} = V \cap W(s). \tag{8}$$

Proof. Let $s \in U$ be an element, such that $U(s)$ is linear subspace of X_1.
1) Suppose that s is solution of the general spline interpolation problem (1) (corresponding to U) and show that the equality (8) is true.

Since $s \in U$ it is obvious that

$$T(s) \in V. \tag{9}$$

Let $\lambda \in [0,1]$ be an arbitrary number and $v_1, v_2 \in V$ be arbitrary elements. From Lemma 2 ii) results that there are the elements $\widetilde{v}_1, \widetilde{v}_2 \in V(s)$ so that $v_1 = T(s) + \widetilde{v}_1, v_2 = T(s) + \widetilde{v}_2$. Consequently, we have

$$(1-\lambda)v_1 + \lambda v_2 = (1-\lambda)(T(s) + \widetilde{v}_1) + \lambda(T(s) + \widetilde{v}_2) =$$

$$= T(s) + ((1-\lambda)\widetilde{v}_1 + \lambda\widetilde{v}_2).$$

Because $U(s)$ is linear subspace of X_1, applying Lemma 2 iii), results that $V(s)$ is linear subspace of X_2, hence $(1-\lambda)\widetilde{v}_1 + \lambda\widetilde{v}_2 \in V(s)$. Therefore, we have $(1-\lambda)v_1 + \lambda v_2 \in T(s) + V(s)$ and using Lemma 2 ii) we obtain

$$(1-\lambda)v_1 + \lambda v_2 \in V,$$

i.e. V is a convex set.

Since $s \in U$ is solution of the general spline interpolation problem (1) (corresponding to U) it follows that

$$\|T(s)\|_2 = \inf_{u \in U} \|T(u)\|_2$$

and seeing the equality $\{T(u) \mid u \in U\} = \{t \mid t \in V\}$ it obtains

$$\|T(s)\|_2 = \inf_{v \in V} \|v\|_2. \tag{10}$$

Let $\widetilde{v} \in V(s)$ be an arbitrary element.

Applying Lemma 2 ii) it follows that there is an element $v \in V$ so that

$$\widetilde{v} = v - T(s). \tag{11}$$

We consider a certain $\alpha \in (0,1)$ and define the element

$$v' = (1-\alpha)T(s) + \alpha v. \tag{12}$$

Because $\alpha \in (0,1)$, $T(s), v \in V$ and taking into account that V is a convex set, from the relation (12) results

$$v' \in V. \tag{13}$$

Therefore, from the relations (10), (13) we deduce

$$\|T(s)\|_2 \leq \|v'\|_2$$

and considering the equality (12) we find

$$\|T(s)\|_2 \leq \|(1-\alpha)T(s) + \alpha v\|_2,$$

which is equivalent to

$$\|T(s)\|_2^2 \leq \|(1-\alpha)T(s) + \alpha v\|_2^2. \tag{14}$$

Using the relation (11) and the properties of the inner product it obtains

$$\|(1-\alpha)T(s) + \alpha v\|_2^2 = \tag{15}$$

$$= \|T(s) + \alpha(v - T(s))\|_2^2 = \|T(s) + \alpha\widetilde{v}\|_2^2 =$$

$$= \|T(s)\|_2^2 + 2\alpha(T(s), \widetilde{v})_2 + \alpha^2\|\widetilde{v}\|_2^2.$$

Substituting the equality (15) in the relation (14) it follows that

$$\|T(s)\|_2^2 \leq \|T(s)\|_2^2 + 2\alpha(T(s), \widetilde{v})_2 + \alpha^2\|\widetilde{v}\|_2^2,$$

i.e.

$$2\alpha(T(s), \widetilde{v})_2 + \alpha^2\|\widetilde{v}\|_2^2 \geq 0$$

and dividing by $2\alpha \in (0, 2)$ we obtain

$$(T(s), \widetilde{v})_2 + \frac{\alpha}{2}\|\widetilde{v}\|_2^2 \geq 0. \tag{16}$$

Because $\alpha \in (0, 1)$ was chosen arbitrarily it follows that the inequality (16) holds $(\forall)\ \alpha \in (0, 1)$ and passing to the limit for $\alpha \to 0$ it obtains

$$(T(s), \widetilde{v})_2 \geq 0.$$

As the element $\widetilde{v} \in V(s)$ was chosen arbitrarily we deduce that the previous relation is true $(\forall)\ \widetilde{v} \in V(s)$, i.e.

$$(T(s), \widetilde{v})_2 \geq 0, \quad (\forall)\ \widetilde{v} \in V(s). \tag{17}$$

Let show that in the relation (17) we have only equality. Suppose that $(\exists)\ \widetilde{v}_0 \in V(s)$ such that

$$(T(s), \widetilde{v}_0)_2 > 0. \tag{18}$$

Using the properties of the inner product, from the relation (18) we find

$$(T(s), -\widetilde{v}_0)_2 < 0. \tag{19}$$

Because $\widetilde{v}_0 \in V(s)$ it results that $-\widetilde{v}_0 \in -V(s)$. But, $U(s)$ being linear subspace of X_1, applying Lemma 2 iii) we deduce that $V(s)$ is linear subspace of X_2, hence $-V(s) = V(s)$. Consequently, $-\widetilde{v}_0 \in V(s)$, i.e.

$$(\exists)\ \widetilde{v}_1 \in V(s) \text{ such that } -\widetilde{v}_0 = \widetilde{v}_1. \tag{20}$$

From the relations (19) and (20) it follows that there is an element $\widetilde{v}_1 \in V(s)$ so that $(T(s), \widetilde{v}_1)_2 < 0$, which is in contradiction with the relation (17).

Therefore, the relation (17) is equivalent to

$$(T(s), \widetilde{v})_2 = 0, \quad (\forall)\ \widetilde{v} \in V(s), \tag{21}$$

i.e.

$$T(s) \in W(s). \tag{22}$$

Consequently, from the relations (9) and (22) it follows that

$$T(s) \in V \cap W(s). \tag{23}$$

Let show that $T(s)$ is the unique element from $V \cap W(s)$. Suppose that $(\exists)\ y \in V \cap W(s)$, with $y \neq T(s)$. Using the properties of the inner product it obtains

$$\|y - T(s)\|_2^2 = (y - T(s), y - T(s))_2 = (y, y - T(s))_2 - (T(s), y - T(s))_2. \tag{24}$$

Because $y \in V$ we deduce that $y - T(s) \in V - T(s)$ and applying Lemma 2 ii) we find $y - T(s) \in V(s)$. Taking into account that $y \in W(s), T(s) \in W(s)$ it results

$$(y, y - T(s))_2 = 0 \tag{25}$$

respectively

$$(T(s), y - T(s))_2 = 0. \tag{26}$$

Substituting the equalities (25), (26) in the relation (24) we obtain

$$\|y - T(s)\|_2^2 = 0,$$

i.e. $y = T(s)$, which is in contradiction with the assumption made before.

Therefore, $T(s)$ is the unique element from $V \cap W(s)$, hence

$$\{T(s)\} = V \cap W(s).$$

2) Suppose that the equality (8) is true and show that s is solution of the general spline interpolation problem (1) (corresponding to U).

Let $v \in V$ be an arbitrary element.

Applying Lemma 2 ii) we deduce that there is an element $\widetilde{v} \in V(s)$ such that $v = T(s) + \widetilde{v}$. Taking into account that $T(s) \in W(s)$ we find

$$(T(s), \widetilde{v})_2 = 0. \tag{27}$$

Using the properties of the inner product, considering the relation (27) and taking into account the properties of the norm, it follows that

$$\|v\|_2^2 = \|T(s) + \widetilde{v}\|_2^2 =$$

$$= \|T(s)\|_2^2 + 2(T(s), \widetilde{v})_2 + \|\widetilde{v}\|_2^2 =$$

$$= \|T(s)\|_2^2 + \|\widetilde{v}\|_2^2 \geq \|T(s)\|_2^2,$$

with equality if and only if $\|\widetilde{v}\|_2^2 = 0$, which is equivalent to $\widetilde{v} = 0_{X_2}$, i.e. $v = T(s)$.

The previous relation implies

$$\|T(s)\|_2 \leq \|v\|_2,$$

with equality if and only if $v = T(s)$.

As the element $v \in V$ was chosen arbitrarily we obtain that the previous inequality is true $(\forall)\ v \in V$, i.e.

$$\|T(s)\|_2 \leq \|v\|_2, \quad (\forall)\ v \in V \tag{28}$$

and the equality is attained in the element $v = T(s)$, which is equivalent to

$$\|T(s)\|_2 = \inf_{v \in V} \|v\|_2.$$

Consequently, s is solution of the general spline interpolation problem (1) (corresponding to U). □

References

[1] A. M. Acu, *Spline quasi-interpolants and quadrature formulas*, Acta Universitatis Apulensis, nr. 13, 2007, pp 21-36.

[2] A. Branga, *Contribuţii la teoria funcţiilor spline şi aplicaţii*, Teză de Doctorat, Universitatea "Babeş-Bolyai", Cluj-Napoca, 2003.

[3] Gh. Micula, *Funcţii spline şi aplicaţii*, Editura Tehnică, Bucureşti, 1978.

[4] Gh. Micula, S. Micula, *Handbook of splines*, Kluwer Acad. Publ., Dordrecht-Boston-London, 1999.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Pre-interpolating Type Quadrature Formulas

Eugen Constantinescu
"Lucian Blaga" University, Department of Mathematics,
Sibiu, Romania
E-mail: egnconst68@yahoo.com

Abstract

The purpose of this paper is to give a more general presentation of the quadrature formulas of preinterpolating type, these problems having the starting point in the work of A. Lupaş, for the weight $\omega(x) = 1$ and equidistant points.

Keywords: quadrature formula, interpolating polynomial.

Let $E = \{x_0, x_1, \ldots, x_n\}$ be a set of $n+1$ distinct points in the real interval $[a, b]$. To such a system we attach the polynomials

$$\omega(x) = \prod_{j=0}^{n}(x - x_j) \text{ the polynomial of knots}$$

$$l_k(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)},\ 0 \le k \le n \ - \text{ fundamental interpolating polynomials.}$$

We denote by $\mathcal{F}$ the linear space of all functions defined in $E \to \mathbb{R}$. We define in this space the metric $\rho(f, g) = \max_{x_k \in E} |f(x_k) - g(x_k)|,\ f, g \in \mathcal{F}$.

The interpolating polynomial of Lagrange $(L_n f)(x) := L_n(x_0, x_1, \ldots, x_n;\ f|\cdot)$ can be defined as being the only one polynomial of degree $\le n$ for which $\delta(f; L_n f) = \min_{k \in \Pi_n} \rho(f, h),\ f \in \mathcal{F}$, where Π_n is the space of polynomials of degree at most n. Let $f : [a, b] \to \mathbb{R}$ a Riemann-integrable functions and $w : (a, b) \to [0, \infty]$ a weight. The quadrature formulas of interpolating type have, as starting-point, the approximative equality: $f(x) \approx L_n(x_0, x_1, \ldots x_n; f|x)$.

Hence:

$$(*) \qquad \int_a^b w(x)f(x)dx \approx \int_a^b w(x)L_n(x_0, x_1, \ldots x_n; f|x)dx.$$

Similarly , the quadrature formulas of interpolating type of the form

$$\int_a^b w(x)f(x)dx = \sum_{k=1}^{n} c_k f(x_k) + R(f)$$

ISSN-1584-5990

where $R(f)$ represents the remainder, are characterized by the fact that the "exactedness-degree" is less equal to $n-1$.

Considering the above-mentioned facts, A.Lupaş[1] studied the possibility of introducing some quadratures of the form: (to be compared (**) with (*))

$$(**) \qquad \int_a^b w(x)f(x)dx \approx \int_a^b w(x)\mathcal{L}_{n-1}(E;f|x)dx$$

where $(\mathcal{L}_{n-1}f)(x) := \mathcal{L}_{n-1}(E;f|x)$ is the polynomial, of degree $n-1$ of the best approximation of the function f, considered as an element of the space $\mathcal{F}$; that is

$$\rho(f;L_{n-1}f) = \min_{p\in\Pi_{n-1}} \rho(f,p);\ p\in\Pi_{n-1}$$

Definition 1 *The quadrature formulas of the form:*

$$\int_a^b w(x)f(x)dx = \int_a^b w(x)\mathcal{L}_{n-1}(E;f|x)dx + R[f]$$

where $R[f]$ is the remainder, are known as quadratures of interpolating-type.

The paper [1] contains this problem for the particular case of the weight $w(x) = 1$ and for equidistant knots.

The aim of this paper is to deal with a more general form of this problem by considering that both the weight and the set E are arbitrary chosen.

In what follows, we need the next auxiliary result.

Lemma 1 *For $k\in\{0,1,\dots n\}$, the following equalities hold:*

$$\begin{gathered}\mathcal{L}_{n-1}(x_0,\dots,x_{k-1},x_{k+1},\dots,x_n;f|x) = \\ = L_n(x_0,x_1,\dots,x_n;f|x) - \tfrac{\omega(x)}{x-z_k}[x_0,x_1,\dots x_n;f]\end{gathered} \tag{1}$$

Proof. Because
$\mathcal{L}_{n-1}(x_0,\dots,x_{k-1}x_{k+1},\dots,x_n;f|x) = [x_0,x_1,\dots,x_{k-1},x_{k+1},\dots,x_n;f]\cdot$
$\cdot\, x^{n-1} + \dots + L_n(x_0,x_1,\dots,x_n;f|x) = [x_0,x_1,\dots,x_n;f]x^n + \dots$
it is easy to see that in the both sides of (1) we have polynomials of degree $\le n-1$. On the other side, then two polynomials take the same values in n distinct points, that means in the set $\{x_0,\dots,x_{k-1},x_{k+1},\dots,x_n\}$.

By using the paper [2], A. Lupaş suggested that $\mathcal{L}(E;f|x)$ is in fact a convex combination of the interpolating polynomials

$\mathcal{L}_{n-1}(x_0,\dots,x_{k+1},x_{k+1},\dots,x_n;f|x),\ k\in\{0,1,\dots n\}$

Theorem 1 *The preinterpolating polynomial $\mathcal{L}_{n-1}(E;f|x)$ of a function $f: E\to\mathbb{R}$ admits the representation:*

$$\mathcal{L}_{n-1}(E;f|x) = \frac{\sum_{k=0}^n c_k(E)L_{n-1}(x_0,\dots,x_{k-1},x_{k+1},\dots,x_n;f|x)}{\sum_{k=0}^n c_k(E)} \tag{2}$$

where $x_k(E) = \frac{1}{|\omega'(x_k)|}$

Remark 1 *Although this representation follows also from the paper of T.S. Matzkin and A.Sharma [2], we will give a more simpler proof.*

Proof. Let $O(x)$ be the polynomial in the right-hand side of the equality (2). We have $Q(x) \in \Pi_{n-1}$, and from (1) we find:

$$Q(x) = L_n(x_0, x_1, \ldots, x_n; f|x) - \frac{[x_0, x_1, \ldots, x_n; f]}{\sum_{k=0}^{n} c_k(E)} \sum_{k=0}^{n} \frac{\omega'(x_k)}{|\omega'(x_k)|} l_k(x)$$

If we denote $\delta = \dfrac{|[x_0, x_1, \ldots, x_n; f]|}{\sum\limits_{k=0}^{n} c_k(E)}$, we have

$$Q(x_j) = f(x_j) - \frac{\omega'(x_j)}{|\omega'(x_j)|} \cdot \delta,$$

hence

$$|Q(x_j) - f(x_j)| = \delta, \quad j \in \{0, 1, \ldots, n\}$$

$$(Q(x_i) - f(x_i))(Q(x_{i+1}) - f(x_{i+1})) = \omega'(x_i)\omega'(x_{i+1}\cdot$$

$$\frac{\delta^2}{|\omega'(x_i)\omega'(x_i + 1)|}) < 0 \qquad i \in \{0, 1, \ldots, n-1\},$$

the last inequality being justified by the fact that, the point in E being distinct, we have $\omega'(x_i)\omega'(x_{i+1}) < 0$. But, based on the theorem of alternance of Chebîşev, we know that the conditions:

$$\begin{cases} |Q(x_j) - f(x_j)| = \delta, & 0 \le j \le n \\ (Q(x_i) - f(x_i))(Q(x_{i+1} - f(x_{i+1}))) < 0, & 0 \le i \le n-1 \end{cases}$$

assure that $Q(x) = \mathcal{L}_{n_1}(E; f|x)$.
From these facts, we keep in mind also the following result:

Corollary 1 *The following equalities hold:*

$$\mathcal{L}_{n-1}(E; f|x) = L_n(x_0, x_1, \ldots, x_n; f|x) - \frac{[x_0, x_1, \ldots, x_n; f]}{\sum_{k=0}^{n} c_k(E)} \cdot$$

$$\sum_{k=0}^{n} \frac{\omega'(x_k)}{|\omega'(x_k)|} l_k(x)$$

$$f(x_j) - \mathcal{L}_{n-1}(E; f|x_j) = \frac{[x_0, x_1, \ldots, x_n; f]}{\sum\limits_{k=0}^{n} c_k(E)} \cdot \frac{\omega'(x_j)}{|\omega'(x_j)|}, x_j \in E$$

$$f(x) - \mathcal{L}_{n-1}(E; f|x) = \omega(x)[x, x_0, x_1, \ldots, x_n; f] + \\ + \frac{[x_0, x_1, \ldots, x_n; f]}{\sum\limits_{k=0}^{n} c_k(E)} \sum_{k=0}^{n} \frac{\omega'(x_k)}{|\omega'(x_k)|} l_k(x), x \notin E \qquad (3)$$

With the above notations, the following result holds:

Theorem 2 *Let $E = \{x_0, x_1, \ldots, x_n\}$ a system of distinct points in the interval $[a, b]$, and $w : (a, b) \to \mathbb{R}$ a weight.*
If

$$w(x) = \prod_{j=0}^{n}(x - x_j), \quad \text{and for } \ 0 \le k, \ \nu \le n$$

we denote

$$d_k = \frac{1}{\omega'(x_k)} \int_a^b w(x) \frac{\omega(x)}{(x - x_k)} dx,$$

$$c_\nu(E) = \frac{1}{|\omega'(x_\nu)|}, \quad \delta_k(E) = \frac{c_k(E)}{\sum_{j=0}^{n} c_j(E)}$$

then the quadrature formulas of pre-interpolating type, can be represented under the form:

$$\int_a^b w(x) f(x) dx = \sum_{k=0}^{n} c(k) f(x_k) + R[f] \tag{4}$$

where

$$c(k) = d_k - \frac{1}{\omega'(x_k)} \sum_{k=0}^{n} \omega'(x_J) \delta_j(E) d_j \tag{5}$$

The remainder in formula (4) has the form

$$R[f] = \int_a^b w(x)\omega(x)[x, x_0, x_1, \ldots, x_n; f]dx + [x_0, x_1, \ldots, x_n; f] \sum_{j=0}^{n} \omega'(x_j)\delta_j(E)d_j$$

Proof. The form of the coefficients follows by integration from Corollary 1. The form of the remainder is obtained from the equalities presented in the same corollary.

References

[1] A. Lupaş, *Asupra unor metode de integrare numerică,* Revista de Analiză Numerică şi Teoria Aproximţiei, vol3, fasc. 1(1974) 85-93.

[2] T.S. Motzkin, A.Sharma, *Next to interpolatory approximation on sets with multiplicities*, Canad.J.Math., 18(1966), 1196–1211.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Concerning the Improvement Results on Adomian Decomposition Method

Dumitru Deleanu, Letitia Ion
Department of Mathematical Sciences
Constantza Maritime University, Romania
dumitrudeleanu@yahoo.com, letiziaion@yahoo.com

Abstract

The paper proposes an improvement to the classical Adomian Decomposition Method (ADM). The time interval, where we search the solution, is splitted into subintervals. The solution given by ADM in first subinterval is extended for the next one only if an error criterion is carried out. If this is not true, then the value for the approximate solution at the first subinterval's end is taken as initial condition for the second subinterval. The process continues till the end of the time interval. In this way, the difference between the approximate solution and the exact one decreases significantly. The procedure is verified through test examples.

Keywords: Adomian decomposition method, approximate solutions, improved results

1 Introduction

In recent years, much attention has been given to develop some analytical methods for solving differential or integral equations including the perturbation methods and decomposition methods. The Adomian Decomposition Method (ADM), which accurately computes the series solution, is of great interest to applied sciences. The main advantage of the method is that it can be applied directly for all types of differential and integral equations, linear or nonlinear, homogeneous or non-homogeneous, with constant coefficients or with variable coefficients. Another advantage is that it provides a direct scheme for solving the problem, without the need of linearization, perturbation, massive computation or any other transformation and, for many cases, the resulting series converge to the exact solution. In other cases, the approximate solution is close to the exact ones when time and space variables are small but far from exact ones when these variables are big enough.

The method consists of splitting the given equation into a linear and nonlinear operator on both sides, identifying the initial and/or boundary conditions and the term involving the independent variables alone as initial approximation,

ISSN-1584-5990

decomposing the unknown functions into a series whose components are to be determined, decomposing the nonlinear function in terms of special polynomials called Adomian's polynomials, and finding the succesive terms of the series solution by recurrent relation using Adomian polynomials.

Over the last twenty years, the ADM has been applied to obtain formal solutions to a wide class of both deterministic and stochastic ODEs and PDEs. Guellal and Cheruault (1995) used Adomian's technique for solving an elliptic boundary value problem with an auxiliary condition. Adomian et al. (1996) solved mathematical models of the dynamic interaction of immune response with a population of bacteria, viruses, antigens or tremor cells. Laffez and Abbaoui (1996) studied a model of thermic exchanges in a drilling well. Abbaoui and Cherruault (1999) used the decomposition method for solving the Cauchy problem. They also gave a proof of convergence by using a new formulation of the Adomian polynomials. Sanchez et al. (2000) investigated the weaknesses of the thin-sheet approximation method and proposed a higher-order development allowing increase in the range of convergence. Lesnic (2002) investigated the convergence of ADM to periodic temperature fields in heat conductors.

Present paper proposes an improvement to the classical ADM. The paper is organized as follows. In Section 2, we give a brief description of the ADM while in Section 3 we present the basic idea for the improvement the results of the method. In the next section we're considered numerical results to demonstrate the efficiency of our idea through two test examples. Conclusion remarks are presented in Section 5.

2 Adomian Decomposition Method

Consider the general equation

$$Fu = g \tag{2.1}$$

where F represents a general nonlinear differential operator involving both linear and nonlinear terms. The linear term is decomposed into $L+R$, where L is easily invertible and R is the remainder of the linear operator. In most of the cases, L may be taken as the highest order derivative to avoid difficult integration involving Green's functions. The equation (2.1) may be written as

$$Lu + Ru + Nu = g \tag{2.2}$$

where Nu represents the nonlinear terms. Solving Lu from (2.2) and operating next with L^{-1} yields

$$u = \Phi + L^{-1}g - L^{-1}Ru - L^{-1}Nu \tag{2.3}$$

where Φ is the integration constant ($L\Phi = 0$). In the case of an initial value problem, the integral operator L^{-1} may be regarded as definite integrals from t_o to t. The decomposition method represents the solution of (2.2) as a series

$$u(t) = \sum_{n=0}^{\infty} u_n(t) \tag{2.4}$$

whereas the nonlinear operator is decomposed as

$$Nu = \sum_{n=0}^{\infty} A_n \tag{2.5}$$

From (2.3) - (2.5) we obtain

$$\sum_{n=0}^{\infty} u_n = u_0 - L^{-1}R\left(\sum_{n=0}^{\infty} u_n\right) - L^{-1}\left(\sum_{n=0}^{\infty} A_n\right) \tag{2.7}$$

where

$$\begin{aligned} u_0 &= \Phi + L^{-1}g \\ u_1 &= -L^{-1}Ru_0 - L^{-1}A_0 \\ u_2 &= -L^{-1}Ru_1 - L^{-1}A_1 \\ &\dots \\ u_n &= -L^{-1}Ru_{n-1} - L^{-1}A_{n-1} \end{aligned} \tag{2.8}$$

and A_n, $n = 0, 1, 2, ..,$ are Adomian's polynomials of $u_0, u_1, .., u_n$. They are obtained from

$$\begin{aligned} A_0 &= Nu_0 \\ A_1 &= u_1 \frac{dN}{du}(u_0) \\ A_2 &= u_2 \frac{dN}{du}(u_0) + \frac{u_1^2}{2!}\frac{d^2N}{du^2}(u_0) \\ \dots &\quad \dots \\ A_n &= \sum_{j=1}^{n} c(j,n) \cdot \frac{d^jN}{du^j}(u_0) \end{aligned} \tag{2.9}$$

where $c(j, n)$ are products or sums of products of j components of u whose subscripts sums to n, divided by the factorial of the number of repeated subscripts. Because the series (2.4) converges rapidly enough we can take the partial sum $S_n = \sum_{k=1}^{n-1} u_k$ as an approximation for the exact solution.

3 How to improve the results given by ADM?

Suppose we search an approximate solution in time interval $[t_0, T]$ in the form $u \cong S_n = \sum_{k=1}^{n-1} u_k$. We divide $[t_0, T]$ into subintervals $[t_{i-1}, t_i], i = 1, 2, ..$ with $\Delta t = t_i - t_{i-1}$ constant. The n-term partial sum S_n is considered as a good approximation of the exact solution on the subinterval $[t_{i-1}, t_i]$ only if the last term $u_{n-1}(t_i) \leq error$ (*), where error is a chosen accurate degree. If this is true, then we extend S_n as a feasible approximation in subinterval $[t_i, t_{i+1}]$. If the condition (*) is not true, we regard $S_n(t_i)$ like an initial condition in the subinterval $[t_i, t_{i+1}]$ and apply Adomian method to obtain the new series solution. The procedure is repeated until $t \geq T$.

4 Test examples

Example 1: Let us consider the first differential order equation involving a nonlinear term

$$\frac{dy}{dt} = y^2 - y, y(0) = 2 \tag{4.1}$$

with the exact solution

$$y_{exact}(t) = \frac{2}{2 - e^t},\ t \in (0, \ln 2).$$

We then have

$$Ly = \frac{dy}{dt},\ L^{-1}y(t) = \int_0^t y(\xi)\, d\xi,\ Ry = y,\ Ny = -y^2,\ g = 0 \tag{4.2}$$

and

$$\begin{array}{l} A_0 = -y_0^2,\ A_1 = -2y_1y_0,\ A_2 = -2y_2y_0 - y_1^2, \\ A_3 = -2y_3y_0 - 2y_2y_1,\ A_4 = -2y_4y_0 - 2y_3y_1 - y_2^2, \\ A_5 = -2y_5y_0 - 2y_4y_1 - 2y_3y_2 \end{array} \tag{4.3}$$

Consequently, we obtain the following approximants

$$y_0(t) = 2\,,\ y_1(t) = 2t\,,\ y_2(t) = 3t^2\,,\ y_3(t) = \frac{13}{3}t^3,$$

$$y_4 = \frac{25}{4}t^4\,,\ y_5(t) = \frac{541}{60}t^5\,,\ y_6(t) = \frac{1561}{120}t^6 \tag{4.4}$$

The series solution with four terms is given by

$$y(t) \cong y^{(3)}(t) = 2 + 2t + 3t^2 + \frac{13}{3}t^3 \tag{4.5}$$

whereas the series solution with seven terms is

$$y(t) \cong y^{(6)}(t) = 2 + 2t + 3t^2 + \frac{13}{3}t^3 + \frac{25}{4}t^4 + \frac{541}{60}t^5 + \frac{1561}{120}t^6 \tag{4.6}$$

If we search the solution with variational iteration method (VIM), after three iterations, we find that

$$y_{VIM}(t) = \frac{2}{3} + 2e^t - 2e^{2t} + \frac{4}{3}e^{3t} \tag{4.7}$$

Table 4.1 exhibits some representative values for the exact solution, the series solution obtained by VIM and for the series solution given by ADM (after three steps and six steps, respectively). It is obvious that the results obtained with VIM are better than those offered by ADM (at the same number of steps) and that the errors can be reduced further and higher accuracy can be obtained by evaluating more components of $y(t)$.

t	**Exact solution**	**VIM** (three iterations)	**ADM** (three iterations)	**ADM** (six iterations)
0.00	2.0000	2.0000	2.0000	2.0000
0.05	2.1081	2.1080	2.1080	2.1081
0.10	2.2351	2.2340	2.2343	2.2351
0.15	2.3862	2.3817	2.3821	2.3861
0.20	2.5687	2.5553	2.5547	2.5684
0.25	2.7934	2.7599	2.7552	2.7916
0.30	3.0763	3.0016	2.9870	3.0690
0.35	3.4427	3.2875	3.2533	3.4184
0.40	3.9356	3.6261	3.5573	3.8624
0.45	4.6330	4.0273	3.9024	4.4331
0.50	5.6935	4.5031	4.2917	5.1673

Table 4.1. Representative values obtained with ADM and VIM

The value listed in Table 4.1 shows us clearly that for t enough big the errors become unacceptable. If we follow the algorithm presented in Section 3 with $n = 3$, $error = 10^{-3}$, $\Delta t = 0.01$ and $T = 0.19$ we get

$$y^{(3)}_{ADMm}(t) = \begin{cases} 2 + 2t + 3t^2 + 4.333333t^3, & t \in [0, 0.06) \\ 1.999812 + 2.007366t + 2.816606t^2 + 6.222551t^3, & t \in [0.06, 0.11) \\ 1.998186 + 2.063827t + 2.186327t^2 + 8.649088t^3, & t \in [0.11, 0.15) \\ 1.991596 + 2.211636t + 1.067514t^2 + 11.491049t^3, & t \in [0.15, 0.19] \end{cases} \tag{4.8}$$

Table 4.2 shows the comparison between exact solution and the series solution (4.5) and (4.8), respectively.

t	$y^{(3)}_{ADM}$ initial	$y^{(3)}_{ADM}$ modified	$y_{exact} - y^{(3)}_{ADM}$ initial	$y_{exact} - y^{(3)}_{ADM}$ modified
0.11	2.2620677	2.2631734	0.0010800	0.0000179
0.15	2.3821255	2.3861431	0.0041200	0.0000191
0.19	2.5180233	2.5291616	0.0112200	0.0000815

Table 4.2. Comparison of numerical errors (three steps), $error = 0.001$

For $error = 10^{-3}$, $\Delta t = 0.01$ and $n = 6$ we get:

$$\begin{aligned} t \in [0; 0.2): \quad & y^{(6)}_{ADMm} = 2 + 2t + 3t^2 + 4.333333t^3 + 6.25t^4 + 9.016667t^5 + \\ & + 13.008333t^6 \\ t \in [0.2; 0.33): \quad & y^{(6)}_{ADMm} = 2.002432 + 1.912246t + 4.238069t^2 - 5.256412t^3 + \\ & 49.291638t^4 - 99.523358t^5 + 140.814510t^6 \end{aligned}$$

When $t = 0.33$, we have:

$$y_{exact}(0.33) = 3.283900\ ,\ y^{(6)}_{ADMm}(0.33) = 3.268634\ ,\ y^{(6)}_{ADMm} = 3.283029$$

Obviously, $y^{(6)}_{ADMm}$ is much closer by y_{exact} than $y^{(6)}_{ADMm}$. Improvement of accuracy requires smaller value for error.
If we choose $error = 10^{-4}$ and $\Delta t = 0.05$ we get the following approximate solutions (see Table 4.3):

$$y^{(3)}_{ADMm} = \begin{cases} 2 + 2t + 3t^2 + 4.333333t^3, & t \in [0; 0.015) \\ 1.999995 + 2.000044t + 2.973485t^2 + 5.018989t^3, & t \in [0.025; 0.05) \\ 1.999942 + 2.004125t + 2.879461t^2 + 5.845817t^3, & t \in [0.05; 0.075) \\ 1.999679 + 2.016199t + 2.689872t^2 + 6.850176t^3, & t \in [0.075; 0.1] \end{cases}$$

t	y_{exact}	$y^{(3)}_{ADMm}$	$y^{(3)}_{ADM}$
0.025	2.0519432	2.0519427	2.0519427
0.050	2.1080838	2.1080781	2.1080417
0.075	2.1689249	2.1689149	2.1687031
0.100	2.2350637	2.2350483	2.2343333

Table 4.3: Comparison of numerical errors (three steps), $error = 0.0001$

Finally, for $n = 6$ and $error = 10^{-4}$, we have:

$$\begin{aligned} t \in [0; 0.14): \quad & y^{(6)}_{ADMm} = 2 + 2t + 3t^2 + 4.333333t^3 + 6.25t^4 + 9.016667t^5 + \\ & \quad +13.008333t^6 \\ t \in [0.14; 0.24]: \quad & y^{(6)}_{ADMm} = 2.000081 + 2.154248t + 3.096002t^2 + 3.247667t^3 + \\ & \quad +13.427342t^4 - 18.101732t^5 + 63.106224t^6 \end{aligned}$$

t	y_{exact}	$y^{(6)}_{ADMm}$	$y^{(6)}_{ADM}$
0.24	2.744422	2.7443709	2.7440224

Table 4.4: Comparison of numerical errors (six steps), $error = 0.0001$

Example 2. In the second example, we solve the following non-linear system of differential equations:

$$\begin{aligned} \frac{dy_1}{dt} &= 2y_2^2, & y_1(0) = 1 \\ \frac{dy_2}{dt} &= e^{-t}y_1, & y_2(0) = 1 \\ \frac{dy_3}{dt} &= y_2 + y_3, & y_3(0) = 0 \end{aligned} \tag{4.9}$$

with the exact solution $y_1(t) = e^{2t}$, $y_2(t) = e^t$, $y_3(t) = te^t$.
Applying the Adomian method, we have

$$\begin{array}{lll} y_1^{(0)} = 1 & y_1^{(1)} = 2t & y_1^{(2)} = 4(t + e^t - 1) \\ y_2^{(0)} = 1 \quad ; & y_2^{(1)} = 1 - e^{-t} \quad ; & y_2^{(2)} = 2 - (2t + 2)e^{-t} \\ y_3^{(0)} = 0 & y_3^{(1)} = t & y_3^{(2)} = t + \dfrac{t^2}{2} + e^{-t} - 1 \end{array}$$

$$\begin{aligned} y_1^{(3)} &= 10t - 19 - e^{-2t} + (8t+20)e^{-t} \\ y_2^{(3)} &= 2 - 4te^{-t} - 2e^{-2t} \\ y_3^{(3)} &= t + \frac{t^2}{2} + \frac{t^3}{6} + e^{-t}(2t+3) - 3 \end{aligned}$$

Combining all these terms, we get

$$\begin{aligned} y_1 &\cong -22 + 16t + (8t+24)e^{-t} - e^{-2t} \\ y_2 &\cong 6 - (6t+3)e^{-t} - 2e^{-2t} \\ y_3 &\cong -4 + 3t + t^2 + \frac{t^3}{6} + (2t+4)e^{-t} \end{aligned} \tag{4.10}$$

Table 4.5 exhibits a comparison between the exact solution and the series solution given by ADM. Higher accuracy can be obtained by evaluating more terms.

t	**Exact solution**			**Approximate solution**		
	y_1	y_2	y_3	y_1	y_2	y_3
0.1	1.2214	1.1052	0.1105	1.2212	1.1051	0.1105
0.2	1.4918	1.2240	0.2448	1.4892	1.2207	0.2437
0.3	1.8221	1.3498	0.4049	1.8088	1.3485	0.4023
0.4	2.2255	1.4918	0.5967	2.1834	1.4816	0.5882
0.5	2.7183	1.6487	0.8244	2.6150	1.6251	0.8035

Table 4.5: Comparison between ADM and exact solution

By taking $error = 10^{-3}$, $n = 3$, $\Delta t = 0.01$ we found that solution (4.10) is valid for $t \in [0.00, 0.10)$. For brevity, we present only the next form of approximate solutions (see Table 4.6).

$t \in [0.10; 0.18]$:
$y_1(t) = -32.40666 + 21.62018t - 1.82152e^{-2t} + e^{-t}(11.93231t + 35.23001)$
$y_2(t) = 7.46789 - 3.29667e^{-2t} - e^{-t}(8.66529t + 3.18248)$
$y_3(t) = -5.12808 + 3.41932t + 1.16352t^2 + 0.2026t^3 + e^{-t}(2.69938t + 5.12882)$

5 Conclusions

In this paper, we proposed an improvement to the classical ADM. The main conclusions of the study are:

t	**Exact solution**			**Approximate solution**		
	y_1	y_2	y_3	y_1	y_2	y_3
0.18	1.4333	1.1972	0.2155	1.4337	1.1976	0.2158

Table 4.6: Comparison between ADM modified solution and exact solution (three steps), $error = 0.001$

a) The Adomian procedure is quite efficient to determine solution in closed form by using only initial condition;

b) Comparison with VIM reveals that the approximate solutions obtained by the ADM converge to their exact solution lower than these of VIM. For the same accuracy it was necessary a higher order of approximation for the ADM compared with VIM;

c) The improvement introduces by our idea permits us to obtain more precise values for the approximate solutions. For a high accuracy we can choose between decreasing the size of interval where the approximate solution is valid or increasing the number of terms in series solution;

d) The next step is to mechanize the computer process by means of high-level soft like Maple package

References

[1] Abbasbandy, S., *A new application of He's variational iteration method for quadratic Riccati differential equation by using Adomian's polynomials*, J. Comput. Appl. Math., **207**(2007), 59-63

[2] Anderson, G., Cherruault, Y., Abbaoui, K., *A nonperturbative analytical solution of immune response with time-delays and possible generalization*, Math. Comput. Modelling, **20**(10), 1996, 89-96

[3] He, J. H., *Homotopy perturbation technique*, Comput. Math. Appl. Mech., Energy (1999), 178-257

[4] Mustafa, I., *On numerical solutions of partial differential equations by the decomposition method*, Kragujivac J. Math., **26**(2004), 153-164

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On the Application of VIM to the Analysis of Weakly Non-linear Van der Pol Oscillator

Dumitru Deleanu
Department of Mathematical Sciences
Constantza Maritime University
104, Mircea cel Batran, Constantza, Romania
dumitrudeleanu@yahoo.com

Abstract

This study is motivated by the desire to investigate the approximate solutions of weakly nonlinear Van der Pol equation in non-resonant case by the implementation of a relatively new analytical method, called Variational Iteration Method (VIM). It is demonstrated that VIM yields a bifurcation equation as an integral part of the procedure. This equation shows that a family of quasi-periodic motions bifurcates from a family of periodic motions at a critical value of the parameters. Numerical results agree very well with the approximate one.

Keywords: Variational iteration method, Van der Pol oscillator

1 Introduction

Mathematical modeling of real-life problems usually results in functional equations, like ordinary or partial differential equations, integral and integral - differential equations or stochastic equations.

In most of cases these equations are difficult to solve analytically so it is required to obtain an efficient approximate solution. There are some valuable efforts that focus on solving the differential equations arising in engineering scientific applications. Among these solution methods we'll mention the variational iteration method (VIM), Adomian Decomposition Method (ADM) and Homotopy Perturbation Method (HPM).

The VIM was developed by Chinese mathematician Ji-Huan He. Over the years, VIM was applied to Klein-Gordon equation, to Helmholtz equation, to differential algebraic equations, to epidemic and prey-predator models, to nonlinear boundary value problems and other fields [1,2,5].

VIM provides the solution is a rapid convergent series which may lead the solution to a closed form. In this technique, the correction functional is developed and the Lagrange multipliers are calculated optimally via variational theory. The use of Lagrange multipliers reduces the successive application of

ISSN-1584-5990

the integral operator and the cumbersome of huge computational work while still maintaining a very high level of accuracy. The initial approximation can be freely chosen.

This study is organized as follows: in Section 2 we'll review the basic concepts of variational iteration method. In Section 3 we'll discuss a little about the Van der Pol's equation. Application of the proposed method to determine the first order approximation for the steady-state solution of weakly non-linear Van der Pol's equation in the non-resonant case will be given (analitically and numerical) in Sections 4 and 5 and a conclusion will be drawn in Section 6.

2 Variational Iteration Method. Basic Ideas

Let us consider the differential equation

$$Lx(t) + Nx(t) = g(t), t \in I, \tag{2.1}$$

where L and N are linear and nonlinear operator, respectively and $g(t)$ is a given function defined for all $t \in I$. According to VIM, we utilize the correction functional:

$$x^{(n+1)}(t) = x^{(n)}(t) + \int_0^t \lambda(\xi) \left[Lx^{(n)}(\xi) + N\bar{x}^{(n)}(\xi) - g(\xi) \right] d\xi \tag{2.2}$$

where λ is the Lagrange multiplier which can be identified optimally via the variational theory. The superscript (n) indicates the n-th approximation and $\bar{x}^{(n)}$ is considered as a restricted variation, i.e., $\delta\bar{x}^{(n)} = 0$. $x(t)$ will be obtained using λ and any initial approximation $x^{(0)}$. Consequently, the solution is found from

$$x(t) = \lim_{n\to\infty} x^{(n)}(t) \tag{2.3}$$

3 The Van der Pol's equation

Van der Pol's equation provides an example of an oscillator with non-linear damping, energy being dissipated at larger amplitudes and generated at low amplitudes. Such systems typically possess limit cycles, sustained oscillations around a state at which energy generation and dissipation balance.

The original application described by engineer Van der Pol, in 1927, models an electrical circuit with a triode valve, the restrictive properties of which change with current, negative resistance becoming positive as current increases.

Limit cycle oscillations, similar to the Van der Pol's system, also occur in models of wind-induced oscillations of building due to vortex shedding, in general aero-elastic flutter problems, in stability studies of both tracked and rubber tired vehicles, as well as in certain models of chemical reactions.
In our study the form chosen for Van der Pol's equation is

$$\ddot{x} + \omega_c^2 x - \epsilon(\eta - x^2)\dot{x} = F \sin \Omega t \tag{3.1}$$

where ω_c is the natural frequency, Ω the external frequency, F the amplitude of the external excitation, η the damping coefficient and $\epsilon > 0$ is a small parameter. That is the case of weakly non-linear oscillator. Obviously, $F = 0$ gives the corresponding autonomous system and $x = 0$ is the equilibrium position. System (3.1) with $F = 0$ exhibits a Hopf bifurcation as ξ passes through zero [3].

4 Variational iteration method for the non-resonant case of Van der Pol's equation

It is assumed that natural frequency ω_c and external frequency Ω satisfy the non-resonance relationship $k_1\omega_c \neq k_2\Omega$, where k_1, k_2 are any positive integers.

In the equation (3.1) the following form of the L operator is used

$$L = \frac{d^2}{dt^2} + \omega_c^2 \tag{4.1}$$

Consequently, the correction functional of equation (3.1) can be written as

$$x^{(n+1)}(t) = x^{(n)}(t) + \int_0^t \lambda \left(\frac{d^2 x^{(n)}}{d\xi^2} + \omega_c^2 x^{(n)} + N\bar{x}^{(n)} - g \right) d\xi \tag{4.2}$$

Making the above correction functional stationary and noticing that $\delta x^{(n)} = 0$, we find that:

$$\begin{aligned} \delta x^{(n+1)}(t) \quad &= \delta x^{(n)}(t) + \lambda(\xi)\delta x'^{(n)}(\xi)\,|_{\xi=t} - \lambda'(\xi)\delta x^{(n)}(\xi)\,|_{\xi=t} + \\ &+ \int_0^t \left(\lambda''(\xi) + \omega_c^2\lambda(\xi)\right) \delta x^{(n)}(\xi)\, d\xi = 0 \end{aligned} \tag{4.3}$$

which yields the following stationary conditions

$$\lambda'' + \omega_c^2\lambda = 0, \quad 1 - \lambda'\,|_{\xi=t} = 0\,, \quad \lambda\,|_{\xi=t} = 0 \tag{4.4}$$

From (4.4), the Lagrange multiplier can be identified as

$$\lambda(\xi) = \frac{1}{\omega_c} \sin\omega_c(\xi - t) \tag{4.5}$$

We start with the initial approximation

$$x^{(0)}(t) = A_{1,0} \sin\omega_c t + B_{1,0} \cos\omega_c t + \frac{F}{\omega_c^2 - \Omega^2} \sin\Omega t \tag{4.6}$$

obtained for $\epsilon = 0$. After the first iteration we obtain that

$$x^{(1)}(t) = x^{(0)}(t) + \frac{1}{\omega_c} \int_0^t \sin\omega_c(\xi - t) \left\{ \ddot{x}^{(0)}(\xi) + \omega_c^2 x^{(0)}(\xi) - \right.$$

$$-F\sin\Omega t-\epsilon\left(\eta-x^{(0)2}(\xi)\right)\dot{x}^{(0)}(\xi)\Big\}\,d\xi=x^{(0)}(t)-\frac{\epsilon}{\omega_c}\cdot$$

$$\cdot\int_0^t\sin\omega_c(\xi-t)\cdot\left[\eta-x^{(0)2}(\xi)\right]\cdot\dot{x}^{(0)}(\xi)d\xi=A_{1,0}\sin\omega_c t+$$

$$+B_{1,0}\cos\omega_c t+\frac{F}{\omega_c^2-\Omega^2}\sin\Omega t+\epsilon\{B_{0,1}\cos\Omega t+$$

$$+A_{2,-1}\sin(2\omega_c-\Omega)t+B_{2,-1}\cos(2\omega_c-\Omega)t+A_{-1,2}\sin(-\omega_c+2\Omega)t+ \quad (4.7)$$

$$+B_{-1,2}\cos(-\omega_c+2\Omega)t+A_{2,1}\sin(2\omega_c+\Omega)t+B_{2,1}\cos(2\omega_c+\Omega)t+$$

$$+A_{1,2}\sin(\omega_c+2\Omega)t+B_{1,2}\cos(\omega_c+2\Omega)t+A_{3,0}\sin 3\omega_c t+$$

$$+B_{3,0}\cos 3\omega_c t+B_{0,3}\cos 3\Omega t\}$$

where

$$B_{0,1}=\frac{\eta\Omega F}{(\omega_c^2-\Omega^2)^2}-\frac{\Omega F}{2(\omega_c^2-\Omega^2)^2}\left(A_{1,0}^2+B_{1,0}^2\right)-\frac{\Omega F^3}{4(\omega_c^2-\Omega^2)^4}$$

$$A_{2,-1}=\frac{(\Omega-2\omega_c)F}{2(\omega_c^2-\Omega^2)(3\omega_c^2-4\omega_c\Omega+\Omega^2)}A_{1,0}B_{1,0}$$

$$B_{2,-1}=\frac{(2\omega_c-\Omega)F}{4(\omega_c^2-\Omega^2)(3\omega_c^2-4\omega_c\Omega+\Omega^2)}(A_{1,0}^2-B_{1,0}^2)$$

$$A_{-1,2}=\frac{(\omega_c-2\Omega)F^2}{16(\omega_c^2-\Omega^2)^2(\omega_c\Omega-\Omega^2)}B_{1,0};\quad B_{-1,2}=\frac{(\omega_c-2\Omega)F^2}{16(\omega_c^2-\Omega^2)^2(\omega_c\Omega-\Omega^2)}A_{1,0}$$

$$A_{2,1}=\frac{(\Omega+2\omega_c)F}{2(\omega_c^2-\Omega^2)(3\omega_c^2+4\omega_c\Omega+\Omega^2)}A_{1,0}B_{1,0} \quad (4.8)$$

$$A_{2,1}=\frac{(\Omega+2\omega_c)F}{2(\omega_c^2-\Omega^2)(3\omega_c^2+4\omega_c\Omega+\Omega^2)}A_{1,0}B_{1,0}$$

$$B_{2,1}=\frac{(\Omega+2\omega_c)F}{4(\omega_c^2-\Omega^2)(3\omega_c^2+4\omega_c\Omega+\Omega^2)}(B_{1,0}^2-A_{1,0}^2)$$

$$A_{1,2}=\frac{(\omega_c+2\Omega)F^2}{16(\omega_c^2-\Omega^2)(\omega_c\Omega+\Omega^2)}B_{1,0};\quad B_{1,2}=-\frac{(\omega_c+2\Omega)F^2}{16(\omega_c^2-\Omega^2)(\omega_c\Omega+\Omega^2)}A_{1,0}$$

$$A_{3,0}=\frac{1}{32\omega_c}(3A_{1,0}^2-B_{1,0}^2)B_{1,0};\quad B_{3,0}=\frac{1}{32\omega_c}\quad(3B_{1,0}^2-A_{1,0}^2)A_{1,0}$$

$$B_{0,3}=\frac{\Omega F^3}{4}(\omega_c^2-\Omega^2)^3(\omega_c^2-9\Omega^2$$

In order to avoid secular terms in $x^{(1)}$ we need to eliminate coefficients of $t\sin\omega_c t$ and $t\cos\omega_c t$:

$$t\sin\omega_c t:\frac{A_{1,0}}{2}\left[\eta-\frac{A_{1,0}^2+B_{1,0}^2}{4}-\frac{F^2}{2(\omega_c^2-\Omega^2)^2}\right]=0$$

$$t\cos\omega_c t:\ \frac{B_{1,0}}{2}\left[\eta-\frac{A_{1,0}^2+B_{1,0}^2}{4}-\frac{F^2}{2(\omega_c^2-\Omega^2)^2}\right]=0 \tag{4.9}$$

Equations (4.9) yields two distinct, steady state solutions:

Solutions I: If

$$\eta\neq\frac{A_{1,0}^2+B_{1,0}^2}{4}+\frac{F^2}{2(\omega_c^2-\Omega^2)^2},$$

then $A_{1,0}=B_{1,0}=0$. It was obtained a periodic solution with frequency Ω, described by

$$x(t)=\frac{F}{\omega_c^2-\Omega^2}\sin\Omega t+\epsilon\left\{\bar{B}_{0,1}\cos\Omega t+B_{0,3}\cos 3\Omega t\right\}+O(\epsilon^2) \tag{4.10}$$

where $\bar{B}_{0,1}$ is $B_{0,1}$ with $A_{1,0}=B_{1,0}=0$.

Solution II: It corresponds to

$$\eta=\frac{A_{1,0}^2+B_{1,0}^2}{4}+\frac{F^2}{2(\omega_c^2-\Omega^2)^2} \tag{4.11}$$

and represents a quasi - periodic motion with frequencies ω_c and Ω (now $A_{1,0}\neq 0, B_{1,0}\neq 0$).

The solution bifurcates from solution (I) at the critical point (see figure 1)

$$\eta_c=\frac{F^2}{2(\omega_c^2-\Omega^2)^2} \tag{4.12}$$

Figure 1: Plot of damping coefficient η versus amplitude a

5 Numerical results

In order to demonstrate the accuracy of our approximate results we have considered some numerical simulations which cover the entire spectrum of possibilities. Thus, if $F=0$ the system (3.1) exhibits Hopf bifurcation as η passes through zero. For $\eta<0, x(t)=0$ is a stable solution (see figure 2a) and for $\eta>0$ the motion becomes periodic of period $T=2\pi/\omega_c$ and amplitude $a=2\sqrt{\mu}$ (see figure 2b).

In considering the external harmonic excitation ($F\neq 0$), we have been chosen the next two set of values:

(1) $F=4$, $\omega_c=1$, $\Omega=0.37$, $\eta=1$, $\epsilon=0.005$;
(2) $F=1$, $\omega_c=1$, $\Omega=0.37$, $\eta=2$, $\epsilon=0.005$.

In first case we get $\eta < \eta_c = 10.739$, then it is expected to obtain a periodic solution described by (4.10), with periodic $T = 2\pi/\Omega = 16.973$ (see figure 2c). In the second case, $\eta > \eta_c = 1.061$ and (4.7) may be used to approximate the real quasi-periodic motion with frequencies ω_c and Ω (see figure 2d). In figure 3, the results predicted by the first order VIM are compared with the numerical results. The two graphs are almost identical.

Figure 2: The time histories of system (3.1)
a) $F = 0, \eta < 0$; b) $F = 0, \eta > 0$; c) $F \neq 0, \eta < \eta_{cr}$; d) $F \neq 0$, $\eta > \eta_{cr}$

We focused then on the accuracy of solution (I) in the neighborhood of the third super harmonic resonance ($\omega_c : \Omega = 3 : 1$).

A comparison between amplitudes obtained by direct numerical integration of equation (3.1) and the approximate amplitudes given by (4.10) is shown in figure 4. It clearly show that outside the strong resonance region $\Omega/\omega_c \in [0.32 \div 0.34]$, the non-resonant theory is valid and the predicted response is in excellent agreement with the numerical response. Within the resonance region, the $B_{0,3}$ term becomes huge and the relation (4.10) fails to give an accurate response. Here, we must consider another approximation.

Finally, it is worth noting that the consistency of solutions (I) and (II) up to $O(\epsilon^2)$ has been verified by substituting these results into equation (3.1).

6 Conclusions

The variational iteration method has been applied successfully to many non-linear differential equations. In this paper, it was utilized to investigate the weakly Van der Pol oscillator under harmonic forcing, in non-resonant condition. The main conclusions of this study are:

a) The VIM yields a bifurcation equation as an integral part of the procedure. This equation shows that a family of quasi-periodic motions bifurcates from a family of periodic motions at a critical value of the parameters;

b) The external harmonic excitation results in a shift of the bifurcation point along the parameter axis, compared to the Hopf bifurcation associated with the corresponding autonomous system;

c) Numerical results agree very well with the approximate results. For computational and plots has been used Matlab;

d) VIM can be seen as a powerful tool for solving non-linear differential equations. Only one iteration it was sufficient to obtain precious information.

Figure 3: Comparison between VIM first order results and numerical results

Figure 4: The steady-state amplitude-frequency relation in the neighborhood of the third super harmonic resonance (anum-numerical amplitudes, aapx - approximate amplitudes)

References

[1] Abbdou, M. A, Soliman, A. A., *New applications of variational iteration method*, Phys. D211(1-2)(2005), 1-8

[2] He, J. H., *Variational iteration method - some recent results and new interpretations*, J. Comput. Appl. Math., **207**(2007), 3-17

[3] Huyseyn, K., Yu, P., *On bifurcations into non-resonant quasi-periodic motions*, Appl.Math. Model., **12**(1988), 189-201

[4] Nayfeh, A.H., Mook D.T., *Nonlinear Oscillations*, John Wiley, New York, 1979

[5] Noor, M. A., Mohyud-Din, S. T., *Variational iteration technique for solving higher order boundary value problems*, Appl. Math. Comput., **189**(2007), 1929-1942

[6] Sensoy, S., Huseyin, K., *On the Application of IHB Technique to the Analysis on Nonlinear Oscillations and Bifurcations*, Journal of Sound and Vibration, **95**(1988), 39-44

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

The Variational Method of Schiffer-Goluzin in an Extremal Problem of Class S

Miodrag Iovanov
"Constantin Brâncuşi", University of Târgu-Jiu, Romania
E-mail: miovanov@utgjiu.ro; iovanovm@yahoo.com

Abstract

Let S be the class of analytic functions of the form

$$f(z) = z + a_2 z^2 + \cdots, \ f(0) = 0, \ f'(0) = 1$$

defined on the unit disk $|z| < 1$. Petru T. Mocanu [2] raised the question of determination $\max_{f \in S} |f(z)|$ which satisfies the conditions $|f'(z)| = 1$, $|z| = r$, $f \in S$, $0 \leq r < 1$, r given. For solving the problem we shall use the variational method of Schiffer-Goluzin [1].

Keywords: holomorphic function, variational method, extremal function.

1. Let S be the class of functions $f(z) = z + a_2 z^2 + \cdots$, $f(0) = 0$, $f'(0) = 1$ which are regular and univalent in the unit disk $|z| < 1$. In [1] Petru T. Mocanu raised the issue of determination

$$\max_{f \in S} |f(z)| \tag{1}$$

which satisfies the conditions

$$|f'(z)| = 1, \ |z| = r, \ f \in \mathbf{S}, 0 \leq r < 1, r \text{ given.} \tag{2}$$

Since S is a compact class (1) is attained. The aim of this paper is to solve the problem by using the variational method of Schiffer-Goluzin [2].

2. Let $f \in \mathbf{S}$ the extremal function for which (1) is attained in conditions (2). In order to solve this problem let us consider a variation $f^*(z)$ of the function $f(z)$ given by the Schiffer-Goluzin's formula :

$$f^*(z) = f(z) + \lambda V(z; \zeta; \psi) + O(\lambda^2), \ |\zeta| < 1, \ \lambda > 0, \psi \text{ real} \tag{3}$$

where

$$\begin{cases} V(z;\zeta;\psi) = e^{i\psi} \dfrac{f^2(z)}{f(z) - f(\zeta)} - e^{i\psi} f(z) \left[\dfrac{f(\zeta)}{\zeta f'(\zeta)} \right]^2 - \\ \\ -e^{i\psi} \dfrac{z f'(z)}{z - \zeta} \zeta \left[\dfrac{f(\zeta)}{\zeta f'(z)} \right]^2 + e^{-i\psi} \dfrac{z^2 f'(z)}{1 - \overline{\zeta} z} \overline{\zeta} \overline{\left[\dfrac{f(\zeta)}{\zeta f'(\zeta)} \right]^2} \end{cases} \tag{4}$$

ISSN-1584-5990

It is known that if λ is small enough, the function $f^*(z)$ belongs to **S** class**.** Let us consider a variation z* of z:

$$z^* = z + \lambda h + O(\lambda^2) \quad , \; h = \frac{\partial z^*}{\partial \lambda} \mid_{\lambda=0}$$

which satisfies the conditions:

$$\mid z^* \mid = r \;\text{ and }\; \left| f^{*'}(z^*) \right| = 1 \tag{5}$$

We notice that

$$\mid z^* \mid^2 = |z|^2 + 2\lambda Re(\overline{z}h) + O(\lambda^2) = r^2.$$

Since $|z| = r$ from the relation above we obtain $Re(\overline{z}h) = 0$. Replacing z with z^* in relation (3) we have

$$f^{*'}(z^*) = f'(z^*) + \lambda V'(z^*;\zeta;\psi) + O(\lambda^2).$$

The condition $\mid f'(z) \mid = 1$ becomes:

$$2Re\left\{\overline{f'(z)}\left[hf''(z) + V'(z;\zeta;\psi)\right]\right\} = 0 \tag{6}$$

Since $f(z)$ is extremal $|f^*(z^*)| \leq \mid f(z) \mid$ which is equivalent with:

$$|f(z) + \lambda h\, f'(z) + \cdots + \lambda V(z;\zeta;\psi) + \cdots| \leq \mid f(z) \mid. \tag{7}$$

By squaring in (7) and using the equality $u\overline{u} = |u|^2$:

$$\begin{aligned} &\{f(z) + \lambda\,[h\, f'(z) + V(z;\zeta;\psi)] + \cdots\} \times \\ &\times \left\{\overline{f(z)} + \lambda\left[\overline{h\, f'(z)} + \overline{V(z;\zeta;\psi)}\right] + \cdots\right\} \leq f(z)\overline{f(z)}. \end{aligned} \tag{8}$$

From relation (8) we obtain the condition:

$$2Re\left\{\overline{f(z)}\,[h\, f'(z) + V(z;\zeta;\psi)]\right\} \leq 0. \tag{9}$$

From $Re(\overline{z}h) = 0$ we note that $overline{h} = -\frac{\overline{z}}{z}h$; having this the relation (6) can also be written as it follows :

$$\overline{f'(z)}\,[hf''(z) + V'(z;\zeta;\psi)] + f'(z)\left[-\frac{\overline{z}}{z}h\overline{f''(z)} + \overline{V'(z;\zeta;\psi)}\right] = 0$$

where

$$h = \frac{z\,\overline{f'(z)}V'(z;\zeta;\psi) + z\, f'(z)\overline{V'(z;\zeta;\psi)}}{\overline{z}f'(z)\overline{f''(z)} - z\overline{f'(z)}f''(z)} \tag{10}$$

Next we use the following notations

$$f = f(z), w = f(\zeta), l = f'(z), m = f''(z), V = V(z;\zeta;\psi), V' = V'_z(z;\zeta;\psi).$$

By using the notations above and h from relation (10) relation (9) can be written as it follows:]

$$Re\ \left[\overline{f}\left(plV'+V\right)\right]\leq 0 \tag{11}$$

where

$$p=\frac{zl-\overline{z}\overline{l}}{\overline{z}l\overline{m}-z\overline{l}m}\quad ,\ p \text{ real.}$$

1°.We assume that $Im(z\overline{l}m)\neq 0$ $(\overline{z}l\overline{m}-z\overline{l}m\neq 0)$. From relation (4) we obtain:

$$V=e^{i\psi}\frac{f^2}{f-w}-e^{i\psi}f\left(\frac{w}{\zeta w'}\right)^2-e^{i\psi}\frac{zl}{z-\zeta}\zeta\left(\frac{w}{\zeta w'}\right)^2+e^{-i\psi}\frac{z^2l}{1-\overline{\zeta}z}\overline{\zeta}\overline{\left(\frac{w}{\zeta w'}\right)^2}$$

and

$$\begin{aligned}V' &= e^{i\psi}\frac{fl(f-2w)}{(f-w)^2}-e^{i\psi}l\left(\frac{w}{\zeta w'}\right)^2-e^{i\psi}\frac{z(z-\zeta)m-\zeta l}{(z-\zeta)^2}\zeta\left(\frac{w}{\zeta w'}\right)^2+\\ &+e^{-i\psi}\frac{z^2(1-\overline{\zeta}z)m+zl(2-\overline{\zeta}z)}{\left(1-\overline{\zeta}z\right)^2}\overline{\zeta}\overline{\left(\frac{w}{\zeta w'}\right)^2}.\end{aligned}$$

By replacing the expresions of V and V' in (11) we obtain the relation

$$Re\ \left[e^{i\psi}\left(E-GF\right)\right]\leq 0 \tag{12}$$

where

$$E=\frac{f\overline{f}\left[\left(-2pl^2-f\right)w+pl^2f+f^2\right]}{\left(f-w\right)^2}\quad ,\quad F=\left(\frac{w}{\zeta w'}\right)^2$$

and

$$\begin{aligned}G &= pl^2\overline{f}+\frac{pl\overline{f}\left[z\left(z-\zeta\right)m-\zeta l\right]}{\left(z-\zeta\right)^2}\zeta-\frac{p\overline{l}f\left[\overline{z}^2\left(1-\zeta\overline{z}\right)\overline{m}+\overline{z}\overline{l}\left(2-\zeta\overline{z}\right)\right]}{\left(1-\zeta\overline{z}\right)^2}\zeta+\\ &+f\overline{f}+\frac{zl\overline{f}}{z-\zeta}\zeta-\frac{\overline{z}^2\overline{l}f}{1-\zeta\overline{z}}\zeta.\end{aligned}$$

Since ψ is arbitrary, from relation (12) it results that the extremal function $w=f(\zeta)$ must satisfy the following differential equation:

$$\left(\frac{\zeta w'}{w}\right)^2\frac{f\overline{f}\left[\left(-2pl^2-f\right)w+pl^2f+f^{\,2}\right]}{\left(f-w\right)^2}=\frac{\sum_{s=0}^{4}t_s\zeta^s}{\left(z-\zeta\right)^2\left(1-\zeta\overline{z}\right)^2} \tag{13}$$

where the coefficients t_s , $s \in \{0,1,2,3,4\}$ have the following expressions

$$\begin{aligned}
t_0 &= z^2\overline{f}\left(pl^2+f\right);\\
t_1 &= -2f\overline{f}z\left(1+2r^2\right)-l\overline{f}z^2-\overline{l}fr^4-2pl^2\overline{f}z\left(1+r^2\right)-\\
&\quad -pl\overline{f}mz^2-p\overline{l}f\left(\overline{z}^2\overline{m}+2\overline{l}zr^2\right);\\
t_2 &= f\ \overline{f}\left(r^4+4r^2+1\right)-2zl\overline{f}\left(2r^2+1\right)+\overline{z}\overline{l}fr^2\left(r^2+2\right)+\\
&\quad +pl^2f\left(r^4+4r^2+1\right)-pl\overline{f}\left[2mz\left(r^2+1\right)+l\right]-2r^2\left(\overline{mz}+2\overline{l}\right);\\
t_3 &= -\overline{z}\left(2r^2+1\right)\left(2f\ \overline{f}+\overline{l}\overline{z}f+2pl^2\overline{f}\right)+l\overline{f}r^2\left(r^2+2\right)+\\
&\quad +pl\overline{f}\left(mr^4+2mr^2+2l\overline{z}\right);\\
t_4 &= f\ \overline{f}\overline{z}^2-l\overline{z}\overline{f}r^2+\overline{l}f\overline{z}^3+pl^2\overline{f}\overline{z}^2-pl\overline{f}\overline{z}^2\left(mz+l\right)+p\overline{l}f\overline{z}^2\overline{mz}+l.
\end{aligned}$$

The extremal function turns the unit disk into a domain without exterior points. In order to justify this it is sufficient to assume that the domain transformed by an extremal function $w=f(\zeta)$ has an exterior point w_0 and to consider the function of variation:

$$f^*(z)=f(z)+\lambda e^{i\psi}\frac{f\ ^2(z)}{f(z)-w_0}\quad ,\ \lambda>0,\ \psi \text{ real}$$

which belongs to class **S.**

3. It is known that the extremal function $w=f(\zeta)$ transforms the unit disk $|\zeta|<1$,onto the entire plane, slit along a finite number of analytic arcs.

Let $q=e^{i\theta}$ be the point on the circle $|\zeta|=1$ which corresponds to the extremity of such slits in which $w\ '(q)=0$ and $\zeta=q$ double root for the polynomial $\sum\limits_{s=0}^{4}t_s\zeta^s$.

Since $\zeta=q$ is a double root for this polynomial, we can write:

$$\sum_{s=0}^{4}t_s\zeta^s=\left(1-\overline{q}\zeta\right)^2\left(a_0+a_1\zeta+a_2\zeta^2\right).$$

From the expressions of the coefficients t_s , $s \in \{0,1,2,3,4\}$ it results that we can consider
$a_0=t_0,\ a_1=-2kq,\ k$ real and $a_2=q^2t_4$.

The differential equation (13) can also be written as it follows:

$$\left(\frac{\zeta w'}{w}\right)^2\frac{f\overline{f}\left[\left(-2pl^2-f\right)w+pl^2f+f^2\right]}{\left(f-w\right)^2}= \tag{14}$$

$$=\frac{\left(1-\overline{q}\zeta\right)^2\left(t_0-2kq\zeta+q^2t_4\zeta^2\right)}{\left(z-\zeta\right)^2\left(1-\overline{z}\zeta\right)^2}$$

4. From the differential equation (14) we obtain:

$$\frac{\sqrt{f\overline{f}\left[(-2pl^2-f)\,w+pl^2f+f^2\right]}}{w\,(f-w)}dw=\frac{(1-\overline{q}\zeta)\,\sqrt{t_0-2kq\zeta+q^2t_4\zeta^2}}{\zeta\,(z-\zeta)\,(1-\overline{z}\zeta)}d\zeta$$

where

$$\int_0^w \frac{\sqrt{f\overline{f}\left[(-2pl^2-f)\,w+pl^2f+f^2\right]}}{w\,(f-w)}dw= \tag{15}$$

$$=\int_0^\zeta \frac{(1-\overline{q}\zeta)\,\sqrt{t_0-2kq\zeta+q^2t_4\zeta^2}}{\zeta\,(z-\zeta)\,(1-\overline{z}\zeta)}d\zeta$$

For the computation of the integral on the left of the relation (15) we note

$$I_1=\int_0^w \frac{\sqrt{f\overline{f}\left[(-2pl^2-f)\,w+pl^2f+f^2\right]}}{w\,(f-w)}dw.$$

We can write

$$I_1=\sqrt{f\overline{f}\,(-2pl^2-f)}\int\frac{\sqrt{w+\alpha^2}}{w\,(f-w)}dw \quad \text{where } \alpha^2=\frac{pl^2f+f^2}{-2pl^2-f}.$$

We notice that $w+\alpha^2=u^2$; I_1 becomes:

$$I_1=\sqrt{f\overline{f}\,(-2pl^2-f)}\int\frac{-2u^2du}{(u^2-\alpha^2)\,(u^2-\beta^2)} \qquad \left(dw=2udu\ ,\ \beta^2=\alpha^2+f\right)$$

We notice that

$$\frac{-2u^2}{(u^2-\alpha^2)\,(u^2-\beta^2)}=\frac{2\alpha^2}{\beta^2-\alpha^2}\,\frac{1}{u^2-\alpha^2}-\frac{2\beta^2}{\beta^2-\alpha^2}\,\frac{1}{u^2-\beta^2} \qquad ,\ \beta^2-\alpha^2=f$$

and the integral above becomes:

$$I_1=\frac{\sqrt{f\overline{f}\,(-2pl^2-f)}}{f}\left(\alpha\ln\frac{u-\alpha}{u+\alpha}-\beta\ln\frac{u-\beta}{u+\beta}\right)$$

or

$$I_1=\frac{\sqrt{f\overline{f}\,(-2pl^2-f)}}{f}\ln\left[\left(\frac{u-\alpha}{u+\alpha}\right)^{\alpha}\left(\frac{u+\beta}{u-\beta}\right)^{\beta}\right] \tag{16}$$

In order to compute the integral on the right of the relation (15) we note

$$I_2=\int\frac{(1-\overline{q}\zeta)\,\sqrt{t_0-2kq\zeta+q^2t_4\zeta^2}}{\zeta\,(z-\zeta)\,(1-\overline{z}\zeta)}d\zeta\ .$$

We notice that $t_0 - 2kq\zeta + q^2t_4\zeta^2 = q^2t_4\left(\zeta - \zeta_1\right)\left(\zeta - \zeta_2\right)$ where $\zeta_{1,2} = = \dfrac{k \pm \sqrt{k^2 - t_0t_4}}{t_4}\overline{q}$.If we note $k - \sqrt{k^2 - t_0t_4} = \rho$, we notice that $\zeta_1 = \frac{\rho}{t_4}\overline{q}$ and $\zeta_2 = \frac{t_0}{\rho}\overline{q}$. Using the above notations, we can write: $\sqrt{t_0 - 2kq\zeta + q^2t_4\zeta^2} =$

$$= \sqrt{q^2t_4}\sqrt{\left(\zeta - \frac{\rho}{t_4}\overline{q}\right)\left(\zeta - \frac{t_0}{\rho}\overline{q}\right)}.$$

In order to compute integral I_2 we make the following substitution:

$$\sqrt{\left(\zeta - \frac{\rho}{t_4}\overline{q}\right)\left(\zeta - \frac{t_0}{\rho}\overline{q}\right)} = v\left(\zeta - \frac{\rho}{t_4}\overline{q}\right). \tag{17}$$

From (17) we obtain:

$$\zeta = \sigma\frac{v^2 - a^2}{v^2 - 1} \quad \text{where } \sigma = \frac{\rho}{t_4}\overline{q} \quad \text{and } a^2 = \frac{t_0t_4}{\rho^2} \tag{18}$$

Next,

$$\begin{cases} d\zeta = \dfrac{2\sigma\left(a^2 - 1\right)}{\left(v^2 - 1\right)^2}dv, \\ z - \zeta = \left(z - \sigma\right)\dfrac{v^2 - b^2}{v^2 - 1} \quad \text{with} \quad b^2 = \dfrac{\sigma a^2 - z}{\sigma - z}, \end{cases} \tag{19}$$

and

$$\begin{cases} 1 - \overline{z}\zeta = \left(1 - \overline{z}\sigma\right)\dfrac{v^2 - c^2}{v^2 - 1} \quad \text{with} \quad c^2 = \dfrac{1 - \overline{z}\sigma a^2}{1 - \overline{z}\sigma} \\ 1 - \overline{q}\zeta = \left(1 - \overline{q}\sigma\right)\dfrac{v^2 - d^2}{v^2 - 1} \quad \text{with} \quad d^2 = \dfrac{1 - \overline{q}\sigma a^2}{1 - \overline{q}\sigma} \\ \sqrt{\left(\zeta - \dfrac{\rho}{t_4}\overline{q}\right)\left(\zeta - \dfrac{t_0}{\rho}\overline{q}\right)} = \dfrac{\sigma\left(1 - a^2\right)v}{v^2 - 1}. \end{cases} \tag{20}$$

By using the relations above we obtain:

$$I_2 = \frac{2\sigma q\left(1 - \overline{q}\sigma\right)\left(1 - a^2\right)^2\sqrt{t_4}}{\left(\sigma - z\right)\left(1 - \overline{z}\sigma\right)}\int P(v)dv, \quad \text{where} \tag{21}$$

$P(v) = \dfrac{v^2\left(v^2 - d^2\right)}{\left(v^2 - 1\right)\left(v^2 - a^2\right)\left(v^2 - b^2\right)\left(v^2 - c^2\right)}$.We are looking for an expansion as it follows:

$$P(v) = \frac{A_1}{v - 1} + \frac{A_2}{v + 1} + \frac{A_3}{v - a} + \frac{A_4}{v + a} + \frac{A_5}{v - b} + \frac{A_6}{v + b} + \frac{A_7}{v - c} + \frac{A_8}{v + c} \tag{22}$$

For the coefficients A_k , $k \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ which occur in the relation (22)

we find the following values:

$$\begin{cases} A_1 = -A_2 = \dfrac{1-d^2}{2\left(1-a^2\right)\left(1-b^2\right)\left(1-c^2\right)} \overset{note}{=} \tau_1 \\ A_3 = -A_4 = \dfrac{a\left(a^2-d^2\right)}{2\left(a^2-1\right)\left(a^2-b^2\right)\left(a^2-c^2\right)} \overset{note}{=} \tau_2 \\ A_5 = -A_6 = \dfrac{b\left(b^2-d^2\right)}{2\left(b^2-1\right)\left(b^2-a^2\right)\left(b^2-c^2\right)} \overset{note}{=} \tau_3 \\ A_7 = -A_8 = \dfrac{c\left(c^2-d^2\right)}{2\left(c^2-1\right)\left(c^2-a^2\right)\left(c^2-b^2\right)} \overset{note}{=} \tau_4 \end{cases} \tag{23}$$

If we note $\mu = \dfrac{2\sigma q\left(1-\overline{q}\sigma\right)\left(1-a^2\right)^2\sqrt{t_4}}{\left(\sigma-z\right)\left(1-\overline{z}\sigma\right)}$, for I_2 we obtain the expression:

$$I_2 = \mu\left(\tau_1 \ln\frac{v-1}{v+1} + \tau_2 \ln\frac{v-a}{v+a} + \tau_3 \ln\frac{v-b}{v+b} + \tau_4 \ln\frac{v-c}{v+c}\right) \tag{24}$$

From relation (17) we note that

$$v\left(\zeta\right) = \sqrt{\frac{\zeta - \dfrac{t_0}{\rho}\overline{q}}{\zeta - \dfrac{\rho}{t_4}\overline{q}}} \tag{25}$$

and from the notation made for the computation of I_1 ,

$$u^2\left(\zeta\right) = w\left(\zeta\right) + \alpha^2 \tag{26}$$

By using the relations (16) and (24) relation (15) becomes:

$$I_1\Big|_0^w = \quad I_2\Big|_0^\zeta \tag{15'}$$

For $\zeta = 0$ (16), (24) and (15′) we obtain the constant which results from the two factors of relations (16) and (24) $\left(\text{corresponding to } \dfrac{u-\alpha}{u+\alpha} \text{ and } \dfrac{v-a}{v+a}\right)$:

$$\ln\left(-1\right)^{\dfrac{\alpha\sqrt{f\overline{f}\left(-2pl^2-f\right)}}{f}+\mu\tau_2}.$$

Thus (15′) will be written as it follows:

$$\left\{\begin{array}{l} \dfrac{\sqrt{f\ \overline{f}\,(-2pl^2-f)}}{f}\ln\left[\left(\dfrac{u(\zeta)-\alpha}{u(\zeta)+\alpha}\right)^{\alpha}\left(\dfrac{u(\zeta)+\beta}{u(\zeta)-\beta}\right)^{\beta}\right]+\dfrac{\sqrt{f\ \overline{f}\,(-2pl^2-f)}}{f}\ln\left(\dfrac{\alpha+\beta}{\alpha-\beta}\right)^{\beta}+ \\ \qquad\qquad +\ln(-1)^{\dfrac{\alpha\sqrt{f\overline{f}\,(-2pl^2-f)}}{f}+\mu\tau_2}= \\ =\mu\left[\ln\left(\dfrac{v(\zeta)-1}{v(\zeta)+1}\right)^{\tau_1}+\ln\left(\dfrac{v(\zeta)-a}{v(\zeta)+a}\right)^{\tau_2}+\ln\left(\dfrac{v(\zeta)-b}{v(\zeta)+b}\right)^{\tau_3}+\ln\left(\dfrac{v(\zeta)-c}{v(\zeta)+c}\right)^{\tau_4}\right]- \\ \qquad -\mu\ln\left[\left(\dfrac{a-1}{a+1}\right)^{\tau_1}\left(\dfrac{a-b}{a+b}\right)^{\tau_3}\left(\dfrac{a-c}{a+c}\right)^{\tau_4}\right] \end{array}\right. \tag{15"}$$

After restrictions and convenient grouping the equation above is written as it follows:

$$\left\{\begin{array}{l} \left[\left(\dfrac{u(\zeta)-\alpha}{u(\zeta)+\alpha}\right)^{\alpha}\left(\dfrac{u(\zeta)+\beta}{u(\zeta)-\beta}\ \dfrac{\alpha+\beta}{\alpha-\beta}\right)^{\beta}\right]^{\dfrac{\sqrt{f\overline{f}\,(-2pl^2-f)}}{f}}\times \\ \qquad \times(-1)^{\dfrac{\alpha\sqrt{f\overline{f}\,(-2pl^2-f)}}{f}+\mu\tau_2}= \\ =\left[\begin{array}{c}\left(\dfrac{v(\zeta)-1}{v(\zeta)+1}\ \dfrac{a+1}{a-1}\right)^{\tau_1}\left(\dfrac{v(\zeta)-a}{v(\zeta)+a}\right)^{\tau_2}\times \\ \times\left(\dfrac{v(\zeta)-b}{v(\zeta)+b}\ \dfrac{a+b}{a-b}\right)^{\tau_3}\left(\dfrac{v(\zeta)-c}{v(\zeta)+c}\ \dfrac{a+c}{a-c}\right)^{\tau_4}\end{array}\right]^{\mu} \end{array}\right. \tag{27}$$

Relation (27) implicitly represents the equation verified by the extremal function $w=w(\zeta)$ which performs $\max\limits_{f\in S}|f(z)|$.

2°. Let us now assume that $Im\left(z\overline{l}m\right)=0$. In this case from the expressin of p there must be $zl-\overline{z}\overline{l}=0$. From the conditions

$z\overline{l}m-\overline{z}l\overline{m}=0,\ zl-\overline{z}\overline{l}=0,\ |f\,'(z)|=1$ and $|z|=r$ we obtain $m=\overline{m}$ and from the expression of p it results that p is complex, which is a contradiction! Therefore this case cannot happen.

5. Next we will show how θ and k can be determined. In relation (14) we perform $\zeta\to z$; and we obtain:

$$-pz^2l^2\overline{f}\left(1-r^2\right)^2=(1-\overline{q}z)^2\left(t_0-2kq\zeta+q^2t_4\zeta^2\right)$$

and by multiplying the resulting equality by $\overline{z}^2\overline{l}^2 f$ we obtain ($|z| = r$) :

$$-pz^2\overline{z}^2 l^2\overline{l}^2 \; f.\overline{f}\left(1-r^2\right)^2 = \left(\overline{z}-\overline{q}r^2\right)^2\left(t_0 - 2kq\zeta + q^2t_4\zeta^2\right) f\; \overline{l}^2 \tag{28}$$

Since the expression on the left of the equality (28) is real, we obtain a system with two equations from which we obtain θ and k :

$$\begin{cases} Re\left[\left(\overline{z}-\overline{q}r^2\right)^2\left(t_0 - 2kq\zeta + q^2t_4\zeta^2\right) f\; \overline{l}^2\right] + \\ +pz^2\overline{z}^2 l^2\overline{l}^2 \; f.\overline{f}\left(1-r^2\right)^2 = 0 \\ \text{and} \\ Im\left[\left(\overline{z}-\overline{q}r^2\right)^2\left(t_0 - 2kq\zeta + q^2t_4\zeta^2\right) f\; \overline{l}^2\right] = 0 \end{cases} \tag{29}$$

With θ and k determined in this way the extremal function $w = w(\zeta)$ from equation (27) is well determined; with its help we find $\max\limits_{f\in S}|f(z)|$ in conditions $|f\,'(z)| = 1,\ \ |z| = r,\ \ f \in \mathbf{S},\ 0 \le r < 1,\ r$ given.

References

[1] G.M. Goluzin, *Geometriceskaia teoria funcţii complecsnogo peremennogo*, Moscova-Leningrad, 1952.

[2] P.T. Mocanu, *An extremal problem for univalent functions*, Babeş-Bolyai University, Faculty of Mathematics, Cluj-Napoca, 1986.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

A Modifies Spline for an Approximation of Singular Integrals of Cauchy Type

Mostefa Nadir
Laboratory of Pure and Applied Mathematics
Department of Mathematics University of M'sila Algeria
E-mail: mostefanadir@yahoo.fr

Abstract

In this work we present an approximation for singular integrals of Cauchy type using a small modification of the splines functions in order to eliminate the singularity. This approximation is destined to resolve numerically the singular integral equations with Cauchy type kernel on a smooth oriented contour.

Keywords: Singular integral, interpolation, Hőlder space and Hőlder condition

1 Introduction

The main considerations of the present work concern the construction and foundation of some numerical schemes which are destined for numerical solution of singular integral equations with Cauchy type kernel.

$$a(t_0)\varphi(t_0) + \frac{b(t_0)}{\pi i}\int_\Gamma \frac{\varphi(t)}{t-t_0}dt + \int_\Gamma k(t,t_0)\varphi(t)dt = f(t_0) \tag{1}$$

where under Γ we designate a smooth contour oriented, t and t_0 are points on Γ, $a(t), b(t), k(t,t_0)$ and $f(t)$ are functions given on Γ.

Let $F(t_0)$ be a singular integral defined by

$$F(t_0) = \frac{1}{\pi i}\int_\Gamma \frac{\varphi(t)}{t-t_0}dt, \quad t, t_0 \in \Gamma. \tag{2}$$

For the existence of the principal value of this integral for a given density $\varphi(t)$, we will need more than mere continuity, in other words, the density $\varphi(t)$ has to satisfy the Hölder condition $H(\mu)$[2].

ISSN-1584-5990

2 The Quadrature

Let $t = t(s) = x(s) + iy(s)$ where $s \in [a, b]$ be the parametric complex equation of the curve Γ with the respect to some parameter s. Consider that N is an arbitrary natural number, generally we take it large enough, and divide the interval $[a, b]$ into N equal subintervals of $[a, b]$

$$[a, b] = \{a = s_0 < s_1 < ..... < s_N = b\},$$

be called I_1 to I_N, so that, we have $I_{\sigma+1} = [s_\sigma, s_{\sigma+1}]$.

$$s_\sigma = a + \sigma\frac{l}{N}\ ,\ l = b - a\ ,\ \sigma = 0, 1, 2, ..., N.$$

Further, fixing a natural number m, and divide each of segments $[s_\sigma, s_{\sigma+1}[$ by points

$$s_{\sigma k} = s_\sigma + hx_k,\quad h = \frac{l}{N},\quad k = 0, 1, ..., m\ ,$$

where the points $\{x_k\}$ represent the increasing sequence belongs to the interval $[0, 1[$.

Denoting by

$$t_\sigma = t(s_\sigma),\ t_{\sigma k} = t(s_{\sigma k});\quad \sigma = 0, 1, 2, ..., N;\quad k = 0, 1, ..., m.$$

Assuming that for the indices $\sigma, \nu = 0, 1, 2, ..., N-1$ the points t and t_0 belong respectively to the arcs $t_\sigma\overset{\frown}{t_{\sigma+1}}$ and $t_\nu\overset{\frown}{t_{\nu+1}}$ where $t_\alpha\overset{\frown}{t_{\alpha+1}}$ designate the smallest arc with ends t_α and $t_{\alpha+1}$ [3], [5] and [6].

For an arbitrary numbers σ, ν from $0, 1, 2, ..., N-1$, we define the function $\beta_{\sigma\nu}(\varphi; t, t_0)$ dependents of φ, t and t_0 by

$$\beta_{\sigma\nu}(\varphi; t, t_0) = U(\varphi; t, \sigma) - V(\varphi; t_0, \nu), \tag{3}$$

where the expression $U(\varphi; t, \sigma)$, $V(\varphi; t_0, \nu)$ designates the approximation of the function density $\varphi(t), \varphi(t_0)$ on the subinterval $[t_\sigma, t_{\sigma+1}[, [t_\nu, t_{\nu+1}[$ respectively of the curve Γ, given by the following formula

$$U(\varphi; t, \sigma) = \frac{t_{\sigma(k+1)} - t}{t_{\sigma(k+1)} - t_{\sigma k}}\varphi(t_{\sigma k})\frac{t - t_0}{t_{\sigma k} - t_0} + \frac{t - t_{\sigma k}}{t_{\sigma(k+1)} - t_{\sigma k}}\varphi(t_{\sigma(k+1)})\frac{t - t_0}{t_{\sigma(k+1)} - t_0},$$

$$V(\varphi; t_0, \nu) = S_1(\varphi; t_0, \nu)\frac{t - t_0}{t_{\sigma(k+1)} - t_{\sigma k}}\left(\frac{t_{\sigma(k+1)} - t}{t_{\sigma k} - t_0} + \frac{t - t_{\sigma k}}{t_{\sigma(k+1)} - t_0}\right),$$

where $S_1(\varphi; t_0, \nu)$ represents the spline function on the subinterval $[t_\nu, t_{\nu+1}[$ on the curve Γ given by

$$S_1(\varphi; t_0, \nu) = \frac{t_{\nu(k+1)} - t_0}{t_{\nu(k+1)} - t_{\nu k}}\varphi(t_{\nu k}) + \frac{t_0 - t_{\nu k}}{t_{\nu(k+1)} - t_{\nu k}}\varphi(t_{\nu(k+1)}),$$

with $t \in t_\sigma \widehat{t_{\sigma+1}}$ and $t_0 \in t_\nu \widehat{t_{\nu+1}}$,$\sigma, \nu = 0, 1, ...N-1$ and $k = 0, 1, ...m-1$.

Theorem 1

The function $\beta_{\sigma\nu}(\varphi; t, t_0)$ constructed above is continuous on the curve Γ and has a definite sense for the values $t_0 = t_{\sigma k}$ or $t_0 = t_{\sigma(k+1)}$.

Proof

Seeing that, the equality $t_{\sigma k} - t_0 = 0$ or $t_{\sigma(k+1)} - t_0 = 0$ is possible only when $\sigma = \nu$, in this case, we can take to the function $\beta_{\sigma\sigma}(\varphi; t, t_0)$ the following form

$$\beta_{\sigma\sigma}(\varphi; t, t_0) = U(\varphi; t, \sigma) - V(\varphi; t_0, \sigma).$$

Therefore,

$$\begin{aligned}\beta_{\sigma\sigma}(\varphi; t, t_0) &= \frac{(t_{\sigma(k+1)} - t)(t - t_0)}{(t_{\sigma(k+1)} - t_{\sigma k})(t_{\sigma k} - t_0)}(\varphi(t_{\sigma k}) - S_1(\varphi; t_0, \sigma)) \\ &+ \frac{(t - t_{\sigma k})(t - t_0)}{(t_{\sigma(k+1)} - t_{\sigma k})(t_{\sigma(k+1)} - t_0)}(\varphi(t_{\sigma(k+1)}) - S_1(\varphi; t_0, \sigma)).\end{aligned}$$

It's easily to see that

$$\beta_{\sigma\sigma}(\varphi; t, t_0) = (t - t_0)Q(\varphi; t, t_0),$$

with

$$\lim_{t_0 \to t_{\sigma k}} Q(\varphi; t, t_0) = Q(\varphi; t, t_{\sigma k}),$$

and

$$\lim_{t_0 \to t_{\sigma(k+1)}} Q(\varphi; t, t_0) = Q(\varphi; t, t_{\sigma(k+1)}).$$

Putting now the function

$$\psi_{\sigma\nu}(\varphi; t, t_0) = \begin{cases} \varphi(t_0) + \beta_{\sigma\nu}(\varphi; t, t_0),\ t \in t_\sigma \widehat{t_{\sigma+1}}\ ;\ t_0 \in t_\nu \widehat{t_{\nu+1}} \\ \sigma = 0, 1, ..., N-1\ ;\ \nu = 0, 1, ..., N-1 \end{cases} . \qquad (4)$$

After this construction, one replaces the singular integral (2)

$$F(t_0) = \frac{1}{\pi i} \int_\Gamma \frac{\varphi(t)}{t - t_0} dt$$

by the following ones

$$S(\varphi; t_0) = \frac{1}{\pi i} \int_\Gamma \frac{\psi_{\sigma\nu}(\varphi; t, t_0)}{t - t_0} dt = \varphi(t_0) + \frac{1}{\pi i} \int_\Gamma \frac{\beta_{\sigma\nu}(\varphi; t, t_0)}{t - t_0} dt. \qquad (5)$$

Theorem 2

Let Γ be a smooth contour oriented and let φ be a density satisfies the Hölder condition $H(\mu)$ then, the following estimation

$$| F(t_0) - S(\varphi; t_0) | \leq \max\left(\frac{C \ln(mN)}{(mN)^\mu}, \frac{C}{N^\mu}\right)\ \ N,\ m > 1$$

holds, where the constant C depends only of the contour Γ.

Proof

For any $t \in t_\sigma \overset{\frown}{t_{\sigma+1}}$ and $t_0 \in t_\nu \overset{\frown}{t_{\nu+1}}$, with $\sigma \neq \nu$, we can write

$$\begin{aligned} & \varphi(t) - \psi_{\sigma\nu}(\varphi; t, t_0) = \varphi(t) - \varphi(t_0) - \beta_{\sigma\nu}(\varphi; t, t_0) \\ = \quad & \varphi(t) - \varphi(t_0) - U(\varphi; t, \sigma) + V(\varphi; t_0, \nu) \end{aligned} \tag{1}$$

If $\sigma = \nu$, we can easily put our expression in the form

$$\begin{aligned} & \varphi(t) - \psi_{\sigma\sigma}(\varphi; t, t_0) = \varphi(t) - \varphi(t_0) \ - \beta_{\sigma\sigma}(\varphi; t, t_0) \\ = \quad & \varphi(t) - \varphi(t_0) - U(\varphi; t, \sigma) + V(\varphi; t_0, \sigma) \end{aligned} \tag{2}$$

Taking into account expressions (7), (8) above, we have

$$\begin{aligned} & \frac{1}{\pi i} \int_\Gamma \frac{\varphi(t) - \psi_{\sigma\nu}(\varphi; t, t_0)}{t - t_0} dt = \frac{1}{\pi i} \sum_{\sigma=0}^{N-1} \int_{t_\sigma t_{\sigma+1}} \frac{\varphi(t) - \varphi(t_0)}{t - t_0} \\ & - \frac{U(\varphi; t, \sigma) - V(\varphi; t_0, \sigma)}{t - t_0} dt. \end{aligned} \tag{3}$$

Passing now to the estimation of the expression (8), we have for $t_0 \in t_\nu \overset{\frown}{t_{\nu+1}}$ and $\sigma \neq \nu$, the relation

$$\left| \begin{array}{l} \sum_{\substack{\sigma=0 \\ \sigma \neq \nu}}^{N-1} \sum_{k=0}^{m-1} \int_{t_{\sigma k}}^{t_{\sigma(k+1)}} \dfrac{\varphi(t) - \varphi(t_0)}{t - t_0} \\ -\{ \dfrac{t_{\sigma(k+1)} - t}{t_{\sigma(k+1)} - t_{\sigma k}} \varphi(t_{\sigma k}) \dfrac{t - t_0}{t_{\sigma k} - t_0} + \dfrac{t - t_{\sigma k}}{t_{\sigma(k+1)} - t_{\sigma k}} \varphi(t_{\sigma(k+1)}) \dfrac{t - t_0}{t_{\sigma(k+1)} - t_0} \\ - \dfrac{t - t_0}{t_{\sigma(k+1)} - t_{\sigma k}} (\dfrac{t_{\sigma(k+1)} - t}{t_{\sigma k} - t_0} + \dfrac{t - t_{\sigma k}}{t_{\sigma(k+1)} - t_0}) S_1(\varphi; t_0, \nu) \} \dfrac{1}{t - t_0} dt. \end{array} \right| = O(\frac{\ln mN}{m^\mu N^\mu}).$$

Naturally, the estimation given above is obtained with the help of which using expressions

$$\begin{aligned} \frac{\mid t_{\sigma(k+1)} - t \mid}{\mid t_{\sigma(k+1)} - t_{\sigma k} \mid} \quad & \mid \quad \simeq \frac{\mid t - t_{\sigma k} \mid}{\mid t_{\sigma(k+1)} - t_{\sigma k} \mid} = O(1), \\ \frac{\mid t_0 - t_{\nu k} \mid}{\mid t_{\nu(k+1)} - t_{\nu k} \mid} \quad & \mid \quad \simeq \frac{\mid t_{\nu(k+1)} - t_0 \mid}{\mid t_{\nu(k+1)} - t_{\nu k} \mid} = O(1), 9 \end{aligned} \tag{4}$$

and the density φ as an element of the Hölder space $H(\mu)$[2].

Besides, it is easy to see that

$$\max_{t_0 \in t_\nu \overset{\frown}{t_{\nu+1}}} \mid \sum_{\substack{\sigma=0 \\ \sigma \neq \nu}}^{N-1} \sum_{k=0}^{m-1} \int_{t_{\sigma k}}^{t_{\sigma(k+1)}} \frac{\varphi(t) - \varphi(t_0)}{t - t_0} \mid = O(\frac{\ln mN}{m^\mu N^\mu})$$

and also, one can estimate the expression

$$\left| \begin{array}{l} \Sigma_{\substack{\sigma=0 \\ \sigma\neq\nu}}^{N-1} \int_{t_\sigma}^{t_{\sigma+1}} \sum_{k=0}^{m-1} -\{\frac{t_{\sigma(k+1)}-t}{t_{\sigma(k+1)}-t_{\sigma k}}\varphi(t_{\sigma k})\frac{t-t_0}{t_{\sigma k}-t_0}+\frac{t-t_{\sigma k}}{t_{\sigma(k+1)}-t_{\sigma k}} \\ \varphi(t_{\sigma(k+1)})\frac{t-t_0}{t_{\sigma(k+1)}-t_0}-\frac{t-t_0}{t_{\sigma(k+1)}-t_{\sigma k}}(\frac{t_{\sigma(k+1)}-t}{t_{\sigma k}-t_0}+\frac{t-t_{\sigma k}}{t_{\sigma(k+1)}-t_0}) \\ S_1(\varphi;t_0,\nu)\}\frac{1}{t-t_0}dt. \end{array} \right|$$

$$\simeq \left| \sum_{\substack{\sigma=0 \\ \sigma\neq\nu}}^{N-1} \sum_{k=0}^{m-1} \int_{t_{\sigma k}}^{t_{\sigma(k+1)}} \frac{\varphi(t_{\nu k})-\varphi(t_{\sigma k})}{t_{\nu k}-t_{\sigma k}}+\frac{\varphi(t_{\nu(k+1)})-\varphi(t_{\sigma(k+1)})}{t_{\nu(k+1)}-t_{\sigma(k+1)}}dt \right| = O(\frac{\ln mN}{m^\mu N^\mu}).$$

Further, for the case where $\sigma=\nu$, using again the condition $\varphi\in H(\mu)$ and the condition of smoothness of Γ, we obtain

$$|\int_{t_\nu t_{\nu+1}} \frac{\varphi(t)-\varphi(t_0)}{t-t_0}dt| \leq A\int_{s_\nu}^{s_{\nu+1}} |s-s_0|^{\mu-1}\,ds = O(N^{-\mu})$$

3 Numerical experiments

Using our approximation, we apply the algorithms to singular integrals and we present results concerning the accuracy of the calculations, in this numerical experiments each table I represents the exact principal value of the singular integral and $\widetilde{I}$ corresponds to the approximate calculation produced by our approximation at points values interpolation.

Example
Consider the singular integral,

$$I = F(t_0) = \frac{1}{\pi i}\int_\Gamma \frac{\varphi(t)}{t-t_0}dt,$$

where the curve Γ designate the unit circle and the function density φ is given by the following expression

$$\varphi(t) = \frac{-2t^2+8t+12}{4t(t^2-t-6)}.$$

N	$\| I-\widetilde{I}\|_1$	$\| I-\widetilde{I}\|_2$	$\| I-\widetilde{I}\|_\infty$
20	1.8246599E - 02	9.1665657E - 03	5.0822943E - 03
40	3.6852972E - 03	1.9270432E - 03	1.5697196E - 03
60	2.3687426E - 03	1.1880117E - 03	6.6070486E - 04

Note Many examples confirm the efficiency of this approximation.

References

[1] J. Antidze, *On the approximate solution of singular integral equations*, Seminar of Institute of Applied Mathematics, 1975, Tbilissi.

[2] N.I. Muskhelishvili, Singular integral equations, Naukah Moscow, 1968, English transl, of 1sted Noordho, 1953; reprint,1972.

[3] M.Nadir, Problèmes aux limites qui se reduisent aux equations intégrales de Fredholm, Seminaire de l'Institut de Mathématiques et Informatique, 1985, Annaba.

[4] M. Nadir, Opérateurs intégraux et bases d'ondelettes, Far East J.Sci. 6(6)(1998), 977-995.

[5] M. Nadir,J. Antidze, On the numerical solution of singular integral equations using Sanikidze's approximation, Comp Math in Sc Tech. 10(1), 83-89 (2004).

[6] M. Nadir, On the approximation of singular integrals of Cauchy types, in Proceedinng Dynamic System and Applications (2004)

[7] M.Nadir, B.Lakehali, An approximation for singular integrals of Cauchy types, in Advance in algebra and analysis (AAA) (1),1, 2006.

[8] J. Sanikidze, Approximate solution of singular integral equations in the case of closed contours of integration, Seminar of Institute of Applied Mathematics, 1971, Tbilissi.

SECTION E

DIFFERENTIAL AND INTEGRAL OPERATORS & EQUATIONS

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On the Concept of Stability of Functional Differential Equations

Haydar Akça
United Arab Emirates University, Faculty of Sciences
Department of Mathematics, P. O. Box 17551, Al Ain, UAE
e-mail: hakca@uaeu.ac.ae

Abstract

We consider different forms of functional differential equations and provide stability definitions and conditions for the solution of the respective equations and discuses an approach to the problem of stability and asymptotic behavior of retarded differential equations

Keywords: Diferential equations, Cauchy operator, stability.

1 Introduction and Definitions

Differential equations are used to model the movement of a certain physical system or experiment. If the behavior of a system or experiment is stable, then a small change in initial data will result in a small change in the behavior for future time. Therefore, by the a statement that "a solution ϕ of a differential equation is stable" we mean that other solutions with initial data close to the solution ϕ will remain close to ϕ for future time.

Stability properties for the linear differential equations considerably simpler than nonlinear differential equations. In the theory of ordinary differential equations the stability by Lyapunov means the continuous (in the matric of space C or Chebyshev metric) dependence of the solution to the Cauchy problem [3]

$$\mathcal{L}x = f, \quad x(a) = \alpha \tag{1.1}$$

in the space D_{loc} which is the linear space of the functions $x : [a, \infty) \to \Re^n$(space of real column vectors), absolutely continuous on every finite interval, i.e. when x is locally absolutely continuous, $\dot{x}$ is locally summable, for every $f \in L_{loc}$ which is a linear space of the classes of the equivalence of the measurable and locally summable functions $z : [a, \infty) \to \Re^n$ and every $\alpha \in \Re^n$. Under these assumptions there exists a $n \times n$ matrix X whose columns are the solutions of semi homogeneous problems

$$\mathcal{L}x = 0, \quad x(a) = E_i$$

ISSN-1584-5990

where the vectors $E_1, E_2, \ldots, E_n$ form the columns of the identity matrix E. For a wide class of equations

$$(\mathcal{L}x)(t) = f(t), \quad t \geq a \tag{1.2}$$

the Cauchy operator is an integral Volterra operator

$$(\mathcal{C}f)(t) = \int_a^t C(t,s)f(s)ds$$

The kernel $C(t,s)$ in this representation of the Cauchy operator is called the *Cauchy matrix* of equation (1.1). The Cauchy formula implies that all asymptotic properties of the solutions to the linear equation, particularly the stability properties, are determined by the properties of the Cauchy matrix $C(t,s)$ and of the fundamental matrix $X(t)$. Generally speaking, differential equations $\mathcal{L}x = f$ may have similar fundamental matrices, and different Cauchy matrices, or similar Cauchy matrices but different fundamental matrices.

In order to formulate the conditions ensuring unique solvability of problem (1.1) and integral representation of the Cauchy operator introducing the Banach space $L[a,b]$ of measurable and summable by Lebesgue functions $z : [a,b] \to \Re^n$ with the norm

$$\|z\|_{L[a,b]} \stackrel{\text{def}}{=} \int_a^t |z(s)|ds,$$

and $D[a,b]$ is the Banach space of the absolutely continuous functions $x : [a,b] \to \Re^n$ with the norm

$$\|x\|_{D[a,b]} \stackrel{\text{def}}{=} \|\dot{x}\|_{L[a,b]} + |x(a)|$$

The solvability of equation (problem) means the existence of at least one solution to the equation. The operator $\mathcal{L} : D_{loc} \to L_{loc}$ is a linear Volterra operator, a natural definition of the solution to the equation $\mathcal{L}x = f$ on the segment $[a,b]$ defined as $\mathcal{L}_b : D[a,b] \to L[a,b]$ for the operator $\mathcal{L}$. For an arbitrary function $x \in D[a,b]$ setting

$$(\mathcal{L}_b x)(t) \stackrel{\text{def}}{=} (\mathcal{L}y_x)(t)$$

almost everywhere on $[a,b]$, where $y_x \in D_{loc}$ and $y_x(t) = x(t)$ everywhere on $[a,b]$. The function $x \in D[a,b]$ for which $(\mathcal{L}y_x)(t) = f(t)$ is called a solution of equation $\mathcal{L}x = f$ almost everywhere on $[a,b]$. The operator $\mathcal{L} : D_{loc} \to L_{loc}$ is Volterra and $x : [a,b] \to \Re^n$. The solutions of various functional differential equations sometimes have similar properties, and vise versa equations of similar form sometimes possess solutions with different properties. This is because of specific features of the operator $\mathcal{L}$. Thus, under the condition of unique solvability of problem (1.1), and by virtue of linearity of the operator $\mathcal{L}$, the general solution of the equation (1.2) has the form

$$x(t) = (\mathcal{C}f)(t) + X(t)x(a)$$

This representation is called the *Cauchy formula* , the matrix X is the *fundamental matrix* and the operator $\mathcal{C} : L_{loc} \to D_{loc}$ is the *Cauchy operator* of the equation (1.2).

Theorem 1.1 [3] Let for every $b > a$ the operator $\mathcal{L}_b : D[a,b] \to L[a,b]$ be bounded and the problem

$$\mathcal{L}_b x = f,$$
$$x(a) = \alpha$$

be uniquely solvable for ever pair $\{f, \alpha\} \in L[a,b] \times \Re^n$. Then the general solution of the equation (1.2) has the representation

$$x(t) = x(a) + \int_a^t C(t,s) f(s) ds$$

The study of stabilities for nonlinear differential equations is difficult because formulas of solutions and eigenvalues are generally not available or applicable. Dealing with the stability of nonlinear differential equation Liapunov method is one of the important and useful tool. The study of stability properties of autonomous differential equations in $\Re^2$ can be extend to general differential equations in $\Re^n$.

Consider differential equation

$$\dot{x}(t) = f(t, x(t)), \tag{1.3}$$

in the space $D = [0,\infty) \times Q$ where $Q \subset \Re^n$ is a domain containing the zero vector. For any $(t_0, x_0) \in D = [0,\infty) \times Q$, equation has a unique solution $x(t, t_0, x_0)$ existing on $[t_0, \infty)$ with $x(t_0) = x_0$.

First to guarantee existence and uniqueness of solutions. Therefore in $\Re^2$ Liapunov functions can be define in the form

$$V(t) = \frac{1}{2}[x_1^2(t) + x_2^2(t)] \tag{1.4}$$

The function V is related to the norm $r(x) = \sqrt{x_1^2 + x_2^2}$ of a solution $(x_1(t), x_2(t))$, that is the difference from $(x_1(t), x_2(t))$ to the origin $(0,0)$. Consequently

$$V'(t) \leq \alpha V(t) = -[-\alpha V(t)], \quad \alpha < 0$$

which enables us to verify that $V(t) \to 0, \quad t \to \infty$.

Regarding stabilities in the sense of Liapunov we have following definitions and theorems.

Definition 1. Let $Q \subset \Re^n$ be a domain containing the zero vector. A continuous function $V : Q \to [0,\infty)$ is called positive definite if $V(x) > 0$, for $x \neq 0$.

Definition 2. Let $Q \subset \Re^n$ be domain containing the zero vector. A function $V : Q \to [0,\infty)$ is called a Liapunov function if $V(0) = 0$, V is called positive definite and has continuous first partial derivatives. The solutions go to the origin $(0,0)$

or the origin $\phi = (0, 0)$ is asymptotically stable [4]. The most important aspect of this approach is that it is done without solving the differential equations explicitly. The method is introduced by Liapunov (1892) and Poincare (1892). Their ideas continue to inspire new research in the area of differential equations and other related areas.

Definition 3 [4]. Let $\phi(t) = \phi(t, t_\phi)$ be a solution of Equation (1.3) on the interval $[t_\phi, \infty), t_\phi \geq 0$. Then;

(**i**). $\phi(t) = \phi(t, t_\phi)$ is said to be *stable* if for any $t_0 \geq t_\phi$ and any $\varepsilon > 0$ there exist a $\delta = \delta(\varepsilon, t_0) > 0$, typically $\delta(\varepsilon, t_0) \leq \varepsilon$, such that $|x_0 - \phi(t_0)| \leq \delta$ implies $|x(t, t_0, x_0) - \phi(t)| \leq \varepsilon$ for $t \geq t_0$.

(**ii**). $\phi(t, t_\phi)$ is said to be *uniformly stable* if it is stable and δ in the definition of "stable" can be chosen to be independent of $t_0 \geq t_\phi$. That is for any $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0, \delta(\varepsilon) \leq \varepsilon)$ such that $t_0 \geq t_\phi$ and $|x_0 - \phi(t_0)| \leq \delta$ imply $x(t, t_0, x_0) - \phi(t)| \leq \varepsilon$ for $t \geq t_0$.

(**iii**). $\phi(t, t_\phi)$ is said to be *asymptotically stable* if it is stable and in addition, for any $t_0 \geq t_\phi$, there exists an $r(t_0) > 0$ such that $|x_0 - \phi(t_0)| \leq r(t_0)$ implies $\lim_{t\to\infty} |x(t, t_0, x_0) - \phi(t)| = 0$.

(**iv**). $\phi(t, t_\phi)$ is said to be *uniformly asymptotically stable* if it is uniformally stable and in addition, there exists an $r > 0$ independent of $t_0 \geq t_\phi$, such that $|x_0 - \phi(t_0)| \leq r$ implies $\lim_{t\to\infty} |x(t, t_0, x_0) - \phi(t)| = 0$ uniformly for $t_0 \geq t_\phi$ in the following sense: For any $\varepsilon > 0$, there exist a $T = T(\varepsilon) > 0$ such that $\{t_0 \geq t_\phi, |x_0 - \phi(t_0)| \leq r, t \geq t_0 + T\}$ imply $|x(t, t_0, x_0) - \phi(t)| \leq \varepsilon$.

(**v**). $\phi(t, t_\phi)$ is said to be *unstable* if it is not stable.

(**vi**). In particular, if $\phi(t) = 0, t \geq 0$, is a solution of Equation (1.1) , or equivalently when $f(t, 0) = 0, t \geq 0$, then the above given the corresponding definitions concerning stability properties for the zero solution $\phi = 0$.

Theorem 1.2 [4] Let $Q \subset \Re^n$ be a domain containing the zero vector. Consider equation

$$x'(t) = f(x(t)), \quad \text{or} \quad x' = f(x)$$

on $[0, \infty) \times Q$ with $f(0) = 0$ so that $\phi = 0$ is a solution of equation. Assume that V is a Liapunov function. Then

(i). If $V'(x) \leq 0$, then $\phi = 0$ is uniformly stable.

(ii). If $V'(x) < 0, x \neq 0$, or $-V'(x)$ is positive definite, then $\phi = 0$ is uniformly asymptotically stable.

(iii). If $V'(x) > 0, x \neq 0$, then $\phi = 0$ is unstable.

This theorem presents a brief coverage of the Liapunov theory concerning stability properties for autonomous differential equations. Roughly speaking, it reduces the study of stability properties to the problem of constructing appropriate Liapunov functions. However, constructing these functions is not an easy task; it requires a great deal of experience and skill.

Consider general linear differential equations of the form

$$\dot{x}(t) = A(t)x(t) + f(t), \qquad x(t_0) = x_0, \quad t \geq t_0 \geq 0, \quad x \in \Re^n, \tag{1.5}$$

where $A(t)$ and $f(t)$ are continuous on $\Re^+ = [0, \infty)$. The unique solution of equation (1.5) is

$$x(t) = U(t, t_0)x_0 + \int_{t_0}^{t} U(t, s)f(s)ds, \quad t \geq t_0 \tag{1.6}$$

where the matrix $U(t, t_0)$ is the fundamental matrix solution of Equation (1.5) when $f = 0$.

Theorem 1.3 [4] Assume that $A(t)$ is continuous on $\Re^n$ and let U be the fundamental matrix solution of

$$\dot{x}(t) = A(t)x(t) \tag{1.7}$$

where the unique solution is given by $x(t) = U(t, t_0)x_0$. The zero solution $\varphi = 0$ of Equation (1.7) is
(**i**). Stable if and only if there is an (independent or generic) constant $C > 1$ such that

$$|U(t, 0)| \leq C, \quad 0 \leq t < \infty$$

(**ii**). Uniformly stable if and only if there is an (independent or generic) constant $C > 1$ such that

$$|U(t, s)| \leq C, \quad 0 \leq s \leq t < \infty$$

(**iii**). Asymptotically stable if and only if

$$|U(t, 0)| \to 0, \quad t \to \infty$$

(**iv**). Uniformly asymptotically (also called exponentially) stable if and only if there are (independent or generic) constants $C > 1$ and $\alpha > 0$ such that

$$|U(t, s)| \leq Ce^{-\alpha(t-s)}, \qquad 0 \leq s \leq t < \infty$$

Theorem 1.4 [4] Assume that $A(t)$ and $f(t)$ are continuous on $\Re^+$. The zero solution of Equation $x'(t) = A(t)x(t)$ is stable if and only if every solution of Equation (1.5) is stable. The same statement is true for uniform stability, asymptotic stability, and uniform asymptotic stability [4].

Theorem 1.5 [4] Assume that $A(t)$ is continuous on $\Re^+$. If the zero solution of equation (1.7) is uniformly stable, and if the $n \times n$ continuous matrix function $B(t)$ satisfies $\int_0^\infty \|B(t)\|dt < \infty$, then the zero solution of

$$x'(t) = A(t)x(t) + B(t)x(t) = \big[A(t) + B(t)\big]x(t) \tag{1.8}$$

is also uniformly stable.

Theorem 1.6 Assume that $A(t)$ is continuous on $\Re^+$. If the zero solution of equation (1.7) is uniformly asymptotically stable, and if the $n \times n$ continuous matrix function $B(t)$ satisfies

$$\int_{t_0}^{t} \big|B(s)\big|ds \leq m(t - t_0) + r, \quad t \geq t_0 \geq 0 \tag{1.9}$$

for some positive constants m and r, then there is an $m_0 > 0$ such that if $m \leq m_0$, the zero solution of equation (1.8) is also uniformly asymptotically stable [4].

2 Main Results, A New Conception of Stability

The modern theory of functional differential equations [1] – [2] are devoted to the modern theory of the equation

$$\dot{x} = Fx \tag{2.1}$$

with the operator F defined on a set of absolutely continuous functions. This equation is a natural generalization of the differential equation of the form:

$$(Fx)(t) = f(t, x(t))$$

We will discuss an approach to the problem of stability and asymptotic behavior of solutions of differential and delay differential equations. Consider the equation

$$(\mathcal{L}x)(t) \stackrel{\text{def}}{=} \dot{x}(t) - \int_0^t x(s)R(t,s)ds = f(t), \quad x(t) \in \Re^n \tag{2.2}$$

Under natural assumptions the "Cauchy formula"

$$x(t) = \int_0^t C(t,s)f(s)d(s) + C(t,0)x(0) \tag{2.3}$$

describes the general solution of this equation. In the case of the differential equation

$$\dot{x}(t) - P(t)x(t) = f(t)$$

the well known equality

$$C(t,s) = X(t)X^{-1}(s) = C(t,\tau)C(\tau,s), \quad (s \leq \tau \leq t)$$

where X is the fundamental matrix of solutions of the homogeneous equation. The Cauchy formula (2.3) directly follows from the fact that there is a one-to-one mapping

$$x(t) = \int_0^t z(s)d(s), \quad z = \dot{x},$$

between the solutions x of the Cauchy problem $\mathcal{L}x = f, \quad x(0) = 0$ and the solutions z of the classical Volterra integral equation

$$z(t) = \int_0^t R(t,s)z(s)d(s) + f(t) \tag{2.4}$$

namely,

$$C(t,s) = E + \int_s^t H(\tau,s)d\tau,$$

where E is the identity $n \times n$ matrix, $H(\tau, s)$ is the kernel at the representation

$$z(t) = f(t) + \int_0^t H(t,s)f(s)ds$$

of the solution (2.4). Let B be a fixed linear space of (locally summable) functions and $z : [0, \infty) \to \Re^n$. Then all the solutions

$$x(t) = \int_0^t C(t,s)f(s)ds + C(t,0)\alpha, \quad \{f, \alpha\} \in B \times \Re^n$$

of the equation $\mathcal{L}x = f, \quad f \in B$ constitute a linear space $D(\mathcal{L}, B)$. Let B be Banach space, then $D(\mathcal{L}, B)$ is also Banach under the norm

$$\|x\|_D = \|\mathcal{L}x\|_B + \|x(0)\|_{\Re^n}.$$

It is relevant to point out that the space $D(\mathcal{L}, B)$ is one and the same for wide class of equations in the case of B is fixed.

Let $\mathcal{L}_0 x = z$ be a "model" equation with the Cauchy matrix $C_0(t, s)$ in explicite form. Then we know some asymptotic property of the elements of $D(L_0, B)$. For instance, if $B = L_\infty$ (the space of measurable and bounded) and $z : [0, \infty) \to \Re^n, \quad \|z\|_{L_\infty} = vraisup_{t>0}\|z(t)\|_{\Re^n}$. Then

$$(\mathcal{L}_0 x)(t) = \dot{x}(t) + Ex(t) = f(t)$$

Thus $x \in D(\mathcal{L}_0, L_\infty)$ has the representation

$$x(t) = \int_0^t e^{-(t-s)}z(s)ds + e^{-t}\alpha, \quad \{z, \alpha\} \in L_\infty \times \Re^n$$

and consequently

$$\sup_{t>0} \|x(t)\|_{\Re^n} < \infty$$

In the case of

$$B = L_\infty^\gamma, \quad 0 < \gamma < 1, \quad z \in L_\infty^\gamma : z(t) = y(t)e^{-\gamma t}, y \in L_\infty)$$

each $x \in D(\dot{x} + x, L_\infty^\gamma)$ has the property

$$\|x(t)\|_{\Re^n} \leq N_z e^{-\gamma t}$$

If we let $B = L_\infty$, the systems of equations

$$(\mathcal{L}_0 x)(t) \stackrel{\text{def}}{=} \left\{ \begin{array}{c} \dot{x}_1(t) \\ \dot{x}_2(t) \end{array} \right\} + \left\{ \begin{array}{cc} \nu & 0 \\ 0 & \mu \end{array} \right\} \left\{ \begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right\} = \left\{ \begin{array}{c} z_1(t) \\ z_2(t) \end{array} \right\} \tag{2.5}$$

has solutions in the form

$$x_1(t) = \int_0^t e^{-\nu(t-s)}z_1(s)ds + \alpha_1 e^{-\nu t}$$

$$x_2(t) = \int_0^t e^{-\mu(t-s)} z_2(s)ds + \alpha_1 e^{-\mu^t}$$

where $|x_1(t)| < \infty$ if $\nu > 0, |x_2(t)| < \infty$ if $\mu < 0$. Now let try to rewrite delay differential equation in the form $\mathcal{L}x = f$. To make matter explicit, it is sufficient to consider following simple example

$$\begin{array}{l} \dot{x}(t) + p(t)x[h(t)] = v(t), \quad h(t) \leq t, \quad t \geq 0 \\ x(\xi) = \varphi(\xi), \quad \text{if} \quad \xi < 0 \end{array} \tag{2.6}$$

Since we consider the unknown x on the interval $[0, \infty)$. The "initial function" φ defines the value of the superposition of $x[h(t)]$ for $h(t) < 0$. Denoting the linear operator S_h by

$$S_h(x) = \begin{cases} x[h(t)], \text{ if } h(t) \geq 0, \\ 0, \text{ if } h(t) < 0, \end{cases} \tag{2.7}$$

Let

$$\varphi^h = \begin{cases} 0, \text{ if } h(t) \geq 0, \\ \varphi\left[x(t)\right], \text{ if } h(t) < 0, \end{cases} \tag{2.8}$$

Then using these definitions (2.6) takes the form

$$(\mathcal{L}x)(t) \stackrel{\text{def}}{=} \dot{x}(t) + \int_0^t x(s)R(t,s)ds = f(t),$$

where $R(t,s) = -p(t)X(t,s)$, where $X(t,s)$ is the characteristic function of the point set $\{t,s\}$ such that $0 \leq s \leq h(t)$.

Let $\mathcal{L}_0 x = z$ be a model equation. Denote

$$(Wz)(t) = \int_0^t C_0(t,s)z(s)ds$$

and let $\mathcal{L}x = f$ be an equation with unknown $C(t,s)$.Then the following assertion gives a criterion of coincidence of the known set $D(\mathcal{L}_0, B)$ and unknown one $D(\mathcal{L}, B)$.

Theorem 2.1 Let $\mathcal{L} : D(\mathcal{L}_0, B) \to B$ be bounded. The equality

$$D(\mathcal{L}_0, B) = D(\mathcal{L}, B)$$

holds if and only if there exists bounded inverse $[\mathcal{L}W]^{-1} : B \to B$.

Define the "Cauchy operator" $C : B \to D(\mathcal{L}_0, B)$ by

$$(Cf)(t) = \int_0^t C(t,s)f(s)ds$$

and the finite dimensional $X : \Re^n \to D(\mathcal{L}_0, B)$ by $(X\alpha)(t) = C(t,0)\alpha$. Because of the assumptions of the theorem these operators are bounded. From the definition of the equation there exists a special case in the form:

$$B = L_\infty, \quad (\mathcal{L}_0 x)(t) \stackrel{\text{def}}{=} \dot{x} + E\omega x(t), \quad \omega > 0$$

The inevertibility of $\mathcal{L}W : L_\infty \to L_\infty$ guarantees the bound of each solution x by virtue of the theorem for the equation

$$\mathcal{L}x = f, \quad f \in L_\infty, \quad (\sup_{t>0} \|x(t)\|_{\Re^n} < \infty)$$

As an example consider

$$\mathcal{L}x(t) \stackrel{\text{def}}{=} \dot{x}(t) + P(t)x(t)$$

the columns of P belong to L_∞ . Since

$$\dot{x} + Px \equiv \dot{x} + E\omega x + Px - E\omega x$$

$$\mathcal{L}Wz = z - (P - E\omega)Wz \stackrel{\text{def}}{=} z - \Omega z$$

The estimation $\|\Omega\|_{L_\infty} < 1$ guarantees the invertibility of $\mathcal{L}W : L_\infty \to L_\infty$. Thus the inequality

$$\|\Omega\|_{L_\infty \to L_\infty} \leq vrai \sup_{t>0} \|P(t) - E\omega\|_{\Re^n} \frac{1}{\omega} < 1$$

where there exist a $\omega > 0$ such that

$$vrai \lim_{t\to\infty} \|P(t) - E\omega\|_{\Re^n} < \omega$$

which is guarantees exponential stability of $\mathcal{L}x = f$ and the estimate

$$\|C(t,s)\|_{\Re^n} \leq Ne^{\gamma(t-s)}$$

as a result of Bohl-Perron Theorem, assuming $B = L_\infty$ and $\mathcal{L}_0$ be defined by (2.5) then the invertibility of $\mathcal{L}W : L_\infty \to L_\infty$ guarantees the stability of the system $\mathcal{L}x = f$ in respect to the first component of x_1 .

Example 2.1 Let $B = L_\infty$,

$$(\mathcal{L}_0 x)(t) \stackrel{\text{def}}{=} \dot{x}(t) - \frac{2t}{t^2+1}x(t)$$

The invertibility of $\mathcal{L}W : L_\infty \to L_\infty$ guarantees the estimate
$\|x(t)\|_{\Re^n} < N(t^2+1)$ for $x \in D(\mathcal{L}, L_\infty)$.
In particular if $n = 1$ we have

$$(\mathcal{L}x)(t) \stackrel{\text{def}}{=} \begin{cases} \dot{x}(t) - \frac{2t}{t^2+1}x(t) - p(t)x[h(t)] \\ x(\xi) = 0, \text{ if } \xi < 0 \end{cases}$$

where $p(t) \geq 0, \quad p(t)\sigma(t)[h^2(t)+1] \arctan h(t) < 0$ and

$$\sigma(t) = \begin{cases} 1, & \text{if } h(t) \geq 0 \\ 0, & \text{if } h(t) < 0 \end{cases}$$

guarantees the invertibility of $\mathcal{L}W$. We can point as a conclusion that the operator $\mathcal{L}W$ has explicit form and the invertibility of $\mathcal{L}W$ is sufficient and necessary for $D(\mathcal{L}_0, B) = D(\mathcal{L}, B)$.

Sometimes it can be useful to replace L_∞ by the "weighted " space $\mathcal{L}^\gamma_\infty$ in such away that $z \in \mathcal{L}^\gamma_\infty$ if $z(t) = e^{\gamma t}y(t), \quad y \in L_\infty$.

Now next let consider quasi-linear equation $\mathcal{L}x = Fx$ with nonlinear Volterra form F.

Theorem 2.2 Assume that $D(\mathcal{L}_0, B) = D(\mathcal{L}, B)$ and $\mathcal{L} : D(\mathcal{L}_0, B) \to B$ be bounded. Let further $F(0) = 0$ and for each $k > 0$ there exists $\delta > 0$ such that

$$\|Fx_2 - Fx_1\|_B \leq k\|x_2 - x_1\|_{D(\mathcal{L}_0,B)}$$

for all $x_1, x_2 \in D(\mathcal{L}_0, B)$ such that $\|x_i\|_{D(\mathcal{L}_0,B)} < \delta, \quad i = 1, 2$. Then the solution of the Cauchy problem $\mathcal{L}x = Fx, \quad x(0) = \alpha$ belongs to $D(\mathcal{L}_0, B)$ for some small $\|\alpha\|_{\Re^n}$.

Example 2.2 Following equations are exponentially stable if $\nu > 0$.

$$\dot{x} + \nu x = \sin \dot{x}^2$$

$$\begin{cases} \dot{x}(t) + \nu x(t) = a(t)\dot{x}^2(t - \tau_1)) + b(t)\dot{x}(t - \tau_2(t))x(t - \tau_3(t)) + c(t)x^2(t - \tau_4(t)), \\ \dot{x}(\xi) = 0 \text{ if } \xi < 0 \end{cases}$$

where $a, b, c \in L_\infty, \quad \tau_i > 0,$ constant $, i = 1, 2, 3, 4$

Acknowledgment This work is dedicated to memory of Nikolai Viktorovich Azbelev. Author gratefully acknowledges that part of the work is come out as a result of personnel communication with Prof. Azbelev earlier in the last couple of years.

References

[1] Azbelev, N. V., Rakhmatullina L., and Maksimov V., Introduction to The Theory of Linear Functional Differential Equations, World Federation Publishers Inc. 1996.

[2] Azbelev, N. V., and Rakhmatullina L., Theory of Abstract FDE and Applications, Memories on Differential Equations Vol. 8, Tbilisi, 1996.

[3] Azbelev N. V. and Simonov P. M., Stability of Differential Equations with Aftereffect, CRC Press, 2003

[4] Liu Hetao J., A First Course in the Qualitative Theory of Differential Equations, Pearson Education, Inc. New Jersey, 2003.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Stability Analysis for a new Difference Scheme Generated by $A(t)$

Allaberen Ashyralyev [1] and Mehmet Emir Koksal [2,3]
[1] Department of Mathematics, Fatih Univ., Buyukcekmece, Istanbul, Turkey
[2] Department of Computer Engineering, Halic Univ., Bomonti Istanbul, Turkey
[3] Department of Mathematics, Gebze Inst. of Technology Gebze, Kocaeli, Turkey
E-mail: mekoksal@halic.edu.tr

Abstract

The initial-value problem for the hyperbolic equation $\frac{d^2u(t)}{dt^2}+A(t)u(t)=f(t) \quad (0\leq t\leq T), u(0)=\varphi, u'(0)=\psi$ in a Hilbert space H with the self-adjoint positive definite operators $A(t)$ is considered. A new second order accurate absolutely stable difference scheme generated by $A(t)$ for approximately solving this abstract initial-value problem is developed. The stability estimates for the solution of this difference scheme and its first and second order difference derivatives are presented.

Keywords: Abstract hyperbolic equation; Initial-value problem; Difference scheme; Stability.

1 Introduction

We consider the difference scheme for the initial value problem

$$\begin{cases} \frac{d^2u(t)}{dt^2}+A(t)u(t)=f(t) \quad (0\leq t\leq T), \\ u(0)=\varphi, u'(0)=\psi \end{cases} \tag{1}$$

where $A(t)$ is the self-adjoint positive definite operator in a Hilbert space H with a t-independent domain $D=D(A(t)):\ A(t)\geq \delta I>0$.

There are many researches on the stability of difference schemes (see [1, 1-8]) for the approximate solution of the problem (1). As a rule, they are based on the assumption that the magnitudes of the grid steps τ and h with respect to the time and space variables are connected.

ISSN-1584-5990

Without using any assumptions with respect to the grid steps τ and h, there are some researches (see [9, 9-14]) on the stability of difference schemes for the approximately solving (1).

For the approximately solving (1), the first order accurate difference scheme

$$\begin{cases} \tau^{-2}(u_{k+1}-2u_k+u_{k-1})+A_k u_{k+1}=f_k, \\ A_k=A(t_k),\ f_k=f(t_k),\ t_k=k\tau,\ 1\le k\le N-1,\ N\tau=1, \\ \tau^{-1}(u_1-u_0)=\psi,\ u_0=\varphi \end{cases}$$

generated by the integer power of $A(t)$ was considered in [9].

In the case $A(t)=A$, the first and two types of the second order accurate difference schemes, generated by integer powers of A were considered in [10]. The stability estimates for the solutions of these difference schemes were established. After this, for the approximately solving (1), the second order accurate difference schemes were investigated in [13, 14]. Unfortunately, these difference schemes are generated by the square root of $A(t)$. For the practical realization of these difference schemes, the operator $A^{1/2}(t)$ is not used. Though the theoretical results are correct, the good results can not be obtained everytime. Therefore, we prefer to study the difference scheme generated by integer powers of $A(t)$.

In the present paper, a new second-order accurate difference scheme generated by integer power of $A(t)$ for approximately solving problem (1) is developed. The stability estimates for the solution of this difference scheme and for the first and second-order difference derivatives are presented.

2 Difference scheme, Stability estimates

Applying the formulas

$$\frac{u(t_{k+1})-2u(t_k)+u(t_{k-1})}{\tau^2}-u''(t_k)=o(\tau^2),$$

$$u(t_k)-\frac{u(t_{k+1})+2u(t_k)+u(t_{k-1})}{4}=o(\tau^2)$$

and equation

$$u''(t_k)=-A(t_k)u(t_k)+f(t_k),$$

we obtain

$$\frac{u(t_{k+1})-2u(t_k)+u(t_{k-1})}{\tau^2}+\frac{1}{4}A(t_k)(u(t_{k+1})+2u(t_k)+u(t_{k-1}))$$
$$=f(t_k)+o(\tau^2).$$

Moreover, we have that

$$(I+\tau^2A(0))\frac{u(\tau)-u(0)}{\tau}=\frac{\tau}{2}(-A(0)u(0)+f(0))+\psi+o(\tau^2).$$

Neglecting small terms $o(\tau^2)$, we obtain the following difference scheme

$$\begin{cases} \frac{u_{k+1}-2u_k+u_{k-1}}{\tau^2}+\frac{1}{2}A_k u_k+\frac{1}{4}A_k\left(u_{k+1}+u_{k-1}\right)=f_k, \\ A_k=A\left(t_k\right),\ f_k=f(t_k),\ t_k=k\tau, 1\le k\le N-1,\ N\tau=T, \\ (I+\tau^2A_0)\tau^{-1}(u_1-u_0)=\frac{\tau}{2}(f_0\ -A_0u_0)+\psi,\ f_0\ =\ f(0),\ u_0\ =\ \varphi. \end{cases} \tag{2}$$

Theorem 2.1. *Let $u(0)\in D(A^{\frac{1}{2}}(0))$. Then, for the solution of the difference scheme (2.1), the stability estimate*

$$\max_{0\le k\le N-1}\left\|\frac{u_{k+1}-u_k}{\tau}\right\|_H+\max_{0\le k\le N-1}\left\|\tau A_k\frac{u_{k+1}+u_k}{2}\right\|_H+\max_{0\le k\le N}\|u_k\|_H$$
$$\le C_1\left[||A^{\frac{1}{2}}(0)u_0||_H+||u_0'||_H+\sum_{s=0}^{N-1}||f_s||_H\tau\right] \tag{3}$$

holds, where C_1 does not depend on $u_0, u_0', f_s\ (0\le s\le N-1)$ and τ.

Theorem 2.2. *Let $u(0)\in D(A(0)), u'(0)\in D(A^{1/2}(0))$. Then, for the solution of the difference scheme (2.1), the stability estimate*

$$\max_{1\le k\le N-1}\left\|\tau^{-2}\left(u_{k+1}-2u_k+u_{k-1}\right)\right\|_H+\max_{1\le k\le N-1}||4^{-1}A_k\left(u_{k+1}+2u_k+u_{k-1}\right)||_H$$
$$\le C_2\left[||A(0)u_0||_H+||A^{\frac{1}{2}}(0)u_0'||_H+\max_{0\le s\le k}\|f_s\|_H+\sum_{s=0}^{N}||f_{s+1}-f_s||_H\right] \tag{4}$$

holds, where C_2 does not depend on $u_0, u_0', f_s\ (0\le s\le N)$, and τ.

The proof of these theorems are based on the discrete analogies of integral inequality and on the following formula

$$u_{k+1}=4^{-1}\left\{[P_k^+(k)B^+ + P_k^-(k)B^-]u_0+[P_k^+(k)C^+ + P_k^-(k)C^-]u_0'\right.$$
$$+[P_k^+(k)D^+ + P_k^-(k)D^-]f_0$$
$$\left.+\sum_{s=0}^{k-1}\left[B_s^+(k)\tau\left(I-\frac{i\tau}{2}A_{k-s}^{1/2}\right)^{-1}-B_s^-(k)\tau\left(I+\frac{i\tau}{2}A_{k-s}^{1/2}\right)^{-1}\right]\psi_{k-s}\right\}$$

for the solution of the difference scheme (2) and on the estimates

$$\left\|\tau^\alpha A_k^{\alpha/2}\left(I+\tau^2A_k\right)^{-1}\right\|_{H\to H}\le 1,\ \alpha=0,2, \left\|\tau A_k^{1/2}\left(I+\tau^2A_k\right)^{-1}\right\|_{H\to H}\le 1/2,$$

$$\left\|\left(I\mp i\tau A_s^{1/2}\right)^{-1}\right\|_{H\to H}\le 1, \left\|\left(I\mp\frac{i\tau}{2}A_s^{1/2}\right)^{-1}\left(I\pm\frac{i\tau}{2}A_s^{1/2}\right)\right\|_{H\to H}\le 1,$$

$$\left\|\tau A_s^{1/2}\left(I\pm\frac{i\tau}{2}A_s^{1/2}\right)^{-1}\right\|_{H\to H}\le 2,$$

and

$$\left\|A^{\rho}(t)A^{-\rho}(s)\right\|_{H\to H}\leq M_{\rho}.$$

Here

$$B^{\pm}=\left(I+\tau^{2}A_{0}\right)^{-1}\left(I-\frac{\tau^{2}}{2}A_{0}\pm i\tau A_{0}^{1/2}\right),C^{\pm}=\left(I+\tau^{2}A_{0}\right)^{-1}\left(\tau\mp iA_{0}^{-1/2}\right),$$

$$D^{\pm}=\left(I+\tau^{2}A_{0}\right)^{-1}\left(\frac{\tau^{2}}{2}\mp i\tau A_{0}^{-1/2}\right),$$

$$\psi_{k}+\psi_{k-1}=-2iA_{k}^{-1/2}f_{k}$$

$$+2i\left(A_{k-1}^{-1/2}-A_{k}^{-1/2}\right)\frac{u_{k}-u_{k-1}}{\tau^{2}}+2^{-1}i\left(A_{k}^{1/2}-A_{k-1}^{1/2}\right)\left(u_{k}+u_{k-1}\right),$$

and

$$P_{k}^{\pm}(k)=X_{k}^{\pm}X_{k-1}^{\pm}\cdots X_{0}^{\pm},\ R_{m}^{\pm}(k)=X_{k}^{\pm}X_{k-1}^{\pm}\cdots X_{m+1}^{\pm},$$

$$X_{k+1}^{\pm}=\left(I\mp\frac{i\tau}{2}A_{k}^{1/2}\right)^{-1}\left(I\pm\frac{i\tau}{2}A_{k}^{1/2}\right),$$

$$B_{s}^{\pm}(k)=X_{k}^{\pm}X_{k-1}^{\pm}\cdots X_{k-s+1}^{\pm},B_{0}^{\pm}(k)=I.$$

Let us denote

$$\psi_{p}=-iA_{k}^{-1/2}f_{k}-2iA_{k}^{-1/2}\left(A_{k}^{1/2}-A_{p}^{1/2}\right)A_{p}^{-1/2}\frac{u_{k}-u_{k-1}}{\tau^{2}}$$

$$+i\left(A_{k}^{1/2}-A_{p}^{1/2}\right)\frac{u_{k}+u_{k-1}}{2},\ \text{for } p=k-1,k.$$

Note that we have different formulas for ψ_k and ψ_{k-1}; nevertheless, we have the same estimate for their norms. Actually,

$$\left\|A_{k}^{1/2}\psi_{k}\right\|_{H}\leq M_{1/2}\left\|f_{k}\right\|_{H}$$

$$\leq M_{1/2}\left\|f_{k}\right\|_{H}+2M_{1/2}^{2}\left\|\frac{u_{k}-u_{k-1}}{\tau}\right\|_{H}+M_{1/2}\left\|\tau A_{k-1}\frac{u_{k}+u_{k-1}}{2}\right\|_{H},$$

$$\left\|A_{k}^{1/2}\psi_{k-1}\right\|_{H}\leq M_{1/2}\left\|f_{k}\right\|_{H}+2M_{1/2}^{2}\left\|\frac{u_{k}-u_{k-1}}{\tau}\right\|_{H}+M_{1/2}\left\|\tau A_{k-1}\frac{u_{k}+u_{k-1}}{2}\right\|_{H}.$$

References

[1] R. K. Mohanty, M. K. Jain and K. George, Fourth-order approximations at first time level, linear stability analysis and the numerical solution of multidimensional second-order nonlinear hyperbolic equations in polar coordinates, Journal of Computational and Applied Mathematics, 93 (1) pp. 1-12, 1998.

[2] R. K. Mohanty, U. Arora and M. K. Jain, Fourth-order approximation for the three space dimensional certain mildly quasi-linear hyperbolic equation, Numerical Methods for Partial Differential Equations, 17 pp. 277-289, 2001.

[3] R. K. Mohanty, U. Arora and M. K. Jain, Linear stability analysis and fourth-order approximations at first time level for the two space mildly quasi-linear hyperbolic equations, Numerical Methods for Partial Differential Equations, 17 pp. 607-618, 2001.

[4] R. K. Mohanty, An unconditionally stable difference scheme for the one-space-dimensional linear hyperbolic equation, Applied Mathematics Letters, 17 pp. 101-105, 2004.

[5] R. K. Mohanty, An operator splitting method for an unconditionally stable difference scheme for a linear hyperbolic equation with variable coefficients in two space dimensions, Applied Mathematics and Computation, 152 pp. 799-806, 2004.

[6] R. K. Mohanty, An operator splitting technique for an unconditionally stable difference method for a linear three space dimensional hyperbolic equation with variable coefficients, Applied Mathematics and Computation, 162 pp. 549-557, 2005.

[7] R. K. Mohanty, An unconditionally stable finite difference formula for a linear second order one space dimensional hyperbolic equation with variable coefficients, Applied Mathematics and Computation, 165 pp. 229-236, 2005.

[8] W. Li, Z. Sun and L. Zhao, An analysis for a high-order difference scheme for numerical solution to $u_{tt} = A(t,x)u_{xx} + F(t,x,u,u_t,u_x)$, Numerical Methods for Partial Differential Equations, 23 (2) pp. 484-498, 2007.

[9] P. E. Sobolevskii and L. M. Chebotaryeva, Approximate solution of the Cauchy problem for an abstract hyperbolic equation by the method of lines, Izv. Vyssh.Uchebn. Zav. Mat. 180 (5) pp.103-116, 1977 (Russian).

[10] A. Ashyralyev and P. E. Sobolevskii, A note on the difference schemes for hyperbolic equations, Abstract and Applied Analysis, 6 (2) pp. 63-70, 2001.

[11] A. A. Samarskii, I. P. Gavrilyuk and V. L. Makarov, Stability and regularization of three-level difference schemes with unbounded operator coefficients in Banach spaces, SIAM Journal of Numerical Analysis, 39 (2) pp. 709-723, 2001.

[12] A. Ashyralyev and P. E. Sobolevskii, Two new approaches for construction of the high order of accuracy difference schemes for hyperbolic differential equations, Discrete Dynamics in Nature and Society, 2 (2) pp. 183-213, 2004.

[13] A. Ashyralyev and M. E. Koksal, On the second order of accuracy difference scheme for hyperbolic equations in a Hilbert space, Numerical Functional Analysis and Optimization, 26 (7-8) pp. 739-772, 2005.

[14] A. Ashyralyev and M. E. Koksal, Stability of a second order of accuracy difference scheme for hyperbolic equation in a Hilbert space, Discrete Dynamics in Nature and Society, 2007 pp. 1-25, 2007.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Gronwall Type Integral Inequalities via Picard Operators

Maria Dobriţoiu
Department of Mathematics-Informatics
University of Petroşani, Romania
e-mail: mariadobritoiu@yahoo.com

Abstract

In this paper we will use the Picard operators technique presented by I. A. Rus in papers [10], [11], [12] and the Abstract Gronwall lemma, to obtain some integral inequalities regarding to the solution of the integral equation

$$x(t) = \int_a^b K(t, s, x(s), x(a), x(b))ds + f(t), \ \ t \in [a, b].$$

An example is also given here.

Keywords: integral equation, modified argument, subsolution, Picard operator, integral inequality.

1 Introduction

In the study of some problems from turbo-reactors industry, in the '70, a Fredholm integral equation with modified argument appears, having the following form

$$x(t) = \int_a^b K(t, s, x(s), x(a), x(b))ds + f(t), \ \ t \in [a, b]. \tag{1}$$

where $K : [a, b] \times [a, b] \times \mathbb{B}^3 \to \mathbb{B}$, $f : [a, b] \to \mathbb{B}$ and $(\mathbb{B}, +, \mathbb{R}, |\cdot|)$ is a Banach space.

This integral equation is a mathematical model from physics, reference with to the turbo-reactors working.

The results obtained by the author regarding to the existence and uniqueness, the existence, the data dependence and an approximation method of the solution of the integral equation (1) have been published in the papers [1], [3], [4], [5], [6] and [7].

The integral inequalities have been studied both by the classical theory and using the abstract Gronwall lemma. We mention several articles which contains

ISSN-1584-5990

Gronwall integral inequalities: A. Buică [2], I. A. Rus [13], V. Mureşan [8], A. Petruşel and I. A. Rus [9], M. A. Şerban [14], M. Zima [15].

In this paper we will use the Picard operators technique, presented by I. A. Rus in the papers [10], [11] and [12], the Abstract Gronwall lemma and the Abstract Comparison lemma in order to obtain some integral inequalities and comparison results regarding to the solution of the integral equation (1). These integral inequalities are new properties of the solution of this integral equation.

In the last section, an example is given.

2 Notations and preliminaries

Let X be a nonempty set, d a metric on X and $A : X \to X$ an operator. In this paper we shall use the following notations:

$$\begin{aligned} F_A &: = \{x \in X |\ A(x) = x\} \text{ - the fixed points set of } A \\ A^{n+1} &: = A \circ A^n, \quad A^0 := 1_X, \quad A^1 := A, \quad n \in \mathbb{N} \\ C([a,b], \mathbb{B}) &: = \{x : [a,b] \to \mathbb{B} \ / \ x \ continuous \ function\} \end{aligned}$$

Definition (I. A. Rus [10] or [11]) Let (X, d) be a metric space. An operator $A : X \to X$ is **Picard operator** (PO) if there exists $x^* \in X$ such that:

(a) $F_A = \{x^*\}$;

(b) the sequence $(A^n(x_0))_{n \in N}$ converges to x^*, for all $x_0 \in X$.

Definition (I. A. Rus [10] or [11]) Let (X, d) be a metric space. An operator $A : X \to X$ is **weakly Picard operator** (WPO) if the sequence $(A^n(x_0))_{n \in N}$ converges for all $x_0 \in X$ and the limit (which may depend on x_0) is a fixed point of A.

If A is a weakly Picard operator, then we consider the following operator

$$A^\infty : X \to X, \quad A^\infty(x) = \lim_{n \to \infty} A^n(x)$$

and we observe that $A^\infty(X) = F_A$.

In order to obtain the proposed integral inequalities, we will need the following results (see [10], [11], [12]).

Theorem (Contraction Principle) *Let (X, d) be a complete metric space and $A : X \to X$ an α-contraction ($\alpha < 1$). In these conditions we have*:

(i) $F_A = \{x^*\}$;

(ii) $x^* = \lim_{n \to \infty} A^n(x_0)$, for all $x_0 \in\in X$;

(iii) $d(x^*, A^n(x_0)) \leq \frac{\alpha^n}{1-\alpha} d(x_0, A(x_0))$.

Let $\leq$ be an order relation on X.

Lemma (I. A. Rus [12]) *Let $(X, d, \leq)$ be an ordered metric space and $A : X \to X$ an operator, such that:*

(i) *the operator A is increasing ;*

(ii) A is WPO.

Then, the operator A^∞ is increasing.

Lemma (Abstract Comparison lemma) *Let $(X, d, \leq)$ be an ordered metric space and $A, B, C : X \to X$ three operators, such that:*

(i) $A \leq B \leq C$;

(ii) A, B, C *are WPOs*;

(iii) *the operator B is increasing.*

Then

$$x \leq y \leq z \implies A(x) \leq B(y) \leq C(z).$$

Remark Let A, B, C be the operators defined in the Abstract Comparison lemma. In addition, we suppose that

$$F_B = \{x_B^*\},$$

i.e. B is a PO. Then we have

$$A^\infty(x) \leq x_B^* \leq C^\infty(x), \quad for \ \ all \ \ x \in X.$$

But

$$A^\infty(X) = F_A,$$

and

$$C^\infty(X) = F_C.$$

Therefore we have

$$F_A \leq x_B^* \leq F_C.$$

Lemma (Abstract Gronwall lemma) (see [11]) *Let $(X, d, \leq)$ be an ordered metric space and $A : X \to X$ an operator. We suppose that:*

(i) *A is a PO;*

(ii) *the operator Ais increasing.*

If we denote with x_A^* the unique fixed point of the operator A, then

(a) $x \leq A(x) \ \Rightarrow \ x \leq x_A^*$;

(b) $x \geq A(x) \ \Rightarrow \ x \geq x_A^*$.

3 The main results

The theorems presented in this section contain the properties of the solution of the integral equation (1). These results are obtained applying the Picard operators technique, the Abstract Gronwall lemma and the Abstract Comparison lemma.

Let $(\mathbb{B}, +, \mathbb{R}, |\cdot|)$ be an ordered Banach space.

We consider the integral equation (1), where $K : [a,b] \times [a,b] \times \mathbb{B}^3 \to \mathbb{B}$, $f : [a,b] \to \mathbb{B}$.

Theorem *We suppose that:*

(i) $K \in C([a,b] \times [a,b] \times \mathbb{B}^3, \mathbb{B})$, $f \in C([a,b], \mathbb{B})$;

(ii) $K(t,s,\cdot,\cdot,\cdot)$ *is increasing for all* $t,s \in [a,b]$;

(iii) *there exists* $L_K > 0$ *such that*

$$|K(t,s,u_1,u_2,u_3) - K(t,s,v_1,v_2,v_3)| \leq$$

$$\leq L_K(|u_1 - v_1| + |u_2 - v_2| + |u_3 - v_3|),$$

for all $t,s \in [a,b]$, $u_i, v_i \in \mathbb{B}$, $i = \overline{1,3}$;

(iv) $3L_K(b-a) < 1$

and let $x^* \in C([a,b], \mathbb{B})$ *be the unique solution of the integral equation* (1). *In these conditions*

(a) *if* $x \in C([a,b], \mathbb{B})$ *is a lower subsolution of the integral equation* (1) *then* $x \leq x^*$;

(b) *if* $x \in C([a,b], \mathbb{B})$ *is an upper subsolution of the integral equation* (1) *then* $x \geq x^*$.

Proof We consider the operator $A : C([a,b], \mathbb{B}) \to C([a,b], \mathbb{B})$, defined by the relation:

$$A(x)(t) := \int_a^b K(t,s,x(s),x(a),x(b))ds + f(t), \quad t \in [a,b]. \tag{2}$$

From the conditions (i), (iii) and (iv) it results that the operator A is an α-contraction with the coefficient $\alpha = 3L_K(b-a)$. Therefore. the operator A is PO. From the condition (ii) it results that the operator A is increasing.

Now, the conditions of the *Abstract Gronwall lemma* being satisfied, it results the conclusions of the theorem, i.e.

$$x \leq A(x) \implies x \leq x^*$$

and

$$x \geq A(x) \implies x \geq x^*$$

and the proof is complete.

Remark The theorem 5 holds true in particular cases $\mathbb{B} = \mathbb{R}$, $\mathbb{B} = \mathbb{R}^m$, $\mathbb{B} = l^2(\mathbb{R})$, if one replaces the conditions (i), (iii) and (iv) with conditions which one assures the existence and uniqueness of the solution of the integral equation (1) in $C[a,b]$ space, in $C([a,b],\mathbb{R}^m)$ space and respectively in $C([a,b],l^2(\mathbb{R}))$ space.

Now, we consider the integral equation (1) corresponding to the functions K_i, f_i, $i = 1,2,3$, i.e.

$$x(t) = \int_a^b K_1(t,s,x(s),x(a),x(b))ds + f_1(t), \ \ t \in [a,b], \tag{3}$$

$$x(t) = \int_a^b K_2(t,s,x(s),x(a),x(b))ds + f_2(t), \ \ t \in [a,b], \tag{4}$$

$$x(t) = \int_a^b K_3(t,s,x(s),x(a),x(b))ds + f_3(t), \ \ t \in [a,b], \tag{5}$$

where $K_i : [a,b] \times [a,b] \times \mathbb{B}^3 \to \mathbb{B}$, $f_i : [a,b] \to \mathbb{B}$, $i = 1,2,3$.

Theorem *We suppose that the functions K_i and f_i, $i = 1,2,3$ satisfies the following conditions:*

(i) $K_i \in C([a,b] \times [a,b] \times \mathbb{B}^3, \mathbb{B})$, $f_i \in C([a,b],\mathbb{B})$, $i = 1,2,3$;

(ii) $K_2(t,s,\cdot,\cdot,\cdot)$ *is increasing for all* $t,s \in [a,b]$;

(iii) $K_1 \leq K_2 \leq K_3$ *and* $f_1 \leq f_2 \leq f_3$;

(iv) *there exists* $L_i > 0$, $i = 1,2,3$ *such that*

$$|K_i(t,s,u_1,u_2,u_3) - K_i(t,s,v_1,v_2,v_3)| \leq$$

$$\leq L_i \left(|u_1 - v_1| + |u_2 - v_2| + |u_3 - v_3|\right),$$

for all $t,s \in [a,b]$, $u_j, v_j \in \mathbb{B}$, $j = \overline{1,3}$;

(v) $3L_i(b-a) < 1$, $i = 1,2,3$.

If we denote by x_1^*, x_2^*, and respectively x_3^* the unique solution of the integral equation (3), (4) *and respectively* (5), *then*

$$x_1^* \leq x_2^* \leq x_3^*.$$

Proof We consider the operators $A_i : C([a,b],\mathbb{B}) \to C([a,b],\mathbb{B})$, $i = 1,2,3$, defined by the relations:

$$A_i(x)(t) := \int_a^b K_i(t,s,x(s),x(a),x(b))ds + f_i(t), \ \ t \in [a,b], \ \ i = 1,2,3. \tag{6}$$

From the condition (ii) it results that the operator A_2 is increasing and from the condition (iii) we have

$$A_1 \leq A_2 \leq A_3.$$

From the conditions (i), (iv) and (v) it results that the operators A_i are α_i-contractions with the coefficients $\alpha_i = 3L_i(b-a)$, $i = 1, 2, 3$. Therefore the operators A_i, $i = 1, 2, 3$ are POs. According to the Contraction Principle it results that each of the integral equations (3), (4) and (5) has a unique solution in $C([a, b], \mathbb{B})$ space. We denote these solutions with x_1^*, x_2^* and x_3^*.

Now, the conditions of the Abstract Comparison lemma being satisfied, it results that

$$x_1 \leq x_2 \leq x_3 \Longrightarrow A_1^\infty(x_1) \leq A_2^\infty(x_2) \leq A_3^\infty(x_3),$$

but A_1, A_2, A_3 are POs and then according to the remark 1 it results the conclusion of this theorem, i.e.

$$x_1^* \leq x_2^* \leq x_3^*.$$

The proof is complete.

4 Example

In the case $\mathbb{B} = \mathbb{R}$, we consider the integral equation with modified argument

$$x(t) = \int_0^1 \left[\frac{t}{7}x(s) + \frac{1}{7}x(0) + \frac{1}{5}x(1)\right] ds + \frac{13}{14}t - \frac{1}{5}, \ \ t \in [0, 1], \tag{7}$$

where

$$\begin{aligned} K &\in C([0,1]\times[0,1]\times\mathbb{R}^3),\ K(t,s,u_1,u_2,u_3) = \frac{t}{7}u_1 + \frac{1}{7}u_2 + \frac{1}{5}u_3, \\ f &\in C[0,1],\ f(t) = \frac{13}{14}t - \frac{1}{5}, \\ x &\in C[0,1], \end{aligned}$$

and one verifies the conditions of the theorem 5.

The solution of the integral equation (7) is $x^*(t) = t$, $t \in [0, 1]$.

We attach to this integral equation the operator $A : C[0, 1] \to C[0, 1]$, defined by the relation:

$$A(x)(t) = \int_0^1 \left[\frac{t}{7}x(s) + \frac{1}{7}x(0) + \frac{1}{5}x(1)\right] ds + \frac{13}{14}t - \frac{1}{5}, \ \ t \in [0, 1]. \tag{8}$$

The solutions set of the integral equation (7) in the $C[0, 1]$ space, coincides with the fixed points set of the operator A, defined above.

Since the function K satifies the Lipschitz condition with the constant $\frac{1}{7}$ with respect to the third and fourth argument and with the constant $\frac{1}{5}$ with respect

to the last argument, it results that the operator A is contraction with the coefficient $\alpha = \frac{17}{35}$. Therefore A is PO. According to the Contraction Principle it results that the integral equation (7) has a unique solution $x^* \in C[0,1]$. This unique solution is $x^*(t) = t$, $t \in [0,1]$.

Since the function $K(t,s,\cdot,\cdot,\cdot)$ is increasing for all $t,s \in [0,1]$, it results that the conditions of the theorem 5 are satisfied ($\mathbb{B} = \mathbb{R}$) and therefore, we have the following integral inequalities:

- if $x \in C[0,1]$ is a lower subsolution of the integral equation (7) then

$$x(t) \leq \int_0^1 \left[\frac{t}{7}x^*(s) + \frac{1}{7}x^*(0) + \frac{1}{5}x^*(1)\right] ds + \frac{13}{14}t - \frac{1}{5}, \quad t \in [0,1]; \qquad (9)$$

- if $x \in C[0,1]$ is a upper subsolution of the integral equation (7) then

$$x(t) \geq \int_0^1 \left[\frac{t}{7}x^*(s) + \frac{1}{7}x^*(0) + \frac{1}{5}x^*(1)\right] ds + \frac{13}{14}t - \frac{1}{5}, \quad t \in [0,1]. \qquad (10)$$

References

[1] Ambro, M., Aproximarea soluţiilor unei ecuaţii integrale cu argument modificat, Studia Univ. Babeş–Bolyai, Mathematica, 2(1978), 26–32

[2] Buică, A., Gronwall-type nonlinear integral inequalities, Mathematica, 44(2002), Nr. 1, 19–23

[3] Dobriţoiu, M., Aproximări ale soluţiei unei ecuaţii integrale Fredholm cu argument modificat, Analele Univesităţii Aurel Vlaicu din Arad, seria Matematică, fascicola Matematică-Informatică, Arad, 28–30 nov. 2002, ISSN 1582-344X, 51–56

[4] Dobriţoiu, M., Formule de cuadratură utilizate în aproximarea soluţiei unei ecuaţii integrale cu argument modificat, Lucrările Ştiinţifice ale Simpozionului Internaţional, Universitaria RoPet 2003, Petroşani, 16-18 oct. 2003, Editura Universitas Petroşani, fascicola Matematică–Informatică–Fizică, ISBN 973-8260-37-X, 53–56

[5] Dobriţoiu, M., The rectangle method for approximating the solution to a Fredholm integral equation with a modified argument, Lucrările ştiinţifice a celei de a XXX-a Sesiuni de comunicări ştiinţifice cu participare internaţională "Tehnologii Moderne in Secolul XXI", Academia Tehnică Militară Bucureşti, secţiunea 16, "Matematică", Bucureşti, 6-7 nov. 2003, ISBN 973-640-012-3, 36–39

[6] Dobriţoiu, M., A Fredholm integral equation – numerical methods, Bulletins for Applied&Computer Mathematics, Budapest, BAM – CVI / 2004, Nr. 2188, ISSN 0133-3526, 285–292

[7] Dobriţoiu, M., Existence and continuous dependence on data of the solution of an integral equation, Bulletins for Applied&Computer Mathematics, Budapest, BAM – CVI / 2005 , Nr. ISSN 0133-3526

[8] Mureşan, V., A Gronwall type inequality for Fredholm operators, Mathematica, 41(1999), Nr. 2, 227–231

[9] Petruşel, A., Rus, I. A., Fixed point theorems in ordered L-spaces, Proc. Amer. Math. Soc. 134 (2005), 411–418

[10] Rus, I. A., Weakly Picard mappings, Comment. Math. Univ. Caroline, 34, 3(1993), 769-773

[11] Rus, I. A., Picard operators and applications, "Babeş-Bolyai" University of Cluj-Napoca, Preprint No.3, 1996

[12] Rus, I. A., Weakly Picard operators and applications, Seminar on Fixed Point Theory, "Babeş-Bolyai" University of Cluj-Napoca, 2, 2001, 41-58

[13] Rus, I. A., Fixed points, upper and lower fixed points: abstract Gronwall lemmas, Carpathian Journal of Mathematics, Baia-Mare, 20(2004), No. 1, 125–134

[14] Şerban, M. A., Application of fiber Picard operators to integral equations, Bul. Ştiinţific Univ. Baia Mare, Seria B, Matematică–Informatică, Vol. XVIII(2002), Nr.1, 119–128

[15] Zima, M., The abstract Gronwall lemma for some nonlinear operators, Demonstratio Math., 31(1998), 325–332

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Linear Systems with Quadratic Criteria

Dragoescu Nina
E-mail: cazanina@yahoo.com

Abstract

This paper contains an example of optimal control problem with quadratic cost function, based on the theory of R.E. Kalman, P.L.Falb and M.A. Arbib. Some programs in C++ and Mathematica 5.2 are used to obtain the optimal solution.

Keywords: optimal, cost, linear, squared, trajectory, command.

Let: $\Sigma =< (T_1, T_2), U, \Omega, X, Y, \varphi, \eta >$ (dynamic smooth linear system). The optimal control problem can be resumed to the determination of the $u(\cdot)$ command which minimize the cost $J(t_0, x_0, u(\cdot))$. It will be considered the version of R.E. Kalman(1963), based on Hamilton theory, in order to resolve the problem. Our contribution is a numerical example for these results.

According to the theory, if a Hamilton-Iacobi solution is found, this solution can be used in order to find the optimal command of the problem. It is proved that $p(\cdot)$ is a Riccati solution; the Riccati equation is formed with the system Σ coefficients:

$$\dot{p}(t) = -p^*(t) \cdot A(t) - A^*(t) \cdot p(t) + p^*(t) \cdot S(t) \cdot p(t) - \rho(t) \tag{1}$$

where: $S(t) = B(t) \cdot \sigma^{-1}(t) \cdot B^*(t)$. Consequently,the optimal command is :

$$u^0(t, x) = -\sigma^{-1}(t) \cdot B^*(t) \cdot p(t) \cdot x(t)$$

with $p(\cdot)$ a unique solution of the Riccati equation which verifies $p(t_1) = \Psi(t_1) \cdot \Psi^{-1}(t_0)$, the optimal trajectory is a solution of the differential equation:

$$\dot{x}(t) = [A(t) - S(t) \cdot p(t)] \cdot x(t) \text{ cu } x(t_0) = x_0, S = B \cdot \sigma^{-1} \cdot B^*.$$

the minim cost is: $J(t_0, x_0, u^0) = \frac{1}{2} < x_0, p(t_0) \cdot x_0 >$ We consider the next Riccati equation coefficients and the initial conditions to be verified:
$A = \begin{pmatrix} 2 & 1 \\ 3 & 5 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}, A^* = \begin{pmatrix} 2 & 3 \\ 1 & 5 \end{pmatrix}, B^* = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}, C(t) = I_2, t_0 = 0, t_1 = 1, z(t) = 0, \rho(t) = 0, \sigma(t) = -\frac{1}{4}, \sigma^{-1}(t) = -4, p(t) = p^*(t), C_0 = e^A, S = B \cdot \sigma^{-1} \cdot B^* = -\begin{pmatrix} 4 & 8 \\ 8 & 16 \end{pmatrix}, A + A^* = \begin{pmatrix} 4 & 4 \\ 4 & 10 \end{pmatrix}, C_1 = \frac{S}{2} \cdot C_0 \cdot e^2, x_0 = e^{C_1}$
Consequently, the equation has the Bernoulli form ($\rho(t) = \mathbf{0}$):

ISSN-1584-5990

$$\dot{p}(t)+\begin{pmatrix}4&4\\4&10\end{pmatrix}\cdot p(t)+\begin{pmatrix}4&8\\8&16\end{pmatrix}\cdot p^2(t)=0_2$$

Dividing with $p^2(t)$:

$$\frac{\dot{p}(t)}{p^2(t)}+\begin{pmatrix}4&4\\4&10\end{pmatrix}\cdot\frac{1}{p(t)}+\begin{pmatrix}4&8\\8&16\end{pmatrix}=0_2$$

Let : $\frac{\dot{p}(t)}{p^2(t)}=\ddot{u}(t)$. Consequently:

$$\ddot{u}(t)-\begin{pmatrix}4&4\\4&10\end{pmatrix}\cdot\dot{u}(t)+\begin{pmatrix}4&8\\8&16\end{pmatrix}=0_2$$

The associate characteristic equation:

$$I_2\cdot r^2-\begin{pmatrix}4&4\\4&10\end{pmatrix}\cdot r+\begin{pmatrix}4&8\\8&16\end{pmatrix}=0_2$$

leads us to the next result: $\begin{pmatrix}r^2-4r+4&-4r+8\\-4r+8&r^2-10r+16\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix}$ with the solution: r=2, so that: $\dot{u}(t)=e^{2t}\cdot C^{'}$ or: $\dot{u}(t)=-\frac{1}{p(t)}\Rightarrow p(t)=C\cdot e^{-2t}$. The initial condition: $p(t_1)=p(1)=C\cdot e^{-2}=\phi(t_1,t_0)=\psi(1)\cdot\psi^{-1}(0)=e^A\Rightarrow C=e^2\cdot e^A$, so that: $p(t)=e^2\cdot e^A\cdot e^{-2t}$ and $p(t_0)=p(0)=e^2\cdot e^A=e^{A+2}$(because $\psi(t)=e^{At}$) As a result of the theory, the optimal command is:

$$u^0(t,x)=-\sigma^{-1}(t)\cdot B^*(t)\cdot p(t)\cdot x(t)$$

The optimal trajectory is a solution of the differential equation:

$$\dot{x}(t)=[A-S\cdot p(t)]\cdot x(t)=$$
$$\begin{pmatrix}2+4\cdot e^2\cdot e^A\cdot e^{-2t}&1+8\cdot e^2\cdot e^{-2t}\\3+8\cdot e^2\cdot e^A\cdot e^{-2t}&5+16\cdot e^2\cdot e^A\cdot e^{-2t}\end{pmatrix}\cdot x(t)=M(t)\cdot x(t);$$

$$M(t)=\begin{pmatrix}2+4\cdot e^2\cdot e^A\cdot e^{-2t}&1+8\cdot e^2\cdot e^{-2t}\\3+8\cdot e^2\cdot e^A\cdot e^{-2t}&5+16\cdot e^2\cdot e^A\cdot e^{-2t}\end{pmatrix}$$

$$x(t)=e^{\int A dt}\cdot e^{-\int S\cdot e^2\cdot e^A\cdot e^{-2t}dt}=e^{At}\cdot e^{-S\cdot e^2\cdot e^A\cdot\int e^{-2t}dt}=e^{At}\cdot e^{S\cdot e^2\cdot e^A\cdot e^{-2t}\cdot\frac{1}{2}}=$$
$$e^{C_1\cdot e^{-2t}+At}$$

with the initial restriction :

$$x(0)=e^{C_1}=x_0$$

In consequence the optimal solution is:

$$4\cdot B^*\cdot e^{A+2}\cdot e^{(A-2)\cdot t}\cdot x_0^{e^{-2t}}$$

The minim cost J is: $J(t_0,x_0,u^0)=\frac{1}{2}<x_0,p(t_0)x_0>=\frac{p(t_0)}{2}\cdot x_0^2=\frac{e^{A+2}}{2}\cdot x_0^2$.

Assumptions

In the ANNEX, we have reproduced the graphics made by the specialized soft. It can be seen that the time evolutions appear near the same for the optimal trajectory and the optimal command for every time interval; the step of the time is equal with 0.2.The Ot-axe, which is horizontal, marks the intervals, for each graphic.

References

[1] R.E. Kalman, P.L., Falb, M.A,Arbib, *Dynamic Systems*, Technical Ed., Bucharest, Romania, 1969.

[2] N.Racoveanu, Gh.Dodescu, I. Mincu, *Numerical methods for parabolic equations with partial derivatives*, Technical Ed., Bucharest, Romania, 1977.

[3] V.I. Arnold, *Ordinary differential equations*, Scientifical and Enciclopedic Ed., Bucharest, Romania, 1978.

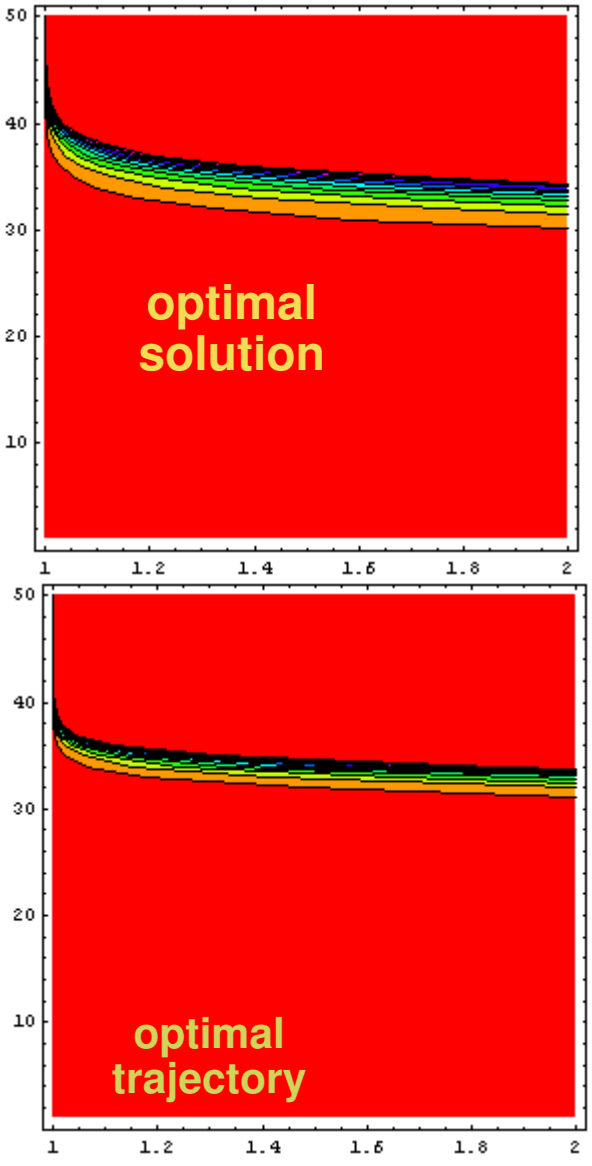
[4] St. Mirica,*Differential equations and equations with partial derivatives*, Bucarest University Ed., Bucharest, Romania,1999.

[5] St. Mirica, *Differential and integral equations*, Bucurest University Ed., Bucharest, Romania, 2000.

[6] S.Sburlan, L.Barbu, C. Mortici, *Differential and integral equations and dynamic systems*, Exponto Ed., Constantza, Romania, 1999.

[7] Soft for applications *Mathematica 5.2, Mathematica 6.0, Latex editor* @UNPUBLISHED Application in Mathematica 6.0,Dragoescu Nina, TITLE=" Riccati equations", Constantza, Romania, 2008 @UNPUBLISHED Application in Mathematica 6.0, Dragoescu Nina, TITLE = "The optimal solution of a dynamic system, regulation according to the state ", Constantza, Romania, 2008

ANNEX
GRAPHICAL COMPARE BETWEEN THE TIME EVOLUTION OF THE OPTIMAL COMMAND AND THE OPTIMAL TRAJECTORY



Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Solution of an Optimal Control Problem with Quadratic Cost Function

Dragoescu Nina
E-mail: cazanina@yahoo.com

Abstract

This paper contains an example of Bernoulli equation which can be used in order to determine the optimal solution for dynamic smooth linear systems, according to R.E. Kalman, P.L.Falb and M.A. Arbib theory. The paper also contains some assumptions in the general case of Bernoulli equation with matrices coefficients.

Keywords: algorithm, Bernoulli, optimal, solution, trajectory, command.

Let: $\Sigma =< (T_1, T_2), U, \Omega, X, Y, \varphi, \eta >$ (dynamic smooth linear system)

$$f(x, u, t) = A(t) \cdot x + B(t) \cdot u \tag{1}$$

$$\eta(t, x) = C(t) \cdot x \tag{2}$$

In the finite dimensional case f is a positive symmetric matrix. The optimal control problem can be stated to the determination of the command $u(t)$ which minimize the quadratic cost $J(t_0, x_0, u(\cdot))$, for any initial pair $(t_0, x_0)\epsilon(T_1, T_2) \times X$. If $p : (T_1, T_2) \to \mathcal{L}(X, X)$ is continuous differentiable then $w(t, x) = \frac{1}{2} < x, p(t)x >$ is also continuous differentiable. According to Kalman theory, if a Hamilton-Iacobi solution $\omega(t, x)$ is found, then the optimal control problem can be solved. It is proved that $p(\cdot)$ is a Riccati solution; the Riccati equation is formed with the system Σ coefficients:

$$\dot{p}(t) + p^*(t) \cdot A(t) + A^*(t) \cdot p(t) - p^*(t) \cdot S(t) \cdot p(t) = \rho(t) \tag{3}$$

with: $S(t) = B(t) \cdot \sigma^{-1}(t) \cdot B^*(t)$. For system Σ problem with **constant** coefficients, the above considerations led to the following :

Proposition

For any fixed initial pair $(t_0, x_0)\epsilon(T_1, t_1] \times X$, with $t_1\epsilon(T_1, T_2)$, the regulation problem of the system Σ has a solution according to the state.

- The optimal command is: $u^0(t, x) = -\sigma^{-1}(t) \cdot B^* \cdot p(t) \cdot x(t)$, with $p(\cdot)$ the unique solution for the Riccati equation which verifies $p(t_1) = \Psi(t_1) \cdot \Psi^{-1}(t_0)$
- The optimal trajectory is a solution of the differential equation: $\dot{x}(t) = [A - S(t) \cdot p(t)] \cdot x(t)$, with: $x(t_0) = x_0, S(t) = B \cdot \sigma^{-1}(t) \cdot B^*$.

ISSN-1584-5990

I. Consider the following Riccati equation coefficients and the initial conditions:
$A = \begin{pmatrix} -2 & 0 \\ 0 & 5 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, A^* = \begin{pmatrix} -2 & 0 \\ 0 & 5 \end{pmatrix}, B^* = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, A, B, C\epsilon$
$\mathcal{M}_2(R), C(t) = I_2, X = Y, t_0 = 0, t_1 = 1, (t_0, t_1) \subset (T_1, T_2) \subset R, \rho(t) = 0, \sigma(t) = -\frac{1}{4}, \sigma^{-1}(t) = -4, x_0 = 1, p(t) = p^*(t), C_0 = e^A, S = B \cdot \sigma^{-1} \cdot B^* = \begin{pmatrix} -4 & 0 \\ 0 & -16 \end{pmatrix}, A + A^* = \begin{pmatrix} -4 & 0 \\ 0 & 10 \end{pmatrix}$
Consequently, we have to solve a Bernoulli equation: $(\rho(t) = \mathbf{0})$:

$$\dot{p}(t) + \begin{pmatrix} -4 & 0 \\ 0 & 10 \end{pmatrix} \cdot p(t) + \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix} \cdot p^2(t) = 0_2$$

with: $p(t_1) = p(1) = \Psi(t_1) \cdot \Psi^{-1}(t_0) = e^A = C_0$ (because $\Psi(t) = e^{At}$), where $p(t)$ is a self adjoint operator and continuous differentiable on (t_0, t_1). If $\det p(t) \neq 0$ and $\det A \neq 0$, we multiply with $[p(t)]^{-2} \Leftrightarrow [(p(t))^2]^{-1}$:

$$\dot{p}(t) \cdot [p(t)]^{-2} + \begin{pmatrix} -4 & 0 \\ 0 & 10 \end{pmatrix} \cdot [p(t)]^{-1} + \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix} = 0_2$$

Let : $\dot{v}(t) = \dot{p}(t) \cdot [p(t)]^{-2}$.

$$\dot{v}(t) - \begin{pmatrix} -4 & 0 \\ 0 & 10 \end{pmatrix} \cdot v(t) + \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix} = 0_2$$

$$\dot{v}(t) - (A + A^*) \cdot v(t) = S \tag{4}$$

$v(t) = e^{\int (A+A^*) \cdot dt} \cdot [C_2 + \int S \cdot e^{-\int (A+A^*) \cdot dt} dt] = e^{(A+A^*) \cdot t} \cdot C_2 - S \cdot (A + A^*)^{-1} = e^{(A+A^*) \cdot t} \cdot C_2 - C_1$, , with: $C_1 = S \cdot (A + A^*)^{-1}$. According to the restriction $p(1) = e^A$ it can be calculated C_2, because $v(1) = -[p(1)]^{-1} = -e^{-A}$:

$$C_2 = -e^{-2 \cdot A - A^*} + e^{-A - A^*} \cdot S \cdot (A + A^*)^{-1}$$

Consequently:
$v(t) = e^{(A+A^*) \cdot t} \cdot C_2 - C_1 \Rightarrow p(t) = [C_1 - e^{(A+A^*) \cdot t} \cdot C_2]^{-1}$ with: $C_1 = S \cdot (A + A^*)^{-1}$ and: $C_2 = -e^{-2 \cdot A - A^*} + e^{-A - A^*} \cdot S \cdot (A + A^*)^{-1}$.

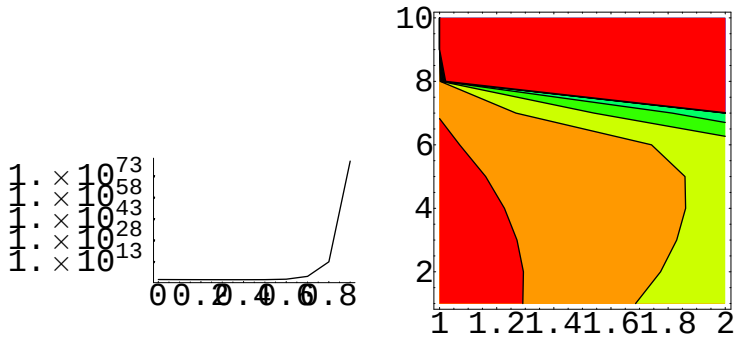
II. As a result of the above proposition, the optimal command is:

$$u^0(t, x) = -\sigma^{-1}(t) \cdot B^* \cdot p(t) \cdot x(t)$$

The optimal trajectory is the solution for the differential equation:

$$\dot{x}(t) = [A - S \cdot p(t)] \cdot x(t) = M(t) \cdot x(t); x(t) = E \cdot e^{\int_0^t M(\nu) d\nu}$$

The optimal trajectory is:
with the initial restriction :

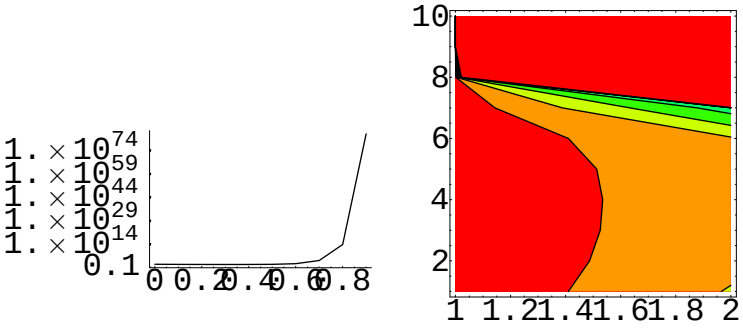


$$x(0) = E = x_0 \rightarrow E = x_0 \Rightarrow x(t) = x_0 \cdot e^{\int_0^t M(\nu)d\nu},$$
$$M(\nu) = A - S \cdot p(\nu) = A - S \cdot [C_1 - e^{(A+A^*)\cdot\nu} \cdot C_2]^{-1}$$

Consequently, the optimal solution is:

$$u^0(t,x) = -\sigma^{-1}(t) \cdot B^* \cdot [C_1 - e^{(A+A^*)\cdot t} \cdot C_2]^{-1} \cdot x_0 \cdot e^{\int_0^t M(\nu)d\nu}$$

The optimal command is:



III. Closing remark

The algorithm does not depend on choice of the A and B matrices. For every pair of diagonal (symmetric) matrices, A and B, and for any initial inputs, with respect to the restrictions in the theory, this algorithm can be used. If $v : (T_1, T_2) \rightarrow \mathcal{L}(X, X)$ is continuous differentiable and the matrices R, S are depending on t, let us try to use together, the classic analytic and numerical technics, in order to solve the following equation:$(X = R^n)$

$$\begin{pmatrix} \dot{v}_{11}(t).. & \dot{v}_{1n}(t) \\ .. & .. \\ \dot{v}_{n1}(t).. & \dot{v}_{nn}(t) \end{pmatrix} = \begin{pmatrix} r_{11}(t).. & r_{1n}(t) \\ .. & .. \\ r_{n1}(t).. & r_{nn}(t) \end{pmatrix} \cdot \begin{pmatrix} v_{11}(t).. & v_{1n}(t) \\ .. & .. \\ v_{n1}(t).. & v_{nn}(t) \end{pmatrix} \pm$$
$$\pm \begin{pmatrix} s_{11}(t).. & s_{1n}(t) \\ .. & .. \\ s_{n1}(t).. & s_{nn}(t) \end{pmatrix}.$$

This equation is equivalent with n linear differential systems, each of them with the system matrix equal with R(t):

$$\begin{pmatrix} \dot{v}_{i1}(t) \\ ... \\ \dot{v}_{in}(t) \end{pmatrix} = \begin{pmatrix} r_{11}(t).. & r_{1n}(t) \\ .. & .. \\ r_{n1}(t).. & r_{nn}(t) \end{pmatrix} \cdot \begin{pmatrix} v_{i1}(t) \\ ... \\ v_{in}(t) \end{pmatrix} \pm \begin{pmatrix} s_{i1}(t) \\ ... \\ s_{in}(t) \end{pmatrix}, i = \overline{1, ..., n}$$

which are to be solved in the usual way: the homogenous systems, by using the eigenvalues which are the same for each system and the linear systems, using the variation of constants method.
It can be demonstrated, using the definition of the differential in every point of interest t_0, that $A(t) \cdot B(t), U^{-1}(t), U^k(t)$ are also differentiable:

$$A(t), B(t), U(t) \epsilon \mathcal{M}_n(R), \forall t \epsilon [T_1, T_2], k \epsilon \mathbf{N}.$$

According to Kalman theory, we have to solve a Bernoulli equation (case $\rho(t) = 0$) :

$$\dot{p}(t) + R(t) \cdot p(t) - S(t) \cdot p^k(t) = 0_n, (R(t) = A(t) + A^*(t))$$

We can determine numerically the inverse matrix $[p(t_0)^k]^{-1}$ in every t_0 of interest. We multiply with $[p(t)^k]^{-1}, \forall t \epsilon [T_1, T_2]$:

$$\dot{p}(t) \cdot [p(t)]^{-k} + R(t) \cdot [p(t)]^{-(k-1)} - S(t) \cdot I_n = 0_n$$

Let : $\dot{v}(t) = \dot{p}(t) \cdot [p(t)]^{-k}$ and we have the same equation in the last paragraph:

$$\dot{v}(t) - (k-1) \cdot R(t) \cdot v(t) - S(t) = 0_n \tag{5}$$

which can be solved according to the above considerations. Consequently, we can generally use this algorithm in order to solve Bernoulli equations with matrices coefficients and determine, according to the theory, the optimal solution of the control problem with quadratic cost. For k=n=2 and constant matrices A, B, S, we are led to the equation (4) from the first paragraph.

References

[1] St. Mirica, *Differential equations and equations with partial derivatives*, Bucurest University Ed., Bucharest, Romania, 1999.

[2] St. Mirica, *Differential and integral equations*, Bucurest University Ed., Bucharest, Romania, 2000.

[3] S. Sburlan, L. Larbu, C. Mortici, *Differential and integral equations and dynamic systems*, Exponto Ed., Constantza, Romania, 1999.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Some generalizations of hyperbolic A-properness

Letitia Ion
Department of Mathematical Sciences
Constantza Maritime University, Romania
Letiziaion@yahoo.com

Abstract

We are concerned with the approximation-solvability of the semilinear operator equation $Lu+Su=f$, where L is a noninvertible linear operator and S is a nonlinear perturbation, whose assumptions interact topological and spectral properties of L. This leads to the so called Fredholm factorization associated with L, due to W. Krawcewitz [2]. By Mawhin's L-convergence [3], the class of nonlinear mappings of type (S_L) and $(S_L)_+$ for which $L+S$ is A_L-proper, extends the monotone-like operators and verifies conditions for the L-approximation solvability. We generalize this class with an adjoint Fredholm factorization and we obtain solution existence results.

Key Words: Approximation solvability, hyperbolic A-property, Fredholm factorization, L-convergence, mappings of modified type (S_L) and $(S_L)_+$

1 Introduction

W.V. Petryshyn [8] studied the elliptic case when $\dim KerL < \infty$. D. Pascali [5] introduced an A-proper technique for almost selfadjoint operators in a real Hilbert space H, which covers also the case when $N(L) = KerL$ is infinite-dimensional. This is called the *hyperbolic A-properness* because L involved inherits the properties of the generalized d'Alembertian with periodic boundary value conditions.

Let $R(L)$ be the range of L and suppose that $L : H \to H$ is almost selfadjoint, that means, L is closed, densely defined and such that $R(L) = (N(L))^{\perp}$. In other words, $H = N(L) + R(L)$ and let P be the orthogonal projection onto $N(L)$.

J. Mawhin [3] said that a sequence $\{u_n\}$ is *L-convergent* and write $u_n \overset{L}{\to} u$ if $Pu_n \to Pu$ and $(I-P)u_n \rightharpoonup (I-P)u$.

Let $\{X_n\}$ be a filtration of $N(L)$, i.e., a sequence of increasing finite - dimensional subspaces of $N(L)$ such that their reunion is dense in $N(L)$ and let P_n be the orthogonal projection onto X_n.

ISSN-1584-5990

We consider now the sequence $\{H_n\}$ of subspaces in H, finite - dimensional in the first component, of the form $H_n = X_n + R(L)$ and the associated orthogonal projections $J_n = P_n + (I - P) : H \to H_n$.

The double sequence $\gamma_0 = \{H_n, J_n\}$ is an admissible approximation scheme for Hilbert spaces, that is, $Pu_n \to Pu$ for each $u \in N(L)$.

The idea of L-approximation solvability is to replace the initial equation

$$Lu + Su = f \tag{1}$$

by a sequence of approximate equations

$$Lu_n + Su_n = J_n f \tag{2}$$

in H_n, to which more classical methods of resolution can be applied, and to extract a solution of the equation (1) from a aequence of solutions of equations (2).

D. Pascali[5] extended the A-properness by

Definition 1 *A mapping $T : H \to H$ is A_L-proper at a point $g \in H$ with respect to γ_0 if the restrictions $T_n : H_n \to H_n$ are continuous for every n and whenever $\{m\}$ is a sequence of natural numbers and $\{u_m\}$ is a corresponding bounded sequence such that $T_m u_m \to g$, it follows that $\{u_m\}$ is L-convergent to $u \in H$, at last on a subsequence, and $Tu = g$.*

The L-approximation solvability of the equation (1) involves two steps:

a) Finite-dimensional solvability of (2), i.e., there exists a solution $u_n \in D(L) \cap \overline{H_n \cap \Omega)}$, Ω a bounded open set in H, of the equation $Lu_n + Su_n = 0$ for n large enough;

b) Prove that $L + S$ is an A_L-proper mapping.

The first step respresents actually an apriori estimate and consists in finding operator assumptions equivalent with an infinity non-vanishing conditions

$$L_n u_n + (1-\lambda) P_n u_n + \lambda S_n u_n \neq 0, \forall u_n \in D(L) \cap \partial(\Omega \cap H_n), \lambda \in (0,1), n \geq n_0 \tag{3}$$

Since the restriction $L|_{D(L)\cap R(L)}$ is one-to-one we may define the right inverse of L associated with P, $K = \left(L|_{D(L)\cap R(L)}\right)^{-1}(I - P)$, which is a continuous operator, by the closed graph theorem.

Regarding the second step, we introduce a large class of nonlinear perturbations, than demicontinuous pseudomonotone operators is stated in the following:

Definition 2 *A mapping $S : \overline{\Omega} \to H$ is said to be L-pseudomonotone if for every sequence $\{u_n\} \subset \overline{\Omega}$ such that $u_n \overset{L}{\to} u$ and $limsup(Su_n, u_n - u) = 0$ it follows that $Su_n \rightharpoonup Su$ in H.*

S. Sburlan [10] pointed out the relationship between hyperbolic A-properness and alternative methods while T. Kaczinski and W. Krawcewitz [1] used it to study hyperbolic inclusions. W. Krawcewitz [2] extended the above theory when L and S are mappings between two real Banach spaces.

2 Fredholm factorization

We shall detail now these generalization following up the ideas in reports of D. Pascali [6], [7].

Let X and Y be real Banach spaces and $D(L)$ a dense subspace of X. Assume that $L : D(L) \to Y$ is a linear operator satisfying the following hypotheses:

(L_1) L is closed operator;

(L_2) $N(L)$ and $R(L)$ are closed and there are two closed subspaces $X_0 \subset X$ and $Y_0 \subset Y$ such that $X = N(L) + X_0$ and $Y = Y_0 + R(L)$;

(L_3) $\dim N(L) = co \dim R(L) = \infty$.

Let $P : X \to X_0$ and $Q : Y \to R(L)$ be the linear projections associated with the direct splittings in (L_2).

Denote by $\{Y_n'\}$ a *filtration* of Y_0, that is, for every $n \in N$, $\{Y_n'\}$ is a finite-dimensional subspace of Y_0 such that $Y_n' \subset Y_{n+1}'$ and the union $\bigcup Y_n'$ is dense in Y_0 . Consider also a filtration $\{X_n'\}$ in $N(L)$ such that $\dim X_n' = \dim Y_n'$, which is always possible by virtue of (L_3).

There are $P_n' : N(L) \to X_n'$ and $Q_n' : Y_0 \to Y_n'$ linear projections such that

$$P_n'x \to x\,, \forall x \in N(L), n \to \infty \quad \text{and} \quad Q_n'y \to y,\ \forall y \in Y_0\,,\ n \to \infty.$$

For each n, we put $X_n = X_n' + X_0$ and $Y_n = Y_n' + R(L)$ and define

$$P_n = P + P_n'(I - P) : X \to X_n\ ,\ Q_n = Q + Q_n'(I - Q) : Y \to Y_n,$$

the linear projections over X_n and Y_n, respectively, and note that

$$P_nx \to x, \forall x \in X\ ,\ Q_ny \to y,\ \forall y \in Y.$$

If $L_n : X_n \cap D(L) \to Y_n$ is the restriction $L\ |_{X\cap D(L)}$ to X_n it can easily show that L_n is a Fredholm operator of index zero. This is the reason why the special admissible approximation scheme $\Gamma = (\{X_n\}\,,\ \{P_n\},\ \{Y_n\},\ \{Q_n\})$ designed above is called the *Fredholm factorization* associated with L, in the W.Krawcewicz's sense.

For Ω an open set in X and $S : \overline{\Omega} \to Y$ a nonlinear mapping we denote

$$\Omega_n = \Omega \cap X_n \quad \text{and} \quad S_n : \overline{\Omega}_n \to Y_n \text{ where } S_n(x) = Q_nS(x),\ \forall x \in \Omega_n$$

and suppose that $\Omega_n \neq \emptyset, \forall n \in N$.

We say S to be an A-proper mapping (weakly A-proper) if the sequence of solutions of the approximant equations

$$L_nu_n + S_nu_n = f_n \to f$$

containes at least a subsequence converging (weakly) to a solution of the original equation

$$Lu + Su = f.$$

The restriction $L\mid_{X_0}: X_0 \cap D(L) \to R(L)$ being injective, we may consider its inverse $K : R(L) \to X_0$, which is a bounded linear operator by virtue of the closed graph theorem.

We recall [2] that a couple $\left\{X_n, \tilde{P}_n\right\}$ is called a Π_α-approximation scheme for a separable Banach space X, if $\left\|\tilde{P}_n\right\| \leq \alpha, \forall n \in N$ and $\tilde{P}_n\tilde{P}_j = \tilde{P}_j$ for $j \geq n$.

3 Mappings of type (S_L) and $(S_L)_+$

We consider a reflexive Banach space X and consider an operator $L : D(L) \subseteq X \to X^*$ satisfying hypotheses (L_1),(L_2),(L_3). We have $Q := P^*$ and $Y_0 := R(P^*)$ and there exists a Fredholm factorization $(\{X_n\}, \{P_n\}, \{X_n^*\}, \{P_n^*\})$ associated with the operator L, so that $X_n^* = R(P_n^*)$ and $P_n = P + P_n'(I - P) : X \to X_n$, $P_n' : N(L) \to X_n'$ where $\{X_n', P_n'\}$ is just an Π_α - approximation scheme for $N(L)$.

An intermediate step in the relationship between the A-properness and L-pseudomonotonicity of S is established in the following

Theorem 1 *Assume that the Banach space X is reflexive and the linear operator $L : X \to X^*$ satisfies the hypotheses (L_1), (L_2), (L_3). Let $S : X \to X^*$ be a bounded mapping for which KQS is compact. Then any sequence $\{u_n\}$, $u_n \in X_n \cap D(L)$ with the properties $u_n \rightharpoonup u$ and $Lu_n + P_n^*Su_n \to f$ contains a subsequence $\{u_{n'}\}$ such that*

$$u_{n'} \overset{L}{\to} u \quad \text{and} \quad (Su_{n'}, u_{n'} - u) \to 0. \tag{4}$$

The theorem show that, in the case of bounded perturbations, the L - pseudomonotonicity of S is also a necessary assumption for the approximation-solvability of equation (2).

For L-pseudomonotone mappings, it can be easily prove:

Corollary 1 *Let $L : X \to X^*$ a bounded mapping such that KQS is compact and S is L-pseudomonotone mapping. Then $L + S$ is weakly A_L-proper.*

Remark Replacing the L - pseudomonotonicity of S with the strong L - monotonicity, that,

$$(Su - Sv, u - v) \geq \beta \left\|(I - P)(u - v)\right\|,$$

with $\beta : R^+ \to R^+$ a continuous function such that $\beta(0) = 0$ and $\beta(t) \to 0$ as $t \to 0$, we have a further consequence,

Corollary 2 *Let $L : X \to X^*$ a bounded mapping such that KQS is compact and S is a demicontinuous strong L-monotone mapping. Then $L + S$ is A_L-proper.*

We introduce now classes of nonlinear mappings for which, the sum $L + S$ is an A_L-proper mapping, generalizing some results of W.V.Petryshyn [9].

Definition 3 *A mapping $S : X \to X^*$ is said to be of type (S_L) if for every sequence $\{u_n\} \subset X$ such that $u_n \overset{L}{\to} u$ and $\lim (Su_n, u_n - u) \to 0$ it follows that $u_n \to u$.*
Likewise, a mapping $S : X \to X^*$ is said to be of type $(S_L)_+$ if for every sequence $\{u_n\} \subset X$ such that $u_n \overset{L}{\to} u$ and $\limsup (Su_n - Su, u_n - u) \leq 0$ it follows that $u_n \to u$.

The definitions generalize the mappings of type (S) and of type $(S)_+$, respectively.

A criterium of L-aproperness of the sum is given by

Theorem 2 *Let X be a reflexive Banach space and $L : X \to X^*$ a linear operator that satisfies the hypotheses (L_1), (L_2), (L_3). If $S : X \to X^*$ is a bounded demicontinuous mapping of type (S_L) (or $(S_L)_+$) such that KQS is compact then $L + S$ is A_L-proper.*

4 A generalization of mappings of type (S_L) and $(S_L)_+$

We consider now the adjoint mapping $L^* : D(L^*) \to X^*$ where $D(L^*) \subset Y^*$. Since $N(L^*) = R(L)^{\perp}$and $R(L^*) = N(L)^{\perp}$, we can put $X_0^* = R(P^*)$ and $Y_0^* = R(Q^*)$. Hence L^* satisfies the hypotheses (L_1), (L_2), (L_3) and $\Gamma^* = (\{Y_n^*\}, \{Q_n^*\}, \{X_n^*\}, \{P_n^*\})$ is a Fredholm factorization associated to L^*, where $X_n^* = R(P_n^*)$ and $Y_n^* = R(Q_n^*)$, the so-called *adjoint Fredholm factorization.*

For what follows, we need first

Definition 4 *A mapping $k : X \to Y^*$ is called $L-$continuous at $u \in X$ if $u_n \overset{L}{\to} u$ implies $k(u_n) \overset{L^*}{\to} k(u)$.*

We are ready to introduce now,

Definition 5 *Let $k : X \to Y^*$ a L - continuous mapping with $k(0) = 0$. S is said to be of modified type (S_L) if for every sequence $\{u_n\} \subset X_n$ such that $u_n \overset{L}{\to} u$ and*

$$(Su_n - SP_n u, k(u_n - P_n u)) \to 0$$

it follows that $u_n \to u$.

Similarly,

Definition 6 *A mapping $S : X \to Y$ is said to be of modified type $(S_L)_+$ if for every sequence $\{u_n\} \subset X_n$ such that $u_n \overset{L}{\to} u$ and*

$$\limsup (Su_n - SP_n u, k(u_n - P_n u)) \leq 0$$

it follows that $u_n \to u$.

Let $k_n = Q_n^* k \mid_{X_n}$ and we suppose that $k : X \to Y^*$ satisfies for every $u \in X_n$, $f \in Y$ the condition $(Q_n f, k_n u) = (f, ku)$, $\forall n \in N$.

In this conditions, the L - pseudomonicity has the the following generalization:

Definition 7 *A continuous mapping $S : X \to Y$ is said to be $k - L$ - pseudomonotone if for every sequence $\{u_n\} \subset X_n$ such that $u_n \overset{L}{\to} u$ and*

$$\limsup \left(Su_n - SP_n u, k\left(u_n - P_n u\right)\right) \leq 0$$

it follows that $\left(Su_n - SP_n u, k\left(u_n - P_n u\right)\right) \to 0$ and $Su_n \rightharpoonup Su$.

A variant of Theorem 1 is,

Theorem 3 *Let $S : X \to Y$ be a bounded continuous mapping for which KQS is compact. Then any sequence $\{u_n\}$, $u_n \in X_n \cap D(L)$ with the properties $u_n \rightharpoonup u$ and $Lu_n + Q_n S u_n \to f$ contains a subsequence $\{u_j\}$ such that*

$$u_j \overset{L}{\to} u \quad \text{and} \quad \left(Su_j - SP_j u, k\left(u_j - u\right)\right) \to 0. \tag{5}$$

Proof. We prove that $u_j \overset{L}{\to} u$. With respect to the complete projection scheme $\left(\{X_n\}, \{P_n\}; \{Y_n\}, \{Q_n\}\right)$, project $Lu_n + Q_n S u_n \to f$ and obtain

$$Lu_j + QSu_j \to Qf \quad \text{and} \quad Q_j' Su_j \to (I - Q)f. \tag{6}$$

Since $K : R(L) \to X_0$ is continuous, we have $Pu_j + KQSu_j \to KQf$. By the hypothesis of compactness of KQS the set $\{KQSu_j\}$ contains a subsequence converging to $KQf - Pu$, because $Pu_j \rightharpoonup Pu$. Hence $Pu_j \to Pu$ and $Pu \in D(L)$. As $(I - P)u_j \rightharpoonup (I - P)u$, we conclude that $u_j \overset{L}{\to} u$.

By the L-continuity of k in zero, $k_j - P_j u \overset{L}{\to} 0$ implies $k\left(u_j - P_j u\right) \overset{L^*}{\to} 0$.

We prove that $\left(Su_j - SP_j u, k\left(u_j - u\right)\right) \to 0$. We derive

$$\left|\left(Su_j - SP_j u, k\left(u_j - P_j u\right)\right)\right| = \left|\left(Q_j Su_j - Q_j SP_j u, k_j\left(u_i - P_j u\right)\right)\right| \leq$$

$$\leq \left|\left(QSu_j - QSP_j u, k_j\left(u_j - P_j u\right)\right)\right| + \left|\left(Q_j' Su_j - Q_j' SP_j u, k_j\left(u_j - P_j u\right)\right)\right| =$$

$$= \left|\left(QSu_j - QSP_j u, Q^* k\left(u_j - P_j u\right)\right) + \left(Q_j' Su_j - Q_j' SP_j u, k_j\left(u_j - P_j u\right)\right)\right|.$$

Since $k\left(u_j - P_j u\right) \overset{L^*}{\to} 0$ then $Q^* k\left(u_j - P_j u\right) \to 0$. Since S is bounded, there is $M \geq 0$ such that $\|QSu_j - QSP_j u\| \leq M$ for every n. Consequently,

$$\left(QSu_j - QSP_j u, Q^* k\left(u_j - P_j u\right)\right) \to 0.$$

On the other side, $Q_j' Su_j = Q_j' Lu_j + Q_j' Su_j = Q_j' f_j \to (I - Q)\, f$ implies $Q_j' SP_j u \to (I - Q)\, Su$ because S is continuous and $P_j u \to u$.

Hence , $Q_j' Su_j - Q_j' SP_j u \to (I - Q)\, f - (I - Q)\, Su$ that implies

$$\left(Q_j' Su_j - Q_j' SP_j u, k_j\left(u_j - P_j u\right)\right) \to 0$$

and furthermore, $\left(Su_j - SP_j u, k\left(u_j - P_j u\right)\right) \to 0$ as $j \to \infty$.□

To Theorem 2 corresponds the following:

Theorem 4 *Let X and Y be reflexive Banach spaces and the linear operator $L : D(L) \to Y$ satisfies the hypotheses (L_1), (L_2), (L_3). If $S : X \to Y$ is a bounded continuous mapping of type (S_L) (or $(S_L)_+$) such that KQS is compact then $L + S$ is A_L-proper*

Proof. It is analogous with the proof of Theorem 2. □

For $k - L-$ pseudomonotone it can be easily proved:

Corollary 3 *Let X and Y be reflexive Banach spaces and the linear operator $L : D(L) \to Y$ satisfies the hypotheses (L_1), (L_2), (L_3). If $S : X \to Y$ is a $k-L-$ pesudomonotone bounded continuous mapping such that KQS is compact then $L + S$ is weak $A_L -$ proper.*

5 Existence theorems

In this section we apply the A_L-proper mappings theory to obtain existence results of solution for the semilinear operator equation $Lu + Su = f$.

Let X be a reflexive Banach space and $L : D(L) \subseteq X \to X^*$ is a self-adjoint mapping satisfying hypotheses (L_1),(L_2),(L_3). We suppose that X^* is a uniformly convex space, that is, for any sequence $\{u_n\} \subset X$ with $u_n \rightharpoonup u$ and $\|u_n\| \to \|u\|$ follows $u_n \to u$. Denote by $J : X \to X^*$, the duality mapping $J(u) := \left\{ f \in X^*; (f,u) = \|u\|^2 \text{ and } \|f\| = \|u\| \right\}$.

We put $M := J(I - P) : X \to X^*$. It can be easily proved that M is a mapping of type $(S_L)_+$.

For now on, $U \subset X$ is a bounded neighborhood of zero.

Theorem 5 *Let $S : X \to X^*$ is a bounded L - pseudomonotone mapping for which KQS is compact. If for every $u \in \partial U \cap D(L)$ and $t \in [0,1)$ we have*

$$Lu + (1-t)Mu + tSu \neq 0, \tag{7}$$

then for every $\lambda \in (0,1)$, there is n_0 such that

$$Lu + (1-t)M_n u + tS_n u \neq 0. \tag{8}$$

for all $n \geq n_0, u \in (\partial U)_n \cap D(L)$ and $t \in [0, 1-\lambda)$.

A further consequence of this theorem is given by

Theorem 6 *Let $U \subset X$ a bounded neighbourhood about 0 and $S : X \to X^*$ is a bounded L-pseudomonotone mapping for which KQS is compact. If for every $u \in \partial U \cap D(L)$ and $t \in [0,1)$ we have*

$$Lu + (1-t)Mu + tSu = 0,$$

then there is $u \in \overline{Conv(U)} \cap D(L)$ such that

$$Lu + Su = 0.$$

We are interested now to obtain existence theorems for mappings of modified type (S_L). For this, we suppose that X and Y are Banach reflexive spaces and the mapping $k : X \to Y^*$ is antisymmetric, i.e. $k(-u) = -ku$ for every $u \in X$. First we need the following results

Theorem 7 *Let $S : X \to Y$ a continuous mapping of modified type (S_L). We suppose that there is a compact mapping $C : X \to Y$ and a functional $l : X \to R$ such that for any sequence $\{u_n\} \subset X_n$, $u_n \xrightarrow{L} u$ we have $l(u_n) \to 0$. If*

$$(Su - Sv, k(u-v)) + (Cu - Cv, k(u-v)) + l(u-v) \geq 0, \forall u, v \in X, \qquad (9)$$

then the mapping

$$h_t(u) := \frac{1}{1+t} Su + \frac{t}{|t|} S(-u), u \in X, t \in [0,1],$$

is of modified type (S_L).

Proof. Let $\{u_n\} \subset X_n$ with $u_n \xrightarrow{L} u$. Hence, we have

$$a_n^t := (h_t(u_n) - h_t(P_n u), k(u_n - P_n u)) \to 0, \ n \to \infty$$

and we put $k_n' = -k_n$, $u' = -u$, $\alpha = \dfrac{1}{1+t}$. Then

$$a_n^t := \alpha (Su_n - SP_n u, k(u_n - P_n u)) + t\alpha (SP_n u' - Su_n', k(P_n u' - u_n')).$$

Since $u_n \xrightarrow{L} u$ and $P_n u \to u$ it follows that $u_n - P_n u \xrightarrow{L} u$ and $u_n' - P_n u' \xrightarrow{L} 0$. From the compactness of C there are the subsequences $\{u_j\} \subset \{u_n\}$ and $\{u_j'\} \subset \{u_n'\}$ such that $C(u_j) \to y \in Y$ and $C(u_j') \to z \in Y$. So, we obtain that $C_j(u_j) \to y$ and $C_j(u_j') \to z$. This implies

$$\begin{array}{l} C_j := (C(u_j) - CP_j u, k(u_j - P_j u)) \to 0 \\ C_j' := \left(C(P_j u') - Cu_j', k\left(P_j u' - u_j'\right)\right) \to 0, j \to \infty \end{array}$$

Likewise, $l_j := l(u_j - P_j u) \to 0$ and $l_j' := \left(P_j u' - u_j'\right) \to 0$. We put

$$r_j := (Su_j - SP_j u, k(u_j - P_j u)) \quad \text{and} \quad r_j' := \left(SP_j u' - Su_j', k\left(P_j u' - u_j'\right)\right).$$

Hence, we have

$$\begin{array}{rcl} g_j^t & : & = a_j^t + \alpha c_j + \alpha l_j + t\alpha c_j' + t\alpha l_j' \\ & : & = \alpha (r_j + c_j + l_j) + t\alpha \left(r_j' + c_j' + l_j'\right) \to 0, \ j \to \infty. \end{array}$$

From (9) it follows that

$$p_j := r_j + c_j + l_j \geq 0 \quad \text{and} \quad p_j' := r_j' + c_j' + l_j' \geq 0, \ \forall j.$$

Since $g_j^t = \alpha p_j + t\alpha p_j' \to 0$ as $j \to \infty$ this implies $p_j \to 0$ and $p_j' \to 0$ as $j \to \infty$.

Because $c_j \to 0$ and $l_j \to 0$ we derive

$$(Su_j - SP_j u, k(u_j - P_j u)) = r_j = p_j - c_j - l_j \to 0,\ j \to \infty.$$

Likewise, we prove $\left(SP_j u' - Su'_j, k\left(P_j - u'_j u\right)\right) \to 0$, $j \to \infty$.

This implies that $u_j \to u$ (respectively, $Su_j \rightharpoonup Su$ and $Su'_j \rightharpoonup Su'$, and so $h_t(u_j) \rightharpoonup h_t u$).□

A further consequence says

Theorem 8 *Let $S : X \to Y$ a continuous mapping. We suppose that there is a compact mapping $C : X \to Y$ and a functional $l : X \to R$ such that $l(0) = 0$ and for any sequence $\{u_n\} \subset X_n$, $u_n \xrightarrow{L} 0$ we have $\limsup l(u_n) \leq l(0) = 0$. If*

$$\begin{gathered}(Su - Sv, k(u-v)) + (Cu - Cv, k(u-v)) + l(u-v) \geq \\ \geq d\left(\|(I-P)(u-v)\|\right), \forall u, v \in X\end{gathered} \tag{10}$$

where $d : R^+ \to R^+$ is a continuous function such that $d(t) \to 0$ implies $t \to 0$, then the mapping

$$h_t(u) := \frac{1}{1+t} Su - \frac{t}{1+t} S(-u),\ u \in X,\ t \in [0,1],$$

is of modified type (S_L).

Proof. We consider the case $t = 0$. Let $\{u_n\} \subset X_n$ with $u_n \xrightarrow{L} u$. In this case $h_0(u) = Su$. We suppose that

$$a_n^0 := (Su_n - SP_n u, k(u_n - P_n u)) \to 0,\ n \to \infty$$

We can suppose like in the theorem 7 that $Cu_n \to g \in Y$. Since $u_n - P_n u \xrightarrow{L} 0$ we derive

$$\begin{gathered}d\left(\|(I-P)(u_n - P_n u)\|\right) \leq (Su_n - SP_n u, k(u_n - P_n u)) + \\ + (Cu_n - CP_n u, k(u_n - P_n u)) + l(u_n - P_n u)\end{gathered}$$

Passing to a limit we have

$$\limsup d\left(\|(I-P)(u_n - P_n u)\|\right) \leq \limsup l(u_n - P_n u) \leq S(0) = 0,$$

and we obtain $limd\left(\|(I-P)(u_n - P_n u)\|\right) = 0$.

The hypotheses of d implies $(I-P)u_n \to (I-P)u$ and , since $u_n \xrightarrow{L} u$, it follows $u_n \to u$.

Furthermore, we consider the case $0 < t \leq 1$. Let $\{u_n\} \subset X_n$ with $u_n \xrightarrow{L} u$ and

$$a_n^t := (h_t(u_n) - h_t(P_n u), k(u_n - P_n u)) \to 0, n \to \infty.$$

Like in the proof of the Theorem 7, we derive,

$$\begin{gathered}a_j^t + \alpha c_j + \alpha l_j + t\alpha l'_j = \alpha(r_j + c_j + l_j) + t\alpha\left(r'_j + c'_j + l'_j\right) \geq \alpha d\left(\|(I-P) \cdot\right. \\ \left. \cdot (u_j - P_j u)\|\right) + t\alpha d\left(\left\|(I-P)\left(u'_j - P_j u'\right)\right\|\right) = d\left(\left\|(I-P)(u_j - P_j u)\right\|\right).\end{gathered}$$

Since $a_j^t \to 0$, $c_j \to 0$ and $u_j - P_j u \xrightarrow{L} 0$, $u'_j - P_j u' \xrightarrow{L} 0$, we can conclude like in the proof of the Theorem 7 that $\|u_j - P_j u\| \to 0$ and, furthermore, $u_n \to u$.□

Theorem 9 *Let $S : X \to Y$ a bounded continuous mapping of modified type (S_L) for which KQS is compact. If there is $r > 0$ such that for every $u \in [0,1]$ and $u \in D(L)$, $\|u\| = r$ we have*

$$(L+S)\,u \neq t\,(L+S)\,(-u)\,.$$

and S satisfies (9) then there exists $u \in \overline{B\,(0,r)} \cap D(L)$ such that $Lu + Su = 0$.

Proof. It obviously follows from the Theorems 2, 6 and 7□

References

[1] Kaczynski, T., Krawcewicz W., *Solvability of boundary value for the inclusion via the theory of multivalued maps*, Z. Anwendungen **7** (1988), 337-346

[2] Krawcewitz, W., *Resolution des equations semilineaires avec la partie lineaire a noyau de dimension infinie via de applications A-propres*, Dissertationes Math. (Rozprawy Mat.) **294** (1990)

[3] Mawhin, J., *Compacite, Monotonie et Convexite dans l'etude de problemes aux limites semi-lineaires*, Seminaire d'Analyse Moderne 19, Univ. de Sherbrooke, 1981

[4] Mawhin, J., Rubakowski, K.P., *Continuation theorems for semilinear equations in Banach spaces: A survey*, in "Nonlinear Analysis" (Th. M. Rassias, Ed.), World Sci.Publ.Co., Singapore (1987), 367-405

[5] Pascali, D., *Approximation-solvability of a semilinear wave equation*, Libertas Math. **4** (1984), 73-79

[6] Pascali, D., *On hyperbolic A-properness in Banach spaces*, The 5-th Congress of Romanian Mathematicians, June 22-28, 2003, Pitesti, Romania

[7] Pascali, D., *About some generalizations of hyperbolic A-properness*, Report 993-47-129, to Joint Math. Meetings, Phoenix, AZ, Jan. 7-10, 2004

[8] Petryshyn, W.V, *Using degree theory for densely defined A-proper maps in the solvability of semilinear equations with unbounded and noninvertible linear part*, Nonlinear Anal., **4** (1980), 259-281

[9] Petryshyn, W.V, *Generalized Topological Degree and Semilinear Equations*, Cambridge Tracts in Math. 117, Cambridge Univ.Press, 1995

[10] Sburlan, S., *Semilinear noncoercive problems*, Workshop on Differential Equations, 45-56, 1986

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

An Admissible Approximation Scheme for a Generalized Divergence Equation

Letitia Ion, Dumitru Deleanu
Department of Mathematical Sciences
Constantza Maritime University, Romania
Letiziaion@yahoo.com, dumitrudeleanu@yahoo.com

Abstract

This article establishes a link between the approximations schemes and an abstract difference method in the study of the Dirichlet problem for a generalized divergence equation. We define a class of A-proper-like operators including the finite difference schemes for divergence equations of 2m order. In this matter, we will develop ideas of R. Schumann and E. Zeidler [3]. More precisely, we are concerned with the following problem:

$$-\sum_{i=1}^{N}\left(D_i\left|D_iu\right|^{p-2}D_iu\right)+su=f \text{ on } G,\ p\geq 2$$
$$u=0 \quad \text{on} \quad \partial G$$

Defining an abstract approximation scheme feasible for finite difference methods, we investigate this problem with the corresponding difference equations. With the aid of Galerkin method, the main result prove the existence and the uniqueness of the solution for equation to whom converges in the discrete sense the sequence solutions of difference equations.

Keywords: External approximation schemes, A-properness, discrete Sobolev spaces, difference equations, Galerkin theorem

1 External approximation scheme

Let us consider the initial problem:

$$Tu=f, u\in X \tag{1}$$

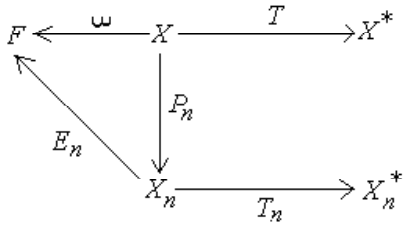
with the corresponding discretized problem:

$$T_nu_n=f_n, u_n\in X_n \tag{2}$$

and the approximation scheme:

Definition 1 *The approximation scheme* $\Gamma=(\{X_n\},\{P_n\},\{X_n^*\},\{E_n\})$ *is called an external admissible approximation scheme if the following hold:*

ISSN-1584-5990



1) X, F, X_n *are real spaces with* $\dim X_n < \infty, \forall n \in N$, F *reflexive*

2) *The operator* $\omega : X \to F$ *is linear, continuous, and injective*

3) *The operators* $P_n : X \to X_n$ *and* $E_n : X_n \to F$ *are linear and continuous such that* $\sup_n \|P_n\| < \infty$ *and* $\sup_n \|E_n\| < \infty$

4) $\forall u \in X$, $E_n P_n u \to \omega(u)$ *in* F, $n \to \infty$, (*compatibility condition*)

5) *If* $E_n u_n \rightharpoonup g$ *in* F, $n \to \infty$ *then* $g \in \omega(X)$ (*synchronization condition*)

The Petrysyn approximation scheme arises as a special case when $F = X$ and ω is the identity.

To describe the convergence of the difference method we use the following discrete convergences.

Definition 2 (a) *The sequence* $\{u_n\} \subset X_n$ *converges discretely to* $u \in X$, *i.e.,* $u_n \xrightarrow{d} u$ *if* $\lim_{n\to\infty} \|u_n - P_n u\|_{X_n} = 0$.

(b) *The sequence of functionals* $\{u_n^*\} \subset X_n^*$ *converges discretely* * *to* $u^* \in X^*$, *i.e.,* $u_n^* \xrightarrow{d^*} u^*$, *if* $\lim_{n\to\infty} \langle u_n^*, u_n \rangle_{X_n} = \langle u^*, u \rangle_X$ *for all sequences* $\{u_n\} \subset X_n$ *with* $\sup_n \|u_n\|_{X_n} < \infty$ *and* $E_n u_n \rightharpoonup \omega(u)$ *in* F, $n \to \infty$.

If there is a strictly monotone increasing and continuous function $c : R^+ \to R$ with $c(0) = 0$ and $c(r) \to \infty$ as $r \to \infty$ such that

$$\|T_n u - T_n v\|_{X_n^*} \geq c\left(\|u - v\|_{X_n}\right), \forall u, v \in X_n, \forall n \geq n_0, \tag{3}$$

we say that T satisfies the *stability condition* with respect to (w.r.t.) Γ.

Likewise, we say that T satisfies the *weak consistency condition* w.r.t. Γ if

$$T_n P_n u \xrightarrow{d^*} Tu\,, \forall u \in X. \tag{4}$$

In what follows we suppose that for each $f \in X^*$, there exists a sequence $\{f_n\} \subset X_n^*$ which converges discretely* to f, for all $n \geq n_0$.We prove a variant of Petryshyn's theorem:

Theorem 1 *Let $T : X \to X^*$ be a stable and weak consistent mapping w.r.t. an admissible approximation scheme Γ whose approximants operators $T_n : X_n \to X_n^*$ are continuous. Then equation $Tu = f, u \in X$ is uniquely approximation solvable for each $f \in X^*$ if and only if T is A-proper w.r.t. Γ and one-to-one.*

To prove this theorem we need the following:

Lemma 1 *The following hold for an admissible external approximation scheme as $n \to \infty$:*

(i) $u_n \xrightarrow{d} u$ *implies* $E_n u_n \to \omega(u)$ *in* F;

(ii) $u_n^* \xrightarrow{d^*} u^*$ *implies* $\sup_n \|u_n^*\|_{X_n^*} < \infty$;

(iii) $u_n^* \xrightarrow{d^*} 0$ *implies* $\|u_n^*\|_{X_n^*} \to 0$;

(iv) $E_n P_n u \to \omega(u)$, $\forall u \in U$ *implies* $E_n P_n u \to \omega(u)$, $\forall u \in X$, U *dense in* X.

Proof of Theorem 1. Let us suppose that T is A-proper w.r.t. Γ and one-to-one. We prove first that the equation (2) has exactly one solution. By (3), T_n is an one-to-one continuous mapping of X_n into X_n^*, $\forall n \geq N$, and hence, by topological degree theorem of invariance of domain, the image $R(T_n)$ is an open set of X_n^* . Meantime, the inequality (3) also implies that this image is closed in X_n^*. Indeed, for any sequence $\{f_k\} \subset R(T_n)$ with $f_k \to f$ there is a sequence $\{u_k\}$ in X_n such that $T_n u_k = f_k$. Then $\|T_n u_j - T_n u_k\| \to 0$ as $j, k \to \infty$, and by (3) and the properties of the function c , we derive that $\|u_k - u_j\| \to 0$ as $k, j \to \infty$, i.e.$\{u_n\}$ is a Cauchy sequence in X_n. Being finite-dimensional, X_n is complete, it follows that $u_k \to u$ and, by the continuity of T_n, $T_n u_k \to T_n u = f$, hence f lies in $R(T_n)$. Since $R(T_n)$ is both open and closed in X_n^*(as well as being nonempty), it follows that $R(T_n) = X_n^*$. Thus T_n is a bicontinuous mapping of X_n onto X_n^*, and $\forall n \geq N$, the equation (2) has a unique solution $u_n \in X_n$, say, for any $f \in X^*$.

We show now that for fixed $f \in X^*$, equation (1) has at most one solution $u \in X$. Let $Tu = Tv$. It follows from the stability condition that

$$c(\|P_n u - P_n v\|) \leq \|T_n P_n u - T_n P_n v\|, \ \forall n.$$

By (4), $T_n P_n u - T_n P_n v \xrightarrow{d^*} 0$. Lemma 1(iii) yields $\|T_n P_n u - T_n P_n v\| \to 0$.

Therefore, (4) implies $\|P_n u - P_n v\| \to 0$ as $n \to \infty$. Hence

$$\|E_n P_n u - E_n P_n v\| \leq \left(\sup_n \|E_n\|\right) \|P_n u - P_n v\| \to 0\,,\ n \to \infty.$$

The compatibility condition yields $\omega(u - v) = 0$; therefore, $u = v$.

We show that, for each $f \in X^*$, the equation (1) has exactly one solution $u \in X$. To this end, we choose a sequence $f_n \xrightarrow{d^*} f$ and u_n with $T_n u_n = f_n$. By

(4), it follows from $T_nP_n(0) \overset{d^*}{\to} T(0)$ and Lemma 1(ii) that $\sup_n \|T_nP_n(0)\| < \infty$. Because $P_n(0) = 0$, the following holds for all n:

$$\|f_n\| = \|T_nu_n\| \geq \|T_nu_n - T_n(0)\| - \|T_nP_n(0)\| \geq c\left(\|u_n\|\right) - \|T_nP_n(0)\|.$$

Therefore, $\sup_n c\left(\|u_n\|\right) < \infty$ and hence $\sup_n \|u_n\| < \infty$. The A-properness of the operator T ensures the existence of a subsequence $\{u_{n'}\}$ with $u_{n'} \overset{d}{\to} u$ and $Tu = f$.

We show now that $f_n \overset{d^*}{\to} f$ and $T_nu_n = f_n$ imply $u_n \overset{d}{\to} u$ and $E_nu_n \to \omega(u)$ in F. We just proved that each subsequence $\{u_{n'}\}$ of $\{u_n\}$ has another subsequence $\{u_{n''}\}$ with $u_{n''} \overset{d}{\to} u$ and $Tu = f$. The limit u is the same for all subsequences since $Tu = f$ has exactly one solution u. This implies the convergence of the entire sequence, i.e., we get $u_n \overset{d}{\to} u$. By lemma 1(i), $u_n \overset{d}{\to} u$ implies $E_nu_n \to \omega(u)$.

Converse. For brevity, we write n instead n'. Let $T_nu_n \overset{d^*}{\to} f$ with $\sup_n \|u_n\| < \infty$. Let u with $Tu = f$. We want to show that $u_n \overset{d}{\to} u$. Then the operator T is A-proper w.r.t. Γ. By (4), $T_nP_nu \overset{d^*}{\to} Tu$. This implies $T_nu_n - T_nP_nu \overset{d^*}{\to} 0$. Lemma 1(iii) yields $\|T_nu_n - T_nP_nu\| \to 0$ as $n \to \infty$. By (3), $c\left(\|u_n - P_nu\|\right) \leq \|T_nu_n - T_nP_nu\| \to 0$ as $n \to \infty$. This implies $\|u_n - P_nu\| \to 0$, i.e., $u_n \overset{d}{\to} u$. □

2 Discrete Sobolev spaces

We recall here some definitions and results about discrete Sobolev spaces useful in next sesions.

We choose a cube lattice in R^N with the grid mesh h. The coordinates of the lattice points P are integer multiples of h. To each lattice point P we assign a cube $c_h(P)$ with edges parallel to the axes having edge length h and P as midpoint. in this connection, we choose $c_h(P)$ to be, say, half-open and in fact so that the union of all cubes yields a disjoint decomposition of R^N. Let G be a bounded region in R^N with sufficiently smooth boundary, i.e., $\partial G \in C^{0,1}$.

Definition 3 *Let G_h be the set of all lattice points P with $\overline{c_h(P)} \subseteq \overline{G}$. These lattice points are called interior lattice points of G. The cubes belonging to G_h approximate G from the interior.*

By the set of *boundary lattice points* ∂G_h we understand all lattice points of G_h that belong to the boundary cubes. These are in the natural way the cubes whose closure does not lie entirely in the interior of the cube set belonging to G_h.

For $m = 1, 2, ..$ by $G_{h,m}$ we understand the set of all lattice points of G_h that have the distance from the boundary lattice points greater than or equal to mh.

A *lattice function* u_h is a function that assigns a real number to each lattice point of R^N.

Definition 4 *We define the difference quotient*

$$\nabla_i^{\pm} u_h(P) = \frac{u_h(P \pm he_i) - u_h(P)}{\pm h}.$$

where e_i is the unit vector in the direction of the ith coordinate.

If P has the coordinates $(hg_1, ..., hg_N)$ there the $g_1, ..., g_N$ are integers, then there results the point $P \pm he_i$ if one replaces g_i by $g_i \pm 1$.

The difference operator $\nabla_i^{\pm}$ corresponds to the differential operator $D_i = \partial/\partial\xi_i$.

Analogous to the differential operator $D^\alpha = D_1^{\alpha_1}....D_N^{\alpha_N}$, we define the difference operator $\nabla_{\pm}^{\alpha} = \left(\nabla_1^{\pm}\right)^{\alpha_1} ... \left(\nabla_N^{\pm}\right)^{\alpha_N}$ for $\alpha = (\alpha_1, ..., \alpha_N)$. For brevity, we agree to write $\nabla_i = \nabla_i^+$, $\nabla^\alpha = \nabla_+^\alpha$.

We now define discrete Lebesque spaces and Sobolev spaces parallel to $L_p(G)$ and $\overset{\circ}{W}{}_p^m(G)$, respectively. Hence, we make use of the discrete integral

$$\int u_h \, dx_h \overset{def}{=} \sum_p u_h(P) h^N.$$

Here, the summation is over all lattice points P of R^N. The sum always exists since the lattice functions to be considered are different from zero only in finitely many lattice points.

Definition 5 *Let $1 \le p < \infty$, $m = 1, 2, ..$ A discrete Lebesque space $L_p(G_h)$ is the set of all lattice functions that vanish identically outside G_h. We choose the norm to be*

$$|u_h|_p = \left(\int |u_h|^p \, dx_h \right)^{1/p}$$

A *discrete Sobolev space* $\overset{\circ}{W}{}_p^m(G_h)$ is the set of all lattice functions u_h that vanish identically outside $G_{m,h}$. We define

$$|u_h|_{m,p} = \left(\int \sum_{|\alpha| \le m} |\nabla^\alpha u_h|^p \, dx_h \right)^{1/p}; |u_h|_{m,p,0} = \left(\int \sum_{|\alpha| = m} |\nabla^\alpha u_h|^p \, dx_h \right)^{1/p}.$$

Proposition 1 *With the given norms, $L_p(G_h)$ and $\overset{\circ}{W}{}_p^m(G_h)$ are real Banach spaces. The norms $|\cdot|_{m,p}$ and $|\cdot|_{m,p,0}$ are equivalent on $\overset{\circ}{W}{}_p^m(G_h)$.*

Parallel to integration by parts there is also a formula for *discrete integration by parts*:

$$\int u \nabla_{\pm}^{\alpha} v \, dx_h = (-1)^{|\alpha|} \int \left(\nabla_{\pm}^{\alpha} u\right) v \, dx_h, \forall u, v \in \overset{\circ}{W}{}_p^m(G_h), \, |\alpha| \le m.$$

In conclusion, we mention an extension method that play an import role in the study of boundary problem in the next section.

Definition 6 *Let u_h be a lattice function. By a lattice function extended to the region G, which we shall also designed by u_h, we mean:*

$$u_h = \begin{cases} u_h(P),\ \forall x \in c_h(P), P \in G_h, \\ 0, \text{ otherwise.} \end{cases}$$

This way u_h is extended in a natural way as a constant to each cube that belongs to an interior lattice point, i.e., u_h is piecewise constant. Obviously,

$$\int\limits_G u_h\, dx = \int u_h\, dx_h.$$

So, the integral appearing on the left-hand side of this equation is always to be understood in this sense.

3 Quasi-eliptic equation

Now we investigate the generalized divergence problem

$$\begin{gathered} -\sum_{i=1}^{N} \left(D_i \left| D_i u \right|^{p-2} D_i u \right) + su = f \text{ on } G,\ p \geq 2 \\ u = 0 \quad \text{on} \quad \partial G \end{gathered} \tag{5}$$

with the corresponding difference equations

$$\begin{gathered} -\sum_{i=1}^{N} \left(\nabla_i^- \left| \nabla_i u_h(P) \right|^{p-2} \nabla_i u_h(P) \right) + s u_h(P) = \overline{g_h}(P), \forall P \in G_h, \\ u_h(P) = 0 \quad , \quad \forall \quad P \in \partial G_h \end{gathered} \tag{6}$$

We suppose that G is a bounded region in R^N, $N \geq 1$, with sufficiently smooth boundary, i.e., $\partial G \in C^{0,1}$, and s is a nonnegative real number. We choose a sufficiently small positive number h_0 so that the set G_h of interior lattice points is not empty for all h, $0 < h \leq h_0$. We understand $\overline{g_h}(P)$ to be the integral mean value of f over the cube $c_h(P)$ belonging to P, i.e.,

$$\overline{g_h}(P) = h^{-N} \int\limits_{c_h(P)} g(x) dx. \tag{7}$$

Definition 7 . *Let $X = \overset{\circ}{W}{}_p^1(G)$, $X_h = \overset{\circ}{W}{}_p^1(G_h)$, $2 \leq p < \infty$, $p^{-1} + q^{-1} = 1$. The generalized problem for* (5) *is:*

For a given $g \in L_q(G)$ we seek a function $u \in X$ such that

$$a(u, v) = f(v),\ \forall v \in X \tag{8}$$

with

$$a(u,v)=\int\limits_G\left(\sum_{i=1}^N|D_iu|^{p-2}\,D_iuD_iv+suv\right)\,dx,$$

$$f(v)=\int\limits_G gv\,dx.$$

Likewise, the generalized problem for (6) *is: we seek a lattice function* $u_h\in X_h$ *such that*

$$a_h(u_h,v_h)=f_h(v_h),\ \forall v_h\in X_h, \tag{9}$$

with

$$a_h(u_h,v_h)=\int\left(\sum_{i=1}^N|\nabla_iu_h|^{p-2}\,\nabla_iu_h\nabla_iv_h+su_hv_h\right)\,dx_h,$$

$$f_h(v_h)=\int\overline{g_h}v_h\,dx_h.$$

Proposition 2 *In the framework specified above we prove:*

(a) *The generalized problem for* (5) *has exactly one solution* u*;*

(b) *Difference equation. The generalized problem for* (6) *has exactly one solution* $u_h,\forall h,0<h\le h_0$*;*

(c) *Convergence. As* $h\to+0$*, the sequence* $\{u_h\}$ *converges to* u *in the sense:*

$$\int\limits_G\left(\sum_{i=1}^N|D_iu-\nabla_iu_h|^p+|u-u_h|^p\right)\,dx\to0, \tag{10}$$

$$\sum_{P\in G_{h,1}}\left(\sum_{i=1}^N|\nabla_i\bar{u}-\nabla_iu_h|^p+|\bar{u}-u_h|^p\right)\,h^N\to0. \tag{11}$$

In this connection, (10) and (11) describe the convergence of the difference method on the region G and on the lattice, respectively. In (10), u_h and ∇_iu_h denote the lattice functions extended to G. In (11), for brevity, we write u_h and $\overline{u}$ instead of $u_h(P)$ and $\overline{u}(P)$, where $\overline{u}(P)$ is the mean value of u at the point P in the sense of (7). Moreover, $\nabla_i\overline{u}$ and ∇_iu_h stand for the difference quotients $\nabla_i\overline{u}(P)$ and $\nabla_iu_h(P)$ of the lattice functions $P\to\overline{u}(P)$ and $P\to u_h(P)$, respectively.

Corollary 1 *The difference equation* (6) *and the corresponding generalized problem* (9) *are mutually equivalent.*

Proof. Multiplication of (6) by v_h, discrete integration, and discrete integration by parts yields (9). Furthermore, (6) follows from (9) by a reversal of this procedure.□

Therefore, instead of (9), we need to solve only the nonlinear system of equations (6). This happens with the aid of the following iteration method for $k = 0, 1, ..$:

$$\begin{aligned} &u_h^{(k+1)}(P) = u_h^{(k)}(P) - tg_P\left(u_h^{(k)}\right),\ P \in G_{h,1}, \\ &u_h^{(k+1)}(P) = 0\,,\ P \in \partial G_h, \end{aligned} \tag{12}$$

with the starting elements $u_h^{(0)}(P) \equiv 0$, where

$$g_P\left(u_h\right) = -\sum_{i=1}^{N} \nabla_i^- \left(\left|\nabla_i u_h(P)\right|^{p-2} \nabla_i u_h(P)\right) + s u_h(P) - \overline{g_h}(P).$$

Can be easily proved that:

Corollary 2 *Let $0 < h \leq h_0$ and $s > 0$. Then, for sufficiently small $t > 0$, the iteration method* (12) *converges as $k \to \infty$ to the solution u_h of the difference equation* (6).

We want to prove now the Proposition 2. To this end, we apply Theorem 1 for the approximation solvability of equation (1).

Proof. We construct the admissible approximation scheme from Figure 1. Let (h_n) be a sequence of positive numbers with $h_n \to 0$ as $n \to \infty$ and $0 < h_n \leq h_0, \forall n \in N$. We set

$$X = \overset{\circ}{W}{}_p^1\left(G\right),\ X_n = \overset{\circ}{W}{}_p^1\left(G_{h_n}\right),\ F = \prod_{i=1}^{N} L_p(G)\,,\ 2 \leq p < \infty.$$

Moreover, we set $u_n = u_{h_n}$. We equip X_n with the norm $\|\cdot\| = |\cdot|_{1,p,0}$. The operator $\omega : X \to F$ is defined by $\omega(u) = (u, D_1 u, ..., D_N u)$. The operator $P_n : X \to X_n$ results from forming the mean value, that is, we set $k = h_n$ and we define

$$P_n(u)(P) = \begin{cases} k^{-N} \int\limits_{c_k(P)} u(x)dx\,,\ P \in G_{k,1}, \\ 0\,,\ P \notin G_{k,1}. \end{cases}$$

The operator $E_n : X_n \to F$ is defined by $E_n u_n = (u_n, \nabla_1 u_n, ..., \nabla_N u_n)$, where the lattice functions extended to the region G occur on the right. This way we obtain $E_n u_n \in F$. By Lemma 1(iv) one needs to verify the compatibility condition $E_n P_n u \to \omega(u)$ in F as $n \to \infty$, $\forall u \in C_0^\infty(G)$ since $C_0^\infty(G)$ is dense in X.

The synchronization condition is obtained as follows. Let $E_n u_n \rightharpoonup g$ in F, $n \to \infty$ with $g = (u, U_1, ..., U_N)$. From this it follows that $u_n \rightharpoonup u$ in $L_p(G)$ and $\nabla_i u_n \rightharpoonup U_i$ in $L_p(G)$ for $n \to \infty$.

It is known the fact that the extended difference quotients converge weakly to the generalized derivatives (cf. Temam [4]). Hence, $U_i = D_i u$ and $g = \omega(u)$.

It is also known, there exists an operator $T : X \to X^*$ with

$$\langle Tu, v\rangle = a(u, v), \forall u, v \in X.$$

So, the generalized problem (8) is equivalent to $Tu = f,\ u \in X$.

Note that $g \in L_q(G)$ with $q^{-1} + p^{-1} = 1$ implies $f \in X^*$. Furthermore, the Galerkin theorem proves that equation $Tu = f, u \in X$ has exactly one solution $u \in X$. In addition, it also follows that

$$a(u, u-v) - a(v, u-v) \geq c \|u-v\|_{1,p,0}^p + s \int_G (u-v)^2 \, dx, \forall u, v \in X, c > 0, \text{fixed.}$$

We prove now (3): from the same Galerkin theorem, there results

$$a_{h_n}(v, u-v) - a_{h_n}(v, u-v) \geq c \|u-v\|_{1,p,0}^p + s \int (u-v)^2 dx_{h_n} \forall u, v \in X_n \quad (13)$$

In this connection, replace the derivative D_i, by the difference quotient ∇_i and the integrals by discrete integrals.

For $u \in X_n$, the mapping $v \to a_{h_n}(u, v)$ is a linear functional on the space X_n and, because $\dim X_n < \infty$, it is also continuous. Therefore, there exists an operator $T_n : X_n \to X_n^*$ with $\langle T_n u, v\rangle = a_{h_n}(u, v), \forall u, v \in X_n$. This implies

$$\|T_n u - T_n v\| \, \|u - v\| \geq |\langle T_n u - T_n v, u - v\rangle| \geq c \|u - v\|^p,$$

and hence, the stability condition

$$\|T_n u - T_n v\| \geq c \|u - v\|^{p-1}, \forall u, v \in X_n.$$

The generalized discrete problem (9) is equivalent to the operator equation $T_n u_n = f_n, u_n \in X_n$, where we write f_n for f_{h_n}.

To prove (4), we must show that $T_n P_n \xrightarrow{d^*} Tu, \forall u \in X$. By definition, this is equivalent to

$$a_{h_n}(P_n u, v_n) \to a(u, v), n \to \infty \quad (14)$$

for all sequences $\{v_n\}$ with the property $v_n \in X_n$, $\forall n \in N$ as well as $\sup_n \|v_n\| < \infty$ and $E_n v_n \rightharpoonup \omega(u)$ in F, $n \to \infty$.

Explicitly, relation (14) means that, as $n \to \infty$,

$$\int \left(\sum_i |\nabla_i \overline{u}_n|^{p-2} \nabla_i \overline{u}_n \nabla_i v_n + s \overline{u}_n v_n \right) dx \to \quad (15)$$

$$\to \int_G \left(\sum_i |D_i u|^{p-2} D_i u D_i v + suv \right) dx,$$

where $\overline{u}_n$ denotes the lattice function arising from the function $u \in X$, for all grid points $P \in G_{h_n,1}$, by forming the mean values according to (7). More precisely, in (12), the symbols $\overline{u}_n$, $\nabla_i \overline{u}_n$, v_n, $\nabla_i v_n$ denote the corresponding extended lattice functions. Recall that $u_n = u_{h_n}$.

Let us prove (12). From $E_n v_n \rightharpoonup \omega(u)$ it follows that, as $n \to \infty$, $v_n \rightharpoonup v$ and $\nabla_i v_n \rightharpoonup D_i v_n$ in $L_p(G)$. The compatibility condition $E_n P_n u \to \omega(u)$ means

that, as $n \to \infty$, $\overline{u}_n \to u$ and $\nabla_i \overline{u}_n \to D_i u$ in $L_p(G)$. Since $q \leq p$, the embedding $L_p(G) \subseteq L_q(G)$ is continuous. This implies that, as $n \to \infty$, $\overline{u}_n \to u$ and $\nabla_i \overline{u}_n \to D_i u$ in $L_q(G)$. The Nemyckii operator $v \to |v|^{p-2} v$ is continuous from $L_p(G)$ to $L_q(G)$, since $p/q = p - 1$. Thus, it follows that, as $n \to \infty$, $|\nabla_i \overline{u}_n|^{p-2} \overline{u}_n \to |D_i u|^{p-2} D_i u$ in $L_q(G)$. We obtain (12), by the convergence theorem [1].

Now we show that $f_n \overset{d^*}{\to} f$. Recall that $f(v) = \int\limits_G gv dx$, $\forall v \in X$. Since we write f_n for f_{h_n}, we have $f_n(v_n) = \int\limits_G \overline{g}_{h_n} v_n dx$, $\forall v_n \in X_n$, where $\overline{g}_h$ denotes the mean value in the sense of (7).

Now, let $\{v_n\}$ be a sequence with $v_n \in X_n$, $\forall n \in N$ with $\sup\limits_n \|v_n\| < \infty$ and $E_n v_n \rightharpoonup \omega(v)$ in F. This implies $v_n \rightharpoonup v$ in $L_p(G)$, $n \to \infty$, and hence $f_n(v_n) \to f(v)$, $n \to \infty$, by a Lemma of Schumann and Zeidler [3] and the convergence theorem. This means that $f_n \overset{d^*}{\to} f$.

The proof of Proposition 2 The unique approximation solvability (identical with the Proposition 2) is enssured by Theorem 1. In particular, the statements (10) and (11) follows from $E_n u_n \to \omega(u)$ in F as $n \to \infty$, and $u_n \overset{d}{\to} u$, i.e., $\|u_n - P_n u\|_{X_n} \to 0$ as $n \to \infty$.

In fact, this implies (10) and $|u_n - \overline{u}|_{1,p,0} \to 0$ as $n \to \infty$, and hence we get (11). Hence, note that $\overline{u} \to u$ in $L_p(G)$ as $n \to \infty$ by Lemma of Schumann and Zeidler, and hence $u_n \to \overline{u}$ in $L_p(G)$ as $n \to \infty$ by (10), i.e., $|u_n - \overline{u}|_p \to 0$ as $n \to \infty$.□

References

[1] Petryshyn, W.V., *Generalized Topological Degree and Semilinear Equations*, Cambridge Tracts in Math. 117, Cambridge Univ.Press, 1995

[2] Popa, C., *Metode Iterative pentru Sisteme Liniare*, Ed. Eurobit, Timisoara, 1996.

[3] Schuman, R., Zeidler E., *The finite difference method for quasilinear elliptic equations of order 2m*, Numer. Funct. Anal. and Optimiz., **1**(2), 1979, 11-194.

[4] Temam R., *Metode numerice de rezolvare a ecuatiilor functionale*, Editura Tehnica, 1973

[5] Zeidler,E., *Nonlinear functional Analysis and its Applications IIA, B*, Springer, 1990

[1] Let $1 < p, q < \infty$ and $p^{-1} + q^{-1} = 1$. From $u_n \to u$ in $L_p(G)$ as $n \to \infty$, and $v_n \rightharpoonup v$ in $L_q(G)$ as $n \to \infty$, it follows that $\int\limits_G u_n v_n dx \to \int\limits_G uv dx$ as $\to \infty$.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

An Existence Result for a Class of Nonlinear Difference Systems

Rodica Luca
Department of Mathematics, Gh. Asachi Technical University
Iasi, Romania
E-mail: rluca@math.tuiasi.ro

Abstract

In this paper we prove an existence and uniqueness result for the solutions of a nonlinear system with generalized second-order differences in a real Hilbert space, subject to some boundary conditions. An application to nonlinear differential systems with monotone operators is also presented.

Keywords: difference system, boundary conditions, monotone operator, differential system.

1 Introduction

Let H be a real Hilbert space with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. We consider in the space H the nonlinear system with generalized second-order differences

$$(S) \qquad \begin{cases} \Delta v_j - \varphi_j \Delta v_{j-1} + c_j A(u_j) \ni f_j^0 \\ -\Delta u_j + \varphi_j \Delta u_{j-1} + d_j B(v_j) \ni g_j^0, \quad j = \overline{1, N}, \end{cases}$$

with the boundary conditions

$$(BC) \qquad \begin{pmatrix} \Delta u_0 \\ -\Delta v_0 \end{pmatrix} \in \Lambda_1 \begin{pmatrix} v_1 \\ u_1 \end{pmatrix}, \quad \begin{pmatrix} -\Delta u_N \\ \Delta v_N \end{pmatrix} \in \Lambda_2 \begin{pmatrix} v_N \\ u_N \end{pmatrix},$$

where $N \in \mathbb{N}$, $N > 1$, $\Delta u_j = u_{j+1} - u_j$, $\Delta v_j = v_{j+1} - v_j$, for $j = \overline{0, N}$, φ_j, c_j, d_j, $j = \overline{1, N}$ are positive real constants, and A, B and Λ_1, Λ_2 are operators on H and H^2, respectively.

The system (S) and the boundary conditions (BC) can be written as

$$\begin{cases} v_{j+1} - (1 + \varphi_j) v_j + \varphi_j v_{j-1} + c_j A(u_j) \ni f_j^0 \\ -u_{j+1} + (1 + \varphi_j) u_j - \varphi_j u_{j-1} + d_j B(v_j) \ni g_j^0, \quad j = \overline{1, N} \end{cases}$$

and

$$\begin{pmatrix} u_1 - u_0 \\ -v_1 + v_0 \end{pmatrix} \in \Lambda_1 \begin{pmatrix} v_1 \\ u_1 \end{pmatrix}, \quad \begin{pmatrix} -u_{N+1} + u_N \\ v_{N+1} - v_N \end{pmatrix} \in \Lambda_2 \begin{pmatrix} v_N \\ u_N \end{pmatrix}.$$

ISSN-1584-5990

The above problem with $\varphi_j = 1,\ j = \overline{1,N}$ has been investigated in [11] (see also [12], [13] for some generalizations). In the last years we have studied a few classes of systems with first- or second-order differences, with finite or infinite numbers of equations, subject to various boundary conditions (see [8], [9], [10], [13]), and also the associated differential systems. The operators Λ_1 and Λ_2 from the boundary condition (BC) are general ones, and cover various classical conditions. For example, if $H = \mathbb{R}^2$ and $\Lambda_1 = \partial l_1$, $\Lambda_2 = \partial l_2$, where $l_1((u,v)^T) = au - bv$, $l_2((u,v)^T) = -cu + dv$, then the boundary conditions (BC) become $u_1 - u_0 = a$, $v_1 - v_0 = b$, $u_{N+1} - u_N = c$, $v_{N+1} - v_N = d$. For other difference equations in abstract spaces we refer the reader to the monographs [2] and [6].

In this paper we shall prove an existence and uniqueness result for the solutions of the problem (S), (BC), and then we shall briefly present an application of the obtained theorems to the study of the existence, uniqueness and asymptotic behaviour of the strong and weak solutions for a differential system in the Hilbert space H, subject to boundary conditions and initial data. In the proofs of our results we shall use some theorems from the theory of monotone operators (see the monographs [4], [5], [7]).

The assumptions that we shall use in the sequel are

(H1) a) The operators $A : D(A) \subset H \to H$, $B : D(B) \subset H \to H$ are maximal monotone, possibly multivalued.

b) There exist $a_0 > 0$, $b_0 > 0$ such that

i) for all $u_1,\ u_2 \in D(A)$, $\gamma_1 \in A(u_1)$, $\gamma_2 \in A(u_2)$ we have

$$\langle \gamma_1 - \gamma_2, u_1 - u_2 \rangle \geq a_0 \|u_1 - u_2\|^2;$$

ii) for all $v_1,\ v_2 \in D(B)$, $\delta_1 \in B(v_1)$, $\delta_2 \in B(v_2)$ we have

$$\langle \delta_1 - \delta_2, v_1 - v_2 \rangle \geq b_0 \|v_1 - v_2\|^2.$$

(H2) The operators $\Lambda_1 : D(\Lambda_1) \subset H^2 \to H^2$, $\Lambda_2 : D(\Lambda_2) \subset H^2 \to H^2$ are maximal monotone, possibly multivalued.

(H3) One of the below assumptions is verified:

a) $(\mathrm{int}D(\Lambda_1)) \cap (D(B) \times D(A)) \neq \emptyset$; $(\mathrm{int}D(\Lambda_2)) \cap (D(B) \times D(A)) \neq \emptyset$.

b) $D(\Lambda_1) \cap [(\mathrm{int}D(B)) \times (\mathrm{int}D(A))] \neq \emptyset$; $D(\Lambda_2) \cap [(\mathrm{int}D(B)) \times (\mathrm{int}D(A))] \neq \emptyset$.

c) $0 \in \mathrm{int}(D(\Lambda_1) - (D(B) \times D(A))) \cap \mathrm{int}(D(\Lambda_2) - (D(B) \times D(A)))$.

(H4) The constants $\varphi_j > 0$, for all $j = \overline{1,N}$.

(H5) The constants $c_j > 0,\ d_j > 0$, for all $j = \overline{1,N}$.

2 The existence and uniqueness result

We shall write the problem (S), (BC) as a first-order difference problem, by using a similar argument as that used in [1]. For this aim, we firstly define the positive numbers $\alpha_0 = 1$, $\alpha_j = 1/(\varphi_1 \varphi_2 \cdots \varphi_j)$, $j = \overline{1,N}$, that satisfy the relations $\alpha_j \varphi_j = \alpha_{j-1}$, for all $j = \overline{1,N}$. We introduce the Hilbert space $X = (H^N_\alpha)^2 (= (H^N)^2)$ with the scalar product

$$\langle (u_1,\dots,u_N,v_1,\dots,v_N)^T,(\widetilde{u}_1,\dots,\widetilde{u}_N,\widetilde{v}_1,\dots,\widetilde{v}_N)^T\rangle_X=\sum_{j=1}^{N}\alpha_j[\langle u_j,\widetilde{u}_j\rangle+\langle v_j,\widetilde{v}_j\rangle]$$

and the corresponding norm $\|\cdot\|_X$.

We now define the operator $\mathcal{A}: D(\mathcal{A})\subset X\to X$, $D(\mathcal{A})=\{(u_1,\dots,u_N,v_1,\dots,v_N)^T$, $(v_1,u_1)^T\in D(\Lambda_1),\ (v_N,u_N)^T\in D(\Lambda_2)\}$,

$$\mathcal{A}((u_1,\dots,u_N,v_1,\dots,v_N)^T)=\{(v_2-(1+\varphi_1)v_1+\varphi_1v_0,\dots,v_{N+1}-(1+\varphi_N)v_N$$
$$+\varphi_Nv_{N-1},-u_2+(1+\varphi_1)u_1-\varphi_1u_0,\dots,-u_{N+1}+(1+\varphi_N)u_N-\varphi_Nu_{N-1})^T,$$
$$(u_1-u_0,-v_1+v_0)^T\in\Lambda_1((v_1,u_1)^T),\ (-u_{N+1}+u_N,v_{N+1}-v_N)^T\in\Lambda_2((v_N,u_N)^T)\},$$

and the operator $\mathcal{B}: D(\mathcal{B})\subset X\to X$, $D(\mathcal{B})=\{(u_1,\dots,u_N,v_1,\dots,v_N)^T;\ u_j\in D(A),\ v_j\in D(B),\ j=\overline{1,N}\}$,

$$\mathcal{B}((u_1,u_2,\dots,u_N,v_1,v_2,\dots,v_N)^T)=\{(c_1\gamma_1,c_2\gamma_2,\dots,c_N\gamma_N,d_1\delta_1,d_2\delta_2,\dots,d_N\delta_N)^T,$$
$$\gamma_j\in A(u_j),\ \delta_j\in B(v_j),\ j=\overline{1,N}\}.$$

Lemma 1. *If the assumptions* (H2) *and* (H4) *hold, then the operator* $\mathcal{A}$ *is maximal monotone.*

Proof. We decompose the operator $\mathcal{A}$ in two operators, namely $\mathcal{A}_1$: $D(\mathcal{A}_1)=X\to X$,

$$\mathcal{A}_1((u_1,\dots,u_N,v_1,\dots,v_N)^T)=(v_2-(1+\varphi_1)v_1+\varphi_1v_0,\dots,v_{N+1}-(1+\varphi_N)v_N$$
$$+\varphi_Nv_{N-1},-u_2+(1+\varphi_1)u_1-\varphi_1u_0,\dots,-u_{N+1}+(1+\varphi_N)u_N-\varphi_Nu_{N-1})^T,$$

with $v_0=v_{N+1}=u_0=u_{N+1}=0$, and $\mathcal{A}_2: D(\mathcal{A}_2)=D(\mathcal{A})\subset X\to X$,

$$\mathcal{A}_2((u_1,\dots,u_N,v_1,\dots,v_N)^T)=\{(\varphi_1v_0,0,\dots,0,v_{N+1},-\varphi_1u_0,0,\dots,0,-u_{N+1})^T,$$
$$(u_1-u_0,-v_1+v_0)^T\in\Lambda_1((v_1,u_1)^T),\ (-u_{N+1}+u_N,v_{N+1}-v_N)^T\in\Lambda_2((v_N,u_N)^T)\}.$$

We evidently have $(\mathcal{A}_1+\mathcal{A}_2)(U)=\mathcal{A}(U)$, for all $U\in D(\mathcal{A})$. The first operator $\mathcal{A}_1$ is single-valued, everywhere defined and linear (so it is continuous). We shall prove that it is also monotone. For this aim, we introduce the new elements $x_j=\alpha_{j-1}(v_j-v_{j-1})$, $j=\overline{1,N+1}$ and $y_j=\alpha_{j-1}(u_j-u_{j-1})$, $j=\overline{1,N+1}$. Then we obtain the relations

$$v_{j+1}-(1+\varphi_j)v_j+\varphi_jv_{j-1}=v_{j+1}-v_j-\varphi_j(v_j-v_{j-1})=\tfrac{1}{\alpha_j}x_{j+1}-\varphi_j\tfrac{1}{\alpha_{j-1}}x_j=\tfrac{1}{\alpha_j}(x_{j+1}-x_j)$$

and

$$u_{j+1}-(1+\varphi_j)u_j+\varphi_ju_{j-1}=u_{j+1}-u_j-\varphi_j(u_j-u_{j-1})=\tfrac{1}{\alpha_j}y_{j+1}-\varphi_j\tfrac{1}{\alpha_{j-1}}y_j=\tfrac{1}{\alpha_j}(y_{j+1}-y_j).$$

The operator $\mathcal{A}$ with these new elements can be equivalently expressed as

$$\mathcal{A}((u_1,\dots,u_N,v_1,\dots,v_N)^T)=\Big\{\Big(\tfrac{1}{\alpha_1}(x_2-x_1),\tfrac{1}{\alpha_2}(x_3-x_2),\dots,\tfrac{1}{\alpha_N}(x_{N+1}-x_N),$$
$$-\tfrac{1}{\alpha_1}(y_2-y_1),-\tfrac{1}{\alpha_2}(y_3-y_2),\dots,-\tfrac{1}{\alpha_N}(y_{N+1}-y_N)\Big)^T,\ (y_1,-x_1)^T\in\Lambda_1((v_1,u_1)^T),$$
$$\Big(-\tfrac{1}{\alpha_N}y_{N+1},\tfrac{1}{\alpha_N}x_{N+1}\Big)^T\in\Lambda_2((v_N,u_N)^T)\Big\}.$$

The operator $\mathcal{A}_1$ is monotone, because

$$\langle\mathcal{A}_1(U),U\rangle_X=\sum_{j=1}^{N}\alpha_j\langle\frac{1}{\alpha_j}(x_{j+1}-x_j),u_j\rangle-\sum_{j=1}^{N}\alpha_j\langle\frac{1}{\alpha_j}(y_{j+1}-y_j),v_j\rangle$$
$$=-\sum_{j=1}^{N}\langle x_{j+1},u_{j+1}-u_j\rangle+\sum_{j=1}^{N}\langle x_{j+1},u_{j+1}\rangle-\sum_{j=1}^{N}\langle x_j,u_j\rangle+\sum_{j=1}^{N}\langle y_{j+1},v_{j+1}-v_j\rangle$$

$-\sum_{j=1}^{N}\langle y_{j+1},v_{j+1}\rangle+\sum_{j=1}^{N}\langle y_j,v_j\rangle=-\sum_{j=1}^{N}\langle\alpha_j(v_{j+1}-v_j),u_{j+1}-u_j\rangle+\sum_{j=1}^{N}[\langle x_{j+1},u_{j+1}\rangle-\langle x_j,u_j\rangle]$
$+\sum_{j=1}^{N}\langle\alpha_j(u_{j+1}-u_j),v_{j+1}-v_j\rangle-\sum_{j=1}^{N}[\langle y_{j+1},v_{j+1}\rangle-\langle y_j,v_j\rangle]=\langle x_{N+1},u_{N+1}\rangle-\langle x_1,u_1\rangle$
$-\langle y_{N+1},v_{N+1}\rangle+\langle y_1,v_1\rangle=0,$
for all $U=(u_1,\dots,u_N,v_1,\dots,v_N)^T\in X$.

Next, the operator $\mathcal{A}_2$ is also monotone, because

$\langle Z-\widetilde{Z},U-\widetilde{U}\rangle_X=\alpha_1\langle\varphi_1(v_0-\widetilde{v}_0),u_1-\widetilde{u}_1\rangle+\alpha_N\langle v_{N+1}-\widetilde{v}_{N+1},u_N-\widetilde{u}_N\rangle$
$+\alpha_1\langle-\varphi_1(u_0-\widetilde{u}_0),v_1-\widetilde{v}_1\rangle+\alpha_N\langle-u_{N+1}+\widetilde{u}_{N+1},v_N-\widetilde{v}_N\rangle=\langle v_0-\widetilde{v}_0,u_1-\widetilde{u}_1\rangle$
$+\alpha_N\langle v_{N+1}-\widetilde{v}_{N+1},u_N-\widetilde{u}_N\rangle-\langle u_0-\widetilde{u}_0,v_1-\widetilde{v}_1\rangle+\alpha_N\langle-u_{N+1}+\widetilde{u}_{N+1},v_N-\widetilde{v}_N\rangle$
$=\langle u_1-u_0-\widetilde{u}_1+\widetilde{u}_0,v_1-\widetilde{v}_1\rangle+\langle-v_1+v_0+\widetilde{v}_1-\widetilde{v}_0,u_1-\widetilde{u}_1\rangle$
$+\alpha_N\langle-u_{N+1}+u_N+\widetilde{u}_{N+1}-\widetilde{u}_N,v_N-\widetilde{v}_N\rangle+\alpha_N\langle v_{N+1}-v_N-\widetilde{v}_{N+1}+\widetilde{v}_N,u_N-\widetilde{u}_N\rangle$
$=\langle\begin{pmatrix}u_1-u_0\\-v_1+v_0\end{pmatrix}-\begin{pmatrix}\widetilde{u}_1-\widetilde{u}_0\\-\widetilde{v}_1+\widetilde{v}_0\end{pmatrix},\begin{pmatrix}v_1\\u_1\end{pmatrix}-\begin{pmatrix}\widetilde{v}_1\\\widetilde{u}_1\end{pmatrix}\rangle_{H^2}$
$+\alpha_N\langle\begin{pmatrix}-u_{N+1}+u_N\\v_{N+1}-v_N\end{pmatrix}-\begin{pmatrix}-\widetilde{u}_{N+1}+\widetilde{u}_N\\\widetilde{v}_{N+1}-\widetilde{v}_N\end{pmatrix},\begin{pmatrix}v_N\\u_N\end{pmatrix}-\begin{pmatrix}\widetilde{v}_N\\\widetilde{u}_N\end{pmatrix}\rangle_{H^2}\geq 0,$
for all $U=(u_1,\dots,u_N,v_1,\dots,v_N)^T$, $\widetilde{U}=(\widetilde{u}_1,\dots,\widetilde{u}_N,\widetilde{v}_1,\dots,\widetilde{v}_N)^T\in D(\mathcal{A}_2)$, $Z\in\mathcal{A}_2(U)$, $\widetilde{Z}\in\mathcal{A}_2(\widetilde{U})$, $(u_1-u_0,-v_1+v_0)^T\in\Lambda_1((v_1,u_1)^T)$, $(\widetilde{u}_1-\widetilde{u}_0,-\widetilde{v}_1+\widetilde{v}_0)^T\in\Lambda_1((\widetilde{v}_1,\widetilde{u}_1)^T)$, $(-u_{N+1}+u_N,v_{N+1}-v_N)^T\in\Lambda_2((v_N,u_N)^T)$, $(-\widetilde{u}_{N+1}+\widetilde{u}_N,\widetilde{v}_{N+1}-\widetilde{v}_N)^T\in\Lambda_2((\widetilde{v}_N,\widetilde{u}_N)^T)$.

The operator $\mathcal{A}_2$ is maximal monotone. To prove this statement, we shall use a generalization of Minty's Theorem to product spaces with various weights (see Theorem 7 from [8]). We choose $C=(\varphi_1,1,\dots,1,\varphi_1,1,\dots,1)^T\in\mathbb{R}^{2N}$ and we shall prove that for any element $Y_0=(f_1^0,\dots,f_N^0,g_1^0,\dots,g_N^0)^T\in X$, the equation

$$(CI+\mathcal{A}_2)(U)\ni Y_0 \tag{1}$$

has a solution $U\in D(\mathcal{A}_2)$.

The equation (1) is equivalent to

$$\begin{cases}\varphi_1u_1+\varphi_1v_0=f_1^0,\ \ u_j=f_j^0,\ \ j=\overline{2,N-1},\ \ u_N+v_{N+1}=f_N^0,\\ \varphi_1v_1-\varphi_1u_0=g_1^0,\ \ v_j=g_j^0,\ \ j=\overline{2,N-1},\ \ v_N-u_{N+1}=g_N^0,\end{cases} \tag{2}$$

with $(v_1,u_1)^T\in D(\Lambda_1)$, $(v_N,u_N)^T\in D(\Lambda_2)$, $(u_1-u_0,-v_1+v_0)^T\in\Lambda_1((v_1,u_1)^T)$, $(-u_{N+1}+u_N,v_{N+1}-v_N)^T\in\Lambda_2((v_N,u_N)^T)$, in the case $N>2$, and only the first and last relations in the case $N=2$.

For the elements u_1, v_1, u_N, v_N we obtain from (2), the system

$$\begin{cases}\begin{pmatrix}v_1\\u_1\end{pmatrix}+\begin{pmatrix}-u_0\\v_0\end{pmatrix}=\begin{pmatrix}\frac{1}{\varphi_1}g_1^0\\\frac{1}{\varphi_1}f_1^0\end{pmatrix}\\ \begin{pmatrix}v_N\\u_N\end{pmatrix}+\begin{pmatrix}-u_{N+1}\\v_{N+1}\end{pmatrix}=\begin{pmatrix}g_N^0\\f_N^0\end{pmatrix},\end{cases}$$

which is equivalent to

$$\begin{cases} \begin{pmatrix} v_1 \\ u_1 \end{pmatrix} + \begin{pmatrix} -u_1 \\ v_1 \end{pmatrix} + \begin{pmatrix} u_1 - u_0 \\ -v_1 + v_0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\varphi_1} g_1^0 \\ \frac{1}{\varphi_1} f_1^0 \end{pmatrix} \\ \begin{pmatrix} v_N \\ u_N \end{pmatrix} + \begin{pmatrix} -u_N \\ v_N \end{pmatrix} + \begin{pmatrix} -u_{N+1} + u_N \\ v_{N+1} - v_N \end{pmatrix} = \begin{pmatrix} g_N^0 \\ f_N^0 \end{pmatrix}. \end{cases}$$

The above system can be written as

$$\begin{cases} \begin{pmatrix} v_1 \\ u_1 \end{pmatrix} + \Phi \begin{pmatrix} v_1 \\ u_1 \end{pmatrix} + \Lambda_1 \begin{pmatrix} v_1 \\ u_1 \end{pmatrix} \ni \begin{pmatrix} \frac{1}{\varphi_1} g_1^0 \\ \frac{1}{\varphi_1} f_1^0 \end{pmatrix} \\ \begin{pmatrix} v_N \\ u_N \end{pmatrix} + \Phi \begin{pmatrix} v_N \\ u_N \end{pmatrix} + \Lambda_2 \begin{pmatrix} v_N \\ u_N \end{pmatrix} \ni \begin{pmatrix} g_N^0 \\ f_N^0 \end{pmatrix}, \end{cases} \tag{3}$$

where the operator $\Phi : H^2 \to H^2$ is defined by $\Phi((x,y)^T) = (-y,x)^T$. Because Φ is single-valued, everywhere defined, linear and monotone, and Λ_1, Λ_2 are maximal monotone, we deduce by [4], Corollary 1.3, Chapter II, that $\Phi + \Lambda_1 : D(\Lambda_1) \subset H^2 \to H^2$ and $\Phi + \Lambda_2 : D(\Lambda_2) \subset H^2 \to H^2$ are also maximal monotone operators. Then $R(I + \Phi + \Lambda_1) = H^2$ and $R(I + \Phi + \Lambda_2) = H^2$, and so the equations $(3)_1$ and $(3)_2$ have solutions $(v_1, u_1)^T \in D(\Lambda_1)$, and $(v_N, u_N)^T \in D(\Lambda_2)$. We conclude that the element $U = (u_1, \ldots, u_N, v_1, \ldots, v_N)^T \in D(\mathcal{A}_2)$, where u_1, v_1, u_N, v_N are defined above and u_j, v_j, $j = \overline{2, N-1}$ are defined by system (2), is solution of the equation (1). Therefore the operator $\mathcal{A}_2$ is maximal monotone.

Because $\mathcal{A}_1$ is single-valued, everywhere defined, continuous and monotone, and $\mathcal{A}_2$ is maximal monotone, we deduce that the operator $\mathcal{A}$ is maximal monotone in X. □

Lemma 2. *If the assumptions* (H1)a, (H4) *and* (H5) *hold, then the operator* $\mathcal{B}$ *is maximal monotone in* X.

The proof of Lemma 2 is similar to that of Theorem 2 from [8].

Theorem 1. *Assume that the assumptions* (H1)a, (H2)-(H5) *are verified. Then the operator* $\mathcal{A} + \mathcal{B}$ *is maximal monotone in* X.

Proof. If the assumption (H3)a holds, then we can easily deduce that $\operatorname{int} D(\mathcal{A}) = \{(u_1, \ldots, u_N, v_1, \ldots, v_N)^T \in X,\ (v_1, u_1)^T \in \operatorname{int} D(\Lambda_1),\ (v_N, u_N)^T \in \operatorname{int} D(\Lambda_2)\}$ and so $\operatorname{int} D(\mathcal{A}) \cap D(\mathcal{B}) \neq \emptyset$. Therefore using Rockafellar's theorem (see [5], Corollaire 2.7) we deduce that $\mathcal{A} + \mathcal{B}$ is maximal monotone.

If the assumption (H3)b holds, then we have $D(\mathcal{A}) \cap (\operatorname{int} D(\mathcal{B})) \neq \emptyset$. Using the same theorem we obtain that $\mathcal{A} + \mathcal{B}$ is maximal monotone.

If the assumption (H3)c holds, it follows that $0 \in \operatorname{int} (D(\mathcal{A}) - D(\mathcal{B}))$. Using Attouch's theorem (see [3]), we deduce that the operator $\mathcal{A} + \mathcal{B}$ is maximal monotone. □

Theorem 2. *Assume that the assumptions* (H1)ab, (H2)-(H5) *are verified and* $F_0 = (f_1^0, \ldots, f_N^0, g_1^0, \ldots, g_N^0)^T \in X$. *Then the problem* (S), (BC) *has a unique solution* $U = (u_1, \ldots, u_N, v_1, \ldots, v_N)^T \in X$, *with* $(v_1, u_1)^T \in D(\Lambda_1)$, $(v_N, u_N)^T \in D(\Lambda_2)$, $u_j \in D(A)$, $v_j \in D(B)$, *for all* $j = \overline{1, N}$.

Proof. By Theorem 1 the operator $\mathcal{A} + \mathcal{B}$ is maximal monotone in X. Besides, it is strongly monotone. Indeed, for all $U = (u_1, \ldots, u_N, v_1, \ldots, v_N)^T$, $\widetilde{U} = (\widetilde{u}_1, \ldots, \widetilde{u}_N, \widetilde{v}_1, \ldots, \widetilde{v}_N)^T \in D(\mathcal{A}) \cap D(\mathcal{B})$, $Z \in (\mathcal{A}+\mathcal{B})(U)$, $\widetilde{Z} \in (\mathcal{A}+\mathcal{B})(\widetilde{U})$ we have

$$\langle Z-\widetilde{Z},U-\widetilde{U}\rangle_X \geq \sum_{j=1}^{N}\alpha_j c_j a_0\|u_j-\widetilde{u}_j\|^2+\sum_{j=1}^{N}\alpha_j d_j b_0\|v_j-\widetilde{v}_j\|^2$$

$$\geq M\left[\sum_{j=1}^{N}(\alpha_j\|u_j-\widetilde{u}_j\|^2+\alpha_j\|v_j-\widetilde{v}_j\|^2)\right]=M\|U-\widetilde{U}\|_X^2,$$

where $M=\min\{c_j a_0,\ d_j b_0,\ j=\overline{1,N}\}>0$. Therefore the operator $\mathcal{A}+\mathcal{B}$ is coercive, and so $R(\mathcal{A}+\mathcal{B})=X$. We deduce that for $F_0=(f_1^0,\dots,f_N^0,g_1^0,\dots,g_N^0)^T\in X$, the equation $(\mathcal{A}+\mathcal{B})(U)\ni F_0$, which is equivalent to problem (S), (BC) has a unique solution $U=(u_1,\dots,u_N,v_1,\dots,v_N)^T\in D(\mathcal{A})\cap D(\mathcal{B})$. □

3 An application to nonlinear differential systems

Let us consider in the space H the nonlinear differential system

$(\widetilde{S})$

$$\begin{cases} u_j'(t)+v_{j+1}(t)-(1+\varphi_j)v_j(t)+\varphi_j v_{j-1}(t)+c_jA(u_j(t))\ni f_j(t)\\ v_j'(t)-u_{j+1}(t)+(1+\varphi_j)u_j(t)-\varphi_j u_{j-1}(t)+d_jB(v_j(t))\ni g_j(t),\ \ j=\overline{1,N},\ \ t>0,\end{cases}$$

with the boundary conditions

$(\widetilde{BC})$

$$\begin{pmatrix} u_1(t)-u_0(t)\\ -v_1(t)+v_0(t)\end{pmatrix}\in\Lambda_1\begin{pmatrix} v_1(t)\\ u_1(t)\end{pmatrix},\quad \begin{pmatrix} -u_{N+1}(t)+u_N(t)\\ v_{N+1}(t)-v_N(t)\end{pmatrix}\in\Lambda_2\begin{pmatrix} v_N(t)\\ u_N(t)\end{pmatrix},\ \ t>0$$

and the initial data

$$(\widetilde{ID})\qquad\qquad u_j(0)=u_{j0},\ \ v_j(0)=v_{j0},\ \ j=\overline{1,N}.$$

Using the operators $\mathcal{A}$ and $\mathcal{B}$ defined in Section 2, the above problem can be expressed as the following Cauchy problem in the space X

$$(P)\qquad\qquad \begin{cases}\dfrac{dU}{dt}(t)+\mathcal{A}(U(t))+\mathcal{B}(U(t))\ni F(t)\\ U(0)=U_0,\end{cases}$$

where $U=(u_1,\dots,u_N,v_1,\dots,v_N)^T$, $U_0=(u_{10},\dots,u_{N0},v_{10},\dots,v_{N0})^T$, $F=(f_1,\dots,f_N,\ g_1,\dots,g_N)^T$.

By using Theorem 1 from Section 2, Theorem 2.2 and Corollary 2.1, Chapter III from [4], and similar arguments as in the proof of Theorem 4 from [12], we obtain for the strong solutions of problem $(P)\equiv(\widetilde{S})$, $(\widetilde{BC})$, $(\widetilde{ID})$ the following existence and uniqueness result.

Theorem 3. *Assume that the assumptions* (H1)a, (H2)-(H5) *hold. If* $(v_{10},u_{10})^T\in D(\Lambda_1)\cap(D(B)\times D(A))$, $u_{j0}\in D(A)$, $v_{j0}\in D(B)$ *for all* $j=\overline{2,N-1}$, $(v_N,u_N)^T\in D(\Lambda_2)\cap(D(B)\times D(A))$, $f_j,\ g_j\in W^{1,1}(0,T;H)$, *for all* $j=\overline{1,N}$, ($T>0$ *fixed*), *then there exist unique functions* $u_j,\ v_j\in W^{1,\infty}(0,T;H)$, $j=\overline{1,N}$, *with* $(v_1(t),u_1(t))^T\in D(\Lambda_1)\cap(D(B)\times D(A))$, $u_j(t)\in D(A)$, $v_j(t)\in$

$D(B)$, *for all* $j = \overline{2, N-1}$, $(v_N(t), u_N(t))^T \in D(\Lambda_2) \cap (D(B) \times D(A))$, *for all* $t \in [0,T]$ *that verify the system* $(\widetilde{S})$ *and the boundary conditions* $(\widetilde{BC})$ *for all* $t \in [0,T]$ *and the initial data* $(\widetilde{ID})$. *Besides* u_j, v_j, $j = \overline{1, N}$ *are everywhere differentiable from right in the topology of* H *and* $\frac{d^+U}{dt}(t) = (F(t) - \mathcal{A}(U(t)) - \mathcal{B}(U(t)))^0$, *for* $t \in [0,T)$.

Remark. Under the assumptions (H1)a, (H2)-(H5), if $U_0 \in \overline{D(\mathcal{A}) \cap D(\mathcal{B})}$ and $F \in L^1(0,T;X)$ ($T > 0$ fixed), then by Corollary 2.2, Chapter III from [4], we deduce that the problem (P) has a unique weak solution $U \in C([0,T];X)$.

Now using Theorem 2 from Section 2 and Theoreme 3.9 from [5], we obtain in the following theorem the asymptotic behaviour of the solutions of (P).

Theorem 4. *Assume that the assumptions* (H1)ab, (H2)-(H5) *hold,* $F \in L^1_{loc}(\mathbb{R}_+;X)$ *verifies the condition* $\lim\limits_{t\to\infty} F(t) = \widetilde{F}$ *strongly in* X, *and* $\widetilde{U}$ *is the unique solution of the problem* (S), (BC) *with* $F_0 = 0$. *Then for any weak solution* $U(t)$, $t \geq 0$ *of the problem* $(P)_1$, *we have* $\lim\limits_{t\to\infty} U(t) = \widetilde{U}$, *strongly in* X.

References

[1] A. R. Aftabizadeh, N. H. Pavel, Nonlinear boundary value problems for some ordinary and partial differential equations associated with monotone operators, *J. Math. Anal. Appl.*, 156 (1991), 535-557.

[2] R. P. Agarwal and P. J. Y. Wong, *Advanced Topics in Difference Equations*, Mathematics and Its applications, vol. 404, Kluwer Academic, Dordrecht, 1997.

[3] H. Attouch, On the maximality of the sum of two maximal monotone operators, *Nonlinear Anal., Theory Methods Appl.*, 5 (1981), 143-147.

[4] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.

[5] H. Brezis, *Operateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[6] W. G. Kelley and A. C. Peterson, *Difference Equations: An Introduction with Applications*, Academic Press, Massachusetts, 1991.

[7] S. Ladde, V. Lakshmikantham and A.S. Vatsalo, *Monotone, Iterative Techniques for Nonlinear Differential Equations*, Monographs, Advanced Texts and Surveys in Pure and Applied Mathematics, 27 Pitman, 1985.

[8] R. Luca, Existence and uniqueness for a nonlinear discrete hyperbolic system, *Nonlinear Anal., Theory Methods Appl.*, 67 (2007), 2433-2446.

[9] R. Luca, Existence and asymptotic behaviour for a discrete hyperbolic system, *J. Math. Anal. Appl.*, 329 (2007), 191-205.

[10] R. Luca, An existence result for a class of nonlinear differential systems, in *Applied Analysis and Differential Equations*, Editors O. Carja and I.I. Vrabie, World Scientific, 2007, 185-198.

[11] R. Luca, Existence and uniqueness results for a class of nonlinear differential systems, *Int. J. Pure Applied Math.*, 40 (2007), No. 3, 367-372.

[12] R. Luca, On a nonlinear differential problem with second-order differences, *Anal. Stiint. Univ. Al.I.Cuza Iasi*, LIV (2008), f.2, 261-277.

[13] R. Luca, Nonlinear differential problems with first- or second-order differences in Hilbert spaces, *Nonlinear Anal., Theory Methods Appl.*, DOI: 10.1016/j.na.2008.10.006, to appear.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

On the existence and the Uuniqueness of Solutions of the Fredholm integral equations of the Second kind on an Interval

Mostefa Nadir
Laboratory of Pure and Applied Mathematics
Department of Mathematics
University of M'sila Algeria
E-mail: mostefanadir@yahoo.fr

Abstract

The idea of this work is the continuation of the study of the paper [5] by the author concerning conditions on the existence and the uniqueness of the solution of the Fredholm integral equations of the second kind on a finite interval. Noting that, except the Banach's theorem where the norm of the integral operator must be less than unity, the existence and uniqueness of the solution of the Fredholm integral equations of the second kind remain an open question.

Keywords: Singular integral operator, Fredholm equation, Algebras theory, Commutator.

1 Introduction

As it is known many problems of mathematical physics can be stated in the form of integral equations. In particular, the domains of ordinary and partial differential equations can be recast as integral equations. Also many existence and uniqueness results can then be derived from results from integral equations.

In this work we try to find conditions concerning the existence and the uniqueness of the solution of the Fredholm integral equation of the second kind on an interval $[a, b]$

$$\varphi(t_0) - \int_a^b k(t_0, t)\varphi(t)dt = f(t_0), \quad a, b \in \mathbb{R}. \tag{1}$$

We note that, the study of these equations is based on the Fredholm alternative where is a fundamental tool for the solvability of certain types of integral equations. For the applications of these equations to different problems, the following relations are known to play a basic part

ISSN-1584-5990

• The homogeneous equation

$$\varphi(t_0) - \int_a^b k(t_0, t)\varphi(t)dt = 0, \tag{2}$$

has only a finite number of linearly independent solutions.

• The adjoint homogeneous equations (2) and

$$\psi(t_0) - \int_a^b k^*(t, t_0)\psi(t)dt = 0, \tag{3}$$

have the same number of linearly independent solutions.

• The non-homogeneous equations (1) is solvable for any second member f if and only if the adjoint homogeneous equation (3) or the equation homogeneous (2) has non solution different from zero.

• The solvability of the non-homogeneous equation (1) is given by the necessary and sufficient conditions

$$\int_a^b f(t)\psi_k(t)dt = 0, \tag{4}$$

where the functions $\psi_k(t)$ form a complete system of linearly independent solutions of the adjoint homogeneous solutions (3) [2].

Lemma 1 *Let $\varphi(t)$ be a function satisfies the Holder condition on the interval $[a, b]$,*

$$\mid \varphi(t) - \varphi(t_0) \mid \leq M \mid t - t_0 \mid^{\alpha},\ 0 < \alpha \leq 1, \tag{5}$$

and equal to zero at the end points,

$$\varphi(a) = \varphi(b) = 0. \tag{5'}$$

Then the principal value of Cauchy integral

$$S\varphi(t_0) = \int_a^b \frac{\varphi(t)}{t - t_0}dt, \tag{6}$$

has the following properties

• *The existence of the limit defined by the integral with weak singularity*

$$\lim_{t_0 \to c} S\varphi(t_0) = S\varphi(c) = \int_a^b \frac{\varphi(t)}{t - c}dt, \tag{6'}$$

when the interior point t_0 tends to its end points $c = a$ or $c = b$.

• *The integral operator S is bounded in all Holder spaces $C^{\alpha}([a, b])$.*

• *The integral operator S belongs to the Holder spaces $C^{\alpha}([a, b])$ for $0 < \alpha < 1$.*

• *The integral operator S belongs to the Holder spaces $C^{1-\varepsilon}([a, b])$ for $\alpha = 1$* with $\varepsilon \geq 0$.

Lemma 2 *Let $a(t)$ be a function in $C^{\alpha}([a,b])$* and the density function $\varphi(t)$ satisfies conditions (5) and (5′). *Then the commutator*

$$(Sa - aS)\varphi(t_0) = \int_a^b \frac{a(t) - a(t_0)}{t - t_0}\varphi(t)dt \tag{7}$$

is compact from $C^{\alpha}([a,b])$ into $C^{\alpha}([a,b])$ [2].

In fact, since the function $a(t) \in C^{\alpha}([a,b])$, then the kernel $\dfrac{a(t) - a(t_0)}{t - t_0}$ has a weak singularity and defines a compact integral operator in all spaces $C^{\alpha}([a,b])$.

Corollary The property of the compactness is enjoyed by the more general operator

$$T\varphi(t_0) = \int_a^b \frac{k(t_0,t) - k(t_0,t_0)}{t - t_0}\varphi(t)dt, \tag{8}$$

if the function $k(t_0,t)$ is of Holder class in both variables $k(t,t_0) \in C^{\alpha}([a,b]\times[a,b])$ $(0 < \alpha < 1)$ and the function $\varphi(t)$ is null at edges [3].

2 Main Results

Theorem *let $k(t_0,t)$ be a function in $C^{\alpha}([a,b]\times[a,b])$ $(0 < \alpha < 1)$ satisfies the Holder condition for both variables with the condition $k(t_0,t_0) \neq 0$* for all t_0 in the interval $[a,b]$ *and the function density $\varphi(t)$ is null at edges, then the equation* (1)

$$\varphi(t_0) - \int_a^b k(t_0,t)\varphi(t)dt = f(t_0), \quad a,b \in \mathbb{R},$$

admits a unique solution in the space $C^{\alpha}([a,b])$ for all second member $f(x)$ in $C^{a}([a,b])$.

Proof

It is clear to see that, the compact operator integral

$$A\varphi(t_0) = \int_a^b k(t_0,t)\varphi(t)dt,$$

has the following representation

$$(Ta - aT)\varphi(t_0) + k(t_0,t_0)(Sa - aS)\varphi(t_0),$$

with the function $a(t) = t$ for all $t \in [a,b]$.

In other words, the operator A has a commutator representation in the Banach algebra $C^{\alpha}([a,b])$.

Also, it is known that, the unit element $I\varphi(t) = \varphi(t)$ of a Banach algebra $C^{\alpha}([a,b])$ is not a commutator

$$MN - NM,$$

for all elements M and N in $C^{\alpha}([a,b])$. Indeed, if

$$I = MN - NM,$$

then the spectrum relation gives

$$sp(MN) = 1 + sp(NM),$$

which is not consistent with the following result

$$sp(MN) \cup \{0\} = sp(NM) \cup \{0\}.$$

Therefore,

$$A\varphi(t_0) = \int_a^b k(t_0,t)\varphi(t)dt \neq \varphi(t_0), \quad \text{for all} \quad \varphi \in C^{\alpha}([a,b]).$$

We see that the homogeneous equation has only the trivial solution in $C^{\alpha}([a,b])$, so the equation (1)

$$\varphi(t_0) - \int_a^b k(t_0,t)\varphi(t)dt = f(t_0),$$

has a unique solution in the space $C^{\alpha}([a,b])$.

References

[1] R. Kadison and J. Ringrose, Fundamentals of the theory of operator algebras, Academic Press 1883.

[2] N. I. Mushelishvili, Singular integral equations, Naukah Moscow, 1968, English transl, of 1sted Noordho, 1953; reprint,1972.

[3] W. Pogorzelski, Integral equations and their applications, Volume 1, Polish Scientific Publishers Warszawa, 1966.

[4] M. Nadir, B. Lakehali, An approximation for singular integrals of Cauchy types, in Advance in algebra and analysis (AAA) (1),1, 2006.

[5] M. Nadir, On the existence and the uniqueness of solutions of the Fredholm integral equations of the second kind on the contour, to appear in Dynamic System and Applications (2007)

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

About Some Fixed Point Result in Space with Perturbated Metric

Ion Marian Olaru[1], Vasilica Olaru, Eugen Constantinescu
[1] Departament of Mathematics,
University "Lucian Blaga" of Sibiu
E-mail: olaruim@yahoo.com

Abstract

In this paper we present a test of convergence for the series with positives terms, with applications in the theory of fixed point in space with perturbated metric.

Keywords: fixed points, test of convergence.

1 Introduction

Let (X, d) be a complete metric space, $f : X \to X$ an operator and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ monotone increasing function such that

$$d(f(x), f(y)) \leq \varphi(d(x, y)),$$

for all $x, y \in X$. We construct the sequence of successive approximations, $(x_n)_{n\in\mathbb{N}}$, $x_{n+1} = f(x_n)$, $x_0 \in X$. We obtain, using the monotonicity of φ, that

$$d(x_n, x_{n+p}) \leq \sum_{k=n}^{n+p-1} \varphi(d(x_0, x_1)), \tag{1}$$

for all $p \geq 1$, $n \geq 1$.

If the series of positive terms

$$\sum_{k=1}^{\infty} \varphi(r)$$

converges, for all $r \in \mathbb{R}_+$ then the sequences $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, hence $(x_n)_{n\in\mathbb{N}}$ is convergent for all $x_0 \in X$.

Definition 1.1 *A function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a comparison functions if :*

(i) φ is monotone increasing;

ISSN-1584-5990

(ii) The sequence $(\varphi^n(t))_{n\in\mathbb{N}}$ *converges to* 0 *for all* $t \geq 0$.

Definition 1.2 *A mapping* $f : X \to X$ *is a* $\varphi-$*contraction if* φ *a comparison function and*

$$d(f(x), f(y)) \leq \varphi(d(x, y)),\ (\forall)x, y \in X.$$

We have the following result

Theorem 1.1 *[6],[7] Let* (X, d) *be a complete metric space and* $f : X \to X$ *a* $\varphi-$ *contraction. Then* f *is a Picard operator.*

In [1], [2] has been given the following generalization of ratio test

Theorem 1.2 *Let* $\sum\limits_{n=1}^{\infty} u_n$ *be an infinite series of positive terms. If there exists convergent series of nonnegative terms,* $\sum\limits_{n=1}^{\infty} u_n$ *and two numbers* k, n_0 *such that*

$$\frac{u_{n+1}}{u_n + v_n} \leq k < 1,$$

for all $n \geq n_0$, *then the series* $\sum\limits_{n=1}^{\infty} u_n$ *is convergent.*

We have the converse of theorem

Theorem 1.3 *[1]A series* $\sum\limits_{n=1}^{\infty} u_n$ *of decreasing positive terms converges if and only if there exists a convergent series of nonnegative terms* $\sum\limits_{n=1}^{\infty} v_n$ *and two numbers* k, n_0 *such that*

$$\frac{u_{n+1}}{u_n + v_n} \leq k < 1,$$

for all $n \geq n_0$.

Definition 1.3 *A function* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *is a c-comparison functions if there exists a convergent series of nonnegative terms* $\sum\limits_{n=1}^{\infty} v_n$ *and two numbers* $k \in (0, 1), n_0$ *such that*

$$\varphi^{n+1}(r) \leq k[\varphi^n(r) + v_n],$$

for all $n \geq n_0$, $r \geq 0$.

Observation 1.1 *(see [1] Theorem 3)If* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *is a c-comparison functions then the series* $\sum\limits_{k=1}^{\infty} \varphi^k(r)$ *is convergent, for all* $r \geq 0$ *and the approximations successive sequence,* $x_{n+1} = f(x_n)$, *converges to the unique fixed point of* f.

The following result is other generalization of ratio test

Theorem 1.4 *[3]Let* $\sum_{n=1}^{\infty} U_n$ *be a series of positives terms. If there exists a convergent series* $\sum_{n=1}^{\infty} V_n$ *and a sequence* $(W_n)_{n\geq 1}$ *satisfying*

(i) $0 < W_n < 1$, *for all* $n \geq 1$;

(ii) $W_{n+m} \leq W_n \cdot W_m$, *for all* $n, m \geq 1$;

(iii) $W_n \cdot U_{n+1} \leq W_{n+1}(U_n + V_n)$, *for all* $n \geq 1$,

then the series $\sum_{n=1}^{\infty} U_n$ *is convergent.*

More result about the generalization of ratio test we find in [4], [5].

2 Main results

In this section we present an result of fixed point in space with perturbated metric. Similarly result we find in [8]. First, we present the following test of convergence for series with positives terms

Proposition 2.1 *We consider the mappings* $f, g : (0, \infty) \to (0, \infty)$ *and* $(u_n)_{n\in\mathbb{N}}$ *a sequence with positives terms, such that:*

(i) there exists $M > 0$ *with the property* $\frac{f(u)}{g(u)} \geq M$, *for all* $u \in (0, \infty)$;

(ii) there exists $K > 0$, *and* $N \in \mathbb{N}$ *with the property* $\frac{f(u_{i+1})}{g(u_i)} \leq K$, *for all* $i \geq N$;

(iii) $\frac{K}{M} < 1$.

Then $\sum_{n\geq 0} f(u_n) < \infty$, $\sum_{n\geq 0} g(u_n) < \infty$

Proof: From the relations (i), (ii) if we take successively $i := N, \cdots, n$ we obtain

$$M^{n-N-1}\frac{f(u_n)}{g(u_N)} \leq \frac{f(u_{N+1})}{g(u_N)}\frac{f(u_{N+2})}{g(u_{N+1})}\cdots\frac{f(u_n)}{g(u_{n-1})} \leq K^{n-N}$$

$$f(u_n) \leq g(u_N)\frac{M^{N+1}}{K^N}\left(\frac{K}{M}\right)^n$$

From the above relations we obtain that $\sum_{n\geq 0} f(u_n) < \infty$. On the other hand, using the ratio test and the relation

$$M\frac{g(u_{n+1})}{g(u_n)} \leq \frac{f(u_{n+1})}{g(u_n)} \leq K,$$

for all $n \geq N$, we obtain that $\sum_{n\geq 0} g(u_n) < \infty$.

Remark 2.1 *For $f, g : (0, \infty) \to (0, \infty)$, $f(u) = g(u) = u$ from the Proposition 2.1 it follow the ratio test.*

Next we consider a complete metric space (X, d), $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ monotone increasing and $g \in P$. P is the class of functions $g : \mathbb{R}_+ \to \mathbb{R}_+$ continuous, strict increasing and surjective.

Remark 2.2 $g(0) = 0$.

Indeed if we suppose that there exists $\alpha > 0$ such that $g(\alpha) = 0$ then we have $g(0) < g(\alpha) = 0$ which is a contradiction with the fact that $g(\mathbb{R}_+) \subseteq \mathbb{R}_+$.

We have the following fixed point result

Proposition 2.2 *Let $f : X \to X$ be an operator such that*

$$g(d(f(x), f(y))) \leq \varphi(g(d(x, y))), \tag{2}$$

for all $x, y \in X$. We suppose that there exists $G : (0, \infty) \to (0, \infty)$ such that the pair (g^{-1}, G) verifying the Proposition 2.1 with $u_n = \varphi^n(r)$, $r \geq 0$. Then f is a Picard operator.

Proof: From the properties of the function g, the relation (2) can be write

$$d(f(x), f(y)) \leq g^{-1} \circ \varphi \circ g(d(x, y)),$$

for all $x, y \in X$.

We denote by

$$\varphi_1 : \mathbb{R}_+ \to \mathbb{R}_+,$$
$$\varphi_1(r) := g^{-1} \circ \varphi \circ g(r).$$

We remark that φ_1 is monotone increasing and $\varphi_1^n(r) = g^{-1} \circ \varphi^n \circ g(r)$. From the Proposition 2.1 we have that the series $\sum\limits_{k \in \mathbb{N}} g^{-1}(\varphi^k(r))$ is converges for all $r > 0$. Then the sequence $\varphi_1^n(r)$ is converges to zero, for all $r > 0$. It follow that the function is a comparison function. Using the Theorem 1.1 we obtain the conclusion.

References

[1] Vasile Berinde, *Error estimates in the approximation of fixed points for a class of $\varphi-$contractions*, Studia Univ. Babes-Bolyai, Mathematica, XXXV, 2, 1990, 86-89

[2] Vasile Berinde, *Une generalization du critere de D'Alembert pour les series positives,* Bul. Stiint. Univ. Baia Mare, Fasc.Mat.-Inf., vol. VII (1991), nr.1-2, 21-26

[3] Vasile Berinde, J Sandor *On the generalized ratio test,* Bul. Stiint. Univ. Baia Mare, Fasc. Matem.-Inf., vol. VIII (1992), 35-40.

[4] Vasile Berinde, *The generalized ratio test revisited,* Bul. Stiint. Univ. Baia Mare, Fasc.Mat.-Inf., vol. XVI (2000), no. 2, 303-306.

[5] Vasile Berinde, *O generalizare a criteriului lui D'Alembert si aplicatii n teoria punctului fix*, Analele Univ. Oradea, Fasc.Matematica, Tome I (1991), 51-58.

[6] I.A.Rus, *Generalized contractions*, Seminar on Fixed Point Theory, Preprint No. 3, 1983, pp 1-131.

[7] I.A.Rus, *Principii 'si aplica'tii ale teoriei punctului fix*, Editura Dacia, Cluj-Napoca, 1979.

[8] Marcel-Adrian Şerban, *Spaces with Perturbated Metrics and Fixed Point Theorems*, the Twelfth International Conference on Applied Mathematics Computer Science and Mechanics, Băişoara, Semptember 10-13, 2008.

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Operators with Single-valued Extension Property on Locally Convex Spaces

Sorin Mirel Stoian
Department of Mathematics and Computer Science,
University of Petroşani, Romania
E-mail: mstoian@upet.ro

Abstract

In this article we extend the definitions of the operators with single-valued extension property and quasi-nilpotent equivalent operators from the Banach spaces to the class of bounded operators on sequentially complete locally convex spaces.

Keywords: single-valued extension property, locally convex space

1 Introduction

The aim of this paper is to define operators with single-valued extension property and to study the class of quasi-nilpotent equivalent operators with such property on sequentially complete locally convex space. The class of quasi-nilpotent equivalent operators on a Banach space was introduced by Colojoară and Foias [4].

Throughout this paper X denotes a sequentially complete locally convex Hausdorff space over the complex space $\mathbb{C}$ and $\mathcal{C}(X)$ denotes the set of all families of seminorms which generate the topology of the space X. If $\mathcal{L}(X)$ is the algebra of linear continuous operators on X, then for every $p, q \in \mathcal{P}$ we define the mixed operator seminorm $m_{pq} : \mathcal{L}(X) \to \mathbb{R} \cup \{\infty\}$, where

$$m_{pq}(T) = \sup_{p(x)\neq 0} \frac{q(Tx)}{p(x)}, \text{ for all } T \in \mathcal{L}(X).$$

If $p = q$ then we denote $\hat{p} = m_{pp}$. From the definition it follows that:
1. $m_{pq}(T) = \sup_{p(x)=1} q(Tx) = \sup_{p(x)\leq 1} q(Tx), (\forall)\, p \in P, (\forall)\, q \in Q$;
2. $q(Tx) \leq m_{pq}(T)\, p(x)$, $(\forall)\, x \in X$, whenever $m_{pq}(T) < \infty$.
3. $m_{pq}(T) = \inf\{M > 0 \,|q(Tx) \leq Mp(x), (\forall)\, x \in X\}$, whenever $m_{pq}(T) < \infty$.

An operator T on a locally convex space X is quotient bounded with respect to a family of seminorms $\mathcal{P} \in \mathcal{C}(X)$ if for every seminorm $p \in \mathcal{P}$ there exists some $c_p > 0$ such that

$$p(Tx) \leq c_p p(x), \text{ for all } x \in X.$$

ISSN-1584-5990

The class of quotient bounded operators with respect to $\mathcal{P} \in \mathcal{C}(X)$ is denoted by $Q_{\mathcal{P}}(X)$ and let $\hat{\mathcal{P}}$ be the family $\{\hat{p} \mid p \in \mathcal{P}\}$. $(Q_{\mathcal{P}}(X), \hat{\mathcal{P}})$ is a sequentially complete multiplicatively convex algebra for all $\mathcal{P} \in \mathcal{C}(X)$ (see [8]). An operator $T \in Q_{\mathcal{P}}(X)$ is a bounded element of the algebra $Q_{\mathcal{P}}(X)$ if it is a bounded element in the sense of G.R.Allan [1], i.e. for some $r > 0$, $r^{-1}T$ generates a bounded semigroup, and the class of the bounded elements of $Q_{\mathcal{P}}(X)$ is denoted by $(Q_{\mathcal{P}}(X))_0$. If $r_{\mathcal{P}}(T)$ is the $\mathcal{P}$-spectral radius of the operator T, i.e. is the radius of boundness of the operator T in $Q_{\mathcal{P}}(X)$ given by the relation

$$r_{\mathcal{P}}(T) = \inf\{\alpha > 0 \ \mid \ \alpha^{-1}T \text{ generates a bounded semigroup in } Q_{\mathcal{P}}(X)\},$$

then in [1] and [8] was proved that the following relations hold

$$r_{\mathcal{P}}(T) = \sup\{ \limsup_{n\to\infty} (\hat{p}(T^n))^{1/n} \mid \ p \in \mathcal{P}\}. \tag{1}$$

$$r_{\mathcal{P}}(T) < +\infty \text{ if and only if } T \in (Q_{\mathcal{P}}(X))_0; \tag{2}$$

$$r_{\mathcal{P}}(T) = \inf\left\{\lambda > 0 \mid \lim_{n\to\infty} \frac{T^n}{\lambda^n} = 0 \right\}. \tag{3}$$

and if $0 < |\lambda| < r_{\mathcal{P}}(T)$, then the set $\left(\frac{T^n}{\lambda^n}\right)_n$ is unbounded.

If $(X, \mathcal{P})$ is a locally convex space and $T \in (Q_{\mathcal{P}}, (X))_0$ then, for every $|\lambda| > r_{\mathcal{P}}(T)$ the Neumann series $\sum\limits_{n=0}^{\infty} \frac{T^n}{\lambda^{n+1}}$ converges to $R(\lambda, T)$(in $Q_{\mathcal{P}}(X)$) and $R(\lambda, T) \in Q_{\mathcal{P}}(X)$ [8]. Moreover, $|\sigma(Q_{\mathcal{P}}, T)| = r_{\mathcal{P}}(T)$.

The Waelbroeck resolvent set $\rho_W(Q_{\mathcal{P}}, T)$ of an operator $T \in (Q_{\mathcal{P}}(X))_0$ is the subset of elements of $\lambda_0 \in \mathbb{C}_\infty = \mathbb{C} \cup \{\infty\}$, for which there exists a neighborhood $V \in \mathcal{V}_{(\lambda_0)}$ such that:

1. the operator $\lambda I - T$ is invertible in $Q_{\mathcal{P}}(X)$ for all $\lambda \in V \backslash \{\infty\}$
2. the set $\{ \ (\lambda I - T)^{-1} \mid \lambda \in V \backslash \{\infty\} \ \}$ is bounded in $Q_{\mathcal{P}}(X)$.

The Waelbroeck spectrum of T, denoted by $\sigma_W(Q_{\mathcal{P}}, T)$, is the complement of the set $\rho_W(Q_{\mathcal{P}}, T)$ in $\mathbb{C}_\infty$. The set $\rho_W(Q_{\mathcal{P}}, T)$ is open and it is obvious that $\rho_W(Q_{\mathcal{P}}, T) \subset \rho(Q_{\mathcal{P}}(X))$. An operator $T \in Q_{\mathcal{P}}(X)$ is regular if $\infty \notin \sigma_W(Q_{\mathcal{P}}, T)$, i.e. there exists some $t > 0$ such that:

1. the operator $\lambda I - T$ is invertible in $Q_{\mathcal{P}}(X)$, for all $\mid \lambda \mid > t$
2. the set $\{R(\lambda, T) \mid\mid \lambda \mid > t\}$ is bounded in $Q_{\mathcal{P}}(X)$.

2 Bounded Operators with SVEP

Lemma If $(X, \mathcal{P})$ is a sequentially complete locally convex space and $T \in (Q_{\mathcal{P}}(X))_0$, then

$$\overset{\circ}{\rho}(Q_{\mathcal{P}}, T) = \rho_W(Q_{\mathcal{P}}, T).$$

Proof Assume that there exists $\lambda_0 \in \rho(Q_{\mathcal{P}}, T) \backslash \rho_W(Q_{\mathcal{P}}, T)$ such that $\lambda_0 \in \overset{\circ}{\rho}(Q_{\mathcal{P}}, T)$. Since $\lambda_0 \notin \rho_W(Q_{\mathcal{P}}, T)$, then for each neighborhood U of λ_0 the set

$$\{ \ (\ \lambda I - T \ )^{-1} \mid \lambda \in U \ \}$$

is not bounded in $Q_{\mathcal{P}}(X)$. Let $U \in \rho(Q_{\mathcal{P}}, T)$ an open set such that $\lambda_0 \in U$. This implies that there exists $\lambda_1 \in U$ and $p \in \mathcal{P}$ such that for every $n \in N$ there exists $x_n \in X$ $(p(x_n) \neq 0)$ with the property

$$p\left(R(\lambda_1, T)x_n\right) > np\left(x_n\right),$$

Therefore, for $y_n = R(\lambda_1, T)x_n$ we have

$$p\left(y_n\right) > np\left((\lambda_1 I - T)y_n\right),$$

which implies that $\lambda_1 \in \sigma_a(Q_{\mathcal{P}}, T) \subset \sigma(Q_{\mathcal{P}}, T)$ (where $\sigma_a(Q_{\mathcal{P}}, T)$ is the approximate spectrum of T [6]). This contradicts the supposition we made, so lemma is proved.

Definition If $(X, \mathcal{P})$ is a sequentially complete locally convex space we say that the operator $T \in (Q_{\mathcal{P}}(X))_0$ has the single-valued extension property (we will write SVEP) if for any analytic function $f : D_f \to X$, where $D_f \subset \mathbb{C}$ is an open set, with the property

$$(\lambda I - T)f(\lambda) \equiv 0_X, \text{ for all } \lambda \in D_f,$$

it results that $f \equiv 0$, for all $\lambda \in D_f$.

Definition Let $(X, \mathcal{P})$ be a sequentially complete locally convex space and $T \in (Q_{\mathcal{P}}(X))_0$. For every $x \in X$ we say that the analytic function $f_x : D_x \to X$ is an analytic extension of the function $\lambda \to R(\lambda, T)$ if D_x is an open set such that $\rho_W(Q_{\mathcal{P}}, T) \subset D_x$ and

$$(\lambda I - T)f(\lambda) \equiv x, (\forall)\, \lambda \in D_x.$$

Definition If $(X, \mathcal{P})$ is a locally convex space and $T \in (Q_{\mathcal{P}}(X))_0$ has SVEP, then we denote by $\rho_T(x)$ the set of all points $\lambda_0 \in \mathbb{C}$ for which there exists an analytic extension of the function $\lambda \to R(\lambda, T)$, defined in some neighborhood of λ_0. The set $\sigma_T(x)$ is the complement of $\rho_T(x)$.

Remark In the case of bounded operators on a Banach space we have the condition $\rho(T) \subset D_x$, but the lemma 2 implies that this conditions in the case of quotient bounded operators on sequentially complete locally convex space is naturally replaced by the condition $\rho_W(Q_{\mathcal{P}}, T) \subset D_x$.

Remark If $T \in (Q_{\mathcal{P}}(X))_0$ has SVEP then for each $x \in X$ there exists an unique maximal analytic extension of the application $\lambda \to R(\lambda, T)$, which will be denoted by $\tilde{x}$. Since $T \in (Q_{\mathcal{P}}(X))_0$ has SVEP the set $\rho_T(x)$ is correctly defined and is unique.

Remark If $T \in (Q_{\mathcal{P}}(X))_0$ has SVEP and $x \in X$, then

1. $\rho_T(x)$ is an open set;

2. $\rho_T(x)$ is the domain of definition for $\tilde{x}$;

3. $\rho_W(Q_{\mathcal{P}}, T) \subset \rho_T(x)$.

We need the following lemma.

Lemma[9] Let $(X, \mathcal{P})$ be a sequentially complete locally convex space. If $T \in (Q_{\mathcal{P}}(X))_0$ then

1. the application $\lambda \to R(\lambda, T)$ is holomorphic on $\rho_W(Q_{\mathcal{P}}, T)$;

2. $\frac{d^n}{d\lambda^n} R(\lambda, T) = (-1)^n n! R(\lambda, T)^{n+1}$, for every $n \in \mathbb{N}$;

3. $\lim\limits_{|\lambda|\to\infty} R(\lambda, T) = 0$ and $\lim\limits_{|\lambda|\to\infty} R(1, \lambda^{-1}T) = \lim\limits_{|\lambda|\to\infty} \lambda R(1, T) = I$.

Lemma If $T \in (Q_{\mathcal{P}}(X))_0$ has SVEP, then $\sigma_T(x) = \varnothing$ if and only if $x = 0_X$.

Proof If $\sigma_T(x) = \varnothing$, then $\tilde{x}$ is an entire function. Since $|\sigma(Q_{\mathcal{P}}, T)| = r_{\mathcal{P}}(T)$, from lema 2 it results that

$$(\lambda I - T)\tilde{x}(\lambda) = x, \text{ for all } |\lambda| > r_{\mathcal{P}}(T), \tag{4}$$

so by lemma 2 we have

$$\lim_{|\lambda|\to\infty} \tilde{x}(\lambda) = \lim_{|\lambda|\to\infty} R(\lambda, T)x = 0.$$

Therefore, from Liouville's theorem it results that $\tilde{x}(\lambda) \equiv 0$. Using the properties of functional calculus presented in [9] and relations (4) we have

$$x = \frac{1}{2\pi i} \int_{r_{\mathcal{P}}(T)+1} R(\lambda, T)x d\lambda = \frac{1}{2\pi i} \int_{r_{\mathcal{P}}(T)+1} x(\lambda) d\lambda = 0$$

It is obvious that if $x = 0_X$, then $\sigma_T(x) = \varnothing$.

3 Quasi-nilpotent Equivalent Operators

For a pair of operators $T, S \in (Q_{\mathcal{P}}(X))_0$, not necessarily permutable, we consider the following notation

$$(T - S)^{[n]} = \sum_{k=0}^{n} (-1)^{n-k} C_n^k T^k S^{n-k},$$

where $C_n^k = \frac{n!}{(n-k)!k!}$, for all $n \geq 1$ and $k = \overline{1, n}$.

Remark [4] If $T, S, P \in (Q_{\mathcal{P}}(X))_0$ then for all $n \geq 1$ we have:

1. $(T - S)^{[n+1]} = T(T - S)^{[n]} - (T - S)^{[n]} S$.

2. $\sum\limits_{k=0}^{n} (-1)^{n-k} C_n^k (T - S)^{[k]} (S - P)^{[n-k]} = (T - P)^{[n]}$.

Definition We say that two operators $T, S \in (Q_{\mathcal{P}}(X))_0$ are quasi-nilpotent equivalent operators if for every $p \in \mathcal{P}$ we have

$$\lim_{n\to\infty} \left(\hat{p}\left((T-S)^{[n]}\right)\right)^{1/n} = 0 \text{ and } \lim_{n\to\infty} \left(\hat{p}\left((S-T)^{[n]}\right)\right)^{1/n} = 0.$$

In this case we write $T \overset{q}{\backsim} S$.

Remark If $T, S \in (Q_{\mathcal{P}}(X))_0$, then $(T-S)^{[n]} \in Q_{\mathcal{P}}(X)$.

Lemma Let $(X, \mathcal{P})$ be a locally convex space and $T, S \in (Q_{\mathcal{P}}(X))_0$, such that $T \overset{q}{\backsim} S$. Then the series $\sum\limits_{n=0}^{\infty}(T-S)^{[n]}$ and $\sum\limits_{n=0}^{\infty}(S-T)^{[n]}$ converges in $Q_{\mathcal{P}}(X)$.

Proof If $T \overset{q}{\backsim} S$, then

$$\lim_{n\to\infty} \hat{p}\left((T-S)^{[n]}\right)^{1/n} = 0, \text{ for all } p \in \mathcal{P},$$

so by root test the series $\sum\limits_{n=0}^{\infty}\hat{p}((T-S)^{[n]})$ converges. Moreover, for each $\varepsilon \in (0,1)$ and every $p \in \mathcal{P}$ there exists some index $n_{\varepsilon,p} \in \mathbb{N}$ such that

$$\hat{p}\left((T_1-T_2)^{[n]}\right) \leq \varepsilon^n, \text{ for all } n \geq n_{\varepsilon,p}$$

which implies that

$$\sum_{k=n}^{m}\hat{p}\left((T_1-T_2)^{[k]}\right) < \sum_{k=n}^{m}\varepsilon^k < \frac{\varepsilon^n}{1-\varepsilon}, \text{ for all } m > n \geq n_{\varepsilon,p},$$

so $\left(\sum\limits_{k=0}^{n}(T_1-T_2)^{[k]}\right)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since the algebra $Q_{\mathcal{P}}(X)$ is sequentially complete it results that the series $\sum\limits_{n=0}^{\infty}(T_1-T_2)^{[n]}$ converges in $Q_{\mathcal{P}}(X)$.

Analogously, we can prove that the series $\sum\limits_{n=0}^{\infty}(S-T)^{[n]}$ converges in $Q_{\mathcal{P}}(X)$

Lemma The relation $\overset{q}{\backsim}$ defined above is a equivalence relation on $(Q_{\mathcal{P}}(X))_0$.

Proof It is obvious that $\overset{q}{\backsim}$ is simetric and reflexive. Now will prove that $\overset{q}{\backsim}$ is transitive. Let $T_1, T_2, T_3 \in (Q_{\mathcal{P}}(X))_0$ such that $T_1 \overset{q}{\backsim} T_2$ and $T_2 \overset{q}{\backsim} T_3$. Then for every $\varepsilon > 0$ and every $p \in \mathcal{P}$ there exists $n_{\varepsilon,p} \in \mathbb{N}$ such that

$$\hat{p}\left((T_1-T_2)^{[n]}\right) \leq \varepsilon^n \text{ and } \hat{p}\left((T_2-T_3)^{[n]}\right) \leq \varepsilon^n,$$

for every $n \geq n_{\varepsilon,p}$. If

$$M_{\varepsilon,p} = \max_{k=\overline{1,n_{\varepsilon,p}-1}}\left\{\frac{\hat{p}\left((T_1-T_2)^{[n]}\right)}{\varepsilon^k}, \frac{\hat{p}\left((T_2-T_3)^{[n]}\right)}{\varepsilon^k}, 1\right\}, \text{ for all } p \in \mathcal{P},$$

then for every $n \in \mathbb{N}$ we have

$$\hat{p}\left((T_1 - T_2)^{[n]}\right) \leq M_{\varepsilon,p}\varepsilon^n \text{ and } \hat{p}\left((T_2 - T_3)^{[n]}\right) \leq M_{\varepsilon,p}\varepsilon^n, \text{ for all } p \in \mathcal{P}\ .$$

The previous relation implies that

$$\hat{p}\left((T_1 - T_3)^{[n]}\right) = \hat{p}\left(\sum_{k=0}^{n}(-1)^{n-k}C_n^k(T_1 - T_2)^{[k]}(T_2 - T_3)^{[n-k]}\right) \leq$$

$$\leq \sum_{k=0}^{n}(-1)^{n-k}C_n^k\hat{p}\left((T_1 - T_2)^{[k]}\right)\hat{p}\left((T_2 - T_3)^{[n-k]}\right) \leq$$

$$\leq \sum_{k=0}^{n}(-1)^{n-k}C_n^k M_{\varepsilon,p}^2\varepsilon^k\varepsilon^{n-k} = (2\varepsilon)^n M_{\varepsilon,p}^2$$

for all $n \in \mathbb{N}$ and every $p \in \mathcal{P}$, so

$$\hat{p}\left((T_1 - T_3)^{[n]}\right)^{1/n} \leq 2\varepsilon \sqrt[n]{M_{\varepsilon,p}^2}, (\forall)\, n \in \mathbb{N}, \text{ for all } p \in \mathcal{P}.$$

Therefore,

$$\lim_{n\to\infty}\hat{p}\left((T_1 - T_3)^{[n]}\right)^{1/n} = 0, \text{ for all } p \in \mathcal{P}$$

Analogously, we can prove that

$$\lim_{n\to\infty}\hat{p}\left((T_3 - T_1)^{[n]}\right)^{1/n} = 0, \text{ for all } p \in \mathcal{P},$$

so $T_1 \overset{q}{\backsim} T_3$.

To study the properties of quasi-nilpotent equivalent operators with SVEP we need the following lemma.

Lemma Let $(X, \mathcal{P})$ be a locally convex space and $T \in (Q_{\mathcal{P}}(X))_0$ such that $r_{\mathcal{P}}(T) < 1$. Then the operator $I - T$ is invertible and $(I - T)^{-1} = \sum\limits_{n=0}^{\infty} T^n$.

Theorem Let $(X, \mathcal{P})$ be a locally convex space. If $T, S \in (Q_{\mathcal{P}}(X))_0$ are quasi-nilpotent equivalent operators, then T has SVEP if and only if S has SVEP.

Proof Assume that T has SVEP. Let $D_f \subset \mathbb{C}$ be an open set such that $\rho_W(Q_{\mathcal{P}}, S) \subset D_f$ and $f : D_f \to X$ be an analytic function on D_f which satisfies the property

$$(\lambda I - S)f(\lambda) = 0, \text{ for all } \lambda \in D_f.$$

Then, for every $n \geq 0$ we have

$$(T - S)^{[n]}f(\lambda) = \sum_{k=0}^{n}(-1)^{n-k}C_n^k T^k S^{n-k}f(\lambda) =$$

$$= \sum_{k=0}^{n} (-1)^{n-k} C_n^k T^k \lambda^{n-k} f(\lambda) = (T - \lambda I)^n f(\lambda) \tag{5}$$

Since $T \overset{q}{\backsim} S$ results that for every $\varepsilon > 0$ and every $p \in \mathcal{P}$ there exists $n_{\varepsilon,p} \in \mathbb{N}$ such that

$$\hat{p}\left((T-S)^{[n]}\right) \leq \varepsilon^n \text{ and } \hat{p}\left((S-T)^{[n]}\right) \leq \varepsilon^n, \text{ for all } n \geq n_{\varepsilon,p}.$$

Let $\mu \neq \lambda$. Then for every $\varepsilon \in (0, |\mu - \lambda|)$ and for every $p \in \mathcal{P}$ there exists $n_{\varepsilon,p} \in \mathbb{N}$ such that

$$\hat{p}\left(\frac{(T-S)^{[n]}}{(\mu-\lambda)^{n+1}}\right) \leq \frac{\varepsilon^n}{|\mu-\lambda|^{n+1}}, \text{ for all } n \geq n_{\varepsilon,p},$$

so the $\left(\sum\limits_{n=0}^{m} \frac{(T-S)^{[n]}}{(\mu-\lambda)^{n+1}}\right)_m$ is a Cauchy sequences. Since $Q_{\mathcal{P}}(X)$ is sequentially complete it results that the series $\sum\limits_{n=0}^{\infty} \frac{(T-S)^{[n]}}{(\mu-\lambda)^{n+1}}$ is absolutely convergent in the topology of $Q_{\mathcal{P}}(X)$, for every $\mu \neq \lambda$. Moreover, if $r_{\mathcal{P}}(T - \lambda I) < |\mu - \lambda|$, then $r_{\mathcal{P}}(\frac{T-\lambda I}{\mu-\lambda}) < 1$ and from lemma 3 it results that

$$\sum_{n=0}^{\infty} \frac{(T-\lambda I)^n}{(\mu-\lambda)^{n+1}} = ((\mu-\lambda)I - (T-\lambda I))^{-1} = R(\mu, T). \tag{6}$$

From the relations (5) and (6) it results that

$$(\mu I - T)\left(\sum_{n=0}^{\infty} \frac{(T-S)^{[n]}}{(\mu-\lambda)^{n+1}}\right) f(\lambda) = (\mu I - T)\left(\sum_{n=0}^{\infty} \frac{(T-\lambda I)^n}{(\mu-\lambda)^{n+1}}\right) f(\lambda) =$$

$$= (\mu I - T) R(\mu, T) f(\lambda) = f(\lambda),$$

for every μ with the property $r_{\mathcal{P}}(T - \lambda I) < |\mu - \lambda|$. Therefore,

$$g_\lambda(\mu) = \sum_{n=0}^{\infty} \frac{(T-S)^{[n]}}{(\mu-\lambda)^{n+1}} f(\lambda)$$

is an analytic function on $\mathbb{C}\backslash\{\lambda\}$ which verifies the relation

$$(\mu I - T) g_\lambda(\mu) = f(\lambda) \tag{7}$$

on the open set

$$\{\mu \in \mathbb{C} \mid r_{\mathcal{P}}(T - \lambda I) < |\mu - \lambda|\} \subset \mathbb{C}\backslash\{\lambda\},$$

so it follows, by analytic extensions, that the function $g_\lambda(\mu)$ verifies the relation (7) for all $\mu \neq \lambda$. This implies that $\mathbb{C}\backslash\{\lambda\} \subset \rho_T(f(\lambda))$, i.e. $\sigma_T(f(\lambda)) \subset \{\lambda\}$.

Let $\lambda_0 \in D_f$ arbitrary fixed and $r > 0$ such that $D_0 = \{\lambda \in \mathbb{C} \mid |\lambda - \lambda_0| \leq r_0\} \subset D_f$. Since $g_\lambda(\mu)$ is analytic on $\mathbb{C}\backslash\{\lambda\}$ from relation (7) it results that

$$(\mu I - T)\frac{1}{2\pi i}\int\limits_{|\xi-\lambda|=r_0} \frac{g_\xi(\mu)}{\xi-\lambda_0}d\xi = \frac{1}{2\pi i}\int\limits_{|\xi-\lambda|=r_0} \frac{f(\mu)}{\xi-\lambda_0}d\xi = f(\lambda_0) \tag{8}$$

for all $\mu \in D_0$, so $\mu \in \rho_T(f(\lambda_0))$, for every $\mu \in D_0$. Hence $\lambda_0 \in \rho_T(f(\lambda_0))$ and since we already proved above that $\sigma_T(f(\lambda_0)) \subset \{\lambda_0\}$ it results that $\sigma_T(f(\lambda)) = \dot{\varnothing}$. Lemma 2 implies that $f(\lambda) \equiv 0$ on D_0 and since $\lambda_0 \in D_f$ is arbitrary chosen, results that $f(\lambda) \equiv 0$ on D_f. Therefore, S has SVEP.

Analogously we can prove that if S has SVEP then T has SVEP.

Theorem Let $(X, \mathcal{P})$ be a locally convex space. If $T, S \in (Q_{\mathcal{P}}(X))_0$ are quasi-nilpotent equivalent operators and T has SVEP, then $\sigma_T(x) = \sigma_S(x)$, for every $x \in X$.

Proof First we remark that from previous theorem it results that S has SVEP. Let $x \in X$ arbitrary chosen and let $x(\lambda)$ be the analytic function on $\rho_T(x)$ which verify the condition

$$(\lambda I - T)x(\lambda) = x,\ \lambda \in \rho_T(x). \tag{9}$$

Let $\lambda_0 \in \rho_T(x)$ arbitrary fixed. Since $\rho_T(x)$ is an open set there exists $0 < r_1 < r_2$ such that $D_i(\lambda_0) \subset \sigma_W(Q_{\mathcal{P}}, T)$, $i \in \{1, 2\}$, where

$$\bar{D}_i(\lambda_0) = \{\mu \in \mathbb{C} |\ |\mu - \lambda_0| \leq r_i\},\ i \in \{1, 2\}.$$

For every $p \in \mathcal{P}$ denote by M_p^1 the maximum of $x(\lambda)$ on $\bar{D}_2(\lambda_0)$. Hence, for $\lambda \in \bar{D}_2(\lambda_0)$ we have

$$p\left(\frac{x^{(n)}(\lambda)}{n!}\right) = p\left(\frac{1}{2\pi i}\int\limits_{|\xi-\lambda_0|=r_2} \frac{x(\xi)}{(\xi-\lambda)^{n+1}}d\xi\right) \leq \frac{M_p^1 r_2}{(r_2-r_1)^{n+1}},\ (\forall)n{\geq}0. \tag{10}$$

In the proof of lemma 3 we proved that for every $\varepsilon > 0$ and each $p \in \mathcal{P}$ there exists $M_{\varepsilon,p} > 0$ such that

$$\hat{p}\left((T-S)^{[n]}\right) \leq M_{\varepsilon,p}\varepsilon^n,\ \text{for all } n{\geq}0. \tag{11}$$

Therefore, the relations (10) and (11) implies that

$$p\left((T-S)^{[n]}\frac{x^{(n)}(\lambda)}{n!}\right) \leq \hat{p}\left((T-S)^{[n]}\right)p\left(\frac{x^{(n)}(\lambda)}{n!}\right) < \frac{M_{\varepsilon,p}M_p^1 r_2}{r_2-r_1}\left(\frac{\varepsilon}{r_2-r_1}\right)^n,$$

for all $n{\geq}0$. Taking $\varepsilon = \frac{r_2-r_1}{2}$ it results that for each $p \in \mathcal{P}$ there exists $M_{\varepsilon,p} > 0$ such that

$$p\left((T-S)^{[n]}\frac{x^{(n)}(\lambda)}{n!}\right) \leq \frac{M_p}{2^n},\ (\forall)n{\geq}0, \tag{12}$$

where $M_p = \frac{M_{\varepsilon,p} M_p^1 r_2}{r_2 - r_1}$ does not depend on $\lambda \in \bar{D}_2(\lambda_0)$. The relation (12) shows that the series

$$\sum_{n=0}^{\infty} p\left((-1)^n (T-S)^{[n]} \frac{x^{(n)}(\lambda)}{n!}\right)$$

converges for every $\lambda \in \bar{D}_2(\lambda_0)$ and every $p \in \mathcal{P}$, so since X is sequentially complete it results that the series $\sum\limits_{n=0}^{\infty} (-1)^n (T-S)^{[n]} \frac{x^{(n)}(\lambda)}{n!}$ converges absolutely and uniformly on $\bar{D}_2(\lambda_0)$. But $\lambda_0 \in \rho_T(x)$ is arbitrary fixed, hence this series converges absolutely and uniformly on every compact $K \subset \rho_T(x)$, which implies that

$$x_1(\lambda) = \sum_{n=0}^{\infty} (-1)^n (T-S)^{[n]} \frac{x^{(n)}(\lambda)}{n!} \tag{13}$$

is analytic on $\rho_T(x)$. Now we prove that

$$(\lambda I - S) x_1(\lambda) = x, \text{ for all } \lambda \in \rho_T(x).$$

If we differentiate $n \geq 1$ times the equality (9), then we have

$$(\lambda I - T) x^{(n)}(\lambda) = -n x^{(n-1)}(\lambda), \ \lambda \in \rho_T(x).$$

From previous relations and remark 3 it results

$$(\lambda I - S) x_1(\lambda) = \sum_{n=0}^{\infty} (-1)^n (\lambda I - S)(S-T)^{[n]} \frac{x^{(n)}(\lambda)}{n!} =$$

$$= \sum_{n=0}^{\infty} (\lambda I - S) \left((\lambda I - S) - (\lambda I - T)\right)^{[n]} \frac{x^{(n)}(\lambda)}{n!} =$$

$$= \sum_{n=0}^{\infty} \{((\lambda I - S) - (\lambda I - T))^{[n+1]} +$$

$$+ ((\lambda I - S) - (\lambda I - T))^{[n]} (\lambda I - T)\} \frac{x^{(n)}(\lambda)}{n!} =$$

$$= \sum_{n=0}^{\infty} (-1)^{n+1} (S-T)^{[n+1]} \frac{x^{(n)}(\lambda)}{n!} + \sum_{n=0}^{\infty} (-1)^n (S-T)^{[n]} (\lambda I - T) \frac{x^{(n)}(\lambda)}{n!} =$$

$$= \sum_{n=0}^{\infty} (-1)^{n+1} (S-T)^{[n+1]} \frac{x^{(n)}(\lambda)}{n!} + (\lambda I - T) x(\lambda)$$

$$- \sum_{n=1}^{\infty} (-1)^n (S-T)^{[n]} \frac{x^{(n-1)}(\lambda)}{(n-1)!} = (\lambda I - T) x(\lambda) = x$$

for all $\lambda \in \rho_T(x)$. This shows that $\rho_T(x) \subset \rho_S(x)$, so $\sigma_S(x) \subset \sigma_T(x)$.

Analogously, it can be proved that $\sigma_T(x) \subset \sigma_S(x)$.

References

[1] Allan G.R., *A spectral theory for locally convex algebras*, Proc. London Math. Soc. **15** (1965), 399-421.

[2] Chilana, A., *Invariant subspaces for linear operators on locally convex spaces*, J. London. Math. Soc., **2** (1970) , 493-503.

[3] Colojoara, I., *Elemente de teorie spectrală*, Editura Academiei Republicii Socialiste România, Bucureşti 1968.

[4] Colojoara, I. and Foias, C., *Theory of Generalized Spectral Operators*, Gordon and Breach, Science Publishers, New York-London-Paris, 1968.

[5] Joseph, G.A., *Boundness and completeness in locally convex spaces and algebras*, J. Austral. Math. Soc., **24** (Series A), (1977), 50-63.

[6] Kramar, E., *On the numerical range of operators on locally and H-locally convex spaces*, Comment. Math. Univ. Carolinae **34**,2(1993), 229-237.

[7] Michael, A., *Locally multiplicatively convex topological algebras*, Mem. Amer. Math. Soc., 11, 1952.

[8] Stoian, S.M., *Spectral radius of a quotient bounded operator*, Studia Univ. Babes-Bolyai, Mathematica, No.**4**, 2004, pg.115-126.

[9] Stoian, S.M., *Holomorphic Functional calculus for Regular Operators*, Proceedings of Functions Theory on Infinite Dimensional Spaces X,11-14 December 2007, 91-110.

[10] Waelbroeck, L., *Etude des algèbres complètes* , Acad. Roy.Belgique Cl. Sci.Mem. coll. in 8, **31**(1960), no.7.

SECTION F

COMPUTER SCIENCE

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Geospatial Analysis via Web Browser: the OGC Web Processing Service (WPS) and its applications within the WRME project

[1]L. Casagrande, [2]A. Pierleoni, [3]M. Bellezza,
[4]S. Casadei
University of Perugia, Department of Civil and Environmental Engineering, Borgo XX Giugno 74 Perugia, Italy
E-mails: [1]luca.casagrande@gmail.com, [2]apierleoni@unipg.it, [3]bellezza@unipg.it, [4] casadei@unipg.it

Abstract

The Open Geospatial Consortium is an international no-profit organization, based on voluntary subscription, that defines specific techniques for geospatial localization services (location based). The OGC is composed of more than 280 members (Governments, Private Industries, Universities and Institutions) and aims to the development and the implementation of standards for contents, services and the exchange of geographical data with these being "open and extendible". In this work, the use of the WPS (Web Processing Service) protocol will be described. This standard makes possible to perform the typical operation of Desktop GIS application, through a Web interface and via an HTTP protocol. Within the WRME project that is a prototypal WEB SDSS for water resources evaluation and management, it has been developed a tool that allows, through a WEB interface, the search of water withdrawal licenses starting from a defined section of the river network. This procedure requires as search parameters both numerical inputs (such as the kind of water use, the maximum allowed withdrawal, etc) and geographical (such as withdrawals upstream of a certain section of the network). For this specific kind of search it is necessary to use the potentialities of a GIS software since it implies the spatial analysis of the river network, together with a database containing the information of the basin. Therefore the system is composed of a relational database with geospatial extension for vectorial data; the GIS engine (GRASS GIS) that performs all the geospatial operations and finally the PyWPS that is a Python implementation of the WPS protocol. The WEB interface offers various thematisms in order to provide a better support to the decisional process. The thematisms are shown by means of a the Web Map Service protocol taking advantage of the javascript framework OpenLayers. Besides all the potentialities described so far, it is also possible to exploit and integrate third party

ISSN-1584-5990

services like Google Maps or others. In this paper, starting from the results achieved within the WRME project, will be clearly shown the wide range of capabilities of the WPS protocol for the analysis of territorial data and how it is easily possible to create "ad hoc" interfaces for any WEB browser. The project is entirely based on and exploits only Open Source technology.

Keywords: geospatial localization, WPS, information retrieval.

1 Introduction

Recent developments in the acquisition of data through remote sensing techniques, the introduction of GIS engines, the ability to share raw data and processed via the Internet and use of increasingly massive simulation models and optimization have provided access to managers to a high amount of information to support decision making.However, as shown by a recent analysis conducted by the U.S. NRC (NRC, 1999), it happens often that the usefulness of this information is limited by the fact that they are not provided in a manner appropriate to the many decision-makers. Decision-making at the basin scale can be interpreted at two opposite approaches: the Top-Down and Bottom-Up. The Top-Down approach provides that the decision to produce a management plan to be submitted to interest groups while the Bottom-Up introduces the input of interest groups in the planning of operations management. It should also be noted that efficient management at the basin scale requires integration of information and knowledge, a solid database, simulation models and expert judgments to solve practical problems and provide a scientific basis for making decisions. In this context it becomes necessary to provide a DSS (Decision Support System) that is complete, integrated, intuitive and easy "approach to all groups of stakeholders to develop, understand and possibly discard alternative strategies, but not very productive. The DSS itself must therefore combine a set of functional components, first among which a DBMS (DataBase Management System), a GIS engine, appropriate simulation models and a refund of the more user friendly that it can make available to different groups interest. The difficulty in developing a DSS with these characteristics is not to be found both in the absence of simulation models, but in the possibility of making it available and usable on a common platform to all stakeholders in the decision-making. Internet offers a great opportunity for the dissemination and sharing of information and applications dedicated. Advances in telecommunications over the Internet, GIS, hydrogeological modeling and their integration in SDSS (Spatial Decision Support System) provides an opportunity to connect research and studies of the technical management of water resources to the political dynamics managerial decision making. The objective of this work is to develop a decision-making system that integrates the technologies available in an "Internet Framework, taking up with the natural evolution of decision making, GIS engine and simulation models to be Web-Based.

2 The Web Processing Service

The specified Web Processing Service (WPS) provides client access to pre-programmed calculations and/or computation models that operate on spatially referenced data. The data required by the service can be delivered across a network, or made available on a server. In this context image data formats or data exchange standards such as Geography Markup Language (GML) could be used. The calculation can be as simple as subtracting one set of spatially referenced numbers from another (e.g. determining the difference in influenza cases between two different seasons), or as complicated as a global climate change model. Enabling geospatial processing on the Internet requires the development of a wide variety of web services to support atomic geospatial operations as well as sophisticated modelling capabilities. It is important to standardize the way these processes are called, in order to reduce the amount of programming required, and to facilitate the implementation and adoption of new services. WPS is meant to help OGC members to achieve these goals.
The WPS interface specifies three operations that can be requested by a client and performed by a WPS server, all mandatory implementation by all servers. Those operations are:

- **GetCapabilities**: This operation allows a client to request and retrieve service metadata (or Capabilities) documents that describe the abilities of the specific server implementation. The GetCapabilities operation provides the names and general descriptions of each of the processes offered by a WPS instance. This operation also supports negotiation of the specification version being used for client-server interactions.

- **DescribeProcess**: This operation allows a client to request and retrieve detailed information about the processes that can be run on the service instance, including the inputs required, their allowable formats, and the outputs that can be produced.

- **Execute**: This operation allows a client to run a specified process implemented by the WPS, using provided input parameter values and returning the outputs produced. These operations have many similarities to other OGC Web Services, including the WMS, WFS, and WCS.

3 The WRME Project

The wrme project consists in a set of tools devoted to the visualization and analysis of hydrometeorological information and that may allow the calculation of indexes for the quantitative evaluation of the available water resource. This tool embeds several hydrological models for the evaluation and the management of water resources via the new calculation, visualization and sharing technologies, but the underlying philosopy is to always bear in mind that the management should be an integrated process and that all the stakeholders have

to reach a certain degree of agreement on the data and the policies adopted. The construction of this system is the first step needed to bring all primary and secondary stakeholders around a common virtual table that can make available a single computational system on which they will build management plans and policies for the exploitation of the water resource at the basin scale. In this way it is possible to avoid all problems deriving from the fact that each stakeholder is always willing to present his/her own data, models and results that most of the time are not even comaparable, so that consensus reaching and/or conflict resolution might be facilitated.

4 The infrastructure

The system uses a relational database to store all geospatial data that are part of the project. For raster data, a system of Tile Index to improve access procedure has been used. Users who have been granted access as administrator, may act directly on single data by changing both the attributes and the geometry property. The structure foresees the adoption of a GRASS GIS processing engine and a pyWPS implementation of the WPS Service. In order to show geographical data inside the web interface the Javascript framework Openlayers has been used.The next paragraphs will provide a more specific description of these components:

4.1 GRASS GIS

GRASS (Geographical Resources Analysis Support System) is a raster/vector GIS combined with integrated image processing and data visualization subsystems. It includes more then 350 modules for managment, processing analysis and visualization of georeferenced data. Starting as a military project in 1984 it adopted the GNU GPL (General Public License) in 1999.
Unlike most proprietary GIS, GRASS provides complete access to its internal structure and algorithms. This is very important in a project like WRME, because there are many operations that are not bundled with standard GIS. Within the WRME project GRASS has been used for all the elaborations involving both raster and vector data with excellent results.

4.2 pyWPS

The pyWPS project is an implementation of the WPS standard written in Python. It has been developed to be used with GRASS GIS, even though the use of other command-line tools like GDAL, R, etc. is still possible. For each specific process implemented inside pyWPS, there's a process file written in Python made of 3 main parts:

1. General information about the process including description, dataset used, etc.

2. A list of all the inputs/outputs used to run the process. Those can be numerical constant, geographic data (vector or raster), or an XML file.

3. The main core of the process with all the elaborations involving the process. To run it succesfully, each output declared in previous part, must own a value.

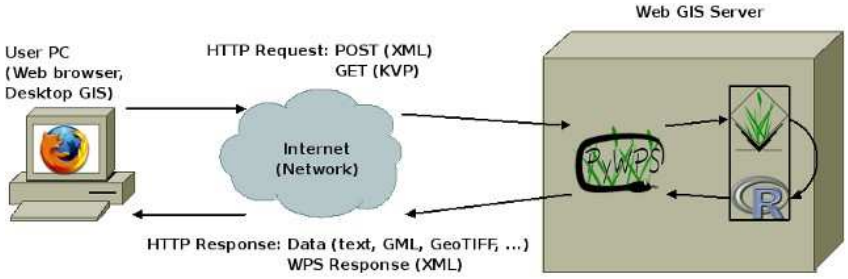


Figure 1: Structure of the pyWPS project

pyWPS works as a CGI application, using POST and GET protocol to handle request. Right now the version 3.0.0 implements the version 1.0.0 of the WPS standard.

4.3 OpenLayers

OpenLayers is a pure JavaScript library for displaying map data in most modern web browsers, with no server-side dependencies. OpenLayers implements a JavaScript API for building rich web-based geographic applications, similar to the Google Maps and MSN Virtual Earth APIs, with one important difference OpenLayers is Free Software, developed for and by the Open Source software community. OpenLayers has been used within the WRME project to visualize all the geographical data inside the web interface using the WMS OGC standard. An advantage of this type of solution is that we could overlay data coming from proprietary source like Google, with specific data related to the Tiber Basin.

5 An application to manage water license

One of the on line applications created with the use of the WPS protocol is the one that allows the management of the withdrawal licences present within the river basin concerned. The purpose of this tool is to enable the water resource managers, to obtain in few simple steps all the information about the state of the water resource in each section of the river together with the features of the withdrawals related to the sub-catchment defined. This operation requires the combined use of a GIS tools (to find all the licences upstream the selected section) and a database management tool (to perform additional filters on the

licences geographically found).
As the user clicks on a point on the hydrographic network, the GetFeatureInfo request of the WMS standard will retrieve the information to be displayed inside a popup created using a spacial object inside the OpenLayers framework. At

Figure 2: GetFeatueInfo output

Figure 3: GetFeatueInfo output visualization

this point the user specifies the types of withdrawals and the attributes to be filtered, hence all these information are sent to the WPS server. Once the

process has been succesfully executed,two output are created:

1. A GML (Geographic Markup Language) showing the withdrawals;
2. A CSV with all the alphanumeric informations on the withdrawals;

A summary of the results is displayed in the popup under the river information described on the previous part. This allows the operator to get all the information in just a few passages. In order to improve the browsing experience these tasks are performed using an AJAX architecture.

6 Conclusions

The use of algorithmic geospatial analysis within the common GIS software, needs a certain know-how ranging from the characteristics of different data formats, to the concept of reference systems. In this paper it has been clearly shown the potentialities of the WPS protocol and its easy implementability in more complex projects. The case study of the WRME project has also proved its capabilities of dealing with different data format and different input-output relation.
This structure provides many fundamental advantages to the users of the system:

- Both the geographical and alphanumeric data are centralized. This means that every update process is immediately available to all the users.
- The use of GIS analysis tools doesn't require specific knowledge other then the one concerning the parameters to be analized.
- The Web interface can be changed according to the needs of the individual user and more over it is cross-platform meaning that it can be used regardlessly the operative system.
- The system is based exclusively on the standards defined by the Open Geospatial Consortium granting compatibility with the most common instruments that deal with geographical data.
- It only uses free and open source software therefore there are no additional costs for software licences.

References

[1] NRC (National Research Council), 1999; New Strategies for America's Watersheds. National Academy Press, Washington, D.C.

[2] Open Source GIS: A GRASS approach - Markus Neteler and Helena Mitasova

[3] GRASS GIS Home page: http://grass.osgeo.org

[4] pyWPS Home Page: http://pywps.wald.intevation.org/

[5] OpenLayers Home Page: http://www.openlayers.org/

[6] OGC WPS Standard Specification: http://www.opengeospatial.org/standards/wps

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Improving Development Process of Information Systems Based Internet

Marinela Lazarica
"Constantin Brâncoveanu" University, Braila, Romania
E-mail: mlazarica@yahoo.com

Abstract

Software engineering is dominated by intellectual activities focused on solving problems with immense complexity and numerous unknowns in competing perspectives. The next generation of software processes is driving toward a more production-intensive approach dominated by automation and economies of scale. In present, it seems likely that the ability to handle complexity, and deliver systems in a timely manner, will ultimately become business critical factors; the "Patterns for e-business" initiative of IBM is a frontal assault on proliferating complexity and integration. For this reason, I described the patterns because these are valuable tools for communicating acquired knowledge and experience to improve software quality and productivity of development information systems.

Keywords: e-business, patterns development process, best practices.

Software engineering is dominated by intellectual activities focused on solving problems with immense complexity and numerous unknowns in competing perspectives. In the 1960s and 1970s, each project used a custom process and custom tools. After that, in 1980s and 1990s, the software industry matured and transitioned to more of an engineering discipline. The next generation of software processes is driving toward a more production-intensive approach dominated by automation and economies of scale.

Two important remarks must be mentioned:

- Software will be the key differentiator for every business in the new economy.
- Reduce complexity and improve processes at all levels of software engineering.

Today, most software organizations are facing the need to integrate their own environment and infrastructure for software development. This typically results in the selection of more or less incompatible tools with different information repositories, from different vendors, on different platforms, using different jargon, and based on different process assumptions.

ISSN-1584-5990

Based on many observations of specialists, these are the three most discriminating approaches for achieving significant process improvements:

1. Transitioning to an iterative process.
2. Attacking the significant risks first through a component-based, architecture-first focus.
3. Using key software engineering best practices, from the outset, in requirements management, visual modeling, change management, and assessing quality throughout the life-cycle.

These trends and new requirements proven *the importance of using best practices in the development of successful information systems based Internet.*

To improve the development process [1] of information systems over time, it need to reuse the experience of the IT developers in such a way that the whole process can be made simpler and faster. A solution for this problem is the patterns. Patterns encapsulate a designer's time skill, and knowledge to solve a software problem.

The specialists defined a pattern as "*a solution to a problem in a context*". Each pattern describes a problem which occurs over and over again in our environment and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice. Patterns begin as an idea or a best practice used in many projects. Patterns can exist at various level - from business to architecture, design and programming to development and run-time administration - and assist in the creation of artefacts according to best practices through all stages of the solution development life cycle. *The Patterns development process* [2] is a live project, ever-evolving and being updated as new products are released and are used in the building of real applications. IBM has compiled the collective wisdom and experience gained from more than 20,000 successful Internet-based engagements and transformed that wisdom into the IBM Patterns for e-business. These patterns provide the best-practice blueprints and tools to facilitate the application development process and enable companies to shorten time to market, reduce risk and, in general, see a more significant return on investment.

No matter the methodology driving an application development project, *the major steps in successful projects are essentially the same.*

The entire patterns development process, in fact the IBM Patterns for e-business layered asset model, is presented in Figure 1, (Source - http://www-128.ibm.com/developerworks/patterns/)

As a development team puts together requirements [4], the Patterns Web site helps match those requirements to the appropriate pattern, *Business pattern.* As the team refines the requirements and determines which existing systems, data stores and infrastructure will be integrated into the system, they can use the *Application pattern* to develop how application components and data within a business solution interact. After choosing the Application pattern, the team

Figure 1: The Patterns for e-business layered asset model

can match *Runtime patterns* topology based on the existing environment and business needs. The Runtime pattern establishes the components needed to support the chosen Application pattern. Developers must now determine which products to use for the actual development. At this point, the Patterns provide a wide range of options and lend the developer significant assistance. Developers can access from the Patterns Web site *Runtime Product Mappings* that identify tested, optimal software implementations for each Runtime pattern. Associated with each Runtime Product Mapping on the Web site are best-practice application, design, development and management guidelines that have been gleaned in the process of developing these patterns. Developers can use them to access a wealth of information about other, similar development efforts.

I exemplified how the singleton pattern solved a problem appeared in an information system based Internet (an e-business application), using the tool Rational Software Architect [3] (Figure 2, Figure 3, Figure 4, Figure 5, Figure 6).
Singleton It is very common to discover classes in applications that should only have one instance. Singleton is a design pattern that shows how to ensure that only one single instance of a class exists at any one time. A common example of utility's singleton is when need to maintain basic information, such as the name, the phone number to call if the customers need support.

These five steps necesaires for the integration of Singleton pattern (using the tool Rational Software Architect) in an e-business application, are presented in the following figures.

Conclusions

Today, most software organizations are facing the need to integrate their own environment and infrastructure for software development. This typically results in the selection of more or less incompatible tools with different information repositories, from different vendors, on different platforms, using different

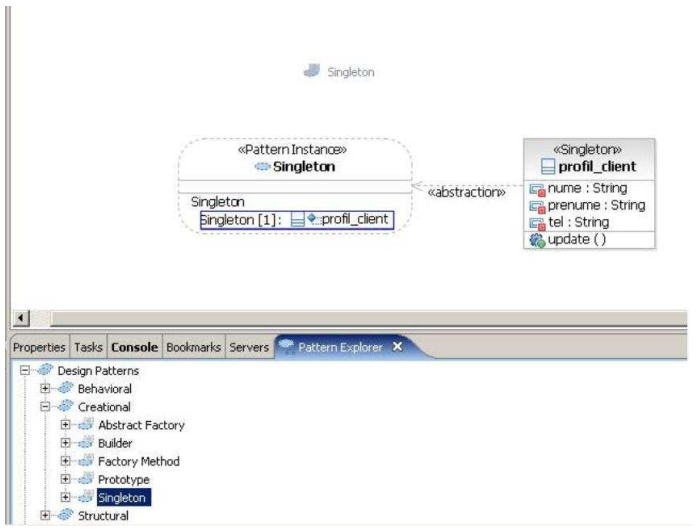


Figure 2: Step 1

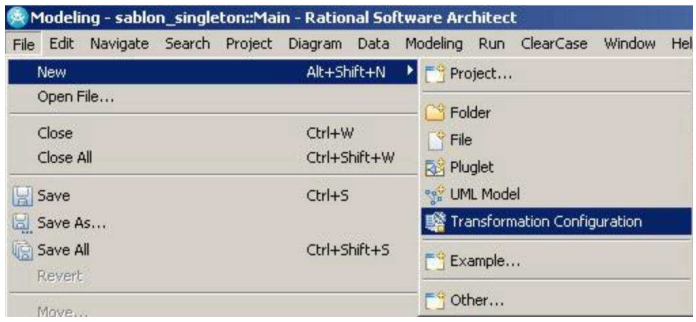


Figure 3: Step 2

jargon, and based on different process assumptions. The most discriminating approaches for achieving significant process improvements are: an iterative process and the best practices, from the outset, in requirements management, visual modeling, change management, and assessing quality throughout the lifecycle. The Patterns development process is a live project, ever-evolving and being updated as new products are released and are used in the building of real applications. IBM has constructed the Patterns and the Patterns Web site to enable development teams to work through the development process using their preferred methodology or the methodology suggested by consultants engaged to assist in the project.

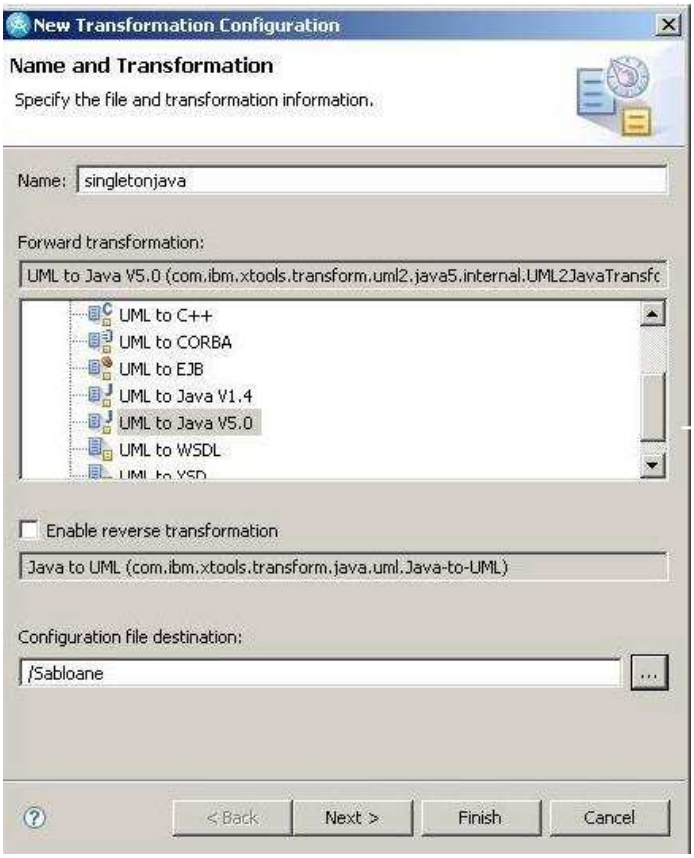


Figure 4: Step 3

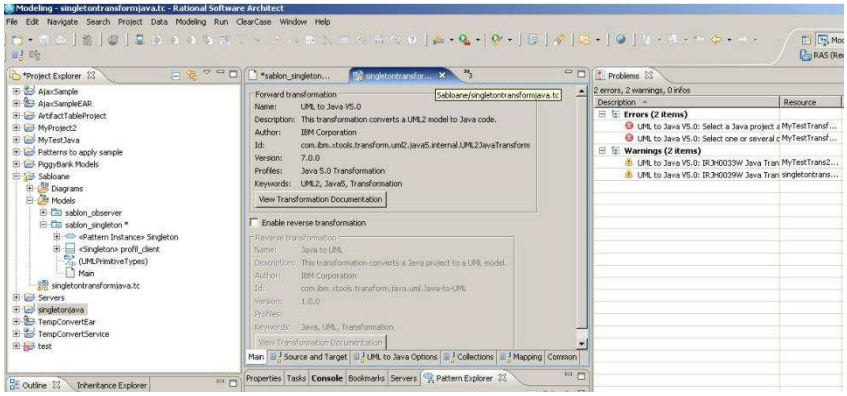


Figure 5: Step 4

References

[1] Lazarica Marinela, *The aspects regarding the design of e-commerce systems*, doctoral thesis, 2007, Bucharest

[2] Lord John G., *Facilitating the application development process using the IBM Patterns for e-business*, 2001

[3] Swithinbank Peter, Chessell Mandy, ş.a - *Patterns: Model-Driven Development Using IBM Rational Software Architect*, IBM Redbooks, 2005

[4] http://www-128.ibm.com/developerworks/patterns/

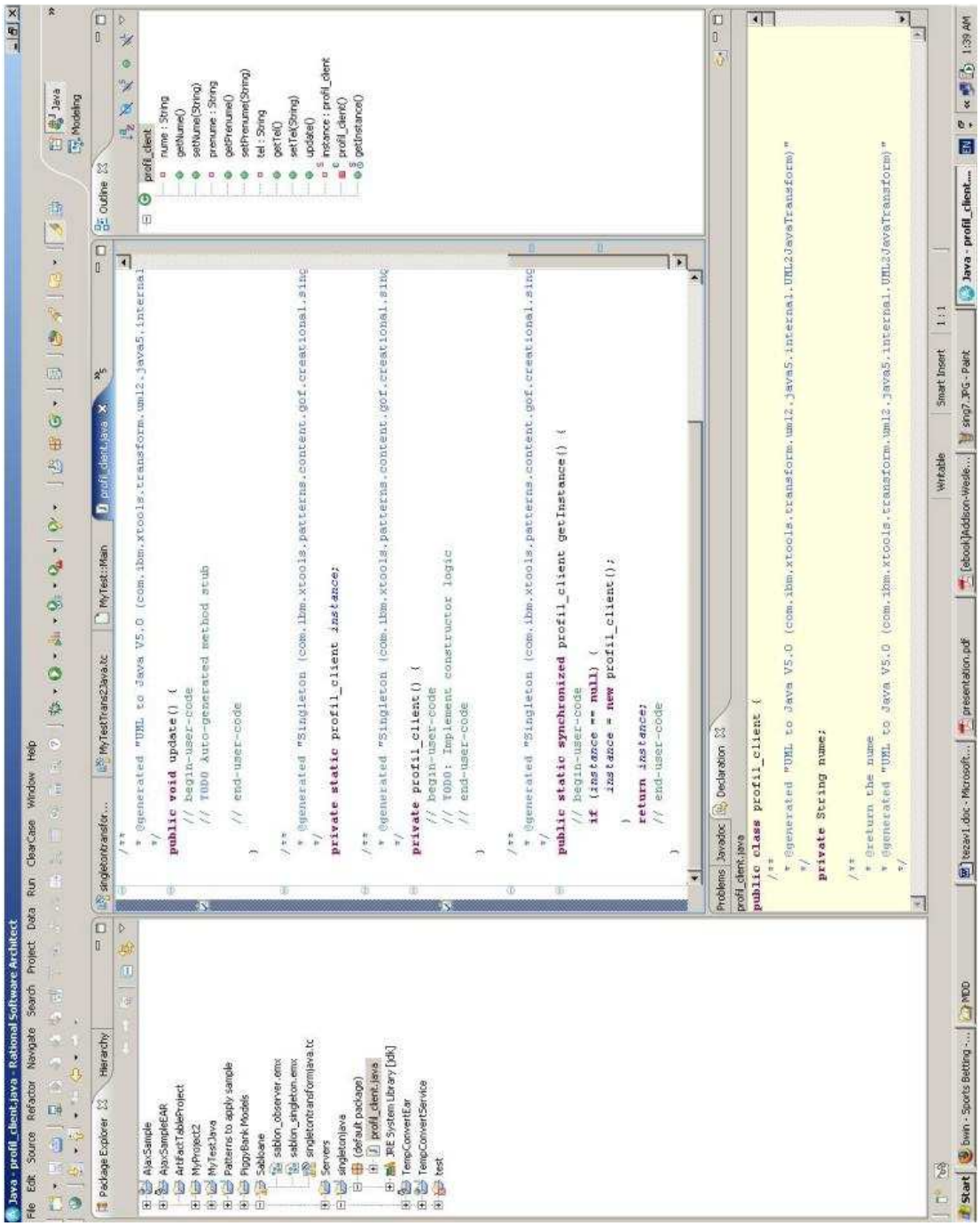


Figure 6: Step 5

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

An Ant Colony Algorithm for Solving the Dynamic Generalized Vehicle Routing Problem

[1]Petrică C. Pop, [2]Camelia-M. Pintea,
[2]D. Dumitrescu
[1]Department of Mathematics and Computer Science,
North University of Baia Mare, Romania
[2] Department of Computer Science, Babeş-Bolyai University
Cluj-Napoca, Romania
E-mail: [1]pop_petrica@yahoo.com

Abstract

The Generalized Vehicle Routing Problem (GVRP) is a generalization of the Vehicle Routing Problem (VRP) introduced by Ghiani and Improta in 2000, in which the nodes of a graph are partitioned into a given number of nodes sets, called clusters, and we are interested in finding the optimal routes from the given depot to the number of predefined clusters which include exactly one node from each cluster. The dynamic GVRP is a variation of the GVRP, in the sense that it is assumed that distances between nodes (cities) are no longer fixed. In this paper we present an ant colony based algorithm for solving the dynamic version of the Generalized Vehicle Routing Problem.

Keywords: optimization problem, metaheuristic algorithms, ant colony - like algorithms.

1 Introduction

After an initial emphasis on static problems, some of the focus is now shifting towards dynamic variants of combinatorial optimization problems. The work done so far deals with static problems where all the data are known in advance, i.e. before the optimization has started.

Problems associated with determining optimal routes for vehicles from one or several depots to a set of locations/customers, subject to various constraints, such as vehicle capacity, route length, time windows, etc., are known as Vehicle Routing Problems (VRP). These problems have a significant economic importance due to the many practical applications in the field of distribution, collection, logistics, etc. A wide body of literature exists on the VRP problem (for an extensive bibliography, see Laporte and Osman [12], Laporte [14], etc).

ISSN-1584-5990

The Generalized Vehicle Routing Problem (GVRP) is a generalization of the Vehicle Routing Problem (VRP) introduced by Ghiani and Improta [6] in 2000, in which the nodes of a graph are partitioned into a given number of nodes sets, called clusters, and we are interested in finding the optimal routes from the given depot to the number of predefined clusters which include exactly one node from each cluster. They proposed as well a solution procedure by transforming the GVRP into a Capacitated Arc Routing problem for which an exact algorithm and several approximate procedures are reported in literature. In 2003, Kara and Bektas [9] proposed an integer programming formulation for GVRP with a polynomially increasing number of binary variables and constraints and in 2008 Kara and Pop [10] presented two integer linear programming formulations for GVRP with $O(n^2)$ binary variables and $O(n^2)$ constraints.

The dynamic GVRP is a variation of the GVRP, in the sense that it is assumed that distances between nodes (cities) are no longer fixed. This situation may appear in real life applications when can appear delays due to maintenance work, accidents, etc. and therefore the travel time may vary. We mention that there are also considered some other dynamic variants for combinatorial optimization problems such as variants resulting from insertion or deletion of nodes.

The GVRP and its dynamic version may arise in real-life applications such as loop material flow design, post-box collection, arc routing, computer operations, manufacturing, logistics, and distribution of goods by sea to a potential number of harbors, etc.

The aim of this paper is to present an effective metaheuristic algorithm based on ant colony optimization for a dynamic version of the generalized vehicle routing problem. Computational results for several benchmarks problems are going to be reported.

2 Definition and complexity of the GVRP

Let $G = (V, A)$ be a directed graph with $V = \{0, 1, 2, ..., n\}$ as the set of vertices and the set of arcs $A = \{(i, j) \mid i, j \in V, i \neq j\}$. A nonnegative cost c_{ij} associated with each arc $(i, j) \in A$. The set of vertices (nodes) is partitioned into $k + 1$ mutually exclusive nonempty subsets, called clusters, $V_0, V_1, ..., V_k$ (i.e. $V = V_0 \cup V_1 \cup ... \cup V_k$ and $V_l \cap V_p = \emptyset$ for all $l, p \in \{0, 1, , k\}$ and $l \neq p$). The cluster V_0 has only one vertex 0, which represents the depot, and remaining n nodes belonging to the remaining k clusters represent geographically dispersed customers. Each customer has a certain amount of demand and the total demand of each cluster can be satisfied via any of its nodes. There exist m identical vehicles, each with a capacity Q.

The generalized vehicle routing problem (GVRP) consists in finding the minimum total cost tours of starting and ending at the depot, such that each cluster should be visited by exactly once, the entering and leaving nodes of each cluster is the same and the sum of all the demands of any tour (route) does not exceed the capacity of the vehicle Q. An illustrative scheme of the GVRP and

a feasible tour is shown in the next figure.

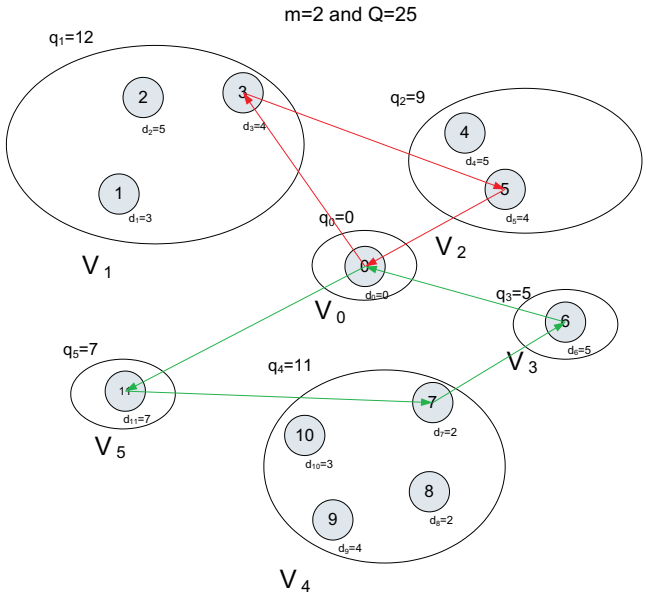


Figure 1: An example of a feasible solution of the GVRP

The GVRP is NP-hard because it includes the generalized traveling salesman problem (GTSP) as a special case when $m = 1$ and $Q = \infty$.

Several real-world situations can be modeled as a GVRP. The post-box collection problem described in Laporte et al. [11] becomes an asymmetric GVRP if more than one vehicle is required. Furthermore, the GVRP is able to model the distribution of goods by sea to a number of customers situated in an archipelago as in Philippines, New Zealand, Indonesia, Italy, Greece and Croatia. In this application, a number of potential harbours is selected for every island and a fleet of ships is required to visit exactly one harbour for every island.

Several applications of the GTSP (Laporte et al. [13]) may be extended naturally to GVRP. In addition, several other situations can be modeled as a GVRP, these include:

- the Traveling Salesman Problem (TSP) with profits (Feillet et al. [4]);
- a number of Vehicle Routing Problem (VRP) extensions: the VRP with selective backhauls, the covering VRP, the periodic VRP, the capacitated general windy routing problem, etc.;
- the design of tandem configurations for automated guided vehicles (Baldacci et al. [1]).

3 Ant Colony System for solving $GVRP$

Ant-systems based techniques are metaheuristics able to solve large NP-hard problems. The Ant Colony System (ACS) is a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation [3]. A metaheuristic based on ant algorithms for the static Generalized Vehicle Routing Problem (GVRP) is given in [16].

In order to solve the $GVRP$, Pop *et al.* [16] used artificial ants to construct vehicle routes by successively choosing exactly one node from each cluster to visit until each cluster has been visited. Whenever the choice of another node from a cluster would lead to an infeasible solution because of vehicles capacity, the depot is chosen and a new route is started.

For the selection of a not yet visited node from a cluster, two aspects were taken into consideration:

- how good was the choice of that node. The artificial ant decides which node is selected based on the pheromone trail $\tau_{ij}(t)$ associated with each edge $e = (i, j)$ at time t;
- how promising is the choice of that node. This measure is called *visibility*, denoted by η_{ij} and is defined as the inverse distance from a node to the next node, $\eta_{ij} = \frac{1}{c_{ij}}$.

To favor the selection of an edge that has a high pheromone level and high visibility, a probability function p_{ij}^k was defined as follows:

$$p_{ij}^k(t) = \frac{[\tau_{ij}^k(t)]^\alpha [\eta_{ij}^k]^\beta}{\sum_{o \in J_i^k} [\tau_{io}^k(t)]^\alpha [\eta_{io}^k]^\beta}$$

where J_i^k is the set of unvisited neighbors of node i by ant k, $j \in J_i^k$, and α and β are parameters used for tuning the relative importance of trails and visibility.

After an artificial ant has constructed a feasible solution, the pheromone trails are laid depending on the objective value L_k. For each edge that was used by ant k, the pheromone trail was updated according to the following rule:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \rho \frac{1}{L_k}$$

where $\rho \in (0, 1)$ is a parameter called evaporation rate.

In order to stop ants visiting the same cluster in the same tour a tabu list was maintained. This prevents ants visiting clusters they have previously visited. The ant tabu list is cleared after each completed tour.

There were used as many ants as there are customers n in $GVRP$ and one ant is placed at each node (customer) at the beginning of an iteration. After initializing the basic ant system algorithm, the two steps: *construction of vehicle routes* and *trail update* were repeated for a given number of iterations.

In the best ant systems to date, pheromone trails are not only modified locally by the artificial ants during or just after the construction of a new solution,

but also globally, considering the best solution generated by all the ants at a given iteration or even the best solution ever constructed, see C. Pintea, P. Pop and D. Dumitrescu [15].

4 The Ant Colony Algorithm for Solving the Dynamic GVRP

The *Dynamic GVRP*, is a variation of *GVRP* in the sense that the total length of the routes may vary due to the fact that it can appear blocked ways due to maintenance work, accidents, etc. We mention that several others variants for combinatorial optimization problems have been considered , such as variants resulting from insertion or deletion of nodes, see [7, 8].

The *Dynamic GVRP* can be stated as follows: we are looking for the optimal routes from a given depot to a number of predefined clusters with the condition that they include exactly one from each distinct and available (unblocked) cluster and the sum of all the demands of any tour (route) does not exceed the capacity of the vehicle Q.

The ant-based introduced algorithm for solving the *Dynamic GVRP* is based on the *Ant Colony System for GVRP* [16] and has the characteristics of a dynamic problem.

In our Ant-based algorithm for the DGTSP a colony of n ants, corresponding to the number of customers, incrementally construct solutions. The choice of the next node is based on two main components: pheromone trails and a heuristic value called visibility.

Initialization phase. First we randomly find a solution of the problem. Let DT^+ be the best routes found and DL^+ the length of the collection of routes. $T^+(i)$ denotes the shortest collection of routes found and $DL^+(i)$ its length for each blocked cluster i. The ants are placed in the depot and all the edges are initialized with a certain amount of pheromone τ_0.

At the beginning there is relatively much exploration, after a while all connections that are not promising will be slowly cut off from the search space because they do not get any positive reinforcement and the associated pheromones have evaporated over time.

Dynamism in GVRP: at each algorithm iteration, a cluster, randomly chosen, is missing from the tour.

Practically, are solved a number of *Generalized Vehicle Routing Problems (GVRP)* equal to the number of clusters. Each *GVRP* problem, for the *blocked* cluster, build a $T^k(blocked)(t)$ collection of routes with the length $L^k(blocked)(t)$. As we already mentioned, the purpose is to find the shortest collection of routes $T^+(blocked)$ and the length $L^+(blocked)$ for each blocked cluster. The solution of *DGVRP* is the shortest tour DT^+ and his length LT^+, found within all these shortest routes.

In the following, for the randomly chosen cluster, *blocked*, we solve *Generalized Vehicle Routing Problems* using an ant-based algorithm, as in [15].

In the dynamic case, solutions that are bad before a change in the environment might be good afterwards. Now, if the ant system has converged to a state where those solutions are ignored, very promising connections will be lost and the result will be a suboptimal solution. That is way we use a new local update rule.

A technique called *shaking* is used in order to smooth all the pheromone levels in a certain way. If the amount of pheromones on an edge becomes much higher than all the other edges, this edge will be always be chosen. That is a way for the static case to ensure that a good connection will always be followed, but it prevents ants from taking another connection when the good connection is blocked. The formula used in shaking is the local update rule:

$$\tau_{ij}(t+1) = \tau_0\Big(1 + \log\Big(\frac{\tau_{ij}}{\tau_0}\Big)\Big).$$

The *Ant-based System for DGVRP* algorithm can be stated as follows:

```
Ant colony algorithm for Dynamic GVRP
begin
initialization phase
for t = 1 to N_iter do
      randomly choose a cluster, blocked
      and set the cluster blocked visited
   for k = 1 to m do
      build a collection of routes T^k(blocked)(t) by applying
      nc − 1 times
         choose the next node j from an unvisited cluster
         update pheromone trails by applying the local rule
   end for
   for k = 1 to n do
      compute the length L^k(blocked)(t) of the collection of routes
      T^k(blocked)(t)
   if an improved collection of routes are found update
T^+(blocked), L^+(blocked)
   and the solutions:  DT^+, DL^+.
   for every edge (i, j) ∈ T^+(blocked) do
      update pheromone trails by applying the global rule
end for
print the shortest tour DT^+ and its length DL^+
end
```

5 Implementation details and computational results

In order to evaluate the performance of the new metaheuristic, the *ACS*-based algorithm for *GVRP* was tested on seven benchmark problems drawn from

TSPLIB library test problems [17]. These problems contain between 51 and 101 customers (nodes), which are partitioned into a given number of clusters, and in addition the depot.

Originally the set of nodes in these problems are not divided into clusters. To divide them into node-sets we used the procedure called CLUSTERING proposed by Fischetti *et al.* [5]. This procedure sets the number of clusters $m = [\frac{n}{5}]$, identifies the m farthest nodes from each other and assigns each remaining node to its nearest center. Obviously, some real-world problems may have different cluster structures. However, the solution proposed in this paper is able to handle any cluster structure.

The ACS-based algorithm for solving the Dynamic GVRP was implemented in *Java*.

The initial value of all pheromone trails, $\tau_0 = 0.1$. The parameters for the algorithm are critical as in all other ant systems. Currently there is no mathematical analysis developed to give the optimal parameter in each situation. In the ant-based algorithm for *GVRP*, the values of the parameters were chosen as follows: $\alpha = 1$, $\beta = 5$, $\rho = 0.0001$ and $q_0 = 0.9$.

In the next two tables, we present the computational results obtained for solving the *Dynamic GVRP* using the ant-based algorithm.

Table 1: Problem average values - *ACS*-based algorithm for *Dynamic GVRP*

Problem	*VR*	*Q*	*Q'*	*Avg. length*	*Avg. time*
11eil51	2	160	320	421.46	1.31
16eil76A	2	140	280	702.61	2.12
16eil76B	3	100	300	686.23	10.74
16eil76C	2	180	360	620.56	7.42
16eil76D	2	220	440	557.88	4.94
21eil101A	2	200	400	717.4	9.12
21eil101B	2	112	224	1015.11	6.52

Table 2: Problem characteristics and Best Solution - *ACS*-based algorithm for *Dynamic GVRP*

Problem	*Best length*	*Best time*	*Number vehicles*	*Number Routes*
11eil51	391.28	4.88	6	3
16eil76A	663.55	4.10	10	5
16eil76B	620.24	1.83	15	5
16eil76C	541.07	29.15	8	4
16eil76D	487.14	2.66	6	3
21eil101A	692.72	1.31	8	4
21eil101B	960.37	7.46	14	7

The columns in Table 1 and Table 2 are as follows:

- Problem: The name of the test problem: the digits at the beginning are the number of clusters and those at the end give the number of nodes.
- VR: The minimal number of vehicles needed for a route in order to cover even the largest capacity of a cluster (VR=Vehicles/Route)
- Q': the capacity $Q \cdot VR$, where Q is the capacity of a vehicle available at the depot.
- Best length: the minimal length of collection routes
- Best time : the time of the minimal collection routes
- Avg. length: the average length of 20 sets of collection routes
- Avg. time: the average time (in seconds) for 20 sets of collection routes
- Number Routes: the number of routes for the best solution
- Number vehicles: the total number of vehicles within the best solution

The computer used for numeric test: AMD 2600, 1.9Ghz and 1024 MB. An example of a best solution for 21*eil*101 is following in table.

Table 3: An example of best solution for 21*eil*101. Successively are: *node*(*cluster*) and the length of a route.

A solution for 21*eil*101 - *total cost 692.72*
14(1) 23(3) 27(5) 29(6) 54(14) - 143.49
35(2) 36(4) 72(11) 8(7) 32(8) 11(16) 20(17) 64(20) - 266.30
15(9) 1(10) 5(12) - 133.02
7(13) 18(18) 47(19) 44(21) - 149.91

The computational values are the result of the average of 20 successively executions of both algorithm. Termination criteria of the *ACS*-based algorithm for *Dynamic GVRP* is given by the number of iteration, $N_{iter} = 25000$.

Acknowledgments: This research is partially supported by the Grant ID 508, New Computational Paradigmes for Dynamic Complex Problems, funded by the Ministry of Education, Research and Innovation, Romania.

References

[1] R. Baldacci, E. Bartolini and G. Laporte, *Some applications of the Generalized Vehicle Routing Problem*, Le Cahiers du GERAD, G-2008-82, 2008.

[2] B. Bullnheimer, R.F. Hartl and C. Strauss, *An improved Ant System algorithm for the Vehicle Routing Problem*, Annals of Operations Research, 89, pp. 319-328, 1999.

[3] M. Dorigo, G. Di Caro, *The ant colony optimization meta-heuristic*, In D. Corne, M. Dorigo, F. Glover (Eds.), New ideas in optimization 11-32, London: McGraw-Hil, 1999.

[4] D. Feillet, P. Dejax and M. Gendreau, *Traveling salesman problems with profits*, Transportation Science, Vol. 39, pp. 188-205, 2005.

[5] M. Fischetti, J.J. Salazar, P. Toth. *A branch-and-cut algorithm for the symmetric generalized traveling salesman problem*, Operations Research, 45:378-394, 1997.

[6] G. Ghiani, G. Improta, *An efficient transformation of the generalized vehicle routing problem*, European Journal of Operational Research, 122, pp. 11-17, 2000.

[7] M. Guntsch, J. Branke, M. Middendorf, H. Schmeck, *ACO strategies for dynamic TSP*, Abstract Proceedings of ANTS'2000, 59-62, 2000.

[8] M. Guntsch, M. Middendorf, H. Schmeck. *An Ant Colony Optimization approach to dynamic TSP*, Proc. of GECCO'2001, 860-867, 2001.

[9] I. Kara, T. Bektas, Integer linear programming formulation of the generalized vehicle routing problem, 5-th EURO/INFORMS Joint International Meeting, July 6-10, 2003, Istanbul, Turkey.

[10] I. Kara, P. Pop, New Mathematical Models of the Generalized Vehicle Routing Problem and Extensions, International Conference on Applied Mathematical Programming and Modelling, Bratislava, Slovacia, May 27-30, 2008.

[11] G. Laporte, S. Chapleau, P.E. Landry, H. Mercure, *An algorithm for the design of mail box collection routes in urban areas*, Transportation Research B, 23, pp. 271-280, 1989.

[12] G. Laporte, I. H. Osman, *Routing Problems: A bibliography*, Annals of Operations Research, Vol. 61, pp. 227-262. 1995.

[13] G. Laporte, A. Asef-Vaziri and C. Sriskandarajah, *Some applications of the generalized traveling salesman problem*, Journal of Operational Research Society, Vol. 47, pp. 1461-1467, 1996.

[14] G. Laporte, *What you should know about the Vehicle Routing Problem*, Naval Research Logistics, 54 (2007) 811-819.

[15] C.M. Pintea, D. Dumitrescu and P.C. Pop, *Combining heuristics and modifying local information to guide ant-based search*, Carpathian Journal of Mathematics, Vol. 24, No. 1, pp. 94-103, 2008.

[16] P.C. Pop, C.M. Pintea, I. Zelina and D. Dumitrescu, *Solving the Generalized Vehicle Routing Problem with an ACS-based Algorithm*, American Institute of Phisics, Vol. 1117, pp. 157-162, 2009, (Conf. Proc. BICS Tg. Mures 5-7 November 2008).

[17] http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/vrp/

Proceedings of the 5th International Conference "Dynamical Systems and Applications"
Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009

Fast Fuzzy Iris Segmentation

Nicolaie Popescu-Bodorin, IEEE Member
nb.popescu.mi@spiruharet.ro

Abstract

This paper presents a new approach to iris segmentation. Here we show that the k-means algorithm proves to be extremely useful in finding the most representative iris segment with almost insignificant computational cost when compared to other classical iris segmentation procedures within Wildes's and Daugman's models (fitting circular contours by solving optimization problems via gradient ascent or by using the Hough transform or by iterating active contours). A new fuzzy iris segmentation procedure based on k-Means and Run Length Encoding is proposed and tested here. It is designed to guarantee that similar segmentation results will be obtained for similar eye images, despite the fact that the degree of occlusion may vary from an image to another (and usually this is the case, even when they are images representing the same eye). The proposed iris segmentation procedure consists in two steps: pupil finding and limbic boundary approximation. Both of them are fuzzy approaches based on k-Means and Run Length Encoding. The result of the procedure is a circular ring (concentric with the pupillary boundary) which approximates the iris segment. It is not the best approximation for one specific iris segment but it is a stable one, meaning that for a class of similar eye images representing the same eye, similar circular rings are obtained (despite the variations introduced by occlusions and illumination conditions). When two similar images of the same eye are compared, each detected circular ring is pointing to the same physical support, possibly occluded by eyelids, eyelashes, specular and lighting reflections.

1 Introduction

This article is meant to illustrate how the iris segmentation can be performed quiqly using only algorithms derived from k-Means and Run Length Encoding.

It is currently accepted the idea that finding the iris boundaries means to identify a collection of pixels as a solution of an optimization problem [1, 3]:

- filtering the edges until the finding of a (nearly) circular contour minimizing the integro-differential Daugman operator [2] ;
- filtering the edges to find circles approximating the iris boundaries using Hough transform [14];
- iterating active contours to find optimal positions matching the iris boundaries [4].

ISSN-1584-5990

The iris segmentation procedure proposed here uses all of these ideas, but without explicitly solving an optimization problem. It is an attempt to reconsider the iris segmentation not as an explicit optimization problem (which tipically has a high computational complexity) but as an implicit one, which can be solved by parsing and classifying chromatic values.

One fundamental notion used very often in this paper is the *k-means echipotential chromatic map*. It was introduced in [8] together with the *Fast k-Means Quantization* (FKMQ) algorithm and its properties were further detailed in [9]. Also, it have to be said that the current proposed solution for finding the pupil has its roots in an attempt to design a real time version for a computer-assisted diagnosis tool which detects the response to hepatits viral infection in an image representing a patient serum [11, 12].

The general idea exploited in this paper is that a target signal can be enhanced against unwanted noise if a suitable filter can be designed for this purpose. On the other hand, the k-means quantization has proved to be suitable for enhancing the area morphology in any image in which the correlation between morphology and chromatic variation is sufficiently strong. Further in this paper we are going to show that three simple operations - *Run Length Encoding / Quantisation* and *Fast K-Means Quantization* - are enough to achieve fast iris segmentation for very large classes of iris images.

The second section of this paper is dedicated to pupil finding. The proposed procedure is meant to enhance the pupil against its surroundings and against the oclusions that may occur in its area (specular and lighting reflections or eyelashes). There are two main operations at this step:

- a k-means quantization of the eye image is applied to obtain its *k-means equipotential chromatic map* in which the actual pupil is the most circular solid object contained in the lowest level (this lowest level will be reffered further as the *pupil cluster*);
- a fuzzy membership assignment of the pixels within the *pupil cluster* to the actual pupil is constructed using Run Length Encoding (RLE) and requantizing run length coefficients in unsigned 8-bit integer domain. The result is a *Run Length Quantization* of the *pupil cluster*. The defuzzification is achieved by running FKMQ once again to determine a threshold above which the membership of a pixel to the actual pupil is guaranteed. The result is a pupil indicator. From each of its pixels a flood-fill operation can be started to determine the available pupillary boundary which can be further approximated through a circle or an ellipse.

The finding of the most representative circular ring of the iris is presented in the third section. The idea is to build three fuzzy membership assignment functions, each of them mapping all the circles concentric with the pupil to the actual pupil, to the actual iris segment and to the actual non-iris area. The defuzzification is achieved by counting the votes that each circle receives from each of those membership functions.

Some benchmark details, comments and conclusions are given in the fourth section in order to illustrate the degree of robustness and the range of applicability of the proposed iris segmentation procedure.

1.1 Computational Routines

This section will discuss the computational routines used in the proposed iris segmentation procedure:

Run Length Encoding (RLE) is one of the most simple and popular data compression algorithms [10]: for a given array, any subarray of redundant values is coded as a pair representing its histogram, as shown in the following example:

if $\mathbf{V} = [1,1,1,1,0,0,1,1,1,1,1,1,1,1]$ is the vector to be encoded, the run length encoding of $\mathbf{V}$ is $rle(\mathbf{V}) = [(1,4),(0,2),(1,8)]$.

Run Length Quantization for Binary Images is defined here as a procedure to replace all ones within a binary image with the corresponding run length coefficients re-quantized in unsigned 8-bit integer domain by some custom quantization function, as illustrated in the following example:

$$[1,1,1,1,0,0,1,1,1,1,1,1,1,1] \Rightarrow [(1,4),(0,2),(1,8)] \Rightarrow$$
$$[(4,4),(0,2),(8,8)] \Rightarrow [(128,4),(0,2),(255,8)] \Rightarrow$$
$$[128,128,128,128,0,0,255,255,255,255,255,255,255,255]$$

The (re-)quantization function used in the above example (and further in this paper) is:

$$\mathrm{rqf}(\mathbf{V}(\mathbf{n})) = \min(255, \max(1, \mathrm{round}(255 * \mathbf{V}(\mathbf{n}) / \max(\mathbf{V}(\mathbf{n}))))),$$

where $\mathbf{V}$ is the vector to be quantized and $\mathbf{n}$ is the index of its non zero components.

If instead being a vector, $\mathbf{V}$ is a matrix (a binary image), it make sense to talk about *Vertical*, or *Horizontal* (or other custom *Directional*) *Run Length Encoding* and quantization. In these cases *Directional Run Length Quantization* procedure will assign for each white (non zero) pixel a quantization value according to the length of the longest continuous segment of ones containing that pixel and lying on a vertical, or horizontal line, or on the other custom direction, as needed. In this way the *(Directional) Run Length Quantization* procedure encodes a morphological property of the input image into a new signal (re-quantized image) by giving the same chromatic meaning to all of those white pixels sharing the same (*Directional*) *Run Length Encoding*.

Fast k-Means Image Quantization (FKMQ) is a variant of k-means algorithm designed for fast chromatic clustering in unsigned 8-bit integer domain. It transforms the input image in an equipotential chromatic map [9] with k levels by replacing each chromatic value with its closest centroid.

A suboptimal (incomplete) variant of FKMQ [7] can be easily derived by imposing termination in a small number of iterations while resetting the first centroid to the minimum available value (or to zero) after each iteration. In

this way, the input image is forced to return a handler to that area covered by lower chromatic values. This is particularly useful in detecting the pupil in those eye images in which the pupil is the darker zone.

2 Finding the pupil

Let us consider an eye image like that in the Fig.1.a (0006-L-0008.j2c file from Iris Database ©University of Bath). Its k-means equipotential chromatic map is revealed (Fig.1.b) by applying the FKMQ algoritm. The pupil cluster (PC) is then defined as the lowest level (cluster) on this map (Fig.1.c). It can be seen in the figure that the pupil cluster PC contains a good indication for the actual pupil perturbed by specular lights and eyelashes.

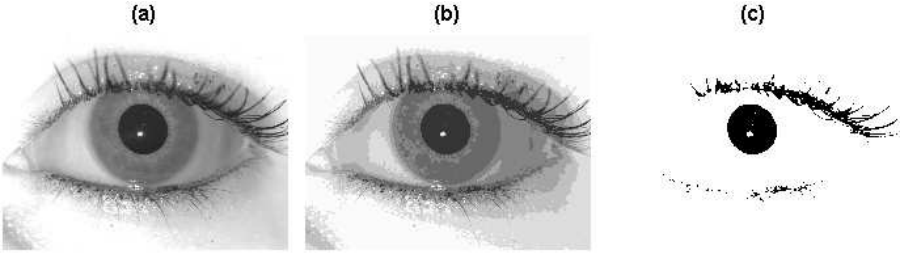


Figure 1: The original eye image (a), its 8-means quantization (b), and the pupil cluster PC (c).

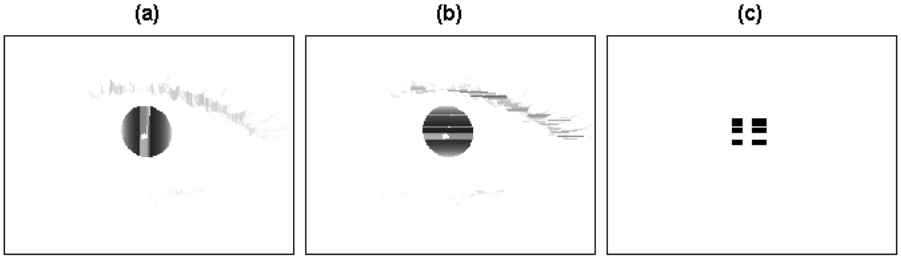


Figure 2: Vertical Run Length quantization of the pupil cluster (a); Horizontal Run Length quantization of the pupil cluster (b); The pupil indicator PI (c).

The first problem to be solved at this stage is finding at least one pixel in PC which belongs for sure to the actual pupil (finding a suitable pupil indicator).

By computing the Haar wavelet decomposition pyramid for the pupil cluster it can be seen that finding a suitable pupil indicator means to erode the pupil cluster up to such a scale above which the details come only from the actual pupil. In such a stage, the subsampled frame of the wavelet pyramid is a suitable pupil indicator itself.

Definig an adaptive erosion procedure is the key for finding a suitable pupil indicator: for each pixel within the pupil cluster, the structuring element of the erosion is adaptively determined by the morphological context in the neighbourhood of that pixel, meaning that the vertical and horizontal Run Length

Encoding is used to requantize (in unsigned 8-bit integer domain) the run length coefficients corresponding to that pixel (i.e. the degree of membership of that pixel to a shorter or a longer continuous segment lying in the pupil cluster). In short, requantized directional run length coefficients encode the degree of membership of each pixel to the actual pupil. The argument is fact that being (or containing) the most circular solid object within the pupil cluster, the actual pupil is the most resilient set to erosion [**?**] to be found in the pupil cluster.

Let us consider the matrices RLV (Fig.2.a) and RLH (Fig.2.b) as being the vertical and horizontal Run Length quantizations of the pupil cluster PC, respectively. The following computational procedure results in finding a pupil indicator for a given pupil cluster:

Function [k,PI] = getpi(RLH, RLV);

k=16; PI = 0*RLH;

While PI is the null matrix do:

Compute the k-means quantizations of RLV and RLH:
RLHQ = fkmq(RLH, k);
RLVQ = fkmq(RLV, k);

Select the logical index of the highest cluster
within RLHQ and RLVQ, respectively:
LIH = (RLHQ == max(RLHQ(:)));
LIV = (RLVQ == max(RLVQ(:)));

Compute the binary matrix PI as logical conjunction of LIH and LIV:
PI = LIH & LIV;
k = k -1;

EndWhile;
END.

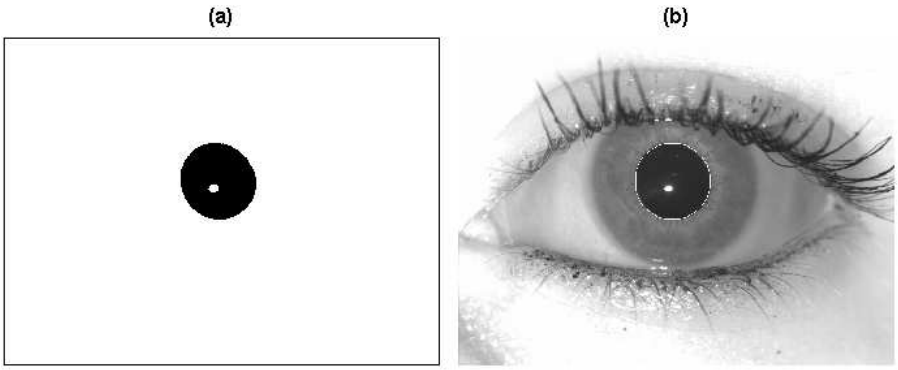


Figure 3: Available pupil segment (a); Ellipse approximating the pupil (b).

Each pixel within the pupil indicator can now be used as a starting point for a flood-fill operation. The most acurate pupil segment available in the pupil cluster is identified (Fig.3.a) this way. Further, the specular lights are filled using Run Length Encoding once again. The result is then fitted into a rectangle and approximated by an ellipse (Fig.3.b).

Summarizing all the operations described above, the proposed Fast Pupil Finder algorithm can be stated as follows:

Fast Pupil Finder Algorithm (N. Popescu-Bodorin):

INPUT: the eye image IM;

1.Extract the pupil cluster
PC = fkmq(IM,16);
PC = (PC == min(PC));

2.Compute horizontal and vertical Run Length quantizations of PC:
RLV(:,j) = vrleq(PC);
RLH(j,:) = hrleq(PC);

3.Compute the pupil indicator PI:
[k, PI] = getpi(RLH, RLV);
PI = find(PI == 1);
PI = PI(1);

4.Extract available pupil segment through a flood-fill operation:
P = imfill(PC, PI);

5.Fill the specular lights:
P = rlefillsl(P);

6.Approximate the pupil by an ellipse;

OUTPUT: The ellipse approximating the pupil;
END.

3 Fast Fuzzy Iris Segmentation

It can be seen in [5] that the most frequent causes of false rejection are the occlusions and the segmentation errors. This is the reason why an accurate segmentation is a must in order to achieve a certain degree of performance in recognition.

The first step in our approach guarantees an acurate pupil localization and enables us to unwrap the eye image (Fig.4.a - 0009-L-0007.j2c file from Iris Database ©University of Bath) in polar coordinates (Fig.4.b) and also to practice the localization of the limbic boundary in the rectangular unwrapped eye image (Fig.4.c), obtaining an iris segment as in Fig.4.e.

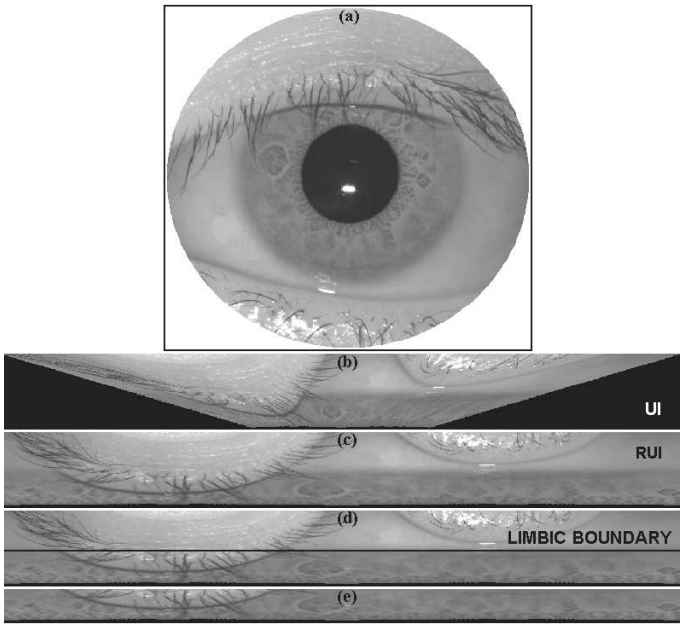


Figure 4: Iris segmentation stages

Fuzzy Iris Segmentation Procedure (N. Popescu-Bodorin):

INPUT: the eye image IM;

1.Apply the Fast Pupil Finder procedure to find pupil radius and pupil center;

2.Unwrap the eye image in polar coordinates (UI - Fig.4.b);

3.Strech the unwrapped eye image UI to a rectangle (RUI - Fig.4.c);

4.Compute three column vectors: A, B, C, where A and B contain the means of the lines within UI and RUI, respectively. C is the mean of the lines within the [A B] matrix;

5.Compute P, Q, R as being 3-means quantizations of A, B, C (Fig.5);

6.For each line of the unwrapped eye image count the votes given by P,Q and R. All the lines receiving at least two positive votes is assumed to belong to the actual iris segment.

7.Find limbic boundary and extract the iris segment (Fig.5, Fig.4.d, Fig.4.e);

OUTPUT: pupil center, pupil radius, index of the line representing limbic boundary and the final iris segment

END.

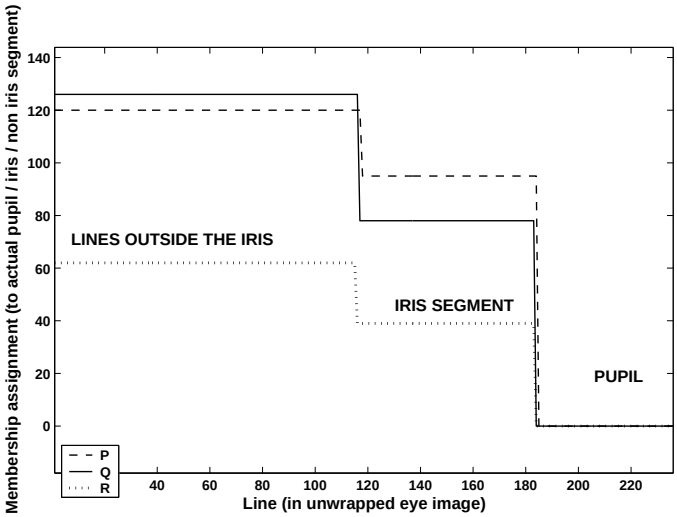


Figure 5: Iris segmentation procedure: Line assignment (step 5)

4 Implementation Details

The proposed Fast Fuzzy Iris Segmentation algorithm is currently implemented [7] using Matlab and calibrated for University of Bath Iris Image Database (the free version [13]).

The accuracy of the iris segmentation algorithm is very good for entire test database (1000 images / 50 unique eyes). For illustration, a demo program is available for download [7].

When a single detector is used for finding the limbic boundary, some wrong segmentation results are observed. For this reason the demo program uses two limbic bondary detectors (complementary to each other), both of them being variants of the above Fast Fuzzy Iris Segmentation procedure. This enables us to obtain the following iris segmentation error rates:
- pupil finder failures: 1/1000;
- limbic boundary detectection failures: 5/999;
- total number of iris segmentation failures: 6/1000.

Also, what is really relevant in our approach, is the fact that the average time spent finding the limbic boundary is just two times greater than the time spent finding pupillary boundary [7]. The speed of the computation is achieved by formulating pupil finding and limbic boundary finding as one-dimensional optimization problems and also by using fast computational procedures: Fast k-Means variants [7, 9] and Run Length Encoding. K-Means and Run Length Quantizations are used to encode the meaningfull information, avoiding any unnecessary traversal of the input image.

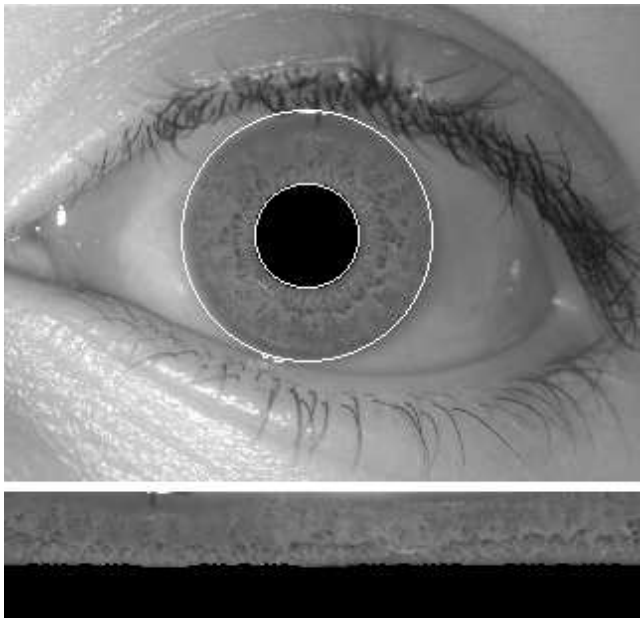


Figure 6: Fast Fuzzy Iris Segmentation Demo Program

Circle finding is usually achieved by solving an optimization problem with three parameters (two center coordinates and radius) varying within a rectangular parallelepiped in $\mathbf{R^3}$. Here, the computation of the pupil indicator depends on a single parameter (a threshold for the requantized horizontal and vertical run length coefficients computed for the pupil cluster, threshold above which the membership of a pixel to the actual pupil is guaranteed). On the other hand, the limbic boundary is determined searching for a line number identifying the first iris line (see the first iris circle in Fig.6, or the first line of the unwrapped iris in the same figure, or the limbic boundary in Fig.4.d).

The range of applicability of the of the proposed algorithm is limited to that class of images which satisfy two working hypotheses:

- the correlation between the area morphology and chromatic variation is sufficiently strong, or else, the initial k-means quantization of the eye image will fail to return an accurate pupil cluster;
- the pupil is (or contains) the biggest, most circular and darker object, or else, other object will prove a higher resiliency to erosion than the actual pupil and a false pupil indicator will be computed in consequence.

In short, the efficiency of the proposed algorithm proves that iris segmentation can be treated as being a one-dimensional optimization problem if there is enough accurate morphological information stored as chromatic variation.

5 Validating Proposed Iris Segmentation Procedure through Recognition Results

A big issue is the fact that numerical segmentation results (numerical values identifying pupil centers and iris radii) for Bath University Iris Database are not publicly available at this time. That is why the accuracy of the proposed iris segmentation procedure is proven here using the iris recognition results presented in figure 7.

Using 4258 different iris images, Daugman [3] shown that the distribution of Hamming distances between different irises matches a binomial distribution with $p = 0.5$ and $N = 249$ degrees-of-freedom, or a normal distribution around the mean $p = 0.499$ with standard deviation $\sigma = 0.0317$. N and σ ($N = p*(1-p)/\sigma^2$) express the the amount of difference between iris codes of different irises as a result of correlated Bernoulli trials with 249 tosses of a fair coin. He also show that if all 2048 bits in an iris code were independent then the standard deviation would be $\sigma = 0.011$.

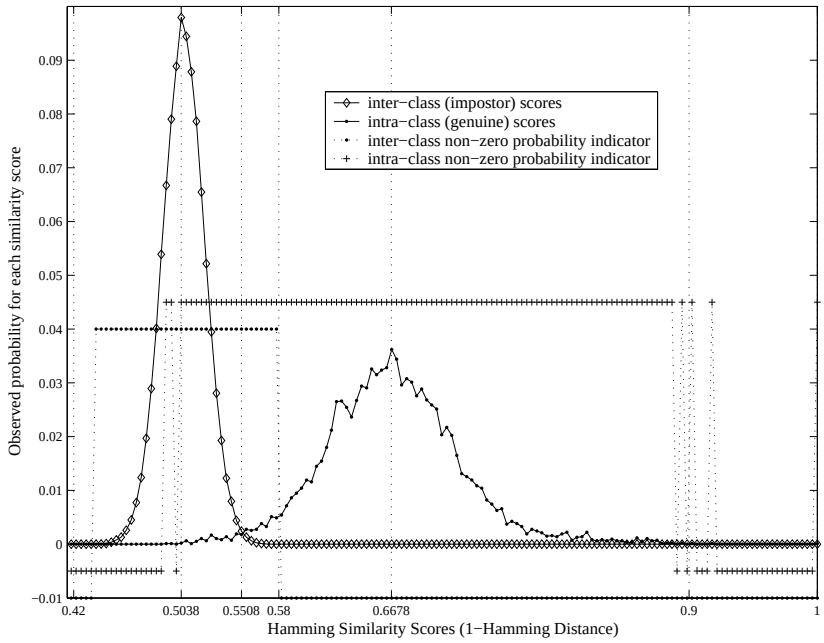


Figure 7: Narrowing the distribution of inter-class matching scores by using Fast Fuzzy Iris Segmentation procedure

In our approach, we use 1000 different images [13] and it can be seen in figure 7 that the distribution of Hamming simmilarity scores between different irises matches a normal distribution around the mean $p = 0.5038$ with standard

deviation $\sigma = 0.0166$ proving more reliable iris texture encoding in terms of statistical independence between the iris codes of different irises.

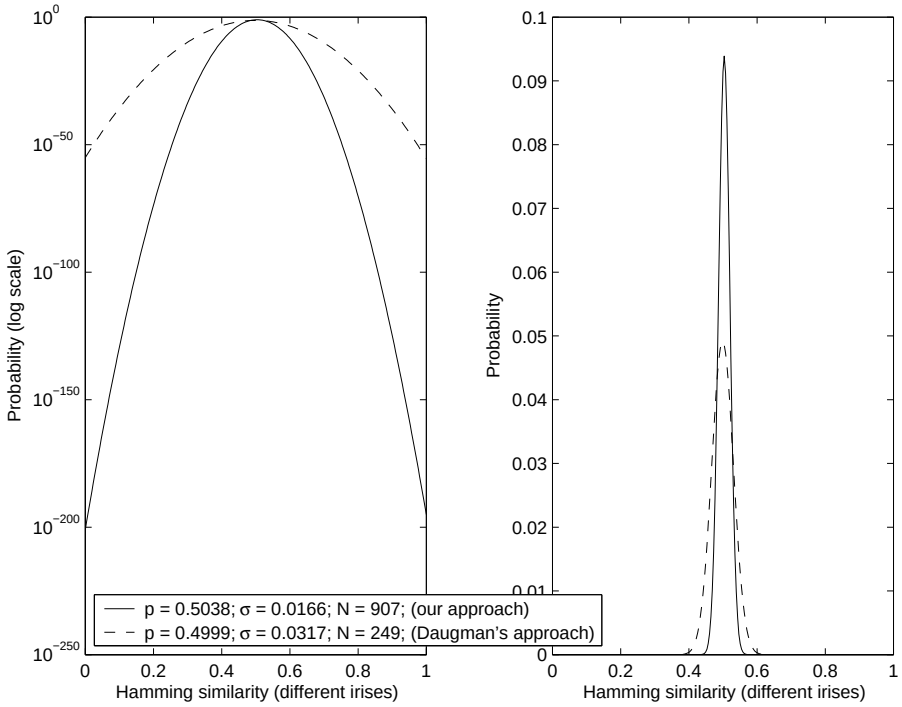


Figure 8: Comparing the distributions of inter-class matching scores

The curve of inter-class Hamming similarity scores in figure 7 matches a binomial distribution with $p = 0.5038$ and $N = 907$ degrees-of-freedom, which is much sharper than the distribution of Hamming distances in Daugman's approach. Consequently, in our approach (based on Fast Fuzzy Iris Segmentation and Gabor Analytic Iris Texture Binary Encoder) a steeper descent of the False Reject Rate is guaranteed (figure 8).

The similitude between theoretical and experimental data is illustrated in figure 7. For a value of 0.58 for the recognition threshold, experimentally observed false accept and false reject rates are $FAR = 0.0$ and $FRR = 0.0397$. Theoretical odds of false accept and false reject are $FAR = 2.1631E - 006$ and $FRR = 0.0497$. All of these results prove the efficiency of the proposed Fast Fuzzy Iris Segmentation procedure.

References

[1] K.W. Bowyer, K. Hollingsworth, P.J. Flynn, *Image Understanding for Iris Biometrics: A survey*, Computer Vision and Image Understanding 110 (2), 281-307, May 2008.

[2] J.G. Daugman, *High Confidence Visual Recognition of Persons by a Test of Statistical Independence*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, November 1993.

[3] , J.G. Daugman, *How Iris Recognition Works*, IEEE Transaction on circuits and Systems for Video Technology, Vol. 14, No. 1, January 2004.

[4] J.G. Daugman, *New Methods in Iris Recognition*, IEEE Transactions on Systems, Man, and Cybernetics - part B: Cybernetics, vol. 37, no. 5, october 2007.

[5] L. Ma, T. Tan, Y. Wang, D. Zhang, *Efficient Iris Recognition by Characterizing Key Local Variations*, IEEE Transactions on Image Processing, vol. 13, no. 6, June 2004.

[6] W. Pegden, *Sets resilient to erosion*, web refference at Department of Mathematics, Rutgers University, June 2008,
http://people.cs.uchicago.edu/˜wes/erosion.pdf

[7] N. Popescu-Bodorin, *Fast Fuzzy Iris Segmentation Demo Program*, http://fmi.spiruharet.ro/bodorin/arch/cffis.zip, June 2009.

[8] N. Popescu-Bodorin, L. State, *Optimal Luminance-Chrominance Downsampling through Fast K-Means Quantization*, Proceedings of the 3rd Annual South East European Doctoral Student Conference, vol. 2, pp. 111-125, ISBN 978-960-89629-7-2, ISSN 1791-3578, SEERC, June 2008.

[9] N. Popescu-Bodorin, *Fast K-Means Image Quantization Algorithm and Its Application to Iris Segmentation*, Scientific Bulletin, No.14/2008, University of Pitesti, ISSN 1453-116x.

[10] D. Salomon, *Data Compression: The Complete Reference*, p. 15, Springer-Verlag, New York, NY, 2006.

[11] L. State, I. Paraschiv-Munteanu, N. Popescu-Bodorin, *Blood corpuscles classification schemes for automated diagnostic of hepatitis*, Scientific Bulletin, No.14/2008, University of Pitesti, ISSN 1453-116x.

[12] L. State, I. Paraschiv-Munteanu, N. Popescu-Bodorin, *Blood corpuscles classification schemes for automated diagnostic of hepatitis using ISODATA algorithm and Run Length Encoding*, Scientific Bulletin, No.16/2009, University of Pitesti, ISSN 1453-116x.

[13] University of Bath Iris Database,
http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/

[14] R. P. Wildes, *Iris Recognition - An Emerging Biometric Technology*, Proceedings of the IEEE, Vol. 85, No. 9, September 1997.

ISSN 1584-5990