siam

4th Annual Meeting of the Bulgarian Section of SIAM
December 21-22, 2009
Sofia

BGSIAM'09

PROCEEDINGS

HOSTED BY THE INSTITUTE OF MATHEMATICS AND INFORMATICS
BULGARIAN ACADEMY OF SCIENCES

PREFACE

The Bulgarian Section of SIAM (BGSIAM) was founded on January 18, 2007 and the accepted Rules of Procedure were officially approved by the SIAM Board of Trustees on July 15, 2007. The activities of BGSIAM follow the general objectives of SIAM, as established in its Certificate of Incorporation.

Being aware of the importance of interdisciplinary collaboration and the role the applied mathematics plays in advancing science and technology in industry, we appreciate the support of SIAM as the major international organization for Industrial and Applied Mathematics in order to promote the application of mathematics to science, engineering and technology in the Republic of Bulgaria.

The 4th Annual Meeting of BGSIAM (BGSIAM'09) was hosted by the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia. It took part on December 21 and 22, 2009. The conference support provided by SIAM is very highly appreciated.

During BGSIAM'09 conference a wide range of problems concerning recent achievements in the field of industrial and applied mathematics were presented and discussed. The meeting provided a forum for exchange of ideas between scientists, who develop and study mathematical methods and algorithms, and researchers, who apply them for solving real life problems.

More than 60 participants from five universities, four institutes of the Bulgarian Academy of Sciences and also from outside the traditional academic departments took part in BGSIAM'09. They represent most of the strongest Bulgarian research groups in the field of industrial and applied mathematics. The involvement of younger researchers was especially encouraged and we are glad to report that 7 from the presented 25 talks were given by Ph.D. students.

LIST OF INVITED LECTURES:

- TODOR GRAMCHEV
  University of Cagliaru, Italy
  REGULARITY AND DECAY ISSUES FOR COMPACTLY SUPPORTED SOLITARY WAVES

- CHRISTO CHRISTOV
  University of Louisiana at Lafayette, Louisiana, USA
  THE CONCEPT OF SPACE AS MECHANICAL CONTINUUM AND ITS IMPLICATIONS IN MATHEMATICAL PHYSICS

- MAYA NEYTCHEVA
  University of Uppsala, Sweden
  CHALLLENGES FOR THE ITERATIVE SOLUTION OF PHASE-SEPARATION FLOW PROBLEMS

- EUGENE NIKOLOV
  National Laboratory on Computer Virology, Bulgarian Academy of Sciences
  COMPUTER NANO VIRUSES

- PANAYOT VASSILEVSKI
  Lawrence Livermore National Laboratory, USA
  APPROXIMATION PROPERTIES OF COARSE FINITE ELEMENT
  SPACES CONSTRUCTED BY ALGEBRAIC MULTIGRID

The present volume contains extended abstracts of the conference talks.

Svetozar Margenov
Chair of BGSIAM Section

Stefka Dimova
Vice-Chair of BGSIAM Section

Angela Slavova
Secretary of BGSIAM Section

Sofia, February 2010

Table of Contents

Acceleration of the Convergence for Interpolated Nonconforming Finite Elements

Andrey Andreev, Milena Racheva

We introduce here nonstandard interpolated finite elements of nonconfor- ming triangular type. These elements provide better accuracy for second-order elliptic boundary value problems. Our aim is to establish a super- convergence result for planar FE approximations by using an interpolated FEM approach (see also [5]). The applied method is a variant of a post- processing procedure when the known finite element solution is used. In our a posteriori error analysis we exploit the ideas known from the FE superconvergence method. Namely, we use local Zienkiewicz-Zhu (ZZ) type estimators [2, 6].

Let Ω be a bounded polygonal domain in $\mathbf{R}^2$ with boundary $\partial\Omega$. Let also $H^m(\Omega)$ be the usual $m-$th order Sobolev space on Ω with a norm $\|\cdot\|_{m,\Omega}$ and seminorm $|\cdot|_{m,\Omega}$. Consider the following second-order model problem in weak form: for $f \in L_2(\Omega)$, find $u \in H_0^1(\Omega)$ such that:

$$a(u,v) = (f,v), \quad \forall v \in V \equiv H_0^1(\Omega), \tag{1}$$

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + a_0 uv) \, dx \, dy \quad \forall u,v \in H_0^1(\Omega)$$

and the notation $(\cdot,\cdot)$ is adopted for the $L_2(\Omega)-$inner product.

Let τ_h be family of regular finite element partitions of Ω which fulfil standard assumptions. The partitions τ_h consist of triangles K and h is mesh parameter. We can define the finite element space V_h by means of linear elements with integral type degrees of freedom which will be introduced.
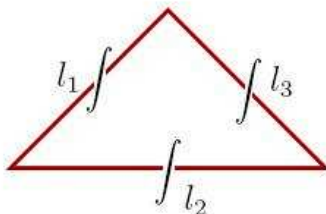

Figure 1:

The corresponding approximate variational problem of (1) is: find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = (f, v_h), \quad \forall \, v_h \in V_h, \tag{2}$$

1

$$a_h(u,v) = \sum_{K \in \tau_h} \int_K (\nabla u \cdot \nabla v + a_0 uv)\, dx\, dy \quad \forall u, v \in V_h.$$

The linear form $(f, \cdot)$ need not be approximated since the inclusion $V_h \subset L_2(\Omega)$ holds. Assume that the mapping $\|v_h\|_{1,h} = \left\{ \sum_{K \in \tau_h} |v_h|^2_{1,K} \right\}^{1/2} \ \forall v_h \in V_h$ is a norm over the space V_h. This space is constructed by means of integral degrees of freedom (fig. 1).

Let $T = \{(t_1, t_2): \ t_1, t_2 \geq 0, \ t_1 + t_2 \leq 1\}$ be the reference element. Then the shape functions of introduced linear element on T are: $\varphi_1(x_1, x_2) = 1 - 2x_1$; $\varphi_2(x_1, x_2) = 1 - 2x_2$; $\varphi_3(x_1, x_2) = -1 + 2x_1 + 2x_2$.

We could prove the following result:

Theorem 1 *Let u and u_h be the solutions of (1) and (2), respectively. If the solution u belongs to space $H^2(\Omega)$, then*

$$\|u - u_h\|_{0,\Omega} \leq Ch^2 |u|_{2,\Omega}. \tag{3}$$

The proof of this theorem is based on the interpolation properties of the conforming and nonconforming linear FE triangles. By comparison of the corresponding interpolants and using standard estimates for the linear case [4] we obtain the estimate (3).

Let $\mathcal{R}_h: \ V \to V_h$ be the elliptic projection operator defined by:

$$a_h(u - \mathcal{R}_h u, v_h) = 0 \quad \forall u \in V, \ \forall v_h \in V_h.$$

It is well-known (see [4]) that the $\mathcal{O}(h)-$convergence is optimal for the linear nonconforming case, i.e. if $u \in H^2 \cap V$, then

$$\|u - \mathcal{R}_h u\|_{1,h} \leq Ch \|u\|_{2,\Omega}. \tag{4}$$

Our aim is to improve the last estimate. For this purpose we construct a nonstandard interpolation operator. So, two principal steps are performed: *(i)* It has to prove a local supercloseness in energy norm by means of special interpolation operator i_h; *(ii)* We construct a higher order interpolation of the original FE solution for global superconvergence to its interpolant $I_{2h}\mathcal{R}_h u$, where I_{2h} verifies certain ultraconvergence properties.

We consider the regular pattern $\tau_h = \cup K$. This partition consists of isosceles right-angled triangles. The operator $i_h: \ C \to V_h$ can be defined using "edges" conditions as degrees of freedom:

$$\forall v \in C(\Omega), \ \forall K \in \tau_h, \quad \int_{l_j} i_h v\, dl = \int_{l_j} v\, dl, \ j = 1, 2, 3.$$

It is evident that $i_h v \in V_h, \ \forall v \in C(\Omega), \ i_h v = v, \ \forall v \in V_h$.

Remark 1 *The basic functions φ_1, φ_2 and φ_3 could be used directly to the finite element implementation and not only to the patch-recovery superconver- gence method.*
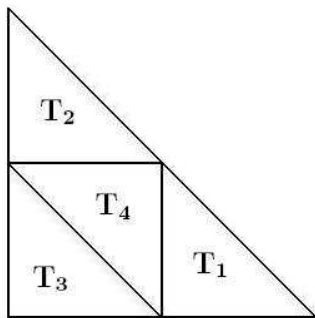
Figure 2:

Now, we can formulate a superclose property concerning i_h.

Theorem 2 *Let $u \in H^3(\Omega)$. Then for any $v_h \in V_h$ the following inequality holds:*

$$a_h(i_h u - u, v_h) \leq C h^2 \|u\|_{3,\Omega} |v_h|_{1,h}.$$

The considered nonconforming element has the following property:

Lemma 1 *If $a_0 = 0$, then the interpolation operator i_h related to the linear noncon-formity triangular element coincides with the projector $\mathcal{R}_h$ of the second order elliptic problem, i.e.*
$$a_h(i_h u - u, v_h) = 0, \quad \forall u \in V, \ \forall v_h \in V_h.$$

Consider an element patch on τ_h containing four adjacent congruent right-angled isosceles triangles (see fig. 2).

So, the patch-recovery FE strategy is fulfilled by coupling every time four triangles as it is designed in fig. 2. This FE partition is denoted by $\widetilde{\tau}_{2h}$. Consequently, the mesh $\widetilde{\tau}_{2h}$ of size $2h$ is obtained as a result of arranging in groups of adjacent elements $K_i \in \tau_h$, $i = 1, 2, 3, 4$. Thus, the nine degrees of freedom are the integral values of a function v on the edges l_j of $K_i \in \tau_h, i = 1, 2, 3, 4$. Then, the interpolation operator is characterized by the edge conditions determined by the degrees of freedom and the basic functions.
Let $\widetilde{V}_{2h}$ be finite element spaces associated with $\widetilde{\tau}_{2h}$. Then $\widetilde{V}_{2h}$ consists of piecewise polynomials from P_3.
I_{2h} is constructed in such a way that:

$$I_{2h} \circ i_h = I_{2h}. \tag{5}$$

3

This operator is also bounded: $I_{2h} : V_h \to \widetilde{V}_{2h}$,

$$\|I_{2h}v_h\|_{r,h} \leq C\|v_h\|_{r,h}, \quad \forall v_h \in V_h, \; r = 0, 1. \tag{6}$$

Finally, taking into account that the interpolation polynomial $I_{2h}v$ is of degree 3, for any $v \in H^4(\Omega) \cap V$ it follows that

$$\|I_{2h}v - v\|_{1,h} \leq Ch^2\|v\|_{4,\Omega}. \tag{7}$$

The next theorem contains the main estimation:

Theorem 3 *Let the solution of (1) belong to $H^4(\Omega) \cap V$. Then the following estimate holds:*

$$\|I_{2h} \circ \mathcal{R}_h u - u\|_{1,h} \leq Ch^2\|u\|_{4,\Omega}. \tag{8}$$

P r o o f. Applying (5), we decompose

$$I_{2h} \circ \mathcal{R}_h u - u = I_{2h} \circ (\mathcal{R}_h u - i_h u) + (I_{2h}u - u).$$

From (6) it follows

$$\|I_{2h} \circ \mathcal{R}_h u - u\|_{1,h} \leq \|I_{2h}\|\|\mathcal{R}_h u - i_h u\|_{1,h} + \|I_{2h}u - u\|_{1,h}).$$

It is easy to see that the first term in the right-hand side in the last inequality vanishes in case $a_0 = 0$, and is of order $\mathcal{O}(h^2)$ otherwise, which completes the proof.

We could apply directly these results to second order elliptic eigenvalue problems. The eigenvalue problem, corresponding to (1) is: find $\lambda \in \mathcal{R}$ and $u \in V$, $u \neq 0$ such that

$$a(u,v) = \lambda(u,v), \quad \forall v \in V. \tag{9}$$

Let us approximate the nonconforming finite element approximation of (9): find $\lambda_h \in \mathcal{R}$ and $u_h \in V_h$, $u_h \neq 0$ such that

$$a_h(u_h, v_h) = \lambda_h(u_h, v_h), \quad \forall v_h \in V_h. \tag{10}$$

Our assumption here is that FE space V_h uses polynomials of degree one. Such being the case, it is well known (see [3]) that the rate of convergence of FE approximation to the eigenvalues and eigenfunctions is given by the following estimations:

$$|\lambda - \lambda_h| \leq C(\lambda)h^2\|u\|_{2,\Omega}^2, \tag{11}$$

$$\|u - u_h\|_{1,h} \leq C(\lambda)h\|u\|_{2,\Omega}. \tag{12}$$

We can improve the estimate (12) concerning the difference $u_h - \mathcal{R}_h u$, namely (see [1]):

Lemma 2 *Let the eigenfunction u belong to $H^4(\Omega) \cap V$ and let u_h be the corresponding FE approximation obtained by (10). Then:*

$$\|u_h - \mathcal{R}_h u\|_{1,h} \leq Ch^2\|u\|_{4,\Omega}.$$

In order to apply the patch-recovery superconvergent procedure to the eigenvalue problem we need an another lemma [5, 1]:

Lemma 3 *Let (λ, u) be any exact eigenpair obtained by (9). Then for every $w \in V$ and $w \neq 0$, the following inequality holds:*

$$|\frac{a(w,w)}{(w,w)} - \lambda| \leq C \frac{\|w - u\|_{1,h}^2}{(w,w)}.$$

The main result concerning patch-recovery technique applied to second order eigenvalue problem and improvement of the estimate (12) by two order (ultraconvergence) is given in the following theorem:

Theorem 4 *Let (λ, u) be an exact eigenpair and (λ_h, u_h) be its FE appro- ximation using triangular nonconforming linear elements. Assume also that the conditions of Theorem 2 are fulfilled. Then*

$$|I_{2h}u_h - u|_{1,h} \leq Ch^2\|u\|_{4,\Omega}, \quad |\frac{a_h(I_{2h}u_h, I_{2h}u_h)}{(I_{2h}u_h, I_{2h}u_h)} - \lambda| \leq Ch^4\|u\|_{4,\Omega}^2.$$

References

[1] A.B. Andreev, Supercloseness between the elliptic projection and the approximate eigenfunction and its application to e postprocessing of finite element eigenvalue problems, LNCS, Springer-Verlag, 3401, 2005, 100-107.

[2] Andreev, A.B., Racheva, M.R.: Superconvergence of the interpolated quadratic finite elements on triangular meshes. Math. Balkanica, New Series. Vol.19 (2005) Fasc. 3-4 385-404

[3] Babuska I., J. Osborn: Eigenvalue Problems., in Handbook of Numerical Analysis, Vol. II, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 641-787.

[4] Ciarlet, P.: Basic Error Estimates for the FEM, Elsevier, Vol. 2 Amsterdam, pp. 17-351 (1991)

[5] Lin, Q., Yan, N., Zhou, A.: A rectangle test for interpolated finite elements. In: Proceedings of Systems Science & Systems Engineering, Culture Publish Co. (1991) 217–229

[6] Zienkiewich, O.C., Zhu, J.Z.: The superconvergence patch-recovery and a posteriori error estimates. Part I: The recovery technique, Int. J. Numer. Methods Eng., **33**, 1992, 1331-1364.

Parallel Quasi-Monte Carlo Quadratures on the Grid

Emanouil Atanassov, Todor Gurov, Mariya Durchova,
Aneta Karaivanova, Sofiya Ivanovska

1 Quasi-Monte Carlo Quadratures

Consider a s dimensional integral $I[f] = \int_{I^s} f(x)dx$, in the unit cube $I^s = [0,1]^s$, of a Lebesgue integrable function $f(x)$, and note that this integral can be expressed as the expectation of the function f, $I[f] = E[f(x)]$, where x is a uniformly distributed vector in the unit cube. Consider the following approximation of the integral:

$$I_N[f] = \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{1}$$

If $\{x_n\}$ is a sequence sampled from uniform distribution, equation (1) *is called Monte Carlo quadrature formula* [12]. The integration error, defined as

$$\epsilon_N[f] = |I[f] - I_N[f]|,$$

has a standard normal distribution, with expectation

$$E[\epsilon_N[f]] = \sqrt{Var(f)}N^{-1/2}.$$

An exact upper bound for (1) is given by Koksma-Hlawka inequality,

$$\epsilon_N[f] \leq V[f]D_N^*, \tag{2}$$

in which V[f] is the variation of f in the Hardy-Krause sense and D_N^* is the star discrepancy of the sequence $\{x_n\}$, [4]. Equation (2) is valid for any function with bounded variation and any choice of sequence, however the best results are obtained with *low discrepancy or quasirandom sequences* for which

$$D_N^* \leq c(logN)^k N^{-1},$$

where c and k are constants depending only on the spatial parameter s. One major advantage of Monte Carlo methods is that they are usually very easy to be parallelized. This is, in principal, also true of quasi-Monte Carlo methods. However, the successful parallel implementation of a quasi-Monte Carlo application depends crucially on various quality aspects of the parallel quasirandom sequences used [5] and [6]. The purpose of using scrambled sequences in quasi-Monte Carlo methods is twofold. Primarily, it provides a practical method to obtain error estimates for QMC based on treating each scrambled sequence as a different and independent random sample from a family of randomly scrambled quasirandom numbers [1, 9]. Thus,

randomized QMC overcomes the main disadvantage of QMC while maintaining the favorable convergence rate of QMC. Secondarily, scrambling gives us a simple and unified way to generate quasirandom numbers for parallel, distributed, and Grid-based computational environments [5, 6, 7].

2   Parallel Implementation on the Grid and Results

2.1   Grid Service for Administration of user reservations

In order to implement efficiently our algorithms on the grid we have developed a new Grid service, called Job Track Service (JTS) for administration of user reservations which allows automatic request and processing. The JTS is a grid middleware component, which facilitates the provision of Quality of Service in grid infrastructures. From the users' point of view the benefit from running a parallel job over a Grid using more CPUs is lost because of the high waiting time, as it was discussed in [10]. On the other side, maintaining and supporting of a medium sized clusters, as ours, meet the challenge to attain more stable load-balancing over all the nodes. This terms, the placement of hard-integrated reservations for the most frequently reserved resources or users is not a proper way for blocking the nodes entirely and only for them. The implemented in most reliable Bulgarian sites JTS does not need a great number of additional requirements than the other already established grid middleware components. It has the advantage to make a reservation for the grid resources in accordance with the users' needs. They are prompted to enter *the site name, number of CPUs and reservation period* and during this period the service check if the requested resources are free and ready for accepting the job. If the request is granted following the special policy a replay is send to the user. This way, decreasing the reservation period leads to spending the waste of time for waiting the job on the queue. This technique has significant advantages especially when the MPI job consists of a huge number of short duration processes, because in fact the waiting time is equal to zero. More over, we prevent the cluster from the hard-integrated reservations, achieving more balanced and time-spending process.

2.2   Parallel quasirandom Generation

There are known several ways of parallelization of a quasirandom sequence but most used of them are *leap-frog, splitting or blocking, independent generation and scrambling.* The first two schemes produce a single quasirandom sequence until the third scheme needs a family of sequences, which could be produced using scrambling techniques, achieving a stochastic family from only one quasirandom sequence. Much of the recent work on parallelizing quasi-Monte Carlo methods has been aimed at splitting a quasirandom sequence into many subsequences which are then used independently on various parallel processes. In this paper we use scrambled Sobol sequence (see [1] for details).

2.3 Numerical Tests

The numerical tests are performed on two Clusters using MPI. The BG03-NGCC grid cluster has 25 worker nodes, 2 x Intel Xeon E5430 2.66 GHz Quad Core CPU (total 200 Cores, > 400 kSI2000) with 16GB RAM per node. The BG04-ACAD has 40 worker nodes with 2 CPU Opteron 2,4 GHz (total 80 cores > 120 KSI2000), 4GB RAM per node, and low-latency Myrinet interconnection for MPI jobs as in [15]. Both clusters use the torque batch system.

Here we present numerical results when computing 100-dimensional integral with the following test function

$$F = \prod_{i=1}^{s=100} \left(x_i{}^3 + \frac{3}{4} \right).$$

This test function is often used as a benchmarking of quadrature algorithms, [11]. First, we present timing results for the two parallel generators - a parallel generator from SPRNG library [16] and ourgenerator of scrambled Sobol sequence [2].

Table 1: Time for test function F with 10^8 points.

	1 CPU BG03-NGCC	1 CPU BG04-ACAD
SPRNG	785.64	533.90
Sobol	65.79	95.65

Obviously, timing results for Sobol sequence are much better (more than 10 times faster generation).

On the next Table 1 we present timing results for parallel solving of 100-dimensional integral - such dimension is typical for financial applications. We use simultaneously

Table 2: Parallel efficiency measurements for $N = 10^9$ points - T_{opt} is the theoretically optimal time and T_m is the measured time.

	T_{opt}	T_m	Efficiency
SPRNG	117.10	129.87	90.17%
Sobol	14.38	15.34	93.74%

20 CPU core on BG04-ACAD and *32 CPU core on BG03-NGCC.*
On Table 2 we show the theoretically estimated time for running MPI jobs on 52 CPU cores of the grid sites BG03-NGCC and BG04-ACAD in parallel and the measured execution time. Here T_{opt} is theoretically optimal time, T_m is measured time and the

efficiency is the ratio $\dfrac{T_{opt}}{T_m}$. We define *theoretically optimal time* as

$$T_{opt} = \frac{1}{\dfrac{c_1}{t_1} + \dfrac{c_2}{t_2}},$$

where t_1 is the measured time for 1 CPU on BG03-NGCC, t_2 is the measured time for 1 CPU on BG04-ACAD, c_1 is the number of CPU cores on BG03-NGCC, and c_2 is the number of CPU cores on BG04-ACAD.

We used the MPICH-G2 implementation of MPI, which is the most general approach for running parallel jobs on computational grids. The necessary reservation was made using JTS in order to ensure simultaneous start-up of the MPI jobs. In this way we successfully utilized machines with different endianness in the same computation.

3   Concluding Remarks

The computing of very high-dimensional integrals (s=100) poses technical challenges on heterogeneous Grid. Using proper strategy for parallelization and JTS for reservation, we succeeded to obtain excellent parallel efficiency. The results show that quasi-Monte Garlo approach outperforms Monte Carlo for the considered problem.

Acknowledgments

References

[1] E. Atanassov, *A New Efficient Algorithm for Generating the Scrambled Sobol' Sequence*, NM&A, Springer, LNCS 2542, pp. 83-90, 2003. E. Atanassov, T. Gurov, D. Slavov *Job Track Service: Architecture and Features*, User Forum-2009 Proceeding, pp. 199-202, 2009.

[2] E. Atanassov, A. Karaivanova, T. Gurov, S. Ivanovska and M. Durchova, Using Sobol Sequence in Grid Environment, Proceeding of 32nd International Convention MIPRO/GVS 2009, pp. 290-294, 2009.

[3] I. Bird, B. Jones, K. Kee, *The organization and management of grid infrastructures*, IEEE Computer, vol. 42(1), pp. 36-46, 2009.

[4] R. Caflisch, *Monte Carlo and quasi-Monte Carlo methods*, Acta Numerica, vol. 7, pp. 1-49, 1998.

[5] H. Chi, E. Jones, *Generating Parallel Quasirandom Sequences by using Randomization*, Journal of distributed and parallel computing, vol. 67(7), pp.876-881, 2007.

[6] H. Chi and M. Mascagni, *Efficient Generation of Parallel Quasirandom Sequences via Scrambling*, Springer, LNCS 4487, pp. 723-730, 2007. I. Foster,*What is the Grid? A Three Point Checklist*, https://www.mcs.anl.gov/ itf/Articles/WhatIsTheGrid.pdf

[7] I. Foster, C. Kesselmann, *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1998.

[8] S. Ivanovska, E. Atanassov, A. Karaivanova, *A Superconvergent Monte Carlo Method for Multiple Integrals on the Grid*, Springer, LNCS 3516, pp. 735-742, 2005.

[9] G. Okten, B. Tuffin and V. Burago, *A central limit theorem and improved error bounds for a hybrid-Monte Carlo sequence with applications in computational Finance*, Journal of Complexity, vol. 22(4), pp. 435-458, 2006.

[10] Tz. Ostromsky and Z. Zlatev, *Parallel and GRID implementation of a large scale air pollution model*, Proc. NM&A'06, LNCS 4310, pp. 475-482, 2007.

[11] W. Schmid, A. Uhl, *Techniques for parallel quasi-Monte Carlo integration with digital sequences and associated problems*, Math. and Comp. in Sim., vol. 55, pp. 249-257, 2001.

[12] I. M. Sobol, *Monte Carlo Numerical Methods*, Nauka, Moscow, 1973.

[13] http://goc.grid.sinica.edu.tw/gstat, Monitoring EGEE Grid Infrastructure.

[14] http://goc.grid.sinica.edu.tw/gstat/seegrid, Monitoring See-Grid Infrastructure.

[15] MPI website, http://www-unix.mcs.anl.gov/mpi/

[16] SPRNG: Scalable Parallel Random Number Generator, http://sprng.cs.fsu.edu/

How Does Clinical Data Fit into Two Leukopoiesis Models?

Gergana Bencheva

Motivation. Blood cells (BCs) perform various vital functions like transporting oxygen to tissues, fighting infections and controlling bleeding. Haematopoiesis is a complex biological process, during which haematopoietic pluripotent stem cells (HSCs) in bone marrow give birth to the three blood cell types. The process of production and regulation of white BCs like T- and B-lymphocytes, NK cells and their subpopulations, is known as *leukopoiesis*. Each type of BCs is a result of the action of specific proteins, called *growth factors* or *Colony Stimulating Factors* (CSF), at specific moments during haematopoiesis. The main growth factors in leukopoiesis are Granulocyte-CSF (G-CSF), Macrophage-CSF (M-CSF), Granulocyte-Macrophage-CSF (GM-CSF) and Interleukins (IL). Those BCs that have not yet matured are called blast cells.

Various haematological diseases (including leukemia) are characterized by abnormal production of particular blood cells (matured or blast). For example, acute leukemia occurs when the blast cells replicate themselves uncontrollably and interfere with the production and activity of normal BCs. If undiscovered or left untreated, it can cause death within few weeks or months.

There are two main stages in the therapy of pathological blood diseases. The first one is total body irradiation (TBI) and chemotherapy, where the "tumour" cells are killed together with (as side effect) the healthy ones. The second stage is bone marrow transplantation (BMT), where HSCs of a donor (collected under special conditions) are put in the peripheral blood. From here HSCs have to: 1) find their way to the stem cell niche in the bone marrow; and afterwards 2) selfrenew and differentiate to regenerate the patient's blood system. The faster the steps 1) and 2) are performed, the shorter is the period in which the patient is missing their effective immune system. The aim in the current work is to compare clinical data and computer simulations with two leukopoiesis models.

Two models of leukopoiesis. The models of blood cells production and regulation, proposed in [1, 2] are used in our investigations. Each of them is a system of ordinary differential equations (ODEs) with delays. Their parameters depend on the particular type of BCs. The first model (see [1]), referred here as GFM, takes into account the CSF action and can be applied to any of the red BCs, white BCs, platelets, or their subpopulations. In the second model (see [2]), designed especially for leukopoiesis and denoted bellow by LM, the intermediate differentiation stages as well as the influence of the CSF are included in the parameter for the so called amplification phase. The data involved in GFM and LM is schematically presented in Figure 1 and has the following meaning. In the bone marrow, HSCs are divided into two groups: proliferating cells and nonproliferating (or quiescent) cells. Their populations at time t for both GFM and LM cases are denoted by $P(t) \geq 0$ and $Q(t) \geq 0$ respectively. The population of the circulating mature BCs is denoted by $M(t) \geq 0$ ($W(t) \geq 0$ for LM) and the growth factor concentration is $E(t) \geq 0$. The proliferating phase duration is denoted by τ (in GFM) or τ_1 (in LM) and it is assumed to be constant. The
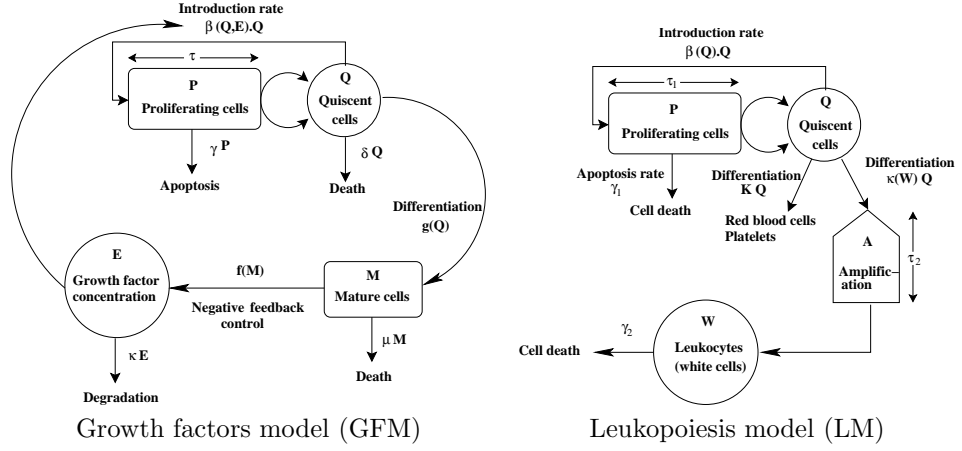
Figure 1: Involved data in GFM and LM models.

amplification phase is characterized by its duration τ_2 (also assumed to be constant) and by the parameter $A = \alpha 2^i$, where $\alpha \in (0,1)$ is the survival rate and i is the number of needed generations (differentiation stages). The apoptosis rate for the proliferating cells is denoted by γ (in GFM) or γ_1 (in LM). The rates at which the quiescent HSCs, the matured cells and the CSF can die are denoted by δ (for Q), μ (for M), γ_2 (for W) and k (for E). The degradation rates μ, γ_2 and k of the mature BCs and of the CSF in the blood are assumed to be positive. Quiescent cells in GFM can either be introduced in the proliferating phase with a rate $\beta(Q(t), E(t))$ or differentiate in mature BCs with a rate $g(Q(t))$. The negative feedback control $f(M(t))$ of the bone marrow production on the CSF production acts by the mean of circulating mature BCs: the more circulating BCs are, the less growth factor is produced. In LM the quiescent cells have three possibilities: to be introduced in proliferating phase with rate $\beta(Q)$, to differentiate in red BCs or platelets with rate K or to differentiate in white BCs with rate $k(W)$. Their degradation rate is included in K.

The GFM is represented by a system of nonlinear ODEs with one constant delay τ:

$$
\begin{cases}
\dfrac{dQ}{dt} &= -\delta Q(t) - g(Q(t)) - \beta(Q(t), E(t))\, Q(t) \\
& \quad + 2e^{-\gamma\tau}\beta(Q(t-\tau), E(t-\tau))\, Q(t-\tau) \\
\dfrac{dM}{dt} &= -\mu M(t) + g(Q(t)) \\
\dfrac{dE}{dt} &= -kE(t) + f(M(t))
\end{cases}
\tag{1}
$$

and initial conditions $Q(t) = Q_0(t)$, $M(t) = M_0(t)$, $E(t) = E_0(t)$ for $t \in [-\tau, 0]$.

The LM is represented by a nonlinear ODE system with two constant delays τ_1, τ_2:

$$
\begin{cases}
\dfrac{dQ}{dt} &= -[K + k(W(t)) + \beta(Q(t))]Q(t) \\
&\quad + 2e^{-\gamma_1\tau_1}\beta(Q(t-\tau_1))Q(t-\tau_1) \\
\dfrac{dW}{dt} &= -\gamma_2 W(t) + Ak(W(t-\tau_2))Q(t-\tau_2)
\end{cases}
\tag{2}
$$

and initial conditions $Q(t) = Q_0(t)$, $W(t) = W_0(t)$, $t \in [-\tau^*, 0]$, for $\tau^* = max\{\tau_1, \tau_2\}$. The trivial steady-state of (1) and (2) is not a biologically interesting equilibrium since it describes a pathological situation that can only lead to death without appropriate treatment. The existence of nontrivial positive steady-state of (1) is ensured by

$$
0 < \delta + g'(0) < \beta\left(0, \frac{f(0)}{k}\right), \ 0 \le \tau < \tau_{max} := \frac{1}{\gamma}\ln\left(\frac{2\beta\left(0, \frac{f(0)}{k}\right)}{\delta + g'(0) + \beta\left(0, \frac{f(0)}{k}\right)}\right).
$$

Nontrivial positive steady-state of (2) exists if $(2^{-\gamma_1\tau_1} - 1)\beta(0) > k(0) + K$ and the function $Q \mapsto Q\beta(Q)$ is decreasing in (Q_0, Q_1), where $Q_0 = \beta^{-1}\left(\dfrac{k(0) + K}{2^{-\gamma_1\tau_1} - 1}\right)$ and $Q_1 = \beta^{-1}\left(\dfrac{K}{2^{-\gamma_1\tau_1} - 1}\right)$.

Solution methods. Various numerical methods and software tools for solution of ODE systems are developed for both stiff and nonstiff problems, as well as for the case of retarded arguments (see e.g. [7, 3] and references therein). *XPPAUT* is "a tool for simulating, animating and analyzing dynamical systems", created by G. B. Ermentrout and freely available through the webpage [6].

The behaviour of the XPPAUT implementations of six widely used ODE solvers is compared in [4] with respect to both computing time and quality of the solution. Three of them are recommended as a toolbox for reliable computer simulations, namely: a) classical Runge-Kutta (RK) – four stage explicit method with fixed step that gives $\mathcal{O}(h^4)$ accuracy; b) Dormand-Prince 5 (DP5) – explicit method based on Runge-Kutta method of order 5 with adaptive step and automatic step size control; c) Rosenbrock (RB2) – implicit method with adaptive step, based on Matlab version of the two stage Rosenbrock algorithms. RB2 is the only one of them appropriate for stiff problems. Retarded differential equations are solved by storing previous data and using cubic polynomial interpolation to obtain the delayed value. More details about the features and the usage of XPPAUT may be found in [5] and in the documentation distributed together with the package.

The clinical data is provided by Dr. M. Guenova and Dr. L. Gartcheva from Laboratory of Haematopathology and Immunology, National Hospital for Active Treatment of Haematological Diseases (NHATHD), Bulgaria. It includes gathered amount of HSCs (CD34+) and measured values for the number of T, B and NK cells and their subpopulations. The data is collected before BMT and 1, 2, 3, 6, 9, 12, 18, 24 months

after BMT for a set of patients with Morbus Hodgkin (MH), Non-Hodgkin's Lymphoma (NHL), Multiple Myeloma (MM), and Acute Myelogeneous Leukemia (AML). To have an idea of the normal amount of cells in the monitored populations, their counts are measured in a control group of 21 healthy individuals and the average, p5 and p95 values are computed from them. The normal amount of particular BCs in a healthy human is considered to be in the range p5-p95. As an illustration of the provided data, lymphocytes and T cells counts of two patients (P1 and P2) with AML after BMT are presented in Figure 2 together with the related values for the healthy controls.

a) Lymphocytes (CD 45 AC)   b) T cells (CD 3 AC)
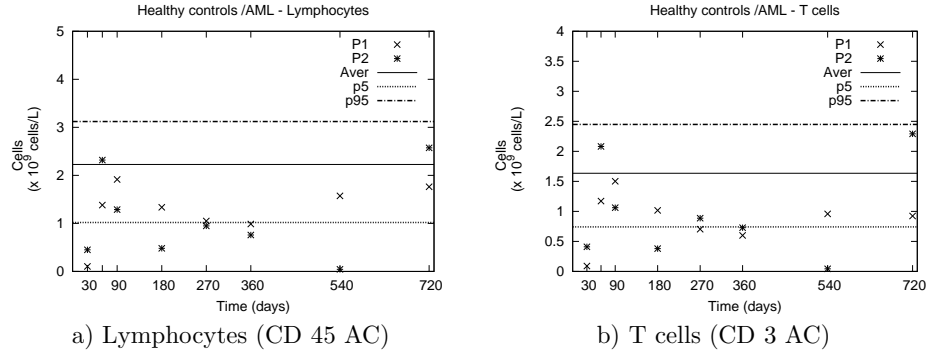
Figure 2: Cell counts for healthy controls and for two patients with AML after BMT.

Medicament like NEUPOGEN – Filgrastim or GRANOCYTE – Lenograstim, containing G-CSF (the main growth factor for white BCs) is applied every day during the first month after BMT. Gathered amount of HSC (CD34+) serves as initial value for Q. Minimal required amount is 2×10^6 cells/kg of body weight and the optimal one is considered to be 5×10^6 cells/kg of body weight. After BMT it is assumed that there is no blood system, i.e. initial values for matured cells are almost equal to 0.

The numerical tests in the current work are performed with each of the methods RK, DP5, RB2. The observed behaviour of the solution does not depend on the solver. First set of experiments are performed with model data given in [1] and [2]. The following functions are used with parameter values, given in Table 1:

$$
\begin{array}{ll}
\text{GFM} & \text{LM} \\
\beta(E) = \beta_0 \dfrac{E}{1+E}, \quad \beta_0 > 0 & \beta(Q) = \dfrac{\beta_0}{1+Q^n}, \quad \beta_0 > 0 \\
g(Q) = GQ, \quad G > 0 & k(W) = \dfrac{k_0}{1+W^m}, \quad k_0 > 0 \\
f(M) = \dfrac{a}{1+KM^r}, \quad a, K > 0, r > 0 & A = \alpha 2^i, \quad \alpha \in (0,1)
\end{array}
$$

For the rest of the tests, some of the functions and parameter values are changed, namely $f(M)$, δ, γ, μ, k, γ_1, γ_2 and A, as well as the values of the delays. The initial conditions are taken for the two patients P1 and P2 with AML. The considered

14

Table 1: Model parameters.

GFM			LM	
Param	Value	Range (day^{-1})	Param	Value
δ	$0.01\ day^{-1}$	$0 - 0.09$	β_0	$1.77\ day^{-1}$
G	$0.04\ day^{-1}$	$0 - 0.09$	k_0	$0.1\ day^{-1}$
β_0	$0.5\ day^{-1}$	$0.08 - 2.24$	n	3
γ	$0.2\ day^{-1}$	$0 - 0.9$	m	2
μ	$0.02\ day^{-1}$	$0.001 - 0.1$	γ_1	$0.1\ day^{-1}$
k	$2.8\ day^{-1}$	—	γ_2	$2.4\ day^{-1}$
a	6570	—	K	$0.02\ day^{-1}$
K	0.0382	—	A	20
r	7	—		

populations are lymphocytes (denoted with W) and the subpopulation T_n of the so called naïve T cells. Degradation rate of T_n is taken from [8], and the one for G-CSF – from the specification of the applied medicaments.

For all the tests with GFM, the model either crashed, or after some pick of the cell population at the beginning, it reached the trivial steady state. The results for the LM model parameters with initial data for each of the patients P1 and P2 are presented in Figure 3 a) and b) respectively. They show that no matter what is the initial condition, i.e. what amount of HSCs is gathered, one and the same steady state is reached. Similar situation is observed with the tests with different values of τ_1 and τ_2. The influence of the amplification parameter A is illustrated with plots c) and d) in Figure 3. For both values of γ_2, i.e. for both lymphocytes and T_n cells, the populations are stabilized at different day and to different cell count.

Concluding remarks. The presented results from numerical experiments pose questions like "Why does the GFM model "crash" with the clinical data?", "How to fix this problem?" and "Which parameters/functions should be changed and how in order to have steady states of LM closer to the clinical data?". Further steps towards the answer of such questions include: more tests with the available clinical data for various diseases and white BCs populations; detailed analysis of the obtained numerical results; sensitivity analysis and parameter estimation of both models with respect to each of the white blood cells and their subpopulations.

References

[1] M. Adimy, F. Crauste, S. Ruan, Modelling Hematopoiesis Mediated by Growth Factors with Applications to Periodic Hematological Diseases, Bulletin of Mathematical Biology, 68 (8), (2006), 2321–2351.

a) P1, $Q(0) = 4.32(\times 10^6)$ cells/kg, $W(0) = 0$ b) P2, $Q(0) = 1.69(\times 10^6)$ cells/kg, $W(0) = 0$

c) P1, W cells, $\gamma_2 = 2.4$, $A = 10, 20$ d) P1, T_n cells, $\gamma_2 = 0.04$, $A = 10, 20$
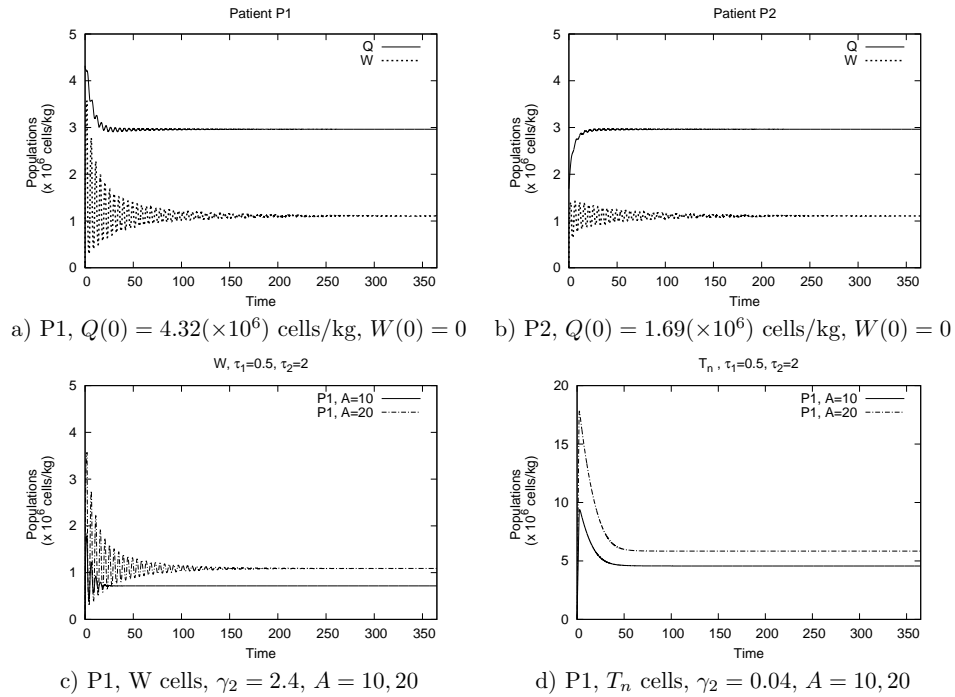
Figure 3: Results with model data from paper [2] for LM and initial data from clinical practice: a) and b) populations Q and W for P1 and P2 respectively; c) and d) populations W and T_n for P1, varying A.

[2] Adimy, M., Crauste, F., Ruan, S.: Periodic oscillations in leukopoiesis models with two delays, Journal of Theoretical Biology 242, 288–299 (2006)

[3] A. Bellen, N. Guglielmi, S. Maset, Numerical methods for delay models in biomathematics, In: A. Quarteroni, L. Formaggia, A. Veneziani (Eds.) Complex Systems in Biomedicine, Springer-Verlag Italia, Milano 2006, 147–185.

[4] G. Bencheva, Comparative Analysis of Solution Methods for Delay Differential Equations in Hematology, In: I. Lirkov, S. Margenov, J. Waśniewski (Eds.) LSSC 2009, Springer LNCS 5910, to appear.

[5] Ermentrout, B.: Simulating, analyzing and animating dynamical systems: a guide to XPPAUT for researchers and students. SIAM (2002)

[6] http://www.math.pitt.edu/~bard/xpp/xpp.html

[7] E. Hairer, (S.P. Norsett), G. Wanner, Solving ordinary differential equations I, II, Springer Ser. in Comp. Math., Springer, 2000 (part I), 2002 (part II)

[8] Moore, H., Li, N.K.: A mathematical model for chronic myelogeneous leukemia (CML) and T cell intercation, Journal of Theoretical Biology 227, 513–323 (2004)

The Concept of Space as Mechanical Continuum and its Implications in Mathematical Physics

Christo I. Christov

Introduction. Is there an Absolute Continuum? The notion that there if there are waves, some material substrate must be 'waving' led 19th century scientists to introduce the concept of the luminiferous continuum (see [1]). Cauchy postulated around 1827 (see the account in [2]) that the luminiferous medium is an elastic continuum, through which light propagates as an elastic shear wave. His model contradicted the natural perception of a particle moving *through* the medium and did not receive much attention, possibly because the notion of elastic liquids was not available at that time. The fluid model for the luminiferous medium was initiated by Lord Kelvin in the first half of the Nineteen Century (see [3]).

The contributions of Oersted, Ampere, and Faraday led to the formulation of the model of electromagnetism. The crucial advance was achieved, however, when Maxwell [4] added the term $\frac{\partial \boldsymbol{E}}{\partial t}$ in Ampere's law, which he termed the 'displacement current'. This new term was similar to the time-derivative term in Maxwell's constitutive relation for elastic gases [4]. One can say that Maxwell postulated an elastic constitutive relation by adding the displacement current. Indeed, the new term added by Maxwell transformed the system of equations, already established in electrostatics, into a hyperbolic system with a characteristic speed of wave propagation.

The Lorentz Transformation (LT) hinted at some superficial symmetry between two inertially moving frames making them equivalent and/or indistinguishable. This led to the 'Relativity principle' according to which the inertial motion of a frame cannot be detected by measurements in the same frame (see the account in [5]). For a system of discrete material points in an empty geometric space, such an assertion may be true (Galileo's RP), but not for the case of continuous frames. Because light *propagates* in the absolute continuum, while a frame is moving with respect to the latter, one should be able to measure the Doppler effect of electromagnetic waves that are emitted from a source that is at rest with respect to the absolute continuum. This is exactly what was reported in the experiments on the so called Local Standard of Rest [6, 7] (LSR). As elucidated by Brillouin [8], the relativity theory is fraught with internal logical inconsistencies and riddled by paradoxes. The paradoxes underscore the fact that while LT and Minkowski space are legitimate mathematical constructs, the relativity principle is still not justified.

In the recent years, the present author proposed a new concept based on the mechanical properties of space (the absolute continuum). We discuss here the material invariance (*frame indifference*) as the true covariance in physics .

Elastic Liquids and Maxwell's Equations *in vacuo* In terms of the velocity vector, the Cauchy balance has the form

$$\rho\frac{D\boldsymbol{v}}{Dt} \overset{\text{def}}{=} \mu\frac{\partial\boldsymbol{v}}{\partial t} + \mu\boldsymbol{v}\cdot\nabla\boldsymbol{v} = \text{div}\sigma = -\nabla\phi - \boldsymbol{E}, \tag{1}$$

where D/Dt stands for the material derivative (called 'convective' or 'total' derivative in the current configuration), $\boldsymbol{v} = \boldsymbol{u}_t$ is the velocity, $\boldsymbol{u}$ is the displacement, and μ is the density of the medium. Respectively we define $\boldsymbol{E} := -div\sigma$, as the negative stress vector (divergence of the stress tensor). Define also the vorticity of the continuum as $\boldsymbol{H} := \nabla\times\boldsymbol{v}$ and $\boldsymbol{B} := \mu\boldsymbol{H}$. Taking the *curl* of Eq. (1), and using the above definitions, we get (see [9, 10, 11]):

$$\nabla\times[\boldsymbol{E} - \boldsymbol{v}\times\boldsymbol{B}] = -\frac{\partial\boldsymbol{B}}{\partial t}, \tag{2}$$

which is the Faraday law with an additional force known as the 'Lorentz force'. In other words, the material invariant Faraday's law automatically accounts for a physical mechanism that can cause the Lorentz force. This is a very important result, because it tells us that the latter is not an additional, empirically observed force that has to be grafted on the Maxwell model in order to complete the electrodynamics, but is connected to the material time derivative, namely to its convective part.

The linear constitutive relationship for an elastic body reads

$$\sigma = (\lambda + \eta)(\text{div}\boldsymbol{u}) + \frac{\eta}{2}(\text{grad }\boldsymbol{u} + \text{grad }\boldsymbol{u}^T), \quad \text{or} \quad \boldsymbol{E}\overset{\text{def}}{=} -\nabla\cdot\sigma = \eta\nabla\times\nabla\times\boldsymbol{u}, \tag{3}$$

where λ and η are the Lamé coefficients (see, e.g., [12, 13]). At this stage we consider a completely incompressible continuum ($\lambda\to\infty$ for which the unknown 'pressure' ϕ in Eq. (1) is am implicit function defined by the requirement $\nabla\cdot\boldsymbol{u} = const$ or $\nabla\cdot\boldsymbol{v} = 0$. Taking the time derivative of Eq. $(3)_2$, one obtains the second of Maxwell's equations (see [9, 10, 11]). The interpretation of the shear elastic modulus is that its inverse defines the electric permittivity.

The cited derivations unequivocally demonstrate that the 'field' described by Maxwell's equations is equivalent to an elastic material space (the metacontinuum). If this is an elastic solid, no *stationary* magnetic fields can exist, since no steady velocities are possible for a solid continuum without discontinuities. In [11] the next decisive step was made by considering elastic liquid *in lieu* of an elastic solid, (see [11]):

$$\frac{\mathfrak{d}\sigma}{\mathfrak{d}t} = \tfrac{1}{2}\eta(\nabla\boldsymbol{v} + \nabla\boldsymbol{v}^T), \quad\Rightarrow\quad \frac{\mathfrak{d}\boldsymbol{E}}{\mathfrak{d}t} = \eta\nabla\times\nabla\times\boldsymbol{v} \quad\text{or}\quad \tau\boldsymbol{E}_t = \zeta\nabla\times\nabla\times\boldsymbol{v}, \tag{4}$$

where $\mathfrak{d}/\mathfrak{d}t$ must be an objective time rate that is frame indifferent. Respectively, ζ is called 'elastic viscosity' [14].

The concept of frame indifference (general covariance of the system) requires that the partial derivative of the stress variable (in our case the stress vector $\boldsymbol{t} = -\boldsymbol{E}$) is

replaced by the appropriate invariant rate. It is argued in [15, 10] that the pertinent invariant rate is the so-called Oldroyd's upper-convected derivative, namely

$$\frac{\eth \boldsymbol{E}}{\eth t} \overset{\text{def}}{=} \frac{\partial \boldsymbol{E}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{E} - \boldsymbol{E} \cdot \nabla \boldsymbol{v} + (\nabla \cdot \boldsymbol{v})\boldsymbol{E}, \tag{5}$$

Upon introducing Eq. (5) in the constitutive law, and after applying a well known vector identity (see, e.g. [16, pg.180]), one gets

$$\boldsymbol{E}_t - \underbrace{\nabla \times (\boldsymbol{v} \times \boldsymbol{E})}_{} + \underline{s\boldsymbol{E}} = -\boldsymbol{v}\widehat{(\nabla \cdot \boldsymbol{E})} + \underbrace{c^2 \widehat{\nabla \times \boldsymbol{B}}}_{}. \tag{6}$$

The terms with the underbraces in Eq. (6) are related to the Biot-Savart law, in the sense that they can be obtained after the operation *curl* is applied to the latter. The sign of the term on the left hand side has a sign opposite to the ubiquitous Biot-Savart law. Similarly to the above outlined argument for the Lorentz force, the sign is different because it refers to the electric field created at a point of the metacontinuum by the presence of a magnetic field.

The important conclusion form the frame-indifferent formulation of the displacement current is that similarly to Lorentz-force law, the convective/ convected terms are related to phenomena that is embodied in Ampere's and Biot–Savart's laws, providing thus their possible unification with Maxwell's model. All three electromagnetic-force laws are manifestations of the inertial forces in the metacontinuum.

Now, Eqs. (2),(6) form the system that is to replace the first two of Maxwell's equations (the dynamic equations). Following, the tradition, one can add also the obvious corollary of the fact that the magnetic field is the *curl* of another vector (velocity), namely, $\nabla \cdot \boldsymbol{B} = 0$. The above derived system *rigorously* fulfills the requirements for 'General Covariance,' because this system is frame-indifferent, i.e., it is invariant when changing to another coordinate frame that can accelerate and can even deform. A very limiting case of the frame indifference is the Galilean invariance. The above described model accomplishes the goal of Cauchy who attempted to explain light waves as shear waves of a material medium

The Mechanical Space in Higher Dimensions Recall that the Schrödinger equation of wave mechanics

$$\boldsymbol{i}\psi_{\partial t} + \hbar\Delta\psi - \chi\psi = 0, \quad \boldsymbol{i} \equiv \sqrt{-1}, \quad \psi \equiv \psi_1 + \boldsymbol{i}\psi_2, \tag{7}$$

where $\hbar$ is the reduced Planck constant and χ is construed to be connected to the potential of any external forces. This equation is not one of Maxwell's equations. As argued by Hinton [17], the thickness of the material world in the direction of the 4th dimension is so minute that it cannot be appreciated directly. Rather, it shows up through additional forces and/or variables. The question that arises here is the following: can one come up with a mechanical construct modeled by a fourth-order (generally nonlinear) dispersive equation that can provide the mechanical interpretation for Schrödinger's equation? The hint is in the original paper of Schrödinger [18]

19

who observed that, in 2D, the equation for the real part of his wave function is the governing equation of a momentum-supporting material surface (an elastic plate or shell).

We begin the demonstration of this idea by rewritng the Schrödinger equation in terms of the real (or imaginary) part of wave function to obtain

$$\psi_{1tt} + \hbar^2 \Delta\Delta\psi_1 - 2\chi\hbar\Delta\psi_1 + \chi^2\psi_1 = 0. \tag{8}$$

For $\chi = 0$, Eq. (8) is nothing else but the Euler–Bernoulli equation for the flexural deformations of thin plates (as mentioned in [18]). Hence, we conjecture that the *metacontinuum* is a thin 3D material layer in 4D space.

Consider a very thin *momentum supporting* elastic structure of a 4D material. A 3D hypershell is the mathematical abstraction for this kind of momentum supporting material structure. The displacement, ζ, along the 4th dimension is the wave function. Contrary to the shell theory of technological applications, one has to consider the limiting case when the deflections are small, the strains (gradients) are of unit-order, and curvatures are large. Such an object is geometrically strongly nonlinear. The *hypershell* of this work is radically different form the notion of a *superstring* because the former can support momenta (the fourth order derivatives) while the latter cannot.

If h is the thickness of the shell, and L is the length scale of the localized deformation, we assume than $h \ll L \ll 1$. Then for the deflection, ζ, along the fourth dimension, the following governing equation can be derived [19] when the flexural deformations are not coupled to the deformations in the middle surface, namely

$$\zeta_{tt} = \beta\left[-\Delta\Delta\zeta - (\Delta\zeta)^3\right] + \Delta\zeta + \mathcal{F}. \tag{9}$$

The dimensionless form is obtained (note that $c_f \neq c$ and hats denote the original variables) by using the scales:

$$\hat{\zeta} = L\zeta, \quad \hat{\boldsymbol{x}} = L\boldsymbol{x}, \quad \hat{t} = Lc_f^{-1}t, \quad c_f^2 = |\sigma|/\mu, \quad \beta = G/(|\sigma|L^2), \tag{10}$$

where G is the stiffness of the shell, σ is the membrane tension, and $\mathcal{F}$ is a 4D body force. Note that the linear part of (9) has the form Eq. (8) if the membrane tension is understood to be related to the heuristic potential.

Particles and Charges as Localized Patterns   In order to understand the interplay between the absoluty of the metacontinuum and the relativity of the rectilinear motion, as discussed by Einstein [20, Ch.7], we stipulate that the particles and charges are localized phase patterns in the material space. Such localized waves are called quasi-particles (or solitons under some additional conditions on integrability). The charges are shear dislocations (see the details in [10, 11]), while the neutral particles are the flexural deformation waves on the surface of hypershell.

Boussinesq [21] derived first an equation containing nonlinearity and dispersion similar to Eq. (9), coming up with a fundamental idea: the dispersion balances the

nonlinearity making the shape of the wave permanent, He found the first solitary wave. It was discovered numerically in [22] that the solitary wave interact very much as particles and were called *solitons* and *quasi-particles* (QP).

This hints at the idea that the QPs of the master equation for the wave function are actually the particles. In other words, a "point particle" is our perception of a localized wave with a very short support and the center of a localized flexural deformation of the shell is perceived as a point particle. We can call with proper justification this kind of a localized wave the *flexons* [23, 24]. Respectively, the equations of electrodynamics admit torsional solitons 'twistons' with topological charge (see [10, 11])

Detection of the Absolute Continuum The first attempt to detect the absolute medium were based on interferometry (the famous experiment of Michelson and Morley [25, 26]) but it produced a nil result. Many were quick to interpret the nil result of the Michelson and Morley experiment as evidence that the absolute continuum did not exist. The problem with the above conclusion is that it presents a fallacy consisting in using an argument that is supposed to prove one proposition but succeeds only in proving a different one. The fallacy here is that the conclusion is overreaching. Actually, the only rigorous conclusion from the nil result is that the absolute continuum cannot be detected by this *particular* experiment. The hypothesis of Lorentz [27] and FitzGerald (see, [28, p.749]) of length contraction in the direction of motion explains the nil effect, and renders the dismissal of a material medium superfluous. This means that all effort must be made to create an experimental set-up that can demonstrate, unequivocally, the relative speed in a laboratory setting. In this connection, the present author proposed a new look at the interferometry approach to the detection of the absolute speed based on beat-frequency measurements [29]

Another celebrated experiment which is believed to have unequivocally supported the theory of relativity was performed by Ives and Stilwell (ISE). It was assumed that the change of frequency of moving atoms in a cathode tube is due solely to time dilation. This is another overreaching conclusion, because there is not one-to-one correspondence between a frequency change and a time dilation. The frequency change can occur for various reasons, and one of them is discussed in a recent paper [30],where it is shown that if the Lorentz contraction is acknowledged in the Bohr-Rydberg formula for the frequency of the emitted light form a moving atom, then the frequency of the propagating wave is $\omega_{\mathrm{mov}} = \omega_{\mathrm{rest}}(1 - w^2/c^2)^{-1/2}$, where w is the speed of the atom relative to the absolute continuum. Using this formula we have shown in [30] that the absolute speed of Earth does not affect the interferometry in the second order. Hence, the absolute continuum cannot be detected by ISE.

Conclusion In the sited works of the present author the case is argued for space being an absolute material continuum (called *metaconrtinuum*) in which the electromagnetic vibrations propagate. The rheology of the mechanical space is of an elastic liquid, which has Maxwell's equations as corollaries. The logical fallacies of the arguments that led to the so-called relativity principle are examined. Particles and charges are interpreted as phase patterns (solitary waves of the metacontinuum. The

deformations in the fourth dimension explain the wave mechanics. Thus a consistent model of space as an elastic material continuum has been formulated and shown to explain the observations from two of the crucial experiments. In a sense, the proposed approach can be called the 'Special Theory of Absoluty.'

References

[1] J. Larmor, *Aether and Matter*, Cambridge, 1990.

[2] E. Whittaker, *A History of the Theories of Aether & Electricity vol. 1 and 2*, Dover, New York, 1989.

[3] M. N. Wise, Archive for History of Exact Sciences **21**, 19 (1981).

[4] J. C. Maxwell, Phil. Trans. Roy. Soc. London **155**, 469 (1865).

[5] R. K. Pathria, *The Theory of Relativity*, Courier Dover, 2003.

[6] B. E. Corey and D. T. Wilkinson, Bull. Astron. Astrophys. Soc. **8**, 351 (1976).

[7] G. F. Smoot, M. V. Gorenstein, and R. A. Miller, Phys. Rev. Lett. **39**, 898 (1997).

[8] L. Brillouin, *Relativity Reexamined*, Academic Press, New York, 1970.

[9] C. I. Christov, Annuaire de L'Universite de Sofia **95**, 109 (2001).

[10] C. I. Christov, Math. Comput. Simul. **74**, 93 (2007).

[11] C. I. Christov, Nonlinear Analysis **71**, e2028 (2009), Appeared online, March 19, 2009.

[12] L. D. Landau and M. Lifschitz, *The Classical Theory of Fields*, Reed Educational and Professional Publishing, New York, 4 edition, 2000.

[13] L. A. Segel, *Mathematics Applied to Continuum Mechanics*, Dover, New York, 1987.

[14] D. D. Joseph, A. Narain, and O. Riccius, J. Fluid Mech. **171**, 28 (1986).

[15] C. I. Christov, Foundations of Physics Letters (2006), Accepted.

[16] A. I. Borisenko and I. E. Tarapov, *Vector and Tensor Analysis with Applications*, Dover, 1979.

[17] C. H. Hinton, *Speculations on the Fourth Dimension. Selected Writings of C. H. Hinton* (Ed. R. v. B. Rucker), Dover, New York, 1980.

[18] E. Schrödinger, Annalen der Physik **79**, 361 (1926).

[19] C. I. Christov, Ann. Univ. Sofia **89**, 129 (1995).

[20] A. Einstein, *Relativity. The Special and the General Theory*, Three Rivers Press, New York, 1961.

[21] J. V. Boussinesq, Journal de Mathématiques Pures et Appliquées **17**, 55 (1872).

[22] N. J. Zabusky and M. D. Kruskal, Phys. Rev. Lett. **15**, 240 (1965).

[23] C. I. Christov, Discrete out of continuous: Dynamics of phase patterns in continua, in *Continuum Models and Discrete Systems – Proceedings of CMDS8*, edited by K. Markov, pages 370–394, Singapore, 1996, World Scientific.

[24] C. I. Christov, Math. Comp. Simul. **80**, 91 (2009).

[25] A. A. Michelson, Am. J. Sci. **22**, 120 (1881).

[26] A. A. Michelson and E. W. Morley, Am. J. Sci. **34**, 333 (1887).

[27] H. A. Lorentz, Zittingsverslagen der Akad. v. Wetenschappen te Amsterdam **1**, 74 (1892-3).

[28] O. Lodge, Phil. Trans. Roy. Soc. London A **184**, 727 (1893).

[29] C. I. Christov, J. Opt. Soc. Am. A **26**, 2292 (2009).

[30] C. I. Christov, Foundations of Physics **40**, doi: 10.1007/s10701-010-9418-2, Published Online Jan. 29, 2010 (Open Access).

Decision Support System with Implementation in Natural Hazards Field Tests

Nina Dobrinkova

1 Intorduction

Between 2000 and 2005 Europe suffered more than 100 floods, including 9 major flood disasters. These major flood events caused 155 casualties and economic losses of more than €35 billion. Furthermore the material flood damage of 2002 is estimated to be higher than in any previous single year (Barredo, 2007). Indeed there is an increase in the frequency of years with very high damage produced by major flood disasters in Europe. Two years of the current decade, i.e. 2000 and 2002, are among the worst concerning losses in the last 36-year period (Barredo, 2007). Despite the relevance of the issue, there is a need for comprehensive, standardised and georeferenced information on floods. Relevant, accurate and up-to-date data is important for political and economic decision-making. In Europe, historical data on flood disasters are neither comprehensive nor standardised, thus making difficult long-term analyses at continental level. A map of the major flood disasters of the last 56 years in the European Union, has been created by the Emergency Events Database (EM-DAT1) and NATHAN2 of Munich. (Figure 1)
One way to approach the problem is by using risk management to reduce the danger to the people living in the potential flood areas. Risk management includes administrative decisions, organization, operational skills and abilities to implement policies, strategies and coping capacities of the society and communities to lessen the impacts of natural hazards and related environmental disasters. Hazard maps are an instrument of hazard assessment and constitute a decisive element in modern risk management. They assist the identification, evaluation and reduction of risks by using an optimal combination of measures. Hazard maps may also support decisions concerning preventive measures againstnatural hazards and mitigation measures to manage disaster events (Kienholz 2005). Thus hazard mapping is an important input for risk analysis. Due to the interdependencies between human action and natural disasters in respect to triggering such extreme events and determining the degree of vulnerability against natural hazards, solving the problem will be possible only by using interdisciplinary cooperation.

2 Main goal

Monitoring is a fundamental observation methodology in risk management, aimed at the continuous surveillance of known hazards or at the detection of previously unknown hazards. Monitoring systems consist of a complex combination of sensors and data processing procedures and software. Monitoring produces information, which
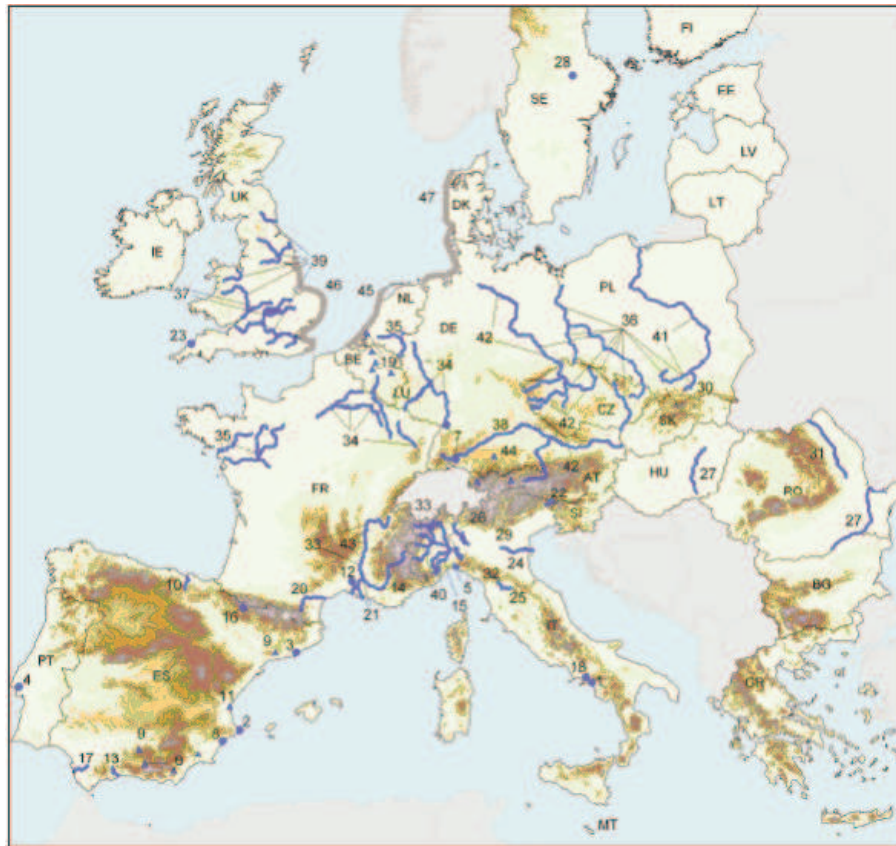
Figure 1: Major flood disasters in Europe: 1950-2005. Numbers from 1 to 23: flash floods, 24 to 44: river floods, 45 to 47 storm surge floods. A triangle feature represents very large regional events.

is valuable for depecting risk scenarios. Monitoring helps to assess hazard scenarios, where, how and when a specific disaster process will take place. From a methodological point of view monitoring can be seen as a complement to simulation models. From a practical point of view monitoring provides the information source of risk management instruments like hazard zonation or early warning systems. As such it serves as an important complement to active risk management measures like protective measures. In comparison to these benefits of monitoring include: the relative cost efficiency, the fast technological development and its positive impact for capacity building. These obvious advantages of monitoring have to be contrasted with some pitfalls: Many new monitoring methods are primarily research oriented and cannot yet offer the complete set of tools needed for practical application. Of these new technologies the potential is often theoretically known but not yet tested in practise.

In addition to that some techniques are often highly innovative but require a high level of expert knowledge. The main goal of project MONITOR II (funded by SEE program) is improvement of the methodology of risk evaluation and communication by applying these innovative methods into a Continuous Situation Awareness system (MONITOR II CSA). This project is the second stage of a project called MONITOR, where a wide range of the methods have been implemented in various test-beds with focus on different types of natural hazards. The suitability of methods was shown to be adequate but related risk communication difficult (short time period of warning). Risk communication procedures were developed as a participatory process in alarm plan development. A regional hazard response has been generated to show the relation between hazard potential and actual land use activities, where risk combinations between several hazards processes and exposition structures is displayed.

2.1 Monitoring methods used into the CSA system

A broad variety of heterogenous monitoring methods exist, that is why we organized criteria which we have used to evaluate their applicability and practical use.

2.1.1 By data acquisition platform

- Remote sensing systems: Remote sensing systems do not require the operation of active sensors directly in the area of interest. A further distinction can be made concerning the operational location of the sensor between spaceborne, aerial and terrestrial systems.

- In-situ sensing systems: In situ sensing systems require the operation of active sensors directly in the area of interest. A further distinction can be made between surface and sub-surface systems.

2.1.2 By data acquisition frequency

- Continuous monitoring: This applies to systems in which data acquisition is automated and the frequency can be varied depending on the user needs, up to a sampling frequency of a few minutes or seconds;

- Discontinuos monitoring: This applies to systems in which data acquisition is not-automated and the frequency depends upon the availability of personnel to carry out the measurements, or on systems in which data acquisition is automated but the frequency cannot be varied depending on the user needs and sampling frequency is in the order of days or weeks.

2.1.3 By data availability timing

- Real-time monitoring: Real time monitoring means acquiring, transmitting and processing data automatically, and reacting to results in an unsupervised manner - e.g. by automatically raising an alarm. This timing is mainly used in forecast, response and recovery phases.

- Near-real-time monitoning: This implies acquiring, transmitting and processing data automatically, and to react quickly to results in a supervised manner. E.g. given by the way of experts who will decide if, for example, an alarm must be raised and/or if decision makers and stakeholders must be informed. This timing is mainly used in forecast, response and recovery phases.

- Non-real-time monitoring: This means acquiring, transmitting and processing data automatically or manually, and to react to results at some point in time in a supervised manner by way of experts who will decide if and when, for example, decision makers and stakeholders must be informed. This timing is mainly used in prevention and only rarely used in forecast, response and recovery.

2.1.4   By spatial extent of data

- locallsed monltormg: e.g. data is collected at specific point-like locations

- distributed monitoring: e.g. data is collected over a more or less large area

Each of the monitoring methods is applicable for one or two of the project partners, but there is no method applicable for all of them. That is why the system architecture is giving opportunity to the user to pick a model incorporating methods applicable for the specific area depending on the hazard maps and hazard analysis provided as an input data.

3   The system

The primary goal of the MONITOR II CSA (Continuous Situation Awareness) is to improve situation awareness and knowledge about those situations, which are relevant for disaster management. This goal has to be achieved for different stakeholders in different phases of the disaster management cycle. This will integrate communication between hazard experts, decision makers and civil protection services with improved flow of information.

3.1   CSA technology

The MONITOR II CSA is a series of software components, which allows the easy integration, presentation and use of disaster management information. The CSA supports the information needs of different phases of the Disaster Management Cycle. The system architecture of the CSA takes into account the existence and well established use of legacy systems. This means that the components of the CSA follow some design rules: they are standards based, supporting OGC standard (like WMS, WFS or Sensor Web) and INSPIRE where ever feasible; they define open service oriented interfaces, allowing to integrate them with other components; their functionality is encapsulated so that they function indepently of specific other components and/or
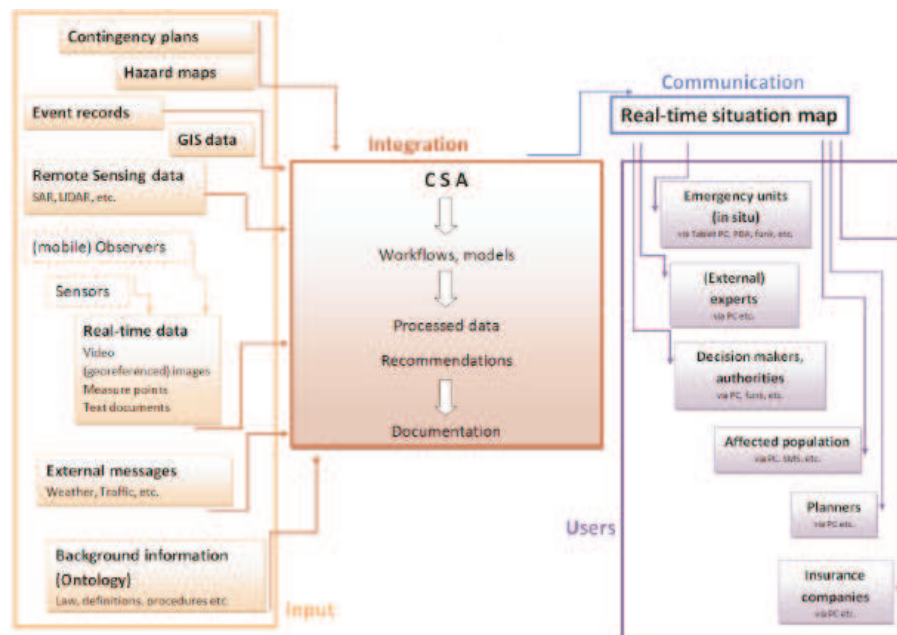
Figure 2: The CSA – Continuous Situation Awareness System architecture

information sources; their modular design is defined on thematic and interoperable units (figure 2).

The CSA is designed to store event data in a special CSA database. Object data - like buildings or roads - are assumed to be stored in the local, regional or national GIS. The CSA can use these object data directly if they conform to the thematically corresponding INSPIRE implementation rules. Otherwise a transformation of data is necessary. The CSA defines different levels of information integration: visual integration by overlaying information sources into one common image (hazad map); functional integration by using external functions (like simulation models) and integrating only the results of this; full integration of data and data processing within CSA.

3.1.1 Scenario modeler

A scenario modeller will be defined and will represent work-flows. It will use the following elements as "process nodes": natural processes; measures (and – depending on the measures defined – the possible processes, which are influenced/changed by these measures); damages. Endangered objects will be linked to processes (exposition). The scenario modeller is intended to describe general models of natural (disaster) processes and shows the resulting damages depending on the measures taken. It

provides a means for communicating the results of hazard assessment (e.g. hazard mapping) to a broad community of non-hazard-experts.

This architecture of the future CSA system is going to be developed for implementation in each partner country, after a series of field test planed in the MONITORII project schedule, prove that this is the correct sequence of measures, which in case of emergency has to be taken.

4 Conclusion

A unified system for early warning concentrated on flood events does not exist in the European Union. Thus the project idea of MONITORII is very innovative and the field test results together with successful development of working CSA system will be the best formula for better Civil Protection measures in future.

5 Acknowledgement

References

[1] Barredo, J.I. (2007) Major Flood Disasters in Europe: 1950-2005. Natural Hazards.

[2] De Roo A, Barredo JI, Lavalle C, Bodis K, Bonk R (2007) Potential flood hazard and risk mapping at pan-European scale. In: Peckham R, Jordan G (eds), Digital Terrain Modelling - Development and Applications in a Policy Support Environment. Springer.

[3] Kienholz, H. (2005): Early warning systems related to mountain hazards; Manuscript hand-out. In: International conference on early warning systems for natural disaster reduction, Potsdam (cited in HUEBL 2008).

[4] Alexander, D.E. (2002): Principles of Emergency Planning and Management. Terra Publishing, Harpenden, Herts, UK, and Oxford University Press, New York.

[5] www.monitor-cadses.org/documents/monitorresults/monitor_ontology_v1.pdf.

Adjoint Estimatoin of Ant Colony Optimization Starts Methods

Stefka Fidanova, Pencho Marinov, Krassimir Atanassov

1 Introduction

Many combinatorial optimization problems are fundamentally hard. This is the most typical scenario when it comes to realistic and relevant problems in industry and science. Examples of optimization problems are Traveling Salesman Problem [6], Vehicle Routing [7], Minimum Spanning Tree [5], Multiple Knapsack Problem [4], etc. They are NP-hard problems and in order to obtain solution close to the optimality in reasonable time, metaheuristic methods are used. One of them is Ant Colony Optimization (ACO) [2].

ACO algorithms have been inspired by the real ants behavior. In nature, ants usually wander randomly, and upon finding food return to their nest while laying down pheromone trails. If other ants find such a path, they are likely not to keep traveling at random, but to instead follow the trail, returning and reinforcing it if they eventually find food. However, as time passes, the pheromone starts to evaporate. The more time it takes for an ant to travel down the path and back again, the more time the pheromone has to evaporate and the path to become less prominent. A shorter path, in comparison will be visited by more ants and thus the pheromone dencity remains high for a longer time.

ACO is implemented as a team of intelligent agents which simulate the ants behavior, walking around the graph representing the problem to solve using mechanisms of cooperation and adaptation. ACO algorithm requires to define the following [1, 3]:

- The problem needs to be represented appropriately, which would allow the ants to incrementally update the solutions through the use of a probabilistic transition rules, based on the amount of pheromone in the trail and other problem specific knowledge. It is also important to enforce a strategy to construct only valid solutions corresponding to the problem definition.

- A problem-dependent heuristic function, that measures the quality of components that can be added to the current partial solution.

- A rule set for pheromone updating, which specifies how to modify the pheromone value.

- A probabilistic transition rule based on the value of the heuristic function and the pheromone value, that is used to iteratively construct a solution.

The structure of the ACO algorithm is shown by the pseudocode below. The transition probability $p_{i,j}$, to choose the node j when the current node is i, is based on the

heuristic information $\eta_{i,j}$ and the pheromone trail level $\tau_{i,j}$ of the move, where $i,j = 1, \ldots, n$.

$$p_{i,j} = \frac{\tau_{i,j}^a \eta_{i,j}^b}{\sum_{k \in Unused} \tau_{i,k}^a \eta_{i,k}^b},$$

where $Unused$ is the set of unused nodes of the graph.

The higher the value of the pheromone and the heuristic information, the more profitable it is to select this move and resume the search. In the beginning, the initial pheromone level is set to a small positive constant value τ_0; later, the ants update this value after completing the construction stage. ACO algorithms adopt different criteria to update the pheromone level.

Ant Colony Optimization
Initialize number of ants;
Initialize the ACO parameters;
while not end-condition **do**
    for k=0 **to** number of ants
        ant k choses start node;
        while solution is not constructed **do**
            ant k selects higher probability node;
        end while
    end for
    Update-pheromone-trails;
end while

Figure 1: Pseudocode for ACO

The pheromone trail update rule is given by:

$$\tau_{i,j} \leftarrow \rho\tau_{i,j} + \Delta\tau_{i,j},$$

where ρ models evaporation in the nature and $\Delta\tau_{i,j}$ is new added pheromone which is proportional to the quality of the solution.

Our novelty is to use estimations of start nodes with respect to the quality of the solution and thus to better menage the search process. On the basis of the estimations we offer several start strategies and their combinations.

2   Start Strategies

The known ACO algorithms create a solution starting from random node. But for some problems, especially subset problems, it is important from which node the search process starts. For example if an ant starts from node which does not belong to the optimal solution, probability to construct it is zero. Therefore we offer several start strategies.

Let the graph of the problem has m nodes. We divide the set of nodes on N subsets. There are different ways for dividing. Normally, the nodes of the graph are randomly enumerated. An example for creating of the nodes subsets, without lost of generality, is: the node number one is in the first subset, the node number two is in the second subset, etc. the node number N is in the $N - th$ subset, the node number $N + 1$ is in the first subset, etc. Thus the number of the nodes in the separate subsets are almost equal. We introduce estimations $D_j(i)$ and $E_j(i)$ of the node subsets, where $i \geq 2$ is the number of the current iteration. $D_j(i)$ shows how good is the jth subset and $E_j(i)$ shows how bad is the jth subset. $D_j(i)$ and $E_j(i)$ are weight coefficients of $j - th$ node subset ($1 \leq j \leq N$), which we calculate by the following formulas:

$$D_j(i) = \phi.D_j(i-1) + (1 - \phi).F_j(i),$$
$$E_j(i) = \phi.E_j(i-1) + (1 - \phi).G_j(i),$$

where $i \geq 1$ is the current process iteration and for each j ($1 \leq j \leq N$):

$$F_j(i) = \begin{cases} \frac{f_{j,A}}{n_j} & \text{if } n_j \neq 0 \\ F_j(i-1) & \text{otherwise} \end{cases}, \tag{1}$$

$$G_j(i) = \begin{cases} \frac{g_{j,B}}{n_j} & \text{if } n_j \neq 0 \\ G_j(i-1) & \text{otherwise} \end{cases}, \tag{2}$$

and $f_{j,A}$ is the number of the solutions among the best $A\%$, and $g_{j,B}$ is the number of the solutions among the worst $B\%$, where $A + B \leq 100$, $i \geq 1$ and

$$\sum_{j=1}^{N} n_j = n,$$

where n_j ($1 \leq j \leq N$) is the number of solutions obtained by ants starting from nodes subset j. Initial values of the weight coefficients are: $D_j(1) = 1$ and $E_j(1) = 0$. The parameter ϕ, $0 \leq \phi \leq 1$, shows the weight of the information from previous iterations and from current iteration. When $\phi = 0$ only the information from current iteration is taken in to account. If $\phi = 0.5$ the influence of previous versus current iterations is equal. When $\phi = 1$ only the information from previous iterations is taken in to account. When $\phi = 0.25$ the weight of the information from previous iterations is three times less than this one of the current iteration. When $\phi = 0.75$ the weight of the previous iterations is three times higher than this one of the current iteration. The balance between the weights of the previous and current iterations is important. At the beginning when the current best solution is far from the optimal one, some of the node subsets can be estimated as good. Therefore, if the value of the parameter ϕ is too high the estimation can be distorted. If the weight of the current iteration is too high than in the estimation are ignored good and bad solutions from previous iterations, which can distort estimation too.

We try to use the experience of the ants from previous iteration to choose the better starting node. Other authors use this experience only by the pheromone, when the

ants construct the solutions. Let us fix threshold E for $E_j(i)$ and D for $D_j(i)$, than we construct several strategies to choose start node for every ant, the threshold E increase every iteration with $1/i$ where i is the number of the current iteration:

1. If $E_j(i)/D_j(i) > E$ then the subset j is forbidden for current iteration and we choose the starting node randomly from $\{j \mid j$ is not forbidden$\}$;

2. If $E_j(i)/D_j(i) > E$ then the subset j is forbidden for current simulation and we choose the starting node randomly from $\{j \mid j$ is not forbidden$\}$;

3. If $E_j(i)/D_j(i) > E$ then the subset j is forbidden for K_1 consecutive iterations and we choose the starting node randomly from $\{j \mid j$ is not forbidden$\}$;

4. Let $r_1 \in [0.5, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ we randomly choose node from subset $\{j \mid D_j(i) > D\}$, otherwise we randomly chose a node from the not forbidden subsets, r_1 is chosen and fixed at the beginning.

5. Let $r_1 \in [0.5, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ we randomly choose node from subset $\{j \mid D_j(i) > D\}$, otherwise we randomly chose a node from the not forbidden subsets, r_1 is chosen at the beginning and increase with r_3 every iteration.

Where $0 \leq K_1 \leq$"number of iterations" is a parameter. If $K_1 = 0$, than strategy 3 is equal to the random choose of the start node. If $K_1 = 1$, than strategy 3 is equal to the strategy 1. If $K_1 =$"maximal number of iterations", than strategy 3 is equal to the strategy 2.

We can use more than one strategy for choosing the start node, but there are strategies which can not be combined. We distribute the strategies into two sets: $St1 = \{strategy1, strategy2, strategy3\}$ and $St2 = \{strategy5, strategy6\}$. The strategies from same set can not be used at once. Thus we can use strategy from one set or combine it with strategies from other set. Exemplary combinations are $(strategy1)$, $(strategy2;\ strategy5)$, $(strategy3;\ strategy6)$.

3 Conclusion

In this paper we address the modelling of the process of ant colony optimization method by using adjoint estimations, combining five start strategies. So, the start node of each ant depends of the goodness of the respective region. In a future we will focus on parameter settings which manage the starting procedure. We will investigate on influence of the parameters to algorithm performance. We will apply our modification of ACO algorithm on various classes of problems. We will investigate the influence of the estimations and start strategies on the achieved results.

ment rules" and TK-"Effective Monte Carlo Methods for large-scale scientific problems".

References

[1] Bonabeau E., Dorigo M., Theraulaz G., Swarm Intelligence: From Natural to Artificial Systems, New York,Oxford University Press, 1999.

[2] Dorigo M., Gambardella L.M., Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. IEEE Transactions on Evolutionary Computation 1,53-66,1997.

[3] Dorigo M., Stutzle T., Ant Colony Optimization, MIT Press, 2004.

[4] Fidanova S., Evolutionary Algorithm for Multiple Knapsack Problem, Int. Conference Parallel Problems Solving from Nature, Real World Optimization Using Evolutionary Computing, ISBN No 0-9543481-0-9,Granada, Spain, 2002.

[5] Reiman M., Laumanns M., A Hybrid ACO algorithm for the Capacitated Minimum Spanning Tree Problem, In proc. of First Int. Workshop on Hybrid Metahuristics, Valencia, Spain, 2004, 1-10.

[6] Stutzle T. Dorigo M., ACO Algorithm for the Traveling Salesman Problem, In K. Miettinen, M. Makela, P. Neittaanmaki, J. Periaux eds., Evolutionary Algorithms in Engineering and Computer Science, Wiley, 163-183, 1999.

[7] Zhang T., Wang S., Tian W., Zhang Y., ACO-VRPTWRV: A New Algorithm for the Vehicle Routing Problems with Time Windows and Re-used Vehicles based on Ant Colony Optimization, Sixth International Conference on Intelligent Systems Design and Applications, IEEE press, 2006, 390-395.

Locally Optimized MIC(0) Preconditioning of Rotated Bilinear FEM Systems

Ivan Georgiev, Svetozar Margenov

We consider the elliptic boundary value problem

$$\begin{aligned} Lu \equiv -\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla \mathbf{u}(\mathbf{x})) &= f(\mathbf{x}) \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where Ω is a polygonal domain in $\mathbf{R}^2$, $f(\mathbf{x}) \in \mathbf{L^2(\Omega)}$. The coefficient matrix $\mathbf{a}(\mathbf{x})$ is symmetric positive definite and uniformly bounded in Ω. The weak formulation of the above problem reads as follows: Given $f \in L^2(\Omega)$ find $u \in \mathcal{V} \equiv H_D^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$, satisfying

$$\mathcal{A}(u, v) = (f, v) \quad \forall v \in H_D^1(\Omega), \quad \text{where} \quad \mathcal{A}(u, v) = \int_\Omega \mathbf{a}(\mathbf{x})\nabla \mathbf{u}(\mathbf{x}) \cdot \nabla \mathbf{v}(\mathbf{x})\mathbf{dx}.$$

The domain Ω is discretized by the partition $\mathcal{T}_h$ which is aligned with the discontinuities of the diffusion coefficients so that over each element $e \in \mathcal{T}_h$ the coefficients of $\mathbf{a}(\mathbf{x})$ are smooth functions. The variational problem is then approximated using the finite element method, i.e., the continuous space $\mathcal{V}$ is replaced by a finite dimensional subspace $\mathcal{V}_h$. Then the finite element formulation is: Find $u_h \in \mathcal{V}_h$, satisfying

$$\mathcal{A}_h(u_h, v_h) = (f, v_h) \quad \forall v_h \in \mathcal{V}_h, \quad \text{where} \quad \mathcal{A}_h(u_h, v_h) = \sum_{e \in \mathcal{T}_h} \int_e \mathbf{a}(e)\nabla u_h \cdot \nabla v_h d\mathbf{x}.$$

Here $\mathbf{a}(e)$ is a piece-wise constant symmetric positive definite matrix, defined by the integral averaged values of $\mathbf{a}(\mathbf{x})$ over the current element from the triangulation $\mathcal{T}_h$. We note that in this way strong coefficient jumps across the boundaries between adjacent finite elements from $\mathcal{T}_h$ are allowed.

Here we will consider the case of discretization by non-conforming rotated bilinear finite elements [7]. The square $\hat{e} = [-1, 1]^2$ is used as a reference element in the definition of the rotated bilinear element $e \in \mathcal{T}_h$. Let

$$\Psi_e : \hat{e} \to e$$

be the corresponding bilinear bijective transformation. The nodal basis functions are determined by the relations

$$\{\phi_i\}_{i=1}^4 = \{\hat{\phi}_i \circ \Psi_e^{-1}\}_{i=1}^4, \qquad \{\hat{\phi}_i\} \in \text{span}\{1, x, y, x^2 - y^2\}$$

where $\circ$ means the superposition of functions $\hat{\phi}_i$ and Ψ_e^{-1}. For the variant MP (midpoint) the shape functions $\{\hat{\phi}_i\}_{i=1}^4$ are found by the point-wise interpolation condition

$$\hat{\phi}_i(b_\Gamma^j) = \delta_{ij},$$

where $b_\Gamma^j, j = 1, 4$ are the midpoints of the boundary $\Gamma_{\hat{e}}$. Alternatively, the variant MV (mean-value) corresponds to the integral mean-value interpolation condition

$$|\Gamma_{\hat{e}}^j|^{-1} \int_{\Gamma_{\hat{e}}^j} \hat{\phi}_i d\Gamma_{\hat{e}}^j = \delta_{ij},$$

where $\Gamma_{\hat{e}}^j$ are the sides of the reference element $\hat{e}$. Here in the presented numerical tests we will consider MV discretization variant.

The preconditioned conjugate gradient (PCG) method is used for solving the resulting linear algebraic system

$$A\mathbf{u} = \mathbf{f},$$

where A and $\mathbf{f}$ are the corresponding global stiffness matrix and global right hand side.

The preconditioner is constructed by modified incomplete Cholesky (MIC(0)) factorization (see [1, 4, 5]). For symmetric and positive definite (SPD) matrices, the set of M-matrices with nonnegative row-sum reads as follows:

$$M_N = \left\{ A \in \mathbf{R}^{N \times N} \; : \; a_{ii} > 0; \; a_{ij} \leq 0, \; i \neq j; \; \sum_{j=1}^{N} a_{ij} \geq 0 \right\}.$$

It is known that a MIC(0) factorization is applicable if the stiffness matrix $A \in M_N$ which is not the case in many applications. Here, the idea is first to construct an M-matrix approximation of the global stiffness matrix and then to apply a MIC(0) factorization. For this purpose we will consider the following local optimization problem: For a given element stiffness matrix $A^{(e)}$ find the symmetric and positive semidefinite matrix $B^{(e)}$, with nonpositive offdiagonal entries, such that the effective spectral condition number

$$\hat{\kappa}\left(B^{(e)}, A^{(e)}\right) = \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}}$$

is as small as possible, where $A^{(e)}\mathbf{v}^{(e)} = \lambda B^{(e)}\mathbf{v}^{(e)}, \quad \mathbf{v}^{(e)} \neq (c, c, \ldots, c)^t, \quad c \in \mathbf{R}$. The arising local optimization problem is solved symbolically following the technique proposed in [6]. Applying then the standard FEM assembling procedure we obtain a locally optimal M-matrix approximation B of the global stiffness matrix A.

Let us consider the elliptic model problem where the coefficient matrix is of the form

$$\mathbf{a}(\mathbf{x}) = a(e) \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix},$$

and for the coefficient ϵ we have the inequality $0 < \epsilon \leq 1$. It is shown in [3] that if we use the diskretization aligned with the coordinate (anisotropy) axes the spectral condition number $\kappa\left(B^{-1}A\right)$ is not uniformly bounded with respect to the anisotropy. We can overcome this disadvantage by using the $45°$ skewed orthogonal square mesh and as it is shown in [3] we get the spectral condition number estimate

$$\kappa\left(B^{-1}A\right) \leq \frac{3}{2},$$

which hold uniformly with respect to the anisotropy and possible coefficient jumps.

We will present numerical tests illustrating the convergence rate of the studied PCG algorithm when the size of the discrete problem and the anisotropy coefficient are varied. A relative stopping criterion $\dfrac{(C^{-1}\mathbf{r}^i, \mathbf{r}^i)}{(C^{-1}\mathbf{r}^0, \mathbf{r}^0)} < \varepsilon^2$ is used, where C is the $MIC(0)$ preconditioner, $\mathbf{r}^i$ stands for the residual at the i-th iteration step, and $\varepsilon = 10^{-6}$. The anisotropy parameter is varied as $\epsilon = 2^{-l}$, where l is a test parameter. Pure homogeneous Dirichlet boundary conditions are imposed. A uniform square mesh with mesh-size parameter $h = 1/n$ is used.

Example 1. The computational domain

$$\Omega = \left\{ |x| \le y \le \sqrt{2} - |x| : \ -\frac{\sqrt{2}}{2} \le x \le \frac{\sqrt{2}}{2} \right\}.$$

The model problem with constant coefficient $a(e) = 1$ is considered. The size of the discrete problem is $N = 2n(n + 1)$. The numerical results are presented in Table 1. The stable behavior of the number of iterations with respect to the anisotropy confirms the obtained uniform condition number estimates.

Table 1: Example 1 - PCG iterations.

h^{-1}	$l = 0$	$l = 2$	$l = 4$	$l = 6$	$l = 8$
128	34	40	39	35	27
256	47	57	55	49	41
512	69	82	78	70	57
1024	100	118	112	100	82

Example 2. The computational domain Ω from Example 1 is divided in two subdomains $\Omega = \Omega_1 \bigcup \Omega_2$, $\Omega_1 \bigcap \Omega_2 = \emptyset$, see Figure 1 (a). We will denote with a_1 and a_2 the problem coefficients corresponding to the subdomains Ω_1 and Ω_2. The coefficient jump relates to the coefficients $a_1 = 1$ and $a_2 = 10000$. The numerical results are presented in Table 2.

Example 3. Here again the domain $\Omega = \Omega_1 \bigcup \Omega_2$ but the geometry of the subdomains Ω_1 and Ω_2 is different, see Figure 1 (b). The related problem coefficients are $a_1 = 1$ and $a_2 = 10000$. The numerical results for Example 3 are presented in Table 3.

The number of iterations in Table 2 and Table 3 increases less than two times when the coefficient jump is of order 10^4. The qualitative analysis of the results for the considered examples confirms that the number of iterations is $O(N^{1/4})$ for large enough
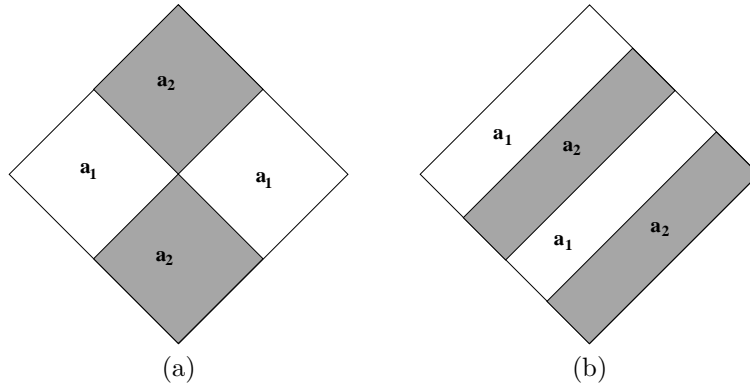
Figure 1: Computational domain: (a) Example 2, (b) Example 3

Table 2: Example 2 - PCG iterations.

h^{-1}	$l = 0$	$l = 2$	$l = 4$	$l = 6$	$l = 8$
128	69	71	65	55	40
256	98	103	98	91	74
512	140	147	142	136	127
1024	198	208	202	199	197

N, which is in agreement with the theoretical expectations, see [4]. The construction of $MIC(0)$ preconditioner C as well as each iteration can be carried out in $O(N)$ arithmetic operations.

The modification of the element stiffness matrices is a natural and very useful procedure for the construction of preconditioners using MIC(0) factorization for second order elliptic problems. It is, however, also an important issue in the separate displacement component preconditioning of elasticity problems [2].

References

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[2] R. Blaheta, *Displacement decomposition - incomplete factorization preconditioning techniques for linear elasticity problems*, Numer. Lin. Alg. Appl., 1(2) (1994), 107–128.

Table 3: Example 3 - PCG iterations.

h^{-1}	$l=0$	$l=2$	$l=4$	$l=6$	$l=8$
128	40	43	50	52	53
256	57	65	72	75	77
512	81	95	103	107	114
1024	117	139	148	153	169

[3] I. Georgiev, J. Kraus, S. Margenov, J. Schicho, *Locally Optimized MIC(0) Preconditioning of Rannacher-Turek FEM Systems*, Applied Numerical Mathematics, 59(10) (2009), 2402–2415

[4] I. Gustafsson, *Stability and rate of convergence of modified incomplete Cholesky factorization methods*, Report 79.02R. Dept. of Comp. Sci., Chalmers University of Technology, Goteborg, Sweden, 1979

[5] I. Gustafsson, *An Incomplete Factorization preconditioning method based on modification of element matrices*, BIT 36(1) (1996), 86–100

[6] U. Langer, S. Reitzinger, J. Schicho, *Symbolic methods for the element preconditioning technique*, SFB-Report No. 02-03, (2002)

[7] R. Rannacher, S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numerical Methods for Partial Differential Equations, 8(2) (1992), 97–112.

Notes on the Numerical Treatment of Sparse Matrices Arising in a Chemical Model

Krassimir Georgiev, Zahari Zlatev

A two-dimensional advection-diffusion-chemistry module of a large-scale environmental model is taken. The module is described mathematically by system of partial differential equations. Sequential splitting is used in the numerical treatment. The non-linear chemistry is most the time-consuming part and it is handled by six implicit algorithms for solving ordinary differential equations. This leads to the solution of very long sequences of systems of linear algebraic equations. It is crucial to solve these systems efficiently. This is achieved by applying four different algorithms.

1 The 2D version of the Danish Eulerian model and rotational test

The Danish Eulerian model (UNI–DEM)([6]) is an model for studying the long-range transport of air pollutants. The model computational domain covers Europe and parts of Asia, Africa and the Atlantic Ocean. The long-range transport of air pollution is usually studied by a system of partial differential equations (PDEs), which can be written as follows (it should be mentioned that similar systems are used in other environmental models):

$$
\begin{aligned}
\frac{\partial c_s}{\partial t} \;=\; & -\frac{\partial (u c_s)}{\partial x} - \frac{\partial (v c_s)}{\partial y} \\
& + \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s}{\partial y}\right) \\
& + E_s + Q_s(c_1, c_2, \ldots c_q) - (k_{1s} + k_{2s}) c_s, \quad s = 1, 2, \ldots q \; ,
\end{aligned}
\tag{1}
$$

where:

(i) $c_s(t, x, y)$ are the concentrations of the chemical species;

(ii) $u(t, x, y)$, $v(t, x, y)$ are the wind components along the coordinate axes;

(iii) $K_x(t, x, y)$, $K_y(t, x, y)$ are the diffusion coefficients;

(iv) $E_s(t, x, y)$ present the emission sources;

(v) $k_{1s}(t, x, y)$, $k_{2s}(t, x, y)$ are correspondingly the dry and wet deposition coefficients, and finally,

(vi) $Q_s((t, x, y, c_1, c_2, \ldots c_q)$ are non-linear functions which describe the chemical reactions between species under consideration.

When some numerical methods shod be tested and tuned it is more convenient to use the following two-dimensional module:

$$\frac{\partial c_s}{\partial t} = -\mu(y - y_0)\frac{\partial c_s}{\partial x} - \mu(x_0 - x)\frac{\partial c_s}{\partial y} + K\left(\frac{\partial c_s}{\partial x^2} + \frac{\partial c_s}{\partial y^2}\right)$$

$$+ E_s(t, x, y) + Q_s(t, x, y, c_1, c_2, \ldots, c_q), \quad s = 1, 2, \ldots q,$$

(2)

where $x \in [a_1, b_1]$, $y \in [a_2, b_2]$, $t \in [a, b]$ $x_0 = \frac{b_1 - a_1}{2}$, $y_0 = \frac{b_2 - a_2}{2}$, $\mu = \frac{2\pi}{b-a}$. Very often the space domain is square, i.e. $a_1 = a_2, b_1 = b_2$ and the length of the time interval is 86 400 s. (24 hours with starting point being 6:00 AM). The problem (2) is considered with a given initial value vector $c(c, y, a)$ and some boundary conditions (Dirichlet boundary conditions which be used hereafter). The most essential difference in (2) according to (1) is the special definition of the wind velocity field. The trajectories of the wind in (2) are concentric circles with center (x_0, y_0) and particles are rotated along these trajectories with a constant angular velocity. Such wind velocity field was first defined in [1, 5]. There, the test in which only the first two terms in the right-hand-side of (2) are kept (the pure advection test) was introduced. Chemical reactions were included to the Crowley-Molenkampf test by Hov et al (see [3]). The module defined in this paper by (2) is a further extension of the original Crowley-Molenkampf test. It is worthwhile to study this module (generalized rotation test) because it is much closer to real environmental model than the previous two modules. By setting some of the coefficients to zero or keeping all of them different from zero, six situations can be studied by (2):

- **No non-zero emissions are specified** (*puff; $E_s = 0$ for all values of s*):

    (A) Pure advection-diffusion process (only the terms containing the spatial derivatives are kept).

    (B) Pure chemical process (only the non-linear functions are kept).

    (C) Combining the advection-diffusion process with the chemistry process (all terms except the emissions are kept).

- **Some emissions are no zero** (*plume; $E_s \neq 0$ for some values of s*):

    (D) Pure advection-diffusion process (only the terms containing the spatial derivatives and the emissions are kept).

    (E) Pure chemical process (only the non-linear functions Q_s and the emissions are kept).

    (F) Combining the advection-diffusion process with the chemistry process (all terms are kept).

40

2 Sequential splitting procedure

A splitting procedure proposed in [4] is used in Danish Eulerian model. In the three dimensional version of the model it leads to five submodels representing the horizontal advection, the horizontal diffusion, the chemistry and the emissions, the depositions (dry and wet) and the vertical exchange. Simple sequential splitting will be used in this paper. Applying this kind of splitting to (2 leads to the following two sub-problems:

$$\frac{\partial g_s}{\partial t} = -\mu(y - y_0)\frac{\partial g_s}{\partial x} - \mu(x_0 - x)\frac{\partial g_s}{\partial y} + K\left(\frac{\partial^2 g_s}{\partial x^2} + \frac{\partial^2 g_s}{\partial y^2}\right), \qquad (3)$$

$$\frac{\partial h_s}{\partial t} = E_s(t, x, y) + Q_s(t, x, y, g_1, g_2, \ldots g_q) \qquad (4)$$

It is worthwhile to emphasize here that:

- The sequential splitting procedure allows different numerical methods to be used in the treatment of the two sub-problems. This is a very useful feature, because the two sub-problems have different properties. The first sub-problem, the advection-diffusion module is a non-stiff problem, while the second sub-problem, the chemistry module is a stiff sub-problem.

- The system (3) consists of q independent PDEs. If the computational space domain is discretized into $N_x \times N_y$ grid-points, then (4) will be decoupled into $N_x \times N_y$ independent systems of ordinary differential equations (ODEs) each of which contains q equations. This observation indicates that in general efficient parallel computations can be achieved in a natural way.

3 Numerical treatment in the chemistry module

The chemical submodel is described by (4). Let the space domain is discretized by using $(N_x + 1)(N_y + 1)$ grid points. Then the non-linear system of PDE's (4) is reduced to $(N_x + 1)(N_y + 1)$ independent systems of ODEs (one per each grid-point in the space domain). The number of equations in every system of ODEs is q. An arbitrary system of ODE's can be written as

$$\frac{\partial \widehat{h}}{\partial t} = f(t, \widehat{h}), \qquad \widehat{h} \in \mathcal{R}^q, \quad f \in \mathcal{R}^q. \qquad (5)$$

The system of ODEs (5) is stiff. Let us denote with $J(t, \widehat{h})$ of the vector function $f \in \mathcal{R}^q$ $(J(t, \widehat{h}) = \partial f(t, \widehat{h})/\partial \widehat{h} \in \mathcal{R}^{q \times q})$. Now, the Newton Iterative Method (see e. g. [2]) is applied to solve the nonlinear system of ODEs (5). At each iteration m of the Newton Iterative Method, the following system of linear algebraic equations has to be solved:

$$B_m z_m = d_m, \quad \widehat{h}_{new}^{[m+1]} = \widehat{h}_{new}^{[m]} + z_m, \tag{6}$$

where

$$B_m = I - \gamma \Delta t J \left(t_{n+1}, \widehat{h}_{new}^{[m+1]} \right), \ z_m = \Delta \widehat{h}_{new}^{[m+1]}, \ d_m = s \left(\widehat{h}_{old}, \widehat{h}_{new}^{[m]} \right), \ m = 0, 1, 2, \dots . \tag{7}$$

Some properties and characteristics of the matrices arising in the chemistry module are:

- **general** (these matrices **ARE NOT** symmetric, diagonally dominant, banded or positive definite);

- **badly scaled**;

- **ill-conditioned**;

- their **elements vary in large intervals** on diurnally basis which cause great diurnal variation of the involved concentrations.

The following several techniques for treatment of the matrices which arise in the chemistry module are used in the performed numerical and computer experiments:

- **Dense matrix technique**: $LAPACK$ – (DGETRF and DGETRS);

- **Regular sparse matrix technique**: $PARASPAR$ – DIR (Direct solution using Gaussian Elimination);

- **Preconditioned sparse matrix technique**: $PARASPAR$ – ORTH (Preconditioned Modified Orthomin);

- **Special sparse matrix technique** – only for the chemical scheme in UNI–DEM.

The main problems which appear when a regular sparse matrix code is applied for small matrices, are related mainly to: (i) *the indirect addressing, the use of many integer arrays* the content of which must very often be updated, *the performance of many short loops* and *the need to search for pivotal elements at every stage of the Gaussian elimination* (in an attempt to preserve both the sparsity and and the stability). Therefore, the special sparse matrix technique has been developed. This technique is based on the following steps:

- A preliminary reordering procedure based on the application of a Markovitz pivotal strategy for general sparse matrices is performed. All small non-zero elements are removed during this preliminary procedure.

- The positions, in which new non-zero elements will be created, are determined and locations for these elements are reserved in the arrays where the LU factorization of the diagonal block under consideration is stored.

Table 1: Computing times measure in CPU hours for different space and time discretizations

Method	Time-steps	$N_x \times N_y$	Dense	Sparse direct	Precond. sparse	Special sparse
Backward	960	1089	0.112	0.067	0.052	0.040
Euler	1920	4225	0.621	0.269	0.157	0.063
Method	3840	16641	4.249	1.732	0.875	0.241
(first	7680	66049	31.383	12.575	6.238	1.506
order)	15360	263169	250.712	98.921	52.775	12.116
Implicit	960	1089	0.243	0.118	0.087	0.039
Midpoint	1920	4225	1.507	0.608	0.360	0.059
Rule	3840	16641	9.464	3.693	2.030	0.200
(second	7680	66049	61.622	22.930	11.772	1.407
order)	15360	263169	384.331	138.737	72.002	11.647
Two-stage	960	1089	0.205	0.088	0.064	0.041
2nd order	1920	4225	1.400	0.402	0.234	0.071
diag.implicit	3840	16641	9.944	2.891	1.552	0.269
Runge-Kutta	7680	66049	75.052	22.665	11.880	1.972
Method	15360	263169	542.686	179.742	93.748	15.479
3-stage	960	1089	0.301	0.322	0.410	0.128
5th order	1920	4225	1.958	2.108	1.152	0.668
fully impl.	3840	16641	13.862	15.184	13.294	4.543
R-K	7680	66049	107.426	118.113	68.278	33.052
Method	15360	263169	657.868	921.914	403.838	268.724
Two-stage	960	1089	0.060	0.067	0.065	0.047
2-nd order	1920	4225	0.321	0.288	0.256	0.095
Rosen-	3840	16641	1.620	2.082	1.724	0.437
brock	7680	66049	12.939	16.976	13.070	2.801
Method	15360	263169	105.220	144.371	99.395	21.083
Trapezo-	960	1089	0.118	0.063	0.051	0.039
idal	1920	4225	0.632	0.233	0.145	0.056
Rule	3840	16641	4.481	1.514	0.817	0.207
(second	7680	66049	33.966	11.445	6.198	1.155
order)	15360	263169	252.372	91.759	49.137	10.369

- A loop-free code for the numerical calculation of the LU factorization of the diagonal block under consideration is prepared.

- A loop-free code for the back-substitution (based on the LU factorization computed in the previous step) is prepared.

The computing time when six different numerical methods for solving systems of ODEs and four different techniques for treatment of the sparse matrices arising in the chemical model discussed above are presented in Table 1.

4 Conclusion

An important component of the large-scale air pollution models, a chemistry module, was studied in this paper. The numerical treatment of this module requires the solution of very long sequences of systems of linear algebraic equations. It has been shown in this paper that the selection of efficient algorithms for solving systems of linear algebraic equations is very important. The computing time can be reduced by a large factor (sometimes by a factor larger than 30) when the proper method is chosen.

Acknowledgments

References

[1] W. P. Crowley: *Numerical advection experiments*, Monthly Weather Review, **Vol. 96** (1968), pp. 1-11.

[2] G. H. Golub, J. M. Ortega: *Scientific computing and differential equations*, Academic Press, (1992)

[3] Ø. Hov, Z. Zlatev, R. Berkowicz, A. Eliassen and L. P. Prahm: *Comparison of numerical techniques for use in air pollution models with non-linear chemical reactions*, Atmospheric Environment **23** (1988), pp. 967-983.

[4] G. I. Marchuk: *Mathematical modelling for the problem of the environment*, North-Holland, Amsterdam, 1985

[5] C. R. Molenkampf: *Accuracy of finite-difference methods applied to the advection equation*, Journal of Applied Meteorology, **Vol. 7** (1968), 160-167.

[6] Z. Zlatev: *Computer treatment of large air pollution models*, Kluwer (1995).

Numerical Study of Sensitivity Analysis Techniques

Rayna Georgieva, Sofiya Ivanovska

1   Introduction

Sensitivity analysis (SA) of model output aims to quantify how a model (describing chemical, physical, biological, or social processes) depends on its input parameters. Sensitivity analysis is increasingly recognized as an important tool for model building and validation. In general, sensitivity analysis is useful for all processes where it is important to know which input factors contribute most to output variability [5, 9].

This paper aims at numerical studies of variance-based techniques for sensitivity analysis - Sobol' approach and Fourier Amplitude Sensitivity Test (FAST). The approaches have been implemented using Monte Carlo algorithms and SIMLAB software tool for sensitivity analysis [11]. The main mathematical problem here is numerical integration of model function to estimate the corresponding numerical indicators of sensitivity. A discussion about reliability of the results have been done after a number of numerical experiments with a model simulating remote transport of air pollutants. The issue of computing small sensitivity indices is very important in this particular case.

2   Mathematical Background

Several variance-based techniques for sensitivity analysis have been compared during the sensitivity study of the concentration levels of ozone to variation of chemical rates of an air pollution model - Unified Danish Eulerian Model (UNI-DEM). The chemical scheme used in the model is the condensed CBM-IV (Carbon Bond Mechanism, [10]).

Consider a scalar model output u $= f(\mathrm{x})$ corresponding to a number of non-correlated model parameters x $= (x_1, x_2, \ldots, x_d)$ with a joint probability density function $p(\mathrm{x}) = p(x_1, \ldots, x_d)$. The issue about parameter importance can be studied via numerical integration in the terms of analysis of variance (ANOVA). A number of unbiased Monte Carlo estimators for global sensitivity indices have been developed applying ANOVA-decomposition of model function [3, 5, 7, 8].

2.1   Sobol' Approach

In Sobol' approach [7] the variance of the square integrable model function is decomposed into terms of increasing dimension. The sensitivity of model output to each parameter or parameter interaction is measured by its contribution to the total variance.

An approach for evaluating *small* sensitivity indices (to avoid loss of accuracy because the analyzed database comes under this case) has been applied - a combined approach

(between approach of reducing of the mean value and correlated sampling) suggested in [8].

2.2 Fourier Amplitude Sensitivity Test

The method is based on Fourier transformation of uncertain model input parameters into a frequency domain [1]. It converts a multidimensional integral over all the uncertain parameters to an one-dimensional integral defined by a set of parametric equations, i.e. the idea is to explore the input space along a curve.

An extension of the FAST (eFAST) method [6] was developed to estimate the total effects. Decomposition of variance in eFAST works by varying different parameters at different frequencies, encoding the identity of parameters in the frequency of their variation. Fourier analysis then measures the strength of each parameter frequency in the model output.

If N denotes the sample size and d - the number of inputs, the original Sobol' method requires $N \times (2d + 1)$ model runs to calculate all first- and total order sensitivity indices. For comparison, the extended FAST requires dN model runs. An enhancement of Sobol' method has been made [4] - this modification provides the first-, second- and total order sensitivity indices using $N \times (2d + 2)$ model runs. As a summary, the computational cost of the sensitivity analysis approaches under consideration to compute all first- and total order sensitivity indices has an optimal rate - it is proportional to the sample size and the number of input parameters (see Table 1).

Table 1: Methods for evaluating global sensitivity indices.

Method	Cost (model runs)	Sensitivity measures
FAST (1973)	$O(d^2)$	$S_i, \forall i$
Sobol' (1993)	$N(2d + 2)$	$S_i, S_{x_i}^{tot}, \forall i$
EFAST (1999)	dN	$S_i, S_{x_i}^{tot}, \forall i$
Saltelli (2002)	$N(d + 2)$	$S_i, S_{x_i}^{tot}, \forall i, S_{-lj}, \forall l, j, l \neq j$
Saltelli (2002)	$N(2d + 2)$	$S_i, S_{x_i}^{tot}, \forall i, S_{lj}, S_{-lj}, \forall l, j, l \neq j$

2.3 Adaptive Monte Carlo Algorithm

The adaptive Monte Carlo algorithm uses a posteriori information about the variance. Its idea consists in the following: the domain of integration is separated initially into sub-domains with identical volume. The subdomain with the largest standard deviation is chosen for the next division. The algorithm stops when the standard deviation at all obtained after division subdomains satisfies the preliminary given accuracy ε.

3 Discussion of Numerical Results

A comparison between results obtained applying plain Monte Carlo approach, two approaches for SA from the software tool SIMLAB and our adaptive Monte Carlo algorithm [2] is given on Table 2. The goal of the performed comparative analysis here is to study and specify the numerical advantages of the proposed adaptive Monte Carlo algorithm for estimation of unknown quantities.

Table 2: First-order and total sensitivity indices of inputs estimated using different approaches of SA applying Plain and Adaptive Monte Carlo algorithms ($S_1 = 0.53234, S_3 = 0.00198, S_{x_1}^{tot} = 0.53782, S_{x_3}^{tot} = 0.00226$).

Est. quant.	Combined: Plain		FAST: SIMLAB	Combined: Adaptive			Sobol': SIMLAB	
	N	Rel. error	Rel. error	$\#_{sub}$	N_{sub}	Rel. error	N	Rel. error
S_1	192	0.2214	0.0296	64	3	0.0473	256	0.0978
	7200	0.0046	0.0011	180	40	0.0031	8192	0.0035
	32000	0.0004	0.0014	64	500	0.0001	16384	0.0051
S_3	192	1.7695	1.9798	64	3	0.1790	256	1.7778
	7200	0.5367	0.1111	64	500	0.3463	8192	0.0606
	32000	0.3250	0.1111	64	500	0.0581	16384	0.1111
$S_{x_1}^{tot}$	192	0.0923	0.0576	64	3	0.0503	256	0.0807
	7200	0.0061	0.0001	180	40	0.0061	8192	0.0006
	32000	0.0020	0.00004	64	500	0.0004	16384	0.0023
$S_{x_3}^{tot}$	192	1.1693	12.628	64	3	1.1013	256	0.5044
	7200	0.7529	0.8142	180	40	0.0895	8192	0.0619
	32000	0.2094	0.6814	64	500	0.0153	16384	0.0619

The following notation is used in the tables: $\#_{sub}$ is the number of subdomains after domain division, N_{sub} is the number of samples in each subdomain, N is the number of samples in the domain of integration (for the plain algorithm). For our experiments the desired estimate of standard deviation is chosen as $\varepsilon \in \{0.0165; 0.1; 0.5\}$. Relative error has been chosen as a main quantitative measure for comparison of approaches under consideration. The number of samples for Monte Carlo techniques has been chosen following the requirement for consistency of obtained results. The number of samples for plain Monte Carlo algorithm is a multiplication of average number of subdomains and number of samples in each subdomain specified for adaptive Monte Carlo algorithm. FAST approach is carried out via SIMLAB for the same corresponding number of samples and Sobol' approach via SIMLAB - for the nearest possible number.

The results for plain and adaptive approach show a numerical advantage for the adaptive algorithm - smaller relative error for a fixed number of samples. It can be expected according to variance reduction property of the adaptive procedure.

Table 3: Higher-order sensitivity indices of inputs estimated using SIMLAB Sobol' approach ($S_{12} = 0.00537$, $S_{13} = 0.000096$, $S_{23} = 0.00018$, $S_{123} = 0.000004$).

Est.	$N = 256$		$N = 8192$		$N = 16384$	
quant.	Est. value	Rel. error	Est. value	Rel. error	Est. value	Rel. error
S_{12}	0.0133	1.4767	0.0068	0.2663	0.0065	0.2104
S_{13}	0.0045	45.875	0.00007	0.2708	0.000072	0.25
S_{23}	0.0067	36.222	0.00021	0.1667	0.0003	0.6667
S_{123}	-0.034	7658.1	-0.0015	339.91	0.0006	135.36

The computational cost of Sobol' approach and FAST is similar and optimal as it has been mentioned before. The results in Table 2 show that the order of relative error is similar for the both cases except for small total sensitivity index - in our case $S_{x_3}^{tot}$. The relative error in computing $S_{x_3}^{tot}$ using FAST is larger than other ones. The adaptive Monte Carlo algorithm has an advantage over Sobol' approach of SIMLAB implementation for more specified data in Table 2. One more advantage of the proposed adaptive algorithm is that we can control number of samples and the accepted level of accuracy for it. In contrast to that number of samples is fixed for SIMLAB algorithms and the maximum number for Sobol' approach is comparatively small - 16384.

The results about estimate value and relative error in computing higher-order sensitivity indices using SIMLAB Sobol' approach are presented in Table 3. It can be seen that this kind of indices have comparatively small values and SIMLAB implementation of Sobol' approach gives acceptable order of approximation only for larger number of samples. This algorithm produces also non acceptable negative value for S_{123}.
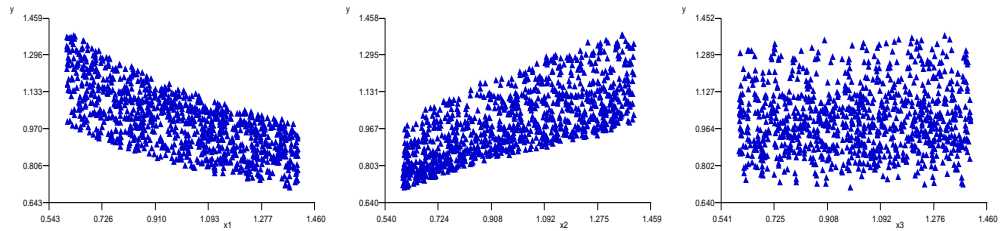

Figure 1: Scatter plots of chemical rate of third and twenty second time-dependent and sixth time-independent reaction (x_3) against ozone concentration (y).

The main conclusion about extent of sensitivity of the mathematical model under consideration according the inputs is that the main effect of input parameters dominates over higher-order interactions.

All scatter plots have been made for $N = 16384$ points. Figure 1a and Figure 1b show the scatter plot of the most important inputs against the model output. The dependence between model inputs and output can be clearly seen by that figures - inversely proportional at the first one and directly proportional at the second one. Figure 1c confirms inessential influence of the third input parameter on the model output.

4   Concluding remarks

The proposed adaptive Monte Carlo algorithm for numerical integration is one more approach to compute the main sensitivity measures - Sobol' global sensitivity indices. The experimental comparative analysis of relative error obtained applying various algorithms shows that the results of Sobol' algorithm from SIMLAB and adaptive Monte Carlo algorithm for all unknown quantities of interest are reliable. The advantages for the adaptive Monte Carlo algorithm can be found according to relative error as well as the possibility to control numerical parameters of the algorithm.

5   Acknowledgment

References

[1] R. Cukier, C. Fortuin, K. Shuler, A. Petschek, J. Schaibly, Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I. Theory, *J. Chem. Phys.* **59** (1973), 3873–3878.

[2] I.T. Dimov, R. Georgieva, Monte Carlo Adaptive Technique for Sensitivity Analysis of a Large-scale Air Pollution Model, *LNCS* **5910**, Springer (to appear).

[3] T. Ishigami, T. Homma, An Importance Quantification Technique in Uncertainty Analysis for Computer Models, in: *Proceedings of the ISUMA '90* (1990).

[4] A. Saltelli, Making Best Use of Model Valuations to Compute Sensitivity Indices, *Computer Physics Communications* **145** (2002), 280–297.

[5] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, Halsted Press, New York (2004).

[6] A. Saltelli, S. Tarantola, K. Chan, A Quantitative Model-independent Method for Global Sensitivity Analysis of Model Output. Source, *Technometrics archive* **41 (1)** (1999), 39–56.

[7] I. M. Sobol', Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates, *Mathematics and Computers in Simulation*, Vol. 55 (1-3) (2001), 271–280.

[8] I. Sobol', E. Myshetskaya, Monte Carlo Estimators for Small Sensitivity Indices, *Monte Carlo Methods and Applications*, Vol. 13 (5-6) (2007), 455–465.

[9] C. Tong, Self-validated Variance-based Methods for Sensitivity Analysis of Model Outputs, *Reliability Engineering & System Safety*, Vol. 95 (3) (2010), 301–309.

[10] Z. Zlatev, I.T. Dimov, *Computational and Numerical Challenges in Environmental Modelling*, Elsevier, Amsterdam (2006).

[11] SIMLAB Software Tool for Sensitivity Analysis: http://simlab.jrc.ec.europa.eu/.

Generalized Zakharov-Shabat Systems and Nonlinear Evolution Equations with Deep Reductions

Vladimir S. Gerdjikov, Nikolay A. Kostov, Tihomir I. Valchev

Abstract

We analyze the nonlinear evolution equations with $\mathbb{Z}_h$-reductions related to classical simple Lie algebras $\mathfrak{g}$, where h is the Coxeter number of $\mathfrak{g}$. We derive their recursion operators Λ and show that they possess factorization properties.

1 Introduction

Many nonlinear evolution equations integrable by means of the inverse scattering method (ISM) correspond to a Lax representation which has additional algebraic symmetries. In some cases, for example, affine Toda field theories [8]

$$\partial_{xt}^2 q_k = e^{q_{k+1}-q_k} - e^{q_k-q_{k-1}}, \qquad k = 1,\dots,n, \qquad e^{q_{n+1}} \equiv e^{q_1} \tag{1}$$

or $\mathbb{Z}_n$-derivative nonlinear Schrödinger equation [3, 10, 5]:

$$i\partial_t q_k + \gamma \coth \frac{\pi k}{n} \partial_{xx}^2 q_k + i\gamma \sum_{p=1}^{n-1} \partial_x (q_p q_{k-p}) = 0, \qquad k = 1,\dots,n \tag{2}$$

these symmetries are induced by a finite group of order bigger than 2 (in both cases above the group is $\mathbb{Z}_n$). Such equations are called deeply reduced. The purpose of this paper is to shed some light on how the basic notions and results of soliton theory are transferred to the case of deeply reduced equations. We stress on the interpretation of the ISM as a generalised Fourier transform which holds true also after the reductions are applied.

2 Direct and Inverse Scattering Problem

All nonlinear evolution equations under consideration here are integrable in the sense of inverse scattering method, i.e. they represent compatibility condition

$$[L(\lambda), M(\lambda)] = 0,$$

for some linear differential operators

$$L(\lambda) = i\partial_x + U(x,\lambda), \qquad M(\lambda) = i\partial_t + V(x,\lambda)$$

The functions U and V take values in a simple complex Lie algebra $\mathfrak{g}$ and they are chosen polynomial in the spectral parameter λ

$$U(x,\lambda) = q(x) - \lambda J, \qquad V(x,\lambda) = \sum_k V_k(x)\lambda^k. \tag{3}$$

The constant matrix J can be chosen as an element of the Cartan's subalgebra $\mathfrak{h} \subset \mathfrak{g}$ while q is expanded over Weyl's generators only

$$q(x) = \sum_{\alpha \in \Delta} q_\alpha(x) E_\alpha,$$

where Δ is the set of roots of $\mathfrak{g}$. We also assume that U and V are $\mathbb{Z}_h$-reduced

$$CU(\omega\lambda)C^{-1} = U(\lambda), \qquad CV(\omega\lambda)C^{-1} = V(\lambda), \tag{4}$$

where C realizes an action of Coxeter's automorphism ($C^h = \mathbb{1}$), $\omega = \exp(2\pi i/h)$ and h is Coxeter's number of $\mathfrak{g}$. Due to the reduction conditions (4) the eigenvalues of J are powers of ω. In the simplest case $\mathfrak{g} \sim \mathfrak{sl}(n)$ we have

$$J = \mathrm{diag}\,(1, \omega, \omega^2, \ldots, \omega^{h-1}).$$

The direct and inverse scattering problems for generalized Zakahrov-Shabat operators with complex-valued J and generic potentials $Q(x)$ has been solved in [1] for $\mathfrak{g} \simeq sl(n)$; for the other simple Lie algebras see [4]. Applying the $\mathbb{Z}_h$-reductions requires additional efforts, see [3, 6, 11].
The continuous spectrum of the $\mathbb{Z}_h$-reduced Lax operator L consists of $2h$ rays l_ν, $\nu = 1, \ldots, 2h$ defined by the requirement

$$\mathrm{Im}\,\lambda\alpha(J) = 0, \qquad \alpha \in \Delta, \qquad \lambda \in l_\nu \equiv |\lambda|e^{\pi i\nu/h}, \tag{5}$$

for some roots α. In the case when $\mathfrak{g} \sim \mathfrak{sl}(n)$ one can check that

$$\mathrm{Im}\,\lambda(J_j - J_k) = \mathrm{Im}\,\lambda\left(\omega^{j-1} - \omega^{k-1}\right) = 0, \qquad j \neq k, \quad j, k = 1, \ldots, n.$$

The subset of all roots to satisfy (5) on some ray l_ν will be denoted by δ_ν. The rays l_ν divide the complex plane $\mathbb{C}$ into $2h$ identical sectors Ω_ν. The discrete eigenvalues of L come in $2h$-tuples $\{\lambda_k^\pm \omega^s, \ s = 0, \ldots, h-1\}_{k=1}^m$ in a symmetric manner, i.e. there is a single eigenvalue in each sector.
There exists an ordering in Δ unique for every sector Ω_ν defined by

$$\Delta_\nu^\pm = \{\alpha \in \Delta; \mathrm{Im}\,\lambda\alpha(J) \gtrless 0, \quad \lambda \in \Omega_\nu.\} \tag{6}$$

The roots in Δ_ν^+ are called positive for Ω_ν and those in Δ_ν^- – negative. As a consequence each sector Ω_ν represents the analyticity region for a fundamental solution $\chi^\nu(x, \lambda)$. With each ray l_ν we also relate the subsets of roots δ_ν such that $\mathrm{Im}\,\lambda\alpha(J) = 0$ for $\lambda \in l_\nu$. The roots $\delta_\nu^+ = \Delta_\nu^+ \cap \delta$ can be viewed as the set of positive roots of a subalgebra $\mathfrak{g}_\nu \subset \mathfrak{g}$. By explicit construction one is able to prove that:

Lemma 1 ([3]) *Let C be the Coxeter automorphism of the simple Lie algebra $\mathfrak{g}$ and let the subalgebras $\mathfrak{g}_\nu$ and their sets of roots δ_ν are defined as above. Then: i) any two non-collinear roots $\alpha, \beta \in \delta_\nu$ are orthogonal, i.e. $(\alpha, \beta) = 0$, and as a consequence ii) $\mathfrak{g}_\nu$ is a direct sum of sl(2)-subalgebras.*

The explicit construction of $\chi^\nu(x, \lambda)$ resembles that in the nonreduced case

$$\chi^\nu(x, \lambda\omega) = \psi(x, \omega\lambda)T_\nu^\mp(\lambda)D_\nu^\pm(\lambda) = \phi(x, \lambda\omega)S_\nu^\pm(\lambda), \tag{7}$$

where the Gauss factors

$$S_\nu^\pm(\lambda) = \exp\left(\sum_{\beta\in\delta_\nu^+} s_{\nu,\beta}^\pm E_{\pm\beta}\right), \qquad D_\nu^+ = \exp\left(\sum_{j=1}^r d_{\nu,j}^+ H_j\right) \tag{8}$$

$$T_\nu^\pm(\lambda) = \exp\left(\sum_{\beta\in\delta_\nu^+} t_{\nu,\beta}^\pm E_{\pm\beta}\right), \qquad D_\nu^- = \exp\left(\sum_{j=1}^r d_{\nu,j}^- w_0(H_j)\right). \tag{9}$$

take values in the corresponding subgroups $\mathfrak{G}_\nu$

Remark 1 *Strictly speaking the Jost solutions ψ and ϕ **do not** admit analytic continuation. The product of the Jost solutions and the corresponding Gauss factors have analytic properties however [1, 4].*

Any pair of fundamental analytic solutions (FAS) with adjacent analyticity regions are connected via a local Riemman-Hilbert problem

$$\chi^{\nu+1}(x, \lambda) = \chi^\nu(x, \lambda)G^\nu(\lambda), \qquad \lambda \in l_\nu, G^\nu(\lambda) = \hat{S}_\nu^-(\lambda)S_\nu^+(\lambda). \tag{10}$$

The Coxeter's automorphism induces a grading in $\mathfrak{g}$.

$$\mathfrak{g} = \overset{h-1}{\underset{k=0}{\oplus}} \mathfrak{g}^{(k)}, \qquad \left[\mathfrak{g}^{(k)}, \mathfrak{g}^{(l)}\right] \subset \mathfrak{g}^{(k+l)} \tag{11}$$

The subspaces $\mathfrak{g}^{(k)}$ are eigenspaces of C

$$\mathfrak{g}^{(k)} = \left\{X \in \mathfrak{g}; CXC^{-1} = \omega^{-k}X\right\}. \tag{12}$$

The special property of the C is in the fact that $\mathfrak{g}^{(0)}$ is isomorphic to the Cartan subalgebra.
The $\mathbb{Z}_h$-symmetry imposes the following constraints on the FAS and on the scattering matrix and its factors:

$$\begin{aligned} C\chi^\nu(x, \lambda\omega)C^{-1} &= \chi^{\nu-2}(x, \lambda), & CT_\nu(\lambda\omega)C^{-1} &= T_{\nu-2}(\lambda), \\ CS_\nu^\pm(\lambda\omega)C^{-1} &= S_{\nu-2}^\pm(\lambda), & CD_\nu^\pm(\lambda\omega)C^{-1} &= D_{\nu-2}^\pm(\lambda), \end{aligned} \tag{13}$$

where the index $\nu - 2$ should be taken modulo $2h$. The independent scattering data is provided on two rays only, e.g. on l_1 and $l_{2h} \equiv l_0$.

3 Generalised Fourier transform

The generalised Fourier transform interpretation of the ISM is based on the notion of "squared solutions". The "squared solutions" are constructed by using FAS and the Cartan-Weyl basis of $\mathfrak{g}$ as follows [4]:

$$e_\beta^\nu(x,\lambda) = \chi^\nu E_\beta \hat{\chi}^\nu(x,\lambda), \qquad \boldsymbol{e}_\beta^\nu(x,\lambda) = P_{0J}(\chi^\nu E_\beta \hat{\chi}^\nu(x,\lambda)), \qquad (14)$$

where $P_{0J} = \mathrm{ad}_J^{-1}\mathrm{ad}_J$ is a projector onto the off-diagonal part of the corresponding matrix-valued function. Obviously $e_\beta^\nu(x,\lambda)$ satisfy the equation:

$$i\frac{de_\beta^\nu}{dx} + [Q - \lambda J, e_\beta^\nu(x,\lambda)] = 0. \qquad (15)$$

The presence of $\mathrm{ad}_J \cdot \equiv [J,\cdot]$ naturally splits the space of functions taking values in $\mathfrak{g}$ into two: the ones that lie in the kernel of ad_J and the rest $\mathfrak{M}_J$, for which $P_{0J}X \equiv X$. The squared solution $\boldsymbol{e}_\beta^\nu(x,\lambda)$ are a complete set of functions in $\mathfrak{M}_J$ [4, 6, 11]. Hence any function $X : \mathbb{R}^2 \to \mathfrak{g}/\ker\mathrm{ad}_J$ can be expanded over them. For example, the potential q and its variation admit the following expansions

$$q(x) = -\frac{i}{\pi}\sum_{\nu=0}^{h-1}(-1)^\nu \int_{l_\nu} d\lambda \sum_{\alpha\in\delta_\nu^+} \left(s_{\nu,\alpha}^+(\lambda)\boldsymbol{e}_\alpha^\nu(x,\lambda) - s_{\nu,\alpha}^-(\lambda)\boldsymbol{e}_{-\alpha}^\nu(x,\lambda)\right)$$

$$\mathrm{ad}_J^{-1}\delta q(x) = \frac{i}{\pi}\sum_{\nu=0}^{h-1}(-1)^\nu \int_{l_\nu} d\lambda \sum_{\alpha\in\delta_\nu^+} \left(\delta s_{\nu,\alpha}^+(\lambda)\boldsymbol{e}_\alpha^\nu(x,\lambda) + \delta s_{\nu,\alpha}^-(\lambda)\boldsymbol{e}_{-\alpha}^\nu(x,\lambda)\right).$$

Thus a one-to-one correspondence between the potential and the scattering data (as well as between their variations) occurs

$$q(x) \leftrightarrows \bigcup_{\nu=0}^{h-1}\{s_\alpha^{\nu,\pm}(\lambda),\ \alpha\in\delta_\nu,\ \lambda\in l_\nu\}, \quad \delta q(x) \leftrightarrows \bigcup_{\nu=0}^{h-1}\{\delta s_\alpha^{\nu,\pm}(\lambda),\ \alpha\in\delta_\nu,\ \lambda\in l_\nu\}.$$

In case the variations of scattering data and the potential are generated by the time flow

$$\delta t \longrightarrow \delta s^\pm(\lambda),\ \delta q$$

one obtains

$$\mathrm{ad}_J^{-1}Q_t = \frac{i}{\pi}\sum_{\nu=0}^{h-1}(-1)^\nu \int_{l_\nu} d\lambda \sum_{\alpha\in\delta_\nu^+} \left(s_{\nu,\alpha;t}(\lambda)\boldsymbol{e}_\alpha^\nu(x,\lambda) + s_{\nu,\alpha;t}(\lambda)\boldsymbol{e}_{-\alpha}^\nu(x,\lambda)\right).$$

4 Recursion operators

Here we introduce the recursion operators $\Lambda_\pm$ as the ones for which the squared solutions $\boldsymbol{e}_\alpha^\nu(x,\lambda)$ are eigenfunctions

$$\boldsymbol{\Lambda}_k^\pm \boldsymbol{e}_{\mp\alpha,k}^\nu(x,\lambda) = \lambda^h \boldsymbol{e}_{\mp\alpha,k}^\nu(x,\lambda), \qquad \alpha\in\Delta_\nu^+. \qquad (16)$$

In order to derive the explicit form of $\Lambda_\pm$ we will make use of equations (14), (15) and of the grading in $\mathfrak{g}$ introduced by C (11). It means that the squared solutions can be represented as the sum of its projections onto $\mathfrak{g}^{(k)}$:

$$e_\alpha^\nu(x,\lambda) = \sum_{k=0}^{h-1} e_{\alpha,k}^\nu(x,\lambda), \qquad e_{\alpha,k}^\nu(x,\lambda) \in \mathfrak{g}^{(k)}. \tag{17}$$

In addition we have to split each of the projections $e_{\alpha,k}^\nu(x,\lambda)$ into diagonal and off-diagonal parts:

$$e_{\alpha,k}^\nu(x,\lambda) = e_{\alpha,k}^{\nu,\mathrm{d}}(x,\lambda) + \boldsymbol{e}_{\alpha,k}^{\nu,\mathrm{f}}(x,\lambda),$$

This requires that we have to establish which of the linear subspaces $\mathfrak{g}^{(k)}$ have non-trivial section with $\mathfrak{h}$. To this end we make use of the explicit form of the Coxeter element C of the Weyl group and its eigenvectors. It is most effective to use the dihedral realization of C in the form:

$$C = w_0 w_1, \qquad w_0^2 = \mathbb{1}, \qquad w_1^2 = \mathbb{1}, \qquad C^h = \mathbb{1}.$$

Here we split the set of simple roots of $\mathfrak{g}$ into two subsets $R_0 \cup R_1$ such that each of the sets R_i contains only roots that are mutually orthogonal; then $w_i = \prod_{\alpha \in R_i} S_\alpha$ where S_α is the Weyl reflection with respect to the root α. Next we evaluate the action of C in the root space $\mathbb{E}^r$ and determine its eigenvectors:

$$C\vec{x}^{(k)} = \omega^{-m_k}\vec{x}^{(k)}, \qquad \omega = \exp(2\pi i/h).$$

The integers m_k, $k = 1,\ldots,r$ are called the exponents of $\mathfrak{g}$. The elements $H^{(k)}$ of the Cartan subalgebra $\mathfrak{h}$ that are dual to $\vec{x}^{(k)}$ obviously satisfy:

$$C(H^{(k)}) = \omega^{m_k} H^{(k)}, \qquad \text{i.e.} \qquad H^{(k)} \in \mathfrak{g}^{(m_k)}.$$

Using the properties of the Killing form $\langle X, Y\rangle$ on $\mathfrak{g}$ (see [7]) and normalizing the non-vanishing $H^{(m_k)}$ so that $\langle H^{(m_k)}, H^{(m_j)}\rangle = \delta_{m_k, h-m_j}$ we conclude that

$$e_{\alpha,m_k}^{\nu,\mathrm{d}}(x,\lambda) = H^{(m_k)} \left\langle H^{(h-m_k)}, e_{\alpha,m_k}^\nu(x,\lambda) \right\rangle, \tag{18}$$

Let for definiteness $\mathfrak{g} \simeq B_r, C_r$. Then $h = 2r$, $m_k = 2k-1$, $k = 1,\ldots,r$ and

$$\dim(\mathfrak{g}^{(2k-1)} \cap \mathfrak{h}) = 1, \qquad \dim(\mathfrak{g}^{(2k)} \cap \mathfrak{h}) = 0.$$

If we choose $J = H^{(m_1)}$, then $H^{(m_k)} = J^{m_k}$ and:

$$e_{\nu,\alpha}^{(2k)}(x,\lambda) \equiv \boldsymbol{e}_{\nu,\alpha}^{(2k)}(x,\lambda), \qquad e_{\alpha,m_k}^\nu(x,\lambda) = e_{\alpha,m_k}^{\nu,\mathrm{d}}(x,\lambda) + \boldsymbol{e}_{\alpha,m_k}^\nu(x,\lambda),$$

Thus from equation (15) we get:

$$\Lambda_{m_k}^\pm \boldsymbol{e}_{\alpha,m_k}^\nu(x,\lambda) = \lambda \boldsymbol{e}_{\alpha,m_k-1}^\nu(x,\lambda), \qquad \Lambda_0 \boldsymbol{e}_{\alpha,m_k+1}^\nu(x,\lambda) = \lambda \boldsymbol{e}_{\alpha,m_k}^\nu(x,\lambda), \tag{19}$$

55

where

$$\Lambda_{m_k}^{\pm} X(x) \equiv \mathrm{ad}_J^{-1} \left(i \frac{dX}{dx} + P_{0J}[Q(x), X(x)] \right.$$

$$\left. + i \left[Q(x), J^{m_k} \right] \int_{\pm\infty}^{x} dy \, \langle J^{h-m_k}, [Q(y), X(y)] \rangle \right), \qquad (20)$$

$$\Lambda_0 X(x) \equiv \mathrm{ad}_J^{-1} \left(i \frac{dX}{dx} + [Q(x), X(y)] \right).$$

Using equation (19) we get that for $\mathbb{Z}_h$-reduced systems the recursion operators factorize as follows:

$$\Lambda_{m_1}^{\pm} \Lambda_0 \Lambda_{m_2}^{\pm} \Lambda_0 \cdots \Lambda_{m_{r-1}}^{\pm} \Lambda_0 \Lambda_{m_r}^{\pm} \Lambda_0 \boldsymbol{e}_{\mp\alpha,0}^{\nu}(x,\lambda) = \lambda^h \boldsymbol{e}_{\mp\alpha,0}^{\nu}(x,\lambda),$$
$$\Lambda_0 \Lambda_{m_2}^{\pm} \Lambda_0 \Lambda_{m_3}^{\pm} \cdots \Lambda_0 \Lambda_{m_r}^{\pm} \Lambda_0 \Lambda_{m_1}^{\pm} \boldsymbol{e}_{\mp\alpha,m_1}^{\nu}(x,\lambda) = \lambda^h \boldsymbol{e}_{\mp\alpha,m_1}^{\nu}(x,\lambda), \qquad (21)$$

i.e.

$$\boldsymbol{\Lambda}_0^{\pm} = \Lambda_{m_1}^{\pm} \Lambda_0 \Lambda_{m_2}^{\pm} \Lambda_0 \cdots \Lambda_{m_r}^{\pm} \Lambda_0, \qquad \boldsymbol{\Lambda}_0^{\pm} = \Lambda_0 \Lambda_{m_2}^{\pm} \Lambda_0 \Lambda_{m_3}^{\pm} \cdots \Lambda_{m_r}^{\pm} \Lambda_0 \Lambda_{m_1}^{\pm} \qquad (22)$$

and similar expressions for the operators $\boldsymbol{\Lambda}_k^{\pm}$ with $k > 1$. These results generalize those presented in [6, 11] for the classical series of Lie algebras B_r and C_r. Similar, but more complicated factorizations exist also for D_r and for the exceptional Lie algebras.

5 Conclusions and perspectives

We have demonstrated that the generalized Fourier transforms hold true also for Lax operators with deep symmetries. The outlined above results could be used in deriving the fundamental properties of the corresponding NLEE with deep reductions. Another potential development of our results includes extending the theory to systems with non-commutative reduction groups (dihedral, octahedral etc., see [9].

Acknowledgements

References

[1] Beals R. and Coifman R., *Scattering and Inverse Scattering for First Order Systems: II*, Inv. Problems **3** (1987) 577–593;

Beals R. and Coifman R., *Linear Spectral Problems, Nonlinear Equations and the Delta-Method*, Inv. Problems **5** (1989) 87–130.

[2] Fordy A. and Gibbons J., *Factorization of Operators I. Miura Transformation*, J. Math. Phys. **21** (1980) 2508–2510.

[3] Gerdjikov V. S., In Nonlinear evolution equations: integrability and spectral methods, Ed. A. P. Fordy, A. Degasperis, M. Lakshmanan, MUP, (1981), p. 367–372.

[4] V. S. Gerdjikov, A. B. Yanovski. *Completeness of the eigenfunctions for the Caudrey–Beals–Coifman system.* J. Math. Phys. **35**, no. 7, 3687–3725 (1994).

[5] V. S. Gerdjikov, G. Vilasi, A. B. Yanovski. *Integrable Hamiltonian Hierarchies. Spectral and Geometric Methods* Lecture Notes in Physics **748**, Springer Verlag, Berlin, Heidelberg, New York (2008).

[6] G. G. Grahovski - On The Reductions and Scattering Data for the CBC System, In: "Geometry, Integrability and Quantization III", Eds: I. Mladenov and G. Naber, Coral Press, Sofia, 2002 pp.262-277;
G. G. Grahovski. On the Reductions and Scattering Data for the Generalized Zakharov-Shabat Systems, In Nonlinear Physics: Theory and Experiment. II, Eds: M.J.Ablowitz, M.Boiti, F.Pempinelli and B.Prinari, World Scientific, Singapore, 2003, pp.71-78.

[7] Helgasson S. , *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978

[8] Mikhailov A., *The Reduction Problem and The Inverse Scattering Method*, Physica D **3** (1981) 73-117.

[9] Mikhailov A., Lombardo S., *Reductions of integrable equations: dihedral group.* J. Phys. A: Math. Gen. **37**, (2004), 7727-7742.

[10] V. Drinfel'd, V. V. Sokolov. *Lie Algebras and equations of Korteweg - de Vries type.* Sov. J. Math. **30**, 1975–2036 (1985).

[11] Valchev T., *On generalized Fourier transform for Kaup-Kupershmidt type equations.* In Proc. 11-th International Conference on Geometry, Integrability and Quantization, June 5-11, 2009, Varna, Eds. I. Mladenov.

The Camassa-Holm Hierarchy and Soliton Perturbations

Georgi G. Grahovski, Rossen I. Ivanov

Introduction. Integrable equations are widely used as model equations in various problems. Such equations are in some sense exactly solvable, e.g. by the inverse scattering method (ISM) and exhibit global regular solutions.

In hydrodynamic context, even though water waves are expected to be unstable in general, they do exhibit certain stability properties in physical regimes where integrable model equations are accurate approximations for the evolution of the free surface water wave. The model equation is not integrable, but is somehow close to an integrable equation, i.e. can be considered as a perturbation of an integrable equation. In such case it is still possible to obtain approximate analytical solutions.

There two main approaches treating the perturbations of integrable equations: a "Direct Approach" and a "Spectral" Approach.

The direct approach is based on expanding the solutions of the perturbed equation around the unperturbed one, then the corrections due to perturbations are to be determined:

$$\tilde{u}(x,t) = u(x,t) + \Delta u(x,t).$$

Here $\tilde{u}(x,t)$ $(u(x,t))$ is the solution of the perturbed (unperturbed) nonlinear equation and $\Delta u(x,t)$ is a (small) perturbation. The strength of the perturbation is measured by a parameter ϵ: $\Delta u(x,t) = \mathcal{O}(\epsilon)$. By small (weak) perturbation one means $0 < \epsilon \ll 1$. Such perturbations can be studied *directly* in the configuration (coordinate) space, The effect of the perturbations on the scattering data can be studied in the *spectral space* of the associated spectral problem.

Several authors had used various versions of the direct approach in the study of soliton perturbations: D. J. Kaup [3] had used a similar approach for the perturbed sine-Gordon equation. Keener and McLaughlin [4] had proposed a direct approach by obtaining the appropriate Green functions for the nonlinear Schrodinger and sine-Gordon equations. For a comprehensive review of the direct perturbation theory see e.g. [2, 5] and the references therein.

In the spectral space, the study of the soliton perturbations is based on the perturbations of the scattering data, associated to the spectral problem. It is based on the use of the expansions of the "potential" $u(x,t)$ of the associated spectral problem over the complete set of "squared solutions" (the eigenfunctions of the corresponding recursion operator). Such methods are used by a number of authors, for studying perturbations of various nonlinear evolutionary equations: the sin-Gordon equation – [6]; the nonlinear Schrödinger equation – [7, 8, 9]; etc.

The completeness of the squared eigenfunctions helps to describe the ISM for the corresponding hierarchy as a GFT. The role of the Fourier modes for the GFT is played by the Scattering data of the associated Lax operator. The GFT provides a natural setting for the analysis of small perturbations to an integrable equation: The leading idea is that starting from a purely soliton solution of a certain integrable equation one can 'modify' the soliton parameters such as to incorporate the changes

caused by the perturbation. There is a contribution to the equations for the scattering data that comes from the GFT-expansion of the perturbation.

2. The Camassa-Holm Hierarchy: An Overview. Closely related to the KdV hierarchy is the hierarchy of the Camassa-Holm (CH) equation [10]. This equation has the form

$$u_t - u_{xxt} + 2\omega u_x + 3uu_x - 2u_x u_{xx} - uu_{xxx} = 0, \qquad (1)$$

where ω is a real constant. It is integrable with a Lax pair [10]

$$\Psi_{xx} = \left(\frac{1}{4} + \lambda(m + \omega)\right)\Psi, \qquad \Psi_t = \left(\frac{1}{2\lambda} - u\right)\Psi_x + \frac{u_x}{2}\Psi + \gamma\Psi, \qquad (2)$$

where $m \equiv u - u_{xx}$ and γ is an arbitrary constant.

Both CH and KdV equations appeared initially as models of the propagation of two-dimensional shallow water waves over a flat bottom. The solitary waves of KdV are smooth solitons, while the solitary waves of CH, which are also solitons, are smooth if $\omega > 0$ and peaked (called "peakons" and representing weak solutions), if $\omega = 0$.

The problem of perturbation of the CH equation arises when one deals with model equations that are in general non-integrable but close to the CH equation. A perturbation could appear for example when one takes into account the viscosity effect.

Another possible scenario comes from the so-called "b-equation" that also is a model of shallow water waves [11] :

$$m_t + b\omega u_x + bmu_x + m_x u = 0.$$

It generalizes the CH equation and is integrable only for: A) $b = 2$ – Camassa-Holm equation; B) $b = 3$ – Degasperis-Procesi equation.

The solutions of the b-equation for values of b close to $b = 2$ can be analyzed in the framework of the CH-perturbation theory. The "b-equation" can be casted into a form of a CH perturbation

$$m_t + 2\omega u_x + 2mu_x + m_x u = (2 - b)(\omega u_x + mu_x) \equiv P[u],$$

for a small parameter $\epsilon = b - 2$.

3. Inverse Scattering Method and Generalised Fourier Transform for the Camassa-Holm Hierarchy. For simplicity we consider the case where m is a Schwartz class function, $m(x) \in \mathcal{S}(\mathbb{R})$. For simplicity we use the notation $q = u - u_{xx} + \omega$; $q(x, t) > 0$ for all t if $q(x, 0) > 0$. Let $k^2 = -\frac{1}{4} - \lambda\omega$, i.e. $\lambda(k) = -\frac{1}{\omega}\left(k^2 + \frac{1}{4}\right)$. The continuous spectrum in terms of k corresponds to k – real. The discrete spectrum consists of finitely many points $k_n = i\kappa_n$, $n = 1, \ldots, N$ where κ_n is real and $0 < \kappa_n < 1/2$. The continuous spectrum vanishes for $\omega = 0$, [12].

The Jost solutions of the spectral problem $f^\pm(x, k)$ are fixed by their asymptotic when $x \to \pm\infty$ for all real $k \neq 0$:

$$\lim_{x \to \infty} e^{\mp ikx} f^\pm(x, k) = 1,$$

59

and moreover at the points of the continuous spectrum

$$f^-(x,k) = a(k)f^+(x,-k) + b(k)f^+(x,k), \qquad \text{Im } k = 0.$$

The scattering coefficient $a(k)$ has an analytic continuation in the upper complex plane. At the points $i\kappa_n$ of the discrete spectrum, $a(k)$ has simple zeroes, $a(i\kappa_n) = 0$. Then $f^-(x, i\kappa_n) = b_n f^+(x, i\kappa_n)$ for some coefficient b_n.

The quantities $\mathcal{R}^\pm(k) = b(\pm k)/a(k)$ are known as reflection coefficients. One can define their analogues $R_n^\pm = \frac{b_n^{\pm 1}}{i\dot{a}_n}$ at the points of the discrete spectrum $k = ik_n$ [13] where $\dot{a}_n = [\frac{\partial}{\partial k}a(k)]_{k=k_n}$.

With the asymptotics of the Jost solutions and (2) one can show that

$$L_\pm F^\pm(x,k) = \frac{1}{\lambda}F^\pm(x,k) \qquad L_\pm F_n^\pm(x) = \frac{1}{\lambda_n}F_n^\pm(x), \tag{3}$$

where $\lambda_n = \lambda(i\kappa_n)$; $F^\pm(x,k) \equiv (f^\pm(x,k))^2$, $F_n^\pm(x) \equiv F(x, i\kappa_n)$ are the squares of the Jost solutions and

$$L_\pm = (\partial^2 - 1)^{-1}\Big[4q(x) - 2\int_{\pm\infty}^x \mathrm{d}y\, m'(y)\Big]\cdot \tag{4}$$

is the recursion operator. The inverse of this operator is also well defined [13]. The completeness relation for the eigenfunctions of the recursion operator is [13]

$$\frac{\omega}{\sqrt{q(x)q(y)}}\theta(x-y) = -\frac{1}{2\pi i}\int_{-\infty}^\infty \frac{F^-(x,k)F^+(y,k)}{ka^2(k)}\mathrm{d}k$$

$$+ \sum_{n=1}^N \frac{1}{i\kappa_n \dot{a}_n^2}\Big[\dot{F}_n^-(x)F_n^+(y) + F_n^-(x)\dot{F}_n^+(y) - \Big(\frac{1}{i\kappa_n} + \frac{\ddot{a}_n}{\dot{a}_n}\Big)F_n^-(x)F_n^+(y)\Big]. \tag{5}$$

where $\dot{F}_n^\pm(x) \equiv [\frac{\partial}{\partial k}F^\pm(x,k)]_{k=k_n}$, etc. Therefore $F^\pm$, $F_n^\pm$ and $\dot{F}_n^\pm$ can be considered as 'generalised' exponents. It is possible to expand $m(x)$ and its variation $\delta m(x)$ over the above mentioned basis, or rather the quantities depending on $q(x)$ (which are determined by $m(x)$) [13]:

$$\omega\Big(\sqrt{\frac{\omega}{q(x)}} - 1\Big) = \pm\frac{1}{2\pi i}\int_{-\infty}^\infty \frac{2k\mathcal{R}^\pm(k)}{\lambda(k)}F^\pm(x,k)\mathrm{d}k + \sum_{n=1}^N \frac{2\kappa_n}{\lambda_n}R_n^\pm F_n^\pm(x); \tag{6}$$

$$\frac{\omega}{\sqrt{q(x)}}\int_{\pm\infty}^x \delta\sqrt{q(y)} \quad \mathrm{d}y = \frac{1}{2\pi i}\int_{-\infty}^\infty \frac{i}{\lambda(k)}\delta\mathcal{R}^\pm(k)F^\pm(x,k)\mathrm{d}k$$

$$\pm \quad \sum_{n=1}^N \Big[\frac{1}{\lambda_n}(\delta R_n^\pm - R_n^\pm\delta\lambda_n)F_n^\pm(x) + \frac{R_n^\pm}{i\lambda_n}\delta\kappa_n\dot{F}_n^\pm(x)\Big] \tag{7}$$

The expansion coefficients are given by the scattering data and their variations. This makes evident the interpretation of the ISM as a generalized Fourier transform. Now

it is straightforward to describe the hierarchy of Camassa-Holm equations. To every choice of the function $\Omega(z)$, known also as the dispersion law we can put into correspondence the nonlinear evolution equation (NLEE) that belongs to the Camassa-Holm hierarchy:

$$\frac{2}{\sqrt{q}} \int_{\pm\infty}^{x} (\sqrt{q})_t \mathrm{d}y + \Omega(L_\pm)\left(\sqrt{\frac{\omega}{q}} - 1\right) = 0. \tag{8}$$

An equivalent form of the equation is

$$q_t + 2q\tilde{u}_x + q_x\tilde{u} = 0, \qquad \tilde{u} = \frac{1}{2}\Omega(L_\pm)\left(\sqrt{\frac{\omega}{q}} - 1\right). \tag{9}$$

The choice $\Omega(z) = z$ leads to $\tilde{u} = u$ and thus to the CH equation (1). Other choices of the dispersion law and the corresponding equations of the Camassa-Holm hierarchy are discussed in [13]. The CH equation is equivalent to the following linear evolution equations for the scattering data:

$$\mathcal{R}_t^\pm(k) \mp ik\Omega(\lambda^{-1})\mathcal{R}^\pm(k) = 0, \tag{10}$$

$$R_{n,t}^\pm \pm \kappa_n\Omega(\lambda_n^{-1})R_n^\pm = 0, \tag{11}$$

$$\kappa_{n,t} = 0. \tag{12}$$

The time-evolution of the scattering data for the CH equation (1) can be computed from the above formulae for $\Omega(z) = z$,

4. Perturbation Theory for for the Camassa-Holm Hierarchy. Let us start with a perturbed equation of the CH hierarchy of the form

$$q_t + 2q\tilde{u}_x + q_x\tilde{u} = P[u], \qquad \tilde{u} = \frac{1}{2}\Omega(L_\pm)\left(\sqrt{\frac{\omega}{q}} - 1\right), \tag{13}$$

where again, $P[u]$ is a small perturbation, by assumption in the Schwartz-class. It is useful to write (13) in the form

$$\frac{2}{\sqrt{q}} \int_{\pm\infty}^{x} (\sqrt{q})_t \mathrm{d}y + \Omega(L_\pm)\left(\sqrt{\frac{\omega}{q}} - 1\right) = \frac{1}{\sqrt{q}} \int_{\pm\infty}^{x} \frac{P(y)}{\sqrt{q(y)}} \mathrm{d}y. \tag{14}$$

With the completeness relation (5) one can deduce the gereralised Fourier expansion for expressions, like the one on the right-hand side of (13).

Theorem: *Assuming that f^+ and f^- are not linearly dependent at $x = 0$ and $g(x) \in \mathcal{S}(\mathbb{R})$, the following expansion formulas hold:*

$$\frac{\omega}{\sqrt{q}} \int_{\pm\infty}^{x} \frac{g(y)}{\sqrt{q(y)}} \mathrm{d}y = \pm \frac{1}{2\pi i} \int_{-\infty}^{\infty} \tilde{g}^\pm(k) F_x^\pm(x,k) \mathrm{d}k$$

$$\mp \sum_{j=1}^{N} \left(g_{1,j}^\pm \dot{F}_{j,x}^\pm(x) + g_{2,j}^\pm F_{j,x}^\pm(x) \right), \tag{15}$$

and the Fourier coefficients are

$$\tilde{g}^{\pm}(k) = \frac{1}{ka^2(k)}\left(g, F^{\mp}\right),$$

$$g_{1,j}^{\pm} = \frac{1}{k_j \dot{a}_j^2}\left(g, F_j^{\mp}\right),$$

$$g_{2,j}^{\pm} = \frac{1}{k_j \dot{a}_j^2}\left[\left(g, \dot{F}_j^{\mp}\right) - \left(\frac{1}{k_j} + \frac{\ddot{a}_j}{\dot{a}_j}\right)\left(g, F_j^{\mp}\right)\right],$$

where $(g, F) \equiv \int_{-\infty}^{\infty} g(x)F(x)\mathrm{d}x$.

The substitution of the expansions (15) for $P[u]$, (6) and (7) into the perturbed equation (14) gives the following expressions for the modified scattering data:

$$\mathcal{R}_t^{\pm} \mp ik\Omega(1/\lambda)\mathcal{R}^{\pm} = \mp\frac{i\lambda(P, F^{\mp})}{2ka^2(k)}, \qquad (16)$$

$$k_{j,t} = \frac{\lambda_j(P, F_j^{\mp})}{2k_j \dot{a}_j^2 R_j^{\pm}} \qquad (17)$$

$$R_{j,t}^{\pm} - R_j^{\pm}\lambda_{j,t} \pm \kappa_j\Omega(1/\lambda_j)R_j^{\pm}$$
$$= -\frac{\lambda_j}{2k_j \dot{a}_j^2}\left[(P, \dot{F}_j^{\mp}) - \left(\frac{1}{k_j} + \frac{\ddot{a}_j}{\dot{a}_j}\right)(P, F_j^{\mp})\right], \qquad (18)$$

From (18) we obtain the following for the coefficient b_j:

$$b_{j,t} + \kappa_j\Omega(1/\lambda_j)b_j = -\frac{\lambda_j}{4\kappa_j \dot{a}_j}\left(P, b_j^2 \dot{F}_j^+ - \dot{F}_j^-\right).$$

The 'perturbed' solution for the hierarchy in the adiabatic approximation can be recovered from the following expansion for $\tilde{u}(x)$ with the 'modified' scattering data keeping the unperturbed 'generalised' exponents:

$$\tilde{u}(x) = \pm\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{k\Omega(1/\lambda(k))}{\omega\lambda(k)}\mathcal{R}^{\pm}(k)F^{\pm}(x, k)\mathrm{d}k + \sum_{n=1}^{N}\frac{\kappa_n\Omega(1/\lambda_n)}{\omega\lambda_n}R_n^{\pm}F_n^{\pm}(x).$$

This formula follows from the second part of (9) and (6). Note that for the CH equation (1) $\tilde{u} \equiv u$.

5. Conclusions. In our derivations we used completeness relations that are valid only given the assumption that the Jost solutions f^+ and f^- are linearly independent at $x = 0$. The case when this condition is not satisfied is quite exceptional, however this is exactly the case when one has purely soliton solution. Then one has to take into account a nontrivial contribution from the scattering data at $k = 0$ and some of the presented results require modification. This means that no shelf is formed behind the soliton.

The presence of shelf for KdV equation is observed e.g. under the perturbation $P[u] = \epsilon u$ [14]. The evaluation of the perturbation terms for the CH hierarchy could

be technically difficult due to the complicated form of the CH multisoliton solutions [15]. However the limit $\omega \to 0$ leads to the relatively simple peakon solutions.

We end up with listing some open problems: 1) Using the presented general formulae to study the perturbations of the peakon parameters; 2) To study the soliton interactions of pure solitons and peakons for CH.

References

[1] G. G. Grahovski, R. I. Ivanov, Discr. Cont. Dyn. Syst. B **12** (2009), no. 3, 579 – 595.

[2] A. Degasperis, *Multiscale Expansion and Integrability of Dispersive Wave Equations*, In: "Integrability" (ed. A.V. Mikhailov), Lect. Notes in Phys. **767**, Springer, Berlin Heidelberg (2009), 215–244.

[3] D.J. Kaup, Phys. Rev. B **27** (1983), 6787–6795.

[4] J. P. Keener and D. W. McLaughlin, J. Math. Phys. **18** (1977) 2008–2013.

[5] R. L. Herman, J. Phys. A **23** (1990), 2327–2362.

[6] D.J. Kaup, *Solitons as particles and the effects of perturbations*, In: "The Significance of Nonlinearity in the Natural Sciences" (eds. B. Kursunoglu, A. Perlmutter and L.F. Scott), Plenum Press, New York (1977), 97–117.

[7] J. P. Keener and D. W. McLaughlin, Phys. Rev. A **16** (1977) 777–790.

[8] V.S. Gerdjikov and M.I. Ivanov, Inv. Problems **8** (1992), 831–847.

[9] V.S. Gerdjikov, Theoret. and Math. Phys. **99** (1994), 593–598.

[10] R. Camassa and D. Holm, Phys. Rev. Lett. **71** (1993), 1661–1664.

[11] R.I. Ivanov, Philos. Trans. R. Soc. Lond. Ser. A: Math. Phys. Eng. Sci. **365** (2007), 2267–2280.

[12] A. Constantin and H.P. McKean, Comm. Pure Appl. Math. **52** (1999), 949–982.

[13] A. Constantin, V. Gerdjikov and R. Ivanov, Inverse Problems **23** (2007), 1565–1597.

[14] V.I. Karpman, Phys. Scripta **20** (1979), 462–478.

[15] A. Constantin, Proc. Roy. Soc. London A **457** (2001), 953–970.

Regularity and Decay Issues for Compactly Supported Solitary Waves

Todor Gramchev

The main goal of the talk is to present some recent results, obtained in collaboration with G. Gaeta (Università di Milano) and S. Walcher (University of Aaachen), on the regularity (smoothness) of solitary waves near non Lipschitz equilibrium points.

We start by recalling some basic facts on solitary (travelling waves). The first documentation of the existence of shallow water waves appeared in 1834 when J. Scott Russell wrote one of the most cited papers about what later became known as soliton theory. Russell observed propagation of a solitary wave in the Glasgow-Edinburgh canal. In 1895 Korteweg and De Vries derived an equation describing shallow water waves, and gave the following interpretation of the solitary wave of Scott Russell. Ignoring some relevant physical aspects and simplifying parameters, we may write for short the KdV equation as

$$v_t + 2vv_x + v_{xxx} = 0, \tag{1}$$

where t is the time variable, x the point in the canal, $v(x,t)$ the height of the water (let us address to [3, 27, 41] for a much more detailed presentation). Looking for a solitary wave solution, travelling forward with velocity $V > 0$, we impose $v(t,x) = u(x - ct)$ in (1) and we obtain

$$\frac{d}{dx}(-cu + u^2 + u'') = 0$$

hence $u(x)$ satisfies $u'' - cu + u^2 = \text{const}$. Assuming further $\text{const} = 0$, we are reduced to solve

$$Pu = u'' - cu + u^2 = 0, \tag{2}$$

sometimes called Newton equation. Equation (2) possesses explicit solutions in terms of special functions. If we impose $u(x) \to 0$ for $x \to \pm\infty$, we obtain simply translations of the function

$$u(x) = \frac{\frac{3}{2}c}{\text{Ch}^2\left(\frac{\sqrt{c}}{2}x\right)}, \tag{3}$$

where

$$\text{Ch}\, t = \frac{e^t + e^{-t}}{2}.$$

We emphasize two properties of $u(x)$ in (3): first, it can be extended as analytic function in a strip of the form $\{z \in \mathbb{C} : |\Im z| < a\}$ in the complex plane. Second property is the exponential decay for $x \to \pm\infty$.

It seems that the first detailed investigations on the analytic regularity on $\mathbb{R}$ of travelling waves have been done by Bona and Li [3] (see also [2], [6], [7] for further studies and generalizations). After KdV equation, several related models were proposed. In particular recently, the theory of the solitary waves had impressive developments,

both concerning applicative aspects and mathematical analysis. Let us mention applications to internal water waves, nerve pulse dynamics, ion-acoustic waves in plasma, population dynamics, etc. In this order of ideas, we observe in particular that during the years 1990-2000, several papers were devoted to 5th order and 7th order generalization of KdV, see [27], Chapter 1. Let us emphasize that, to reach the exponential decay, the boundedness of $u(x)$ is not sufficient as initial assumption. We shall express later a precise threshold in terms of Sobolev estimates; as counter-example, consider here the celebrated Burger's equation (1948):

$$v_t + v_{xx} + 2vv_x = 0. \tag{4}$$

Imposing $v(t, x) = u(x - ct)$ and arguing as before we obtain the Verhulst equation

$$u' - cu + u^2 = 0. \tag{5}$$

It admits the bounded solution

$$u(x) = \frac{c}{1 + e^{-cx}}. \tag{6}$$

Assuming $c > 0$, we have exponential decay only for $x \to -\infty$, whereas $u(x) \to c \neq 0$ as $x \to +\infty$.

Next, we consider travelling waves for the KdV-Burgers (KdVB) equation. We follow the notation of Jeffrey and Mohamad [23], namely, the KdVB equation is written as

$$v_t + 2avv_x + 5bv_{xx} + cv_{xxx} = 0,$$

with $b < 0$, i.e. we have a dissipative term. The authors obtain the travelling wave type solution

$$v(x, t) = (3b^2/(2ac))\{\operatorname{sech}^2(\xi/2) \pm 2\tanh(\xi/2) \pm 2\},$$

where $\xi = \pm(b/c)(x \mp (6b^2/c)t)$.

Next, we describe briefly the findings of Gaeta, Gramchev and Walcher [14].
Solitons – both dynamical and topological – are among the main concepts organizing our understanding of nonlinear phenomena [5, 22, 11, 16, 21, 28, 42].
Solitons are highly local objects, i.e. they differ substantially from a trivial solution (a vacuum in field theoretic language) only in a small region of space. They are however smooth solutions, and as such the approach to trivial states can only be asymptotic, albeit very fast (exponential).
It was shown by Rosenau and Hyman [34] that completely localized (non-smooth) solitons can also exist; they have compact support – i.e. they differ from trivial solutions only in a compact region of space – and were hence termed **compactons**.
In the same way as one can have solitary waves which are not necessarily solitons in proper mathematical sense (see e.g. [5] for a discussion), but in many cases are physically equally interesting, one can have **compact solitary waves** which are not necessarily compactons in proper mathematical sense, but are equally interesting

physically (a large part of physical literature on compactons deals, in strict mathematical sense, with compact solitary waves). We also note that in the same way as one can have *topological solitons* [22, 11] and topological solitary waves, one can have **topological compactons** (that is, topological solitons with compact support; for theories with degenerate vacua, the support is the region in which the field is not in any of them) and topological compact solitary waves.

The possible relevance of compactons for applications as well as for theoretical developments need not be emphasized; here we only quote some papers dealing with their theory and applications [1, 9, 10, 12, 15, 17, 18, 19, 29, 30, 31, 32, 33, 34, 35, 38, 39]. We also mention that certain theories exhibit an intriguing duality between solitons and compactons [24].

We can consider and classify the regularity of **multi-compactons**; from the mathematical point of view, these display an intriguing phenomenon of non-uniqueness in the reduction from the second order Newton equation to the first order ODE expressing conservation of energy.

$$\dot{x}^2(t) \equiv 2(E - U(x(t)), \qquad\qquad (CE)$$

In $U \in C^2$ the point of equilibrium x_0,

$$U(x_0) = E, \qquad U'(x_0) = 0$$

is the unique solution of

$$\dot{x} = \pm W(x) = \pm\sqrt{E - U(x)}$$

and may give rise to homoclilinic or heteroclinic solutions connecting points of equilibrium for infinite time. We observe the the C^2 smoothness of the potential implies that under the non degeneracy assumption $U''(x_0) > 0$, that $E - U(x) \approx \frac{1}{2}U''(x_0)(x-x_0)^2$. The final part of the talk will deal with the regularity of the solutions of we dwell upon the regularity of the solutions (compactons) of (CE) near x_0, provided the $\sqrt{E - U(x)}$ is not Lipschitz near the equilibrium point x_0 which ma lead to nonuniqness of the solution (cf. [26, 40] for more details).

On physical basis, solutions with strictly compact sup- port are quite attractive in many respects both when dealing with very long polymers (as DNA, which was the first motivation for our investigation in this field), where one would not like the solution at a point x to affect (no matter how little) regions of the polymer too far away, and when dealing with field theory, where Relativity re- quires far away regions are not interacting before signals travelling at speed of light had time to connect them. On the other hand, in many physical situations one would like to deal with C^∞ smooth solutions or field con- figurations.

It is thus natural albeit apparently overly optimistic to wonder if within the class of problems studied in [9] one could have potentials leading to SWS which are both compactly supported and C smooth. Quite surprisingly, this turns out to be possible (need- less to say, for rather special potentials), as we can show by explicit example.

We stress these special models and solutions could be of substantial applicative interest, e.g. in transmission lines or optical fibers engineering: in this framework compactly supported signals have obvious advantages (including avoiding any interference between tails of different signal packets), and on the other hand infinitely smooth signals are easier to handle when they must be detected, analyzed and transformed. On the other hand, they would also present a purely theoretical interest from the point of view of Field Theory, being smooth field configurations with particle properties (and, in an obvious sense, stronger particle properties than ordinary solitons).

References

[1] H. Arodz, "Topological compactons", *Acta Phys. Polon.*, **B 33** (2002), no. 5, 1241–1252

[2] H.A. Biagioni and T. Gramchev, "Fractional derivative estimates in Gevrey spaces, global regularity and decay for solutions to semilinear equations in $\mathbb{R}^n$", *J. of Differential Equations*, **194** (2003), 140-165.

[3] J.L. Bona and Y.A. Li, "Decay and analyticity of solitary waves", *J. Math. Pures Appl.* **76** (1997), 377-430

[4] M. Cadoni, R. De Leo and G. Gaeta, "A composite model for DNA torsion dynamics", preprint *q-bio.BM/0604014* (2006)

[5] F. Calogero and A. Degasperis, *Spectral transform and solitons*, North Holland (Amsterdam) 1982

[6] M. Capiello, T. Gramchev and L. Rodino, "Super-exponential decay and holomorphic extensions for semilinear equations with polynomial coefficients", *J. Funct. Anal.* **237** (2006), 634-654

[7] M. Cappiello, T. Gramchev and L. Rodino, "Semilinear pseudo-differential equations and travelling waves", Selected papers of the ISAAC workshop, Toronto, Canada, December 11–15, 2006. Providence, RI: American Mathematical Society (AMS); Toronto, *Fields Institute Communications*, **52**, 213–238 (2007)

[8] M. Destrade, G. Gaeta and G. Saccomandi, "Weierstrass's criterion and compact solitary waves", *Phys. Rev. E* **75** (2007), 047601 (4 pp)

[9] M. Destrade and G. Saccomandi, "Solitary and compactlike shear waves in the bulk of solids", *Phys. Rev. E* **73** (2006), 065604(R)

[10] M. Destrade and G. Saccomandi, "Nonlinear transverse waves in deformed dispersive solids", preprint 2006

[11] A. Dubrovin, S.P. Novikov and A. Fomenko, *Modern geometry*, Springer (Berlin) 1984

[12] S. Dusuel, P. Michaux and M. Remoissenet, "From kinks to compactonlike kinks", *Phys. Rev. E* **57** (1998), 2320ï¿½2326

[13] G. Gaeta: "Compact travelling waves and non-smooth potentials", *EuroPhys. Lett.* **79** (2007), 20003 (6 pp)

[14] G. Gaeta, T. Gramchev and S. Walcher, "Compact solitary waves in linearly elastic chains with non-smooth on-site potential", *J. Phys. A* **40** (2007), 4493-4509

[15] A.V. Gorbach and S. Flach, "Compactlike dicrete breathers in systems with nonlinear and nonlocal dispersive terms", *Phys. Rev. E* **72** (2005), 056607

[16] I.D. Iliev, E.K. Khristov and K.P. Kirchev, *Spectral methods in soliton theory* (with an appendix by A. Yanovski), Wiley (New York) 1994

[17] P.G. Kevrekidis, V.V. Konotop, A.R. Bishop and S. Takeno, "Discrete compactons: some exact results", *J. Phys. A* **35** (2002), L641-L652

[18] Y.S. Kivshar, "Solitons in a nonlinear elastic medium", *Phys. Rev. B*, **43** (1991), 3493-3499

[19] Yuri S. Kivshar, "Intrinsic localized modes as solitons with a compact support", *Phys. Rev. E* **48** (1993), R43-R45

[20] L.D. Landau and E.M. Lifshitz, *Mechanics*, Pergamon, London 1960

[21] T.D. Lee and Y. Pang, "Nontopological solitons", *Phys. Rep.* **221** (1992) 251-350

[22] T. Dauxois and M. Peyrard, *Physique de solitons*, Editions CNRS (Paris) 2004; *Physics of Solitons*, Cambridge UP (Cambridge) 2006

[23] A. Jeffrey and M.N.B. Mohamad, *Exact solutions to the KdV-Burgers' equation*, Wave Motion **14** (1991) n. 4, 369–375.

[24] P.J. Olver and P. Rosenau, "Tri-Hamiltonian duality between solitons and solitary-wave solutions having compact support", *Phys. Rev. E* **53** (1996), 1900ï¿½1906

[25] M. Peyrard, "Nonlinear dynamics and statistical physics of DNA", *Nonlinearity* **17** (2004) R1-R40

[26] I.G. Petrovski, *Ordinary differential equations*, Dover, New York 1973

[27] A.V. Porubov, *Amplification of nonlinear strain waves in solids*, World Scientific, Singapore, 2003.

[28] M. Remoissenet, *Waves called solitons: concepts and experiments*, Springer, Berlin 1996.

[29] P. Rosenau, "Nonlinear dispersion and compact structures", *Phys. Rev. Lett.* **73** (1994), 1737ï¿½1741

[30] P. Rosenau, "On solitons, compactons, and Lagrange maps", *Phys. Lett. A* **211** (1996), 265-275

[31] P. Rosenau, "On a class of nonlinear dispersive-dissipative interactions", *Physica D* **123** (1998), 525-546

[32] P. Rosenau, "On a model equation of traveling and stationary compactons", *Phys. Lett. A*, **356** (2006), 44-50

[33] P. Rosenau, "Compactification of patterns by a singular convection or stress", *Phys. Rev. Lett.* **99** (2007), 234102 (4 pages)

[34] P. Rosenau and J.M. Hyman, "Compactons: Solitons with finite wavelength", *Phys. Rev. Lett.* **70** (1993), 564-567

[35] P. Rosenau and A. Pikovsky, "Phase compactons in chains of dispersively coupled oscillators", *Phys. Rev. Lett.* **94** (2005), 174102

[36] G. Saccomandi, "Elastic rods, Weierstrass' theory and special travelling waves solutions with compact support", *Int. J. Nonlin. Mech.* **39** (2004), 331-339

[37] G. Saccomandi and I. Sgura, "The relevance of nonlinear stacking interactions in simple models of double-stranded DNA", *J. Royal Soc. Interface* **3** (2006), 655-667

[38] S. Takeno, "Compacton-like modes in model DNA systems and their bearing on biological functioning", *Phys. Lett. A* **339** (2005), 352-360

[39] P. Tchofo Dinda and M. Remoissenet, "Breather compactons in nonlinear Klein-Gordon systems", *Phys. Rev. E* **60** (1999), 6218ï¿½6221

[40] W. Walter, *Ordinary differential equations*, Springer, Berlin 1998

[41] G.B. Whitham, *Linear and nonlinear waves*, J. Wiley and Sons Inc., New York, 1974

[42] V. Zakharov, *What is integrability ?*, Springer, Berlin 1991

Nonlinear Dynamics and Complete Stability of Cellular Nonlinear Networks

Victoria Ivanova

An oscillatory Cellular Nonlinear Network (CNN) with two cells using $M \times N = 1 \times 2$ with feedback synaptic weights $a_{0,-1} = \beta; a_{0,0} = \alpha$ and $a_{0,1} = \beta$, can be represented by the signal flow graph show in Fig.1
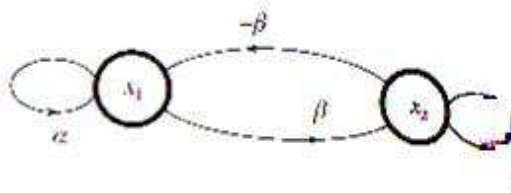
Fig.1.

The state equation for this CNN is given by:

$$\begin{aligned} \dot{x}_1 &= -x_1 + \alpha y_1 - \beta y_2 \\ \dot{x}_2 &= -x_2 + \alpha y_2 + \beta y_1, \end{aligned} \tag{1}$$

where we neglect the row index for simplicity. Here the output y_i is related to the state x_i by the standard nonlinearity.

$$y_i = f(x) = 0.5|x_i + 1| - 0.5|x_i - 1|, \tag{2}$$

which is shown graphically in Fig. 2. The solution waveforms of Eq.(1) corresponding to $\alpha = 2$ and $\beta = 2$ and initial condition: $x_1(0) = 0, 1$ and $x_2(0) = 0, 1$ are shown in examples 1a) and 1a)Since the trajectory from (0.1, 0.1) does not converge to an equilibrium point (x_{1Q}, x_{2Q}). This CNN is not completely stable.
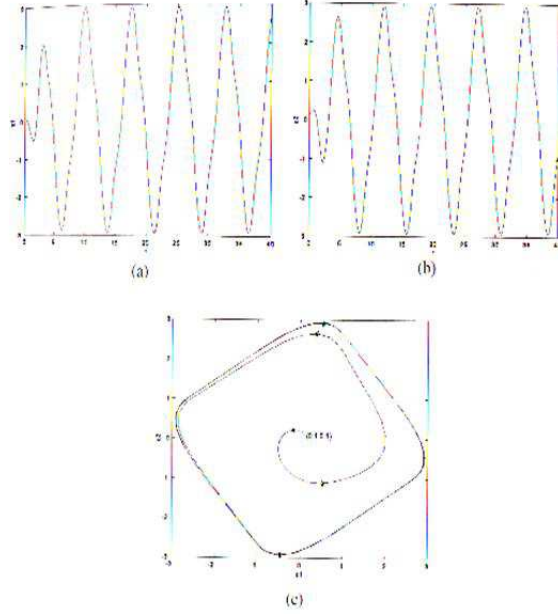
Fig.2.Example 1: Periodic solution wavefront of $x_1(t)$ and $x_2(t)$ and the corresponding trajectory for $\alpha = 2$, $\beta = 2$, $x_1(0) = 0.1$ and $x_2(0) = 0.1$

For this simple example, we can prove that all trajectories starting from any initial state except the origin will converge to a limit cycle. Hence, upon setting Equation (1) to 0, the equilibrium points of this two-cell CNN are the solutions of:

$$-x_1 + 2f(x_1) - 2f(x_2) = 0 \qquad (3)$$
$$-x_2 + 2f(x_2) + 2f(x_1) = 0$$

To determine the dynamical behavior near the origin, we examine the associated linear equation:

$$\dot{x}_1 = x_1 - 2x_2 \qquad (4)$$
$$\dot{x}_2 = 2x_1 + x_2$$

obtained by setting $f(x_1) = x_1$ and $f(x_2) = x_2$ in Eq.(1). Since the eigenvalues are given by $\lambda_1 = 1 + j_2$ and $\lambda_2 = 1 - j_2$, the solution of Eq.(4) has the form:

$$x_1(t) = ke^t \cos(2t + \theta) \qquad (5)$$
$$x_2(t) = ke^t \sin(2t + \theta)$$

71

Consider now a chaotic CNN with only two cells and one sinusoidal input. Suppose we apply a sinusoidal input $u_{11}(t) - 4.04\sin(\frac{\pi}{2})$ to cell $C(1,1)$ of the two-cell CNN shown in Fig.1 and choose $\alpha = 2$ and $\beta = 1.2$ as its parameters. In this case, under the same "zero" boundary conditions as before, the state equation (1) generalizes to the following non-autonomous of two nonlinear differential equations:

$$\dot{x}_1 = -x_1 + 2y_1 - 1.2y_2 + 4.04\sin(\frac{\pi}{2}t) \tag{6}$$
$$\dot{x}_2 = -x_2 + 1.2y_1 + 2y_2,$$

where $y_i = f(x_i)$ is defined by Eq.(1). Equation (6) is the state equation of a 1×2 CNN with templates:

A=	0	0	0		B=	0	0	0		z=	0
	1.2	2	-1.2			0	1	0			
	0	0	0			0	0	0			

Fig.3.

zero boundary conditions, a sinusoidal input $u_{11}(t)$ to cell $C(1,1)$, and a zero input $u_{12}(t) = 0$ to cell $C(1,2)$. The solution waveforms $x_1(t)$ and $x_2(t)$ corresponding to the initial condition $x_1(0) = 0.1$ and $x_2(0) = 0.1$, are shown in Figs 2(a) and 2(b), respectively
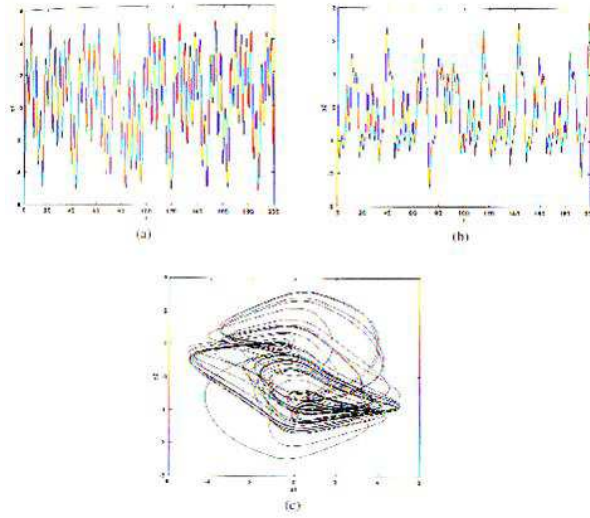

Fig.4.Example 2: Chaotic solution waveforms of $x_1(t)$ and $x_2(t)$ and the corresponding trajectory for $\alpha = 2$; $\beta = 1.2$; $x_1(0) = 0.1$ and $x_2(0) = 0.1$

The resulting set of points is called a Poincare cross section, or by an abuse of language, simply a Poincare map because it was first introduced by the famous French physics and mathematician Poincare. In this example, the periodic of the sinusoidal input is $T = 4$.

72
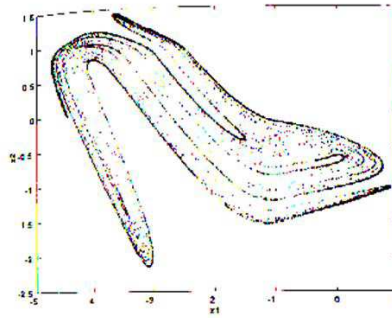
Fig.5.Example 3: The Poincare map extracted from the strange attractor in Example 2 (c) is called the "Lady's shoe attractor" in view of its striking resemblance to a high-heel lady's pump.
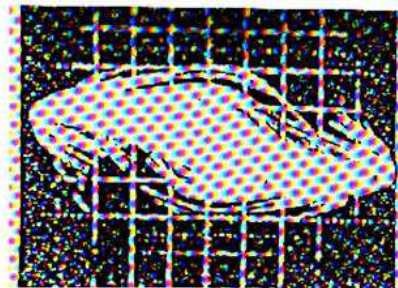
(a)
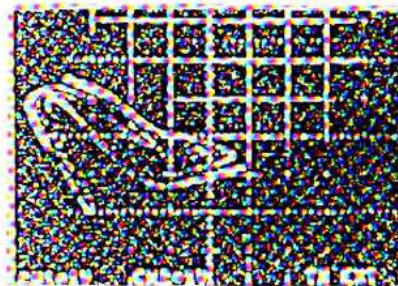
(b)

Fig.6.Example 4: (a) Strange attractor obtained experimentally from the circuit in Fig.3 (b) The "Lady's shoe" Poincare map extracted experimentally from the attractor in (a).

Symmetric A template implies complete stability

73

Complete Stability Theorem 1. *Any $M \times N$ space-invariant CNN of arbitrary neighborhood size with constant inputs and constant threshold is completely stable if the following three hypotheses are satisfied:*

(i) the A template is symmetric $A(i,j;k,l) = A(k,l;i,j)$;

(ii) the nonlinear function $y_{ij} = f(x_{ij})$ is differentiable, bounded and $f'(x_{ij}) > 0$ for all $-\infty < x_{ij} < +\infty$;

(iii) All equilibrium points are isolated.

Positive and sign-symetric A template implies complete stability.

Complete Stability Theorem 2 *An $M \times N$ CNN with a $(2r+1) \times (2r+1)A$ template is completely stable, for arbitrary B template and arbitrary threshold z if the following three conditions are satisfied:*

- The A template is sign symmetric;

- The template satisfied any one of the four synaptic weight conditions;

- All the equilibrium points are isolated.

Corollary to Complete Stability Theorem 2. *An $M \times N$ CNN with $3 \times 3A$ template, for arbitrary B template and arbitrary threshold z is completely stable if the A template possesses any one of the six synaptic weight patterns shown in Fig.7.*

синаптичен
тегловен шаблон 1

синаптичен
тегловен шаблон 2

синаптичен тегловен
шаблон 3

синаптичен
тегловен шаблон 4

синаптичен
тегловен шаблон 5

синаптичен тегловен
шаблон 6
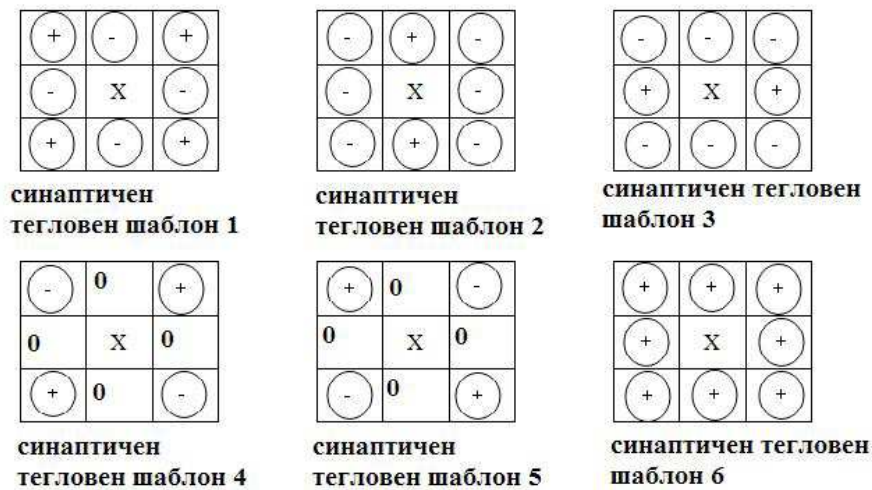
Fig.7.

where: 0 denotes a "zero" synaptic weight + denotes a "positive" or "zero" synaptic weight - denotes a "negative" or "zero" synaptic weight x may assume any value
Thus the stability of synaptic patterns 4 and 5 can be deduced from the stability of synaptic pattern 2.

A Second-Order Numerical Method for a Free-Boundary Problem in Price Formation

Juri D. Kandilarov

In this work we develop and analyze a second-order numerical method for a mean field model in Economics and Finance, which was introduced in a series of papers by J. M. Lasry and P. L. Lions [4]. Given a large group of *buyers* and a large group of *vendors* the non-linear free boundary evolution model describes the dynamical formation of the price of trading good under negotiation between the two groups. The groups are described by two non-negative density functions f_B and f_V. In the model, x denotes a possible value of the price. We denote by $p(t)$ the price resulting from a dynamical equilibrium and we assume that there is the same friction measured by a positive parameter. After substitution $f = f_B - f_V$ this situation can be described by the following free-boundary problem, see also [2, 3]:

$$
\begin{align}
f_t - f_{xx} &= \left(\delta_{p(t)+a} - \delta_{p(t)-a} \right) f_x(p(t), t) \ \text{in} \ (-1, 1) \times (0, \infty) \tag{1}\\
f_x(-1, t) &= f_x(1, t) = 0, \qquad t \in [0, \infty), \tag{2}\\
f(x, 0) &= f_0(x), \qquad x \in (-1, 1). \tag{3}
\end{align}
$$

where

$$
p(t) = \{x : f(x, t) = 0\}. \tag{4}
$$

The symbol δ denotes the Dirac delta function, concentrated respectively on the curves $x = p(t) \pm a$, $a \in (0, 1)$. For the moving interface $p(t)$ one can obtain the equation

$$
\dot{p}(t) = -f_{xx}(p(t), t)/f_x(p(t), t). \tag{5}
$$

1   DISCRETIZATION

Let us introduce the uniform grids in space and time directions:

$$
\omega_h = \{x_i = ih, \ i = 0, 1, ..., M, \ h = 2/M\}; \ \omega_\tau = \{\tau_n = n\tau, \ n = 0, 1, ..., N, \ \tau = T/N\}.
$$

In view of (5) the problem is **nonlinear** and we need of some iteration process. It consists of the following stages:

- Using (5) and the solution on the previous time layer n we obtain an initial guess of $p(t^{n+1})$;

- Then we solve the problem (1)-(3) and obtain the solution $f(x, t^{n+1})$;

- From (4) we update the free boundary $p(t^{n+1})$.

To discretize the problem (1) - (3) we use the idea of the IIM [1, 5]. Let denote the numerical solution at mesh point (x_i, t_n) by y_i^n. We use Crank-Nicolson scheme. It is standard at regular grid points and is modified at irregular grid points, i.e. near the interface. The difference scheme looks as follows:

$$
\begin{aligned}
\frac{y_i^{n+1} - y_i^n}{\tau} - Q_i^{n+1/2} &= \frac{1}{2}\left(\frac{y_{i-1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}}{h^2} + P_i^{n+1}\right) \\
&\quad + \frac{1}{2}\left(\frac{y_{i-1}^n - 2y_i^n + y_{i-1}^n}{h^2} + P_i^n\right)
\end{aligned} \tag{6}
$$

for $i = 1, ..., M - 1$ and $n = 0, 1, ..., N - 1$, where the correction terms $Q_i^{n+1/2}$ and P_i^{n+1}, P_i^n are chosen to improve the local truncation error (LTE) at irregular grid points and will be specified later. To preserve the second order we discretize the boundary conditions in the following way:

$$
\left(\frac{1}{h} + \frac{h}{2\tau}\right) y_0^{n+1} - \frac{1}{h} y_1^{n+1} = \frac{h}{2\tau} y_0^n; \tag{7}
$$

$$
-\frac{1}{h} y_{M-1}^{n+1} + \left(\frac{1}{h} + \frac{h}{2\tau}\right) y_M^{n+1} = \frac{h}{2\tau} y_M^n. \tag{8}
$$

Let us consider the correction term P_i^n. Let for some $t_n = n\tau$ the interfaces $\alpha_1(t_n) = p(t_n) - a$ and $\alpha_2(t_n) = p(t_n) + a$ are located as follows: $x_{I_1} \leq \alpha_1(t_n) < x_{I_1+1}$ and $x_{I_2} \leq \alpha_2(t_n) < x_{I_2+1}$. Then $P_i^n = 0$ for $i \in \{1, 2, ..., M - 1\} \setminus \{I_1, I_1 + 1, I_2, I_2 + 1\}$. At irregular nodes for $j = 1, 2$ we have

$$
\begin{aligned}
P_{I_j}^n &= -\frac{(x_{I_j+1} - \alpha_j(t_n))}{h^2} f_x(p(t_n), t_n) \\
&= -\frac{(x_{I_j+1} - \alpha_j(t_n))}{h^2}(\hat{a}_1 y_{J-1}^n + \hat{b}_1 y_J^n + \hat{c}_1 y_{J+1}^n) + O(h), \tag{9} \\
P_{I_j+1}^n &= \frac{(x_{I_j} - \alpha_j(t_n))}{h^2} f_x(p(t_n), t_n) \\
&= \frac{(x_{I_j} - \alpha_j(t_n))}{h^2}(\hat{a}_1 y_{J-1}^n + \hat{b}_1 y_J^n + \hat{c}_1 y_{J+1}^n) + O(h), \tag{10}
\end{aligned}
$$

where J satisfies $x_J \leq p(t_n) \leq x_{J+1}$.
The coefficients $\hat{a}_1$, $\hat{b}_1$, $\hat{c}_1$ are respectively:

$$
\hat{a}_1 = \frac{2\rho_J - 1}{2h}, \quad \hat{b}_1 = \frac{-2\rho_J}{h}, \quad \hat{c}_1 = \frac{1 + 2\rho_J}{2h}, \quad \text{and} \quad \rho_J = \frac{p(t_n) - x_J}{h}.
$$

I_1, I_2 and J depend on the time t_n, but for simplicity the upper index n is omitted. The correction term $Q_i^{n+1/2} = 0$ if the curve $y = p(t) \pm a$ does not intersect the the

grid line $x = x_i$ for some $t \in [t_n, t_{n+1})$. If it is happened, then the correction is:

$$
\begin{aligned}
Q_i^{n+1/2} &= \frac{t_{n+1/2} - \tilde{t}_n}{\tau} |p'(\tilde{t}_n)| [f_x]_{p(\tilde{t}_n) \pm a} \\
&= \frac{t_{n+1/2} - \tilde{t}_n}{\tau} |p'(\tilde{t}_n)| f_x(p(t_{n+1})), t_{n+1}) + O(h + \tau) \qquad (11) \\
&= \frac{t_{n+1/2} - \tilde{t}_n}{\tau} |p'(\tilde{t}_n)| (\hat{a}_1 y_{J-1}^{n+1} + \hat{b}_1 y_J^{n+1} + \hat{c}_1 y_{J+1}^{n+1}) + O(h + \tau),
\end{aligned}
$$

where $\tilde{t}_n \in [t_n, t_{n+1})$ is the time, for which $p(\tilde{t}_n) \pm a = x_i$.

To summarize, the grid points are divided in three groups:

- **regular**, if the interface curve does not intersect the six point stencil of the Crank-Nicolson scheme and all the correction terms are zero: $P_{I_j}^n = P_{I_j+1}^n = P_{I_j}^{n+1} = P_{I_j+1}^{n+1} = Q_i^{n+1/2} = 0$;

- **irregular**, if there is intersection only in space direction and then the corrections $P_{I_j}^n$, $P_{I_j+1}^n$, $P_{I_j}^{n+1}$, $P_{I_j+1}^{n+1}$ are not zero and $Q_i^{n+1/2} = 0$;

- **irregular** with intersections in both directions and then the corrections $P_{I_j}^n$, $P_{I_j+1}^n$, $P_{I_j}^{n+1}$, $P_{I_j+1}^{n+1}$ and $Q_i^{n+1/2}$ are not zero.

Theorem 1 *Let the solution $f(x,t) \in C^{4,3}((-1,1) \backslash \{p(t) \pm a\}, (0,T))$. Then the scheme (6)-(11) approximates the problem (1)-(3) and the LTE at regular grid points is $O(h^2 + \tau^2)$ and at irregular grid points it is $O(h + \tau)$.*

2   ALGORITHM DESCRIPTION

We denote the solution vector $Y_{n+1} = (y_0^{n+1}, ..., y_i^{n+1}, ..., y_M^{n+1})^T$ on the new time layer t_{n+1} and rewrite the difference scheme in the following matrix form:

$$
A_{n+1} Y_{n+1} = B_n \qquad n = 0, 1, ..., N-1. \qquad (12)
$$

The matrix A_{n+1} can be decomposed as $A_{n+1} = \overline{A} + \widetilde{A}_{n+1}$, where $\overline{A}$ is three-diagonal matrix and $\widetilde{A}_{n+1}$ is a result of the correction terms $P_{I_j}^n$, $P_{I_j+1}^n$, $Q_i^{n+1/2}$, see (9)-(11). The entries of the matrix $\overline{A}$ are:

$$
c_0 = c_M = \frac{1}{h} + \frac{h}{2\tau}, \quad b_0 = a_M = -\frac{1}{h}, \quad a_i = b_i = -\frac{1}{2h^2}, \quad c_i = \frac{1}{\tau} + \frac{1}{h^2}.
$$

The entries of the matrix $\tilde{A}$ are result of the correction terms $P_{I_j}^{n+1}$ $P_{I_j+1}^{n+1}$ and $Q_i^{n+1/2}$:

$$\tilde{A}_{n+1} = \begin{pmatrix} & & \cdots & & & \\ \vdots & \tilde{a}_1 & \tilde{b}_1 & \tilde{c}_1 & \vdots & \\ \vdots & \tilde{\tilde{a}}_1 & \tilde{\tilde{b}}_1 & \tilde{\tilde{c}}_1 & \vdots & \\ & & \cdots & & & \\ \vdots & \tilde{a}_2 & \tilde{b}_2 & \tilde{c}_2 & \vdots & \\ \vdots & \tilde{\tilde{a}}_2 & \tilde{\tilde{b}}_2 & \tilde{\tilde{c}}_2 & \vdots & \\ & & \cdots & & & \end{pmatrix} \begin{matrix} \\ \leftarrow \quad I_1 \\ \leftarrow \quad I_1+1 \\ \\ \leftarrow \quad I_2 \\ \leftarrow \quad I_2+1 \\ \\ \end{matrix} .$$

The matrix B_n is decomposed as $B_n = \overline{B}_n + \widetilde{B}_n$:

$$\overline{B}_n = \begin{pmatrix} \frac{h}{2\tau}y_0^n \\ \vdots \\ \left(\frac{1}{\tau}-\frac{1}{h^2}\right)y_i^n + \frac{1}{2h^2}\left(y_{i-1}^n + y_{i+1}^n\right) \\ \vdots \\ \frac{h}{2\tau}y_M^n \end{pmatrix}, \quad \widetilde{B}_n = \begin{pmatrix} \vdots \\ P_{I_1}^n \\ P_{I_1+1}^n \\ \vdots \\ P_{I_2}^n \\ P_{I_2+1}^n \\ \vdots \end{pmatrix} \begin{matrix} \\ \leftarrow \quad I_1 \\ \leftarrow \quad I_1+1 \\ \\ \leftarrow \quad I_2 \\ \leftarrow \quad I_2+1 \\ \\ \end{matrix} .$$

The matrix $\widetilde{B}_n$ is a result of the corrections $P_{I_j}^n$ and $P_{I_j+1}^n$ on the old time layer t_n. To solve (12) we suggest the next algorithm.

First stage. We seek the numerical solution in the linear form

$$y_i^n = u_i^n + v_i^n y_{J-1}^n + w_i^n y_J^n + z_i^n y_{J+1}^n. \tag{13}$$

Let us fix the time level $t = t_n$ and omit for simplicity the upper index n. We insert (13) into (12). For u_i^n we obtain:

$$\begin{aligned} c_0 u_0 - b_0 u_1 &= \phi_0, & -a_M u_{M-1} + c_M u_M &= \phi_M, \\ -a_i u_{i-1} + c_i u_i - a_i u_{i+1} &= \phi_i, & i &= 1,...,J-2, J+2,...M-1 \\ u_{J-1} = u_J = u_{J+1} &= 0, & & \end{aligned} \tag{14}$$

where ϕ_i is the i-th element of B_n. For v_i^n we obtain:

$$\begin{aligned} c_0 v_0 - b_0 v_1 &= \phi_0^v, & -a_M v_{M-1} + c_M v_M &= \phi_M^v, \\ -a_i v_{i-1} + c_i v_i - a_i v_{i+1} &= \phi_i^v, & i &= 1,...,J-2, J+2,...M-1 \\ v_{J-1} = 1, \quad v_J = v_{J+1} &= 0, & & \end{aligned} \tag{15}$$

where $\phi_i^v = 0$ for $i = 0,...,I_1-1, I_1+2,...,J-2, J+2,...,I_2-1, I_2+2,...,M$, and $\phi_{I_j}^v = -\tilde{a}_j$, $\phi_{I_j+1}^v = -\tilde{\tilde{a}}_j$, $j = 1,2$. Analogously, for w_i^n we obtain:

$$\begin{aligned} c_0 w_0 - b_0 w_1 &= \phi_0^w, & -a_M w_{M-1} + c_M w_M &= \phi_M^w, \\ -a_i w_{i-1} + c_i w_i - a_i w_{i+1} &= \phi_i^w, & i &= 1,...,J-2, J+2,...M-1 \\ w_{J-1} = 0, \quad w_J = 1, \quad w_{J+1} &= 0, & & \end{aligned} \tag{16}$$

78

where $\phi_i^w = 0$ for $i = 0, ..., I_1 - 1, I_1 + 2, ..., J - 2, J + 2, ..., I_2 - 1, I_2 + 2, ..., M$, and $\phi_{I_j}^w = -\widetilde{\widetilde{b}}_j$, $\phi_{I_j+1}^w = -\widetilde{\widetilde{b}}_j$, $j = 1, 2$. For z_i^n we obtain:

$$
\begin{aligned}
c_0 z_0 - b_0 z_1 &= \phi_0^z, & -a_M z_{M-1} + c_M z_M &= \phi_M^z, \\
-a_i z_{i-1} + c_i z_i - a_i z_{i+1} &= \phi_i^z, & i &= 1, ..., J - 2, J + 2, ... M - 1 \quad (17) \\
z_{J-1} = 0, \quad z_J = 0 \quad z_{J+1} &= 1,
\end{aligned}
$$

where $\phi_i^z = 0$ for $i = 0, ..., I_1 - 1, I_1 + 2, ..., J - 2, J + 2, ..., I_2 - 1, I_2 + 2, ..., M$, and $\phi_{I_j}^z = -\widetilde{c}_j$, $\phi_{I_j+1}^z = -\widetilde{\widetilde{c}}_j$, $j = 1, 2$.

Second stage. Plugging (14)-(17) into the $J - 1$-th, J-th and $J + 1$-th equation of the difference scheme, we obtain for y_{J-1}, y_J, y_{J+1} the system:

$$
\begin{aligned}
(c_{J-1} - a_{J-1} v_{J-2}) y_{J-1} - (a_{J-1} w_{J-2} + b_{J-1}) y_J &- a_{J-1} z_{J-2} y_{J+1} \\
&= \phi_{J-1} + a_{J-1} u_{J-2}, \\
-a_J y_{J-1} + c_J y_J - b_J y_{J+1} &= \phi_J, \quad\quad (18) \\
-b_{J+1} v_{J+2} y_{J-1} - (a_{J+1} + b_{J+1} w_{J+2}) y_J &+ (c_{J+1} - b_{J+1} z_{J+2}) y_{J+1} \\
&= \phi_{J+1} + b_{J+1} u_{J+2}.
\end{aligned}
$$

Third stage. Having y_{J-1}, y_J, y_{J+1}, from (13) we find the solution Y_{n+1} on the new time level $t = t_{n+1}$.

3   NUMERICAL EXPERIMENTS

We consider the following model example

$$
\begin{aligned}
f_t - f_{xx} &= g(x,t) + \left(\delta_{p(t)+a} - \delta_{p(t)-a} \right) f_x(p(t), t) \quad \text{in} \quad (-1, 1) \times (0, \infty), \\
f_x(-1, t) &= f_x(1, t) = 0, \quad t \in [0, \infty), \\
f(x, 0) &= f_0(x), \quad\quad x \in (-1, 1)
\end{aligned}
$$

with an exact solution

$$
f(x, t) = \exp(-t) \begin{cases} \lambda a, & x \le \alpha t^2 - a; \\ -\lambda(x - \alpha t^2), & \alpha t^2 - a \le x \le \alpha t^2 + a; \\ -\lambda a, & \alpha t^2 + a \le x. \end{cases}
$$

The functions $g(x,t)$ and $f_0(x)$ are found from the exact solution. The parameters are: $a = 0.3$, $\alpha = .25$, $\lambda = 1$. The moving interface is $p(t) = \alpha t^2$. In Table 1 we present the mesh refinement analysis. With $\|E_{M,N}\|_\infty$ we denote the infinity norm of the error when the mesh parameter $h = 1/M$ and $\tau = 1/N$. With $\|E_{M,N}\|_\infty^p$ we denote the corresponding error in maximum norm of the computed moving boundary $p(t)$. The results with correction $Q^{n+1/2}$ confirm second order of the proposed method. As the scheme is of order $O(h^2 + \tau^2)$ we keep the mesh parameters proportional in all experiments.

Table 1: Grid refinement analysis

		with correction $Q^{n+1/2}$		without correction $Q^{n+1/2}$	
M	N	$\|E_{M,N}\|_\infty$	$\|E^p_{M,N}\|_\infty$	$\|E_{M,N}\|_\infty$	$\|E^p_{M,N}\|_\infty$
30	100	6.1844e-05	8.4295e-05	4.6585e-04	6.2108e-04
60	200	1.6008e-05	4.1981e-05	1.3017e-04	2.5773e-04
120	400	4.1124e-06	5.4424e-06	4.9524e-05	2.0341e-05
240	800	1.0232e-06	1.3462e-06	2.0302e-05	1.3410e-05
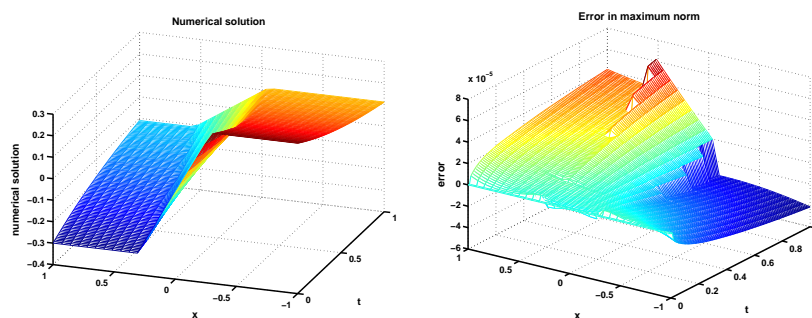
Figure 1: (a) The numerical solution; (b) The error in maximum norm.

References

[1] R. P. Beyer, R. J. Leveque, Analysis of a one-dimensional model for the immersed boundary method, *SIAM J. Numer. Anal.* **29** (1992) 332-364.

[2] M. Gonzalez, M. Gualdani, Asymptotics for a free-boundary model in price formation, arXiv:0811.1636v1 [math.AP].

[3] M. Gonzalez, M. Gualdani, Asymptotics for a Symmetric Equation in Price Formation, *Appl. Math. Optim.*, **59** (2009) 233ï¿½$\frac{1}{2}$246.

[4] J.M. Lasry and J.P. Lions, Large investor trading impacts on volatility. *Ann. Inst. H. Poincare (C) Non Linear Anal.* **24**(2) (2007) 311-323.

[5] Z. Li, Immersed interface method for moving interface problems, *Numer. Algorithms* **14** (1997) 269-294.

Fast Two-Grid Algorithms for the Elliptic Monge-Ampère Equation

Miglena N. Koleva, Lubin G. Vulkov

1. Introduction. Nonlinear partial differential equation (PDEs) arise in a quite natural way in many physical context such as kinetic theory, meteorology and semi-geostrophic fluid dynamics, material science, differential geometry, image processing and computer vision, see [3, 5] and references therein. Recently, new models of finance mathematics was described by fully nonlinear PDEs of Monge-Ampère type [11, 12]. This paper concerns with numerical solution of the Dirichlet problem for the Monge-Ampère equation

$$\det D^2 u = f \ \ \text{in} \ \ \Omega \in \mathbb{R}^2, \quad u = g \ \ \text{on} \ \ \partial\Omega, \tag{1}$$

where $D^2 u$ is the Hessian of the function u and f, g are given functions, $f > 0$ (or, later measure). Since we will be restricting to domains $\Omega \in \mathbb{R}^2$, we can rewrite the PDE in (1) as

$$\left[u_{xx}u_{yy} - u_{xy}^2\right](x,y) = f(x,y), \ \ (x,y) \in \Omega \in \mathbb{R}^2. \tag{2}$$

For the fully nonlinear PDEs, tremendous progress has been made in the past three decades [3, 10]. A revolutionary viscosity solution theory has been established (cf. [3]) and wealthy amount of efficient and robust numerical methods and algorithms have been developed and implemented (cf. [2, 4, 5, 8, 9]). However, for fully nonlinear second order PDEs, the situation is strikingly different. On one hand, there have been enormous advances on PDE analysis in the past two decades after the introduction of the notion of viscosity solutions by M.Crandall and P.L. Lions in 1983 (cf. [3]). On the other hand, in contrast to success of the PDEs analysis, numerical solutions for general fully nonlinear second order PDEs (except in the case of Bellman type PDEs, see [5] details) is mostly an untouched area, and computing viscosity solutions of fully nonlinear second order PDEs has been impracticable. There are several reasons for this lack of progress. *Firstly*, the strong nonlinearity is an obvious one. *Secondly*, the conditional uniqueness (i.e.uniqueness holds only in certain class of functions) of solutions is difficult to handle numerically. *Lastly* and most importantly, the notion of viscosity solutions, which is not variational, has no equivalence at the discrete level. We will emphasize the contribution of J. Benamou, B. Froese, A. Oberman in [2]. They propose two different monotone methods for solving (1), which have advantageous of simplicity and performance. The first one is an explicit, Gauss-Seidel iteration method that is obtained by solving a quadratic equation at each grid point and choosing the smaller root to ensure selection of the convex solution. The second method is an iterative method which requires the solution of a Poisson equation at each iteration. The source term in the Poisson equation involves the source function of (2), f and the Hessian of the current iterate. The authors have been studied in

details the performance of the two methods on solutions ranging from smooth to moderate to very singular. The backs of these methods is *slow convergence*. Even in the simplest cases (of smooth solution) this algorithms require a great number iterations. Loeper and Rapetti [8] solve the equation (with pereodic boundary conditions) by linearization and iterate using Newton's method. They prove convergence of the Newton algorithm for the linearization of the continious problem to the solution of (1). However, they do not address the issure of the convergence of the discretised solution to the solution of the equation in the limit of the discretisation parameter going to zero.

In this paper we propose a two-grid (=two-level) method, which first computes an approximate solution by solving the problem (1) on a coarse grid. The coarse mesh approximation is interpolated on the fine mesh and then a discretization of the linearized problem (1) is computed. The two-grid method was first proposed by Axelsson [1] and J. Xu [13], independently of each other, for a linearization of the nonlinear problems. The two-grid finite element method was also used by J.Xu [13] and many others scientists for discretizing nonsymmetric indefinite elliptic and parabolic equations.

The advantage of the two-grid algorithm is clear: we need to solve the nonlinear system corresponding to problem (1) only on the coarse mesh. On the fine mesh we solve the corresponding *linearized* system. For an appropriate choice of mesh parameters, the two-grid algorithm computes an approximate solution with the same order of accuracy as the one obtained by solving the full nonlinear system on the fine mesh at a fraction of the computational cost, see for example [6, 7].

2. Difference schemes. In this section we present two difference schemes for solving the two-dimensional Monge-Ampère problem (1) in rectangle $\overline{\Omega} = [a, b] \times [c, d]$.

In the domain $\overline{\Omega} = \Omega \cup \partial\Omega$ we introduce the uniform mesh $\overline{\omega}^h$:

$$\overline{\omega}^h = \{(x_i, y_j) \mid x_i = a + (i-1)h_x, \ i = 0, \ldots, n_x, \ h_x = |b-a|/n_x,$$
$$y_j = c + (j-1)h_y, \ j = 0, \ldots, n_y, \ h_y = |d-c|/n_y\},$$
$$\overline{\omega}^h = \omega^h \cup \{a, y_j\} \cup \{b, y_j\} \cup \{x_i, c\} \cup \{x_i, d\}, \quad i(j) = 0, \ldots, n_x(n_y).$$

The first scheme involves simply discretising the second derivatives in (1) using standard central differences on uniform mesh $\overline{\omega}^h$. The result is the following consistent with (1) and second order accurate approximation [2]

$$(\mathcal{D}_{xx}^2 u_{i,j}^h)(\mathcal{D}_{yy}^2 u_{i,j}^h) - (\mathcal{D}_{xy}^2 u_{i,j}^h)^2 = f_{i,j}, \tag{3}$$

where

$$u_{i,j}^h = u^h(x_i, y_j), \quad f_{i,j} = f(x_i, y_j), \quad i = 0, \ldots n_x, \quad j = 0, \ldots n_y,$$

$$\mathcal{D}_{xx}^2 u_{i,j}^h = \frac{1}{h_x^2}(u_{i+1,j}^h - 2u_{i,j}^h + u_{i-1,j}^h), \quad \mathcal{D}_{yy}^2 u_{i,j}^h = \frac{1}{h_y^2}(u_{i,j+1}^h - 2u_{i,j}^h + u_{i,j-1}^h),$$

$$\mathcal{D}_{xy}^2 u_{i,j}^h = \frac{1}{4h_x h_y}(u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1})$$

The next seven-point scheme is obtained by mixed finite element approximation [5]. The only difference with the previous scheme is that in (3) for $\mathcal{D}^2_{xy}u^h$ we have

$$\mathcal{D}^2_{xy}u^h_{i,j} = \frac{1}{2h_x h_y}(u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + 2u_{i,j}).$$

3. Two-grid Newton's algorithms. To solve the system of nonlinear algebraic equations (3) we use Newton's method, i.e. we consider the following iterative process with initial guess $u^h_{i,j}{}^{(0)}$:

$$(\mathcal{D}^2_{xx}u^h_{i,j}{}^{(k)})(\mathcal{D}^2_{yy}\delta u^h_{i,j}{}^{(k+1)}) + (\mathcal{D}^2_{yy}u^h_{i,j}{}^{(k)})(\mathcal{D}^2_{xx}\delta u^h_{i,j}{}^{(k+1)})$$
$$-2(\mathcal{D}^2_{xy}u^h_{i,j}{}^{(k)})(\mathcal{D}^2_{xy}\delta u^h_{i,j}{}^{(k+1)}) = f_{i,j} - (\mathcal{D}^2_{xx}u^h_{i,j}{}^{(k)})(\mathcal{D}^2_{yy}u^h_{i,j}{}^{(k)}) + (\mathcal{D}^2_{xy}u^h_{i,j}{}^{(k)})^2 \quad (4)$$

and find $u^h_{i,j}{}^{(k+1)} = u^h_{i,j}{}^{(k)} + \delta u^h_{i,j}{}^{(k+1)}$, $k = 0, 1, \ldots$.

Analogically to $\overline{\omega}^h$, we define a new mesh $\overline{\omega}^H$ with step sizes $H_x >> h_x$, $H_y >> h_y$ and $N_x + 1(N_y + 1)$ grid nodes in $x(y)$ direction. This mesh we will call *coarse mesh*, while the first one $\overline{\omega}^h$ we will call *fine mesh*. In this section we propose two-grid algorithms based on the Newton's method. Let $u_I = u_I(x)$ is the interpolant of the continuous solution u. Now, we present the *basic steps* of our algorithm.

$1°$ Solve the discrete problem (3) on the coarse grid $\overline{\omega}^H$ and interpolate the result to obtain $u^H_I(x, y)$, $(x, y) \in \overline{\Omega}$;

$2°$ Solve the linear discrete problem

$$(\mathcal{D}^2_{xx}u^H_I(x_i, y_j))(\mathcal{D}^2_{yy}u^h(x_i, y_j)) + (\mathcal{D}^2_{yy}u^H_I(x_i, y_j))(\mathcal{D}^2_{xx}u^h(x_i, y_j))$$
$$-2(\mathcal{D}^2_{xy}u^H_I(x_i, y_j))(\mathcal{D}^2_{xy}u^h(x_i, y_j))$$
$$= f_{i,j} + (\mathcal{D}^2_{xx}u^H_I(x_i, y_j))(\mathcal{D}^2_{yy}u^H_I(x_i, y_j)) - (\mathcal{D}^2_{xy}u^H_I(x_i, y_j))^2$$

to find the fine mesh numerical solution $u^h(x_i, y_j)$, $(x_i, y_j) \in \overline{\omega}^h$.

Algorithm 1 (A1). The iteration process (4) performs as follows:

- For $m = 0$ - do $1°$ with mesh step sizes H_x, H_y;

- For $m = 1, \ldots, M$ - repeat $2°$ with *optimal* mesh step sizes $h_x \lesssim H_x^{2^M}$ and $h_y \lesssim H_y^{2^M}$ fixed for all iterations $m > 0$ and $y^H_I(x_i, y_j) := y^h(x_i, y_j)$ for $m > 1$.

Algorithm 2 (A2). The iteration process (4) performs as follows:

- For $m = 0$ - do $1°$ with mesh step sizes H_x, H_y;

- For $m = 1, 2, \ldots$ - repeat $2°$ with *optimal* mesh step sizes $h_x \lesssim H_x^{2^m}$ and $h_y \lesssim H_y^{2^m}$ for each iteration $m > 0$ and $y^H_I(x_i, y_j) := y^h_I(x_i, y_j)$ for $m > 1$.

Theorem 1 *Suppose that the problem (1)-(2) is classical solvable and $\|u - u^{h^{(0)}}\|_H \leq CH^2$, $H = \max\{H_x, H_y\}$. Then for the error of Algorithm 1 and Algorithm 2 at m-th iteration, with optimal choice of the fine mesh step size $h = \max\{h_x, h_y\}$, we have*

$$\|u^h - u\|_H \leq C(H^2)^{2^m}.$$

4. Numerical results. Numerical results verify the theoretical statements and demonstrate the efficiency of the two-grid algorithms. We will test our method for *smooth example* to compare the results with those, given in [2]. The solution is computed on an uniform quadratic meshes with mesh step sizes $H = H_x = H_y$ ($N = N_x = N_y$) and $h = h_x = h_y$ ($n = n_x = n_y$). The results are given in max discrete and L_2 discrete norms. The convergence rate is calculated, using the formulas: $CR_N^H = \log_2[E_{N/2}^H / E_N^H]$ and $CR_N^h = \log_2[E_{N/2}^h / E_N^h]$ for the coarse and fine mesh solution, respectively. Here, with E_N^H and E_N^h the *errors* in max discrete norms ($E_N^H = \|y_i^H - u(x_i)\|_H$ and $E_N^h = \|y_i^h - u(x_i)\|_h$, $i = 0, \dots, N$) are denoted.
The initial guess for Newton's iteration process is the solution of $u_{xx} + u_{yy} = 2\sqrt{f}$ in Ω, $u = g$ on $\partial\Omega$ [5]. For $m = 0$ iteration of Newton's method continued until the maximum difference between two subsequent iterations is less than 10^{-12}.
The computations are performed with *optimal* chosen fine mesh step size h, i.e.

$$h \lesssim \begin{cases} H^{2^M}, & \text{fixed for all iterations } 0 < m \leq M, \text{ computing up to } M \text{ with A1}, \\ H^{2^m}, & \text{at each iteration } m > 0, \text{ computing with A2}. \end{cases}$$

Example Let $\overline{\Omega} = [-1, 1] \times [-1, 1]$. An exact radial solution is

$$u(x, y) = \exp\left[(x^2 + y^2)/2\right], \quad f(x, y) = (1 + x^2 + y^2)\exp(x^2 + y^2).$$

Errors and convergence rates in corresponding discrete norms are given in Table 1 for A1 and A2 based on the nine-point scheme (3). Also, we compare CPU times for one-grid procedure ($m = 0$) and two-grid algorithms ($m = 1, 2$) when approximately one and the same accuracy is reached. On Figure 1 numerical solution and errors are plotted following the two-grid algorithm A2 based on the nine-point stencil scheme and optimal chosen fine mesh step sizes. In Table 2 errors in max discrete norm for different stages of A1 with seven-point scheme (3) are compared. Thus, the validation of Theorem 1 is established. The convergence rate in max discrete norm of A1 and A2 is $\mathcal{O}(H^2)$ for $m = 0$, $\mathcal{O}(h^2 + H^4) = \mathcal{O}(H^4)$ for $m = 1$ and $\mathcal{O}(h^2 + H^8) = \mathcal{O}(H^8)$ for $m = 2$.

References

[1] O. Axelsson, On mesh independence and Newton methods, Appl. Math. 38(4-5) (1993) 249 - 265.

Table 1: Error and convergence rate in *max discrete norm* and comparison results for CPU time, nine-point stensil

H	$m = 0$		$m = 1$		$m = 2$ (A1)		$m = 2$ (A2)	
	E_N^H	CR_N^H	E_N^h	CR_N^h	E_N^h	CR_N^h	E_N^h	CR_N^h
2^{-2}	4.75e-2		1.29e-2		8.40e-4 ($CPU = 0.48$)		7.81e-4 ($CPU = 0.43$)	
2^{-3}	1.30e-2	*1.866*	7.80e-4	*4.044*	3.30e-6 ($CPU = 1854.23$)	*7.994*	2.91e-6 ($CPU = 1301$)	*8.068*
2^{-4}	3.35e-3	*1.960*	**7.77e-5** ($CPU = 0.42$)	*4.034*				
2^{-5}	8.43e-4 ($CPU = 0.65$)	*1.969*						
2^{-6}	2.11e-4	*1.997*						
2^{-7}	**5.27e-5** ($CPU = 19.25$)	*1.999*						
2^{-8}	1.32e-5	*2.000*						
2^{-9}	3.30e-6 ($CPU = 4455.33$)	*2.000*						

Table 2: Error and convergence rate in *max discrete norm*, seven-point stencil

H	$m = 0$		$m = 1$		$m = 2$ (A1)		$m = 2$ (A2)	
	E_N^H	CR_N^H	E_N^h	CR_N^h	E_N^h	CR_N^h	E_N^h	CR_N^h
2^{-2}	4.58e-2		1.83e-4		8.40e-4		7.81e-4	
2^{-3}	1.30e-2	*1.8220*	1.10e-5	*4.0579*	3.52e-6	*7.8986*	3.03e-6	*8.0099*
2^{-4}	3.33e-3	*1.9539*	6.80e-7	*4.0176*				

[2] J.-D. Benamou, B. Froese, A. Oberman, Two numerical methods for the elliptic Monge Ampère equation, Preprint 2009.

[3] L. Caffarelli, X. Cabre, Fully Nonlinear Elliptic Equations, Amer. Math. Soc. Coll. Publ. 43, American Mathematical Society, Providence, RI, 1995.

[4] X. Feng, M. Neilan, Galerkin methods for the fully nonlinear Monge Ampère equation. http://arxiv.org/abs/0712.1240v1.

[5] R. Glowinski, Numerical methods for fully nonlinear elliptic equations. In R. Jeltsch and G. Wanner (eds), 6th International Congress on Industrial and Applied mathematics, ICIAM'07, Invated Lectures (2009) 155-192.

[6] M. Koleva, L. Vulkov, A Fast Numerical Method for a Nonlinear Black-Scholes Equation, AIP CP 1184 (2009) 64-71

[7] M. Koleva, L. Vulkov, Two-grid quasilinearization approach to ODEs with applications to model problems in physics and mechanics, Comp. Phys. Commun. 181(3) (2010) 663-670.
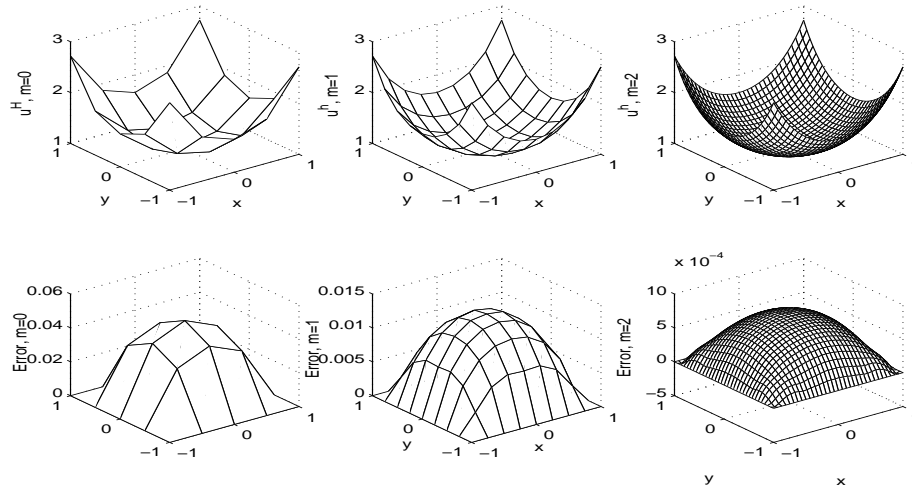
Figure 1: Numerical solutions and errors ($u^h - u$), Algorithm A2 with nine-point stencil for $m = 0$ ($H = 2^{-2}$), $m = 1$ ($h = H^2$), $m = 2$ ($h = H^4$)

[8] G. Loeper, F. Rapetti, Numerical solution of Monge Ampère equation by a Newton's algorithms, C. R. Math. Acad. Sci. Paris 340(4) (2006) 879-895.

[9] A. Oberman, Wide stensil finite difference schemes for the elliptic Monge Ampère equation and functions of the eigenvalues of the Hessian, Discrete Contin. Dyn. Syst. Ser. B 10(1) (2008) 221-238.

[10] P. R. Popivanov, M.D. Kutev, Interionr and exterior boundary value problems for the degenerate Monge-Ampere operator, Hiroshima Math. J. 19 (1989) 167-179.

[11] L. Songzhe, Existence of solutions to initial value problem for a parabolic Monge-Ampère equation and application, Nonl. Anal. 65 (2006) 59-78.

[12] S. Stojanovic, Risk premium and fair option under stochastic volatility: the HARA Solution, C./ R. Acad. Ser. 1340 (2005) 551-556.

[13] J. Xu, A novel two-grid method for semilinear elliptic equations, SIAM J. Sci. Comput. 15(1) (1994) 231-237.

Appliance of Artificial Intelligence Forms for Examination of the Resolution's Radiolocational Station Influence on the Recognition Accuracy of Dynamic Object

Milena Kostova, Valerij Djurov

1 Introduction

The modern radiolocational systems working with cvasioptic probing signals give an opportunity for receiving radioholograms of observing objects in accepting trackt for given radiolocational station. The recreation of the received radiohologram leads to receiving of a radiolocational portrait or also called radioholographic image (RHI) of the object. It is a scaled model of the real object and its quality depends on the count of reflecting points by the surface of the object. They are determined to distance and azimuth by the accomplished resolution. The recognition of the dynamic object by radioholographic image can be realized by comparing with a standard image [3]. This method requires centering of the image in the frame which creates extra difficulties when it is processed. The recognition of dynamic objects (aircrafts) by geometric causes does not require such centering. The use of forms of artificial intelligence - neural networks, fuzzy logic, computer vision, expert systems and so on, helps for processing the information in real time.

2 Model of a system for processing and classification of radioholographic image of an aircraft

A system for processing and classification of radioholographic image of aircrafts is modeled. It includes filtration of the image by a cellular neural network (CNN) [1,4], receiving of an outline, taking of peculiar points and determination of the values of image's informational causes (Fig. 1). The classification is based on the geometrical characteristic of the objects and it is made by two classifiers which take decision based on the fuzzy logic and probability neural network [2].

A vector of geometrical causes (A, α, L_R, L_K) is formed for recognition of dynamic object's image, where

$$A = \frac{Width\ of\ wings}{Length\ of\ fuselage\ axis}, \quad \alpha = arctg\frac{wing}{fuselage\ axis}$$

L_R = geometrical center, geometrical middle, L_K = width of the aircraft's wings.

An algorithm for receiving the values of the causes is developed and applied at resolution 0.07; 0.15 and 0.5 m. The results are given in Table 1. Based on a given criterions are accepted the causes A and α, which are the same when the resolution is changed, i.e invariant the scale.
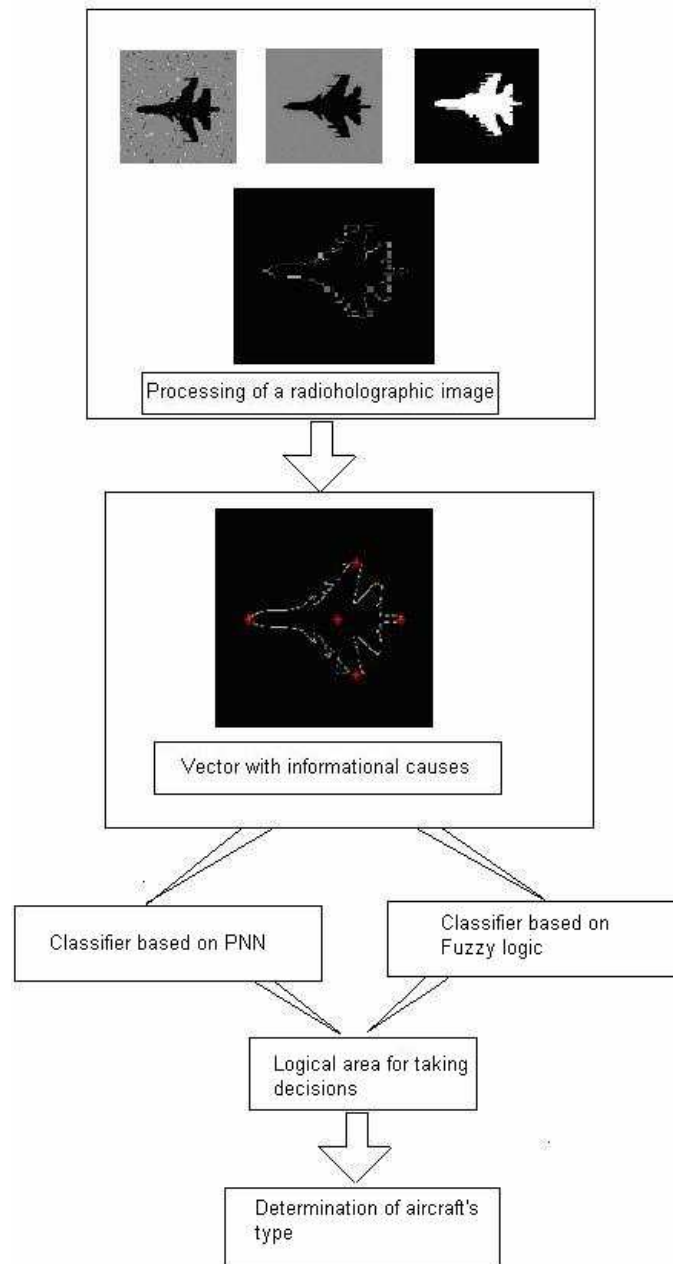
Figure 1: Model of a system for processing and classification of RHI of an aircraft

88

Table 1: Causes at resolution - 0.07; 0.15 and 0.5 m

Values of the causes for the separated types of aircrafts.												
Resolution												
Types of aircrafts	0,07				0,15				0,5			
	L_K	α	A	L_R	L_K	α	A	L_R	L_K	α	A	L_R
Military aircrafts												
Mig- 29	152	72	0,713	46	70	71,11	0,74	20,5	22	76	0,71	4,5
Mirag2000	122	58	0,629	24	58,03	56,80	0,65	9,5	14,04	54,49	0,54	2
Su- 27	198	64	0,634	38	100,05	66,28	0,69	18,5	28,07	63,47	0,72	6,5
Su- 35	203	76	0,645	46	100,84	66,28	0,84	16,5	30,02	68,23	0,79	6
Su- 37	211	63	0.670	39	100,84	66,28	0,73	17,5	31	69,48	0,72	4,5
F-15	173	72	0,706	46	84,05	72,84	0,72	21,5	24	69	0,71	5
Mig -25	198	59	0,626	58	92	60,55	0,62	28	28	63,47	0,7	7
Mig MF1	171	58	0,775	43	84	55,40	0,72	20,5	22	50,74	0,67	4,5
Rafale	153	63	0,714	24	72,06	57,88	0,73	10,5	18,03	71,60	0,75	3
Yak- 38	103	78	0,492	18,0	48,01	73,10	0,49	9	11	68,23	0,48	0,5
Jaguar	123	64	0,574	27	58	62,68	0,55	11,5	17,03	77,55	0,57	1
Su- 34	212	69	0,687	29	100,5	68,18	0,61	14	27,02	63,47	0,68	5
B-58	237	57	0,650	43	112,11	57,75	0,68	19	30,1	55,52	0,57	5,5
Su- 24	254	66	0,717	53	119	66,25	0,73	24,5	31	59,07	0,67	7
Tu- 22	341	54	0,580	76	161,20	52,46	0,61	40	49,16	51,98	0,65	12
Tornado	196	64	0,828	23	93	65,49	0,88	16	26,02	58,42	0,87	5
Buccaneer	181	70	0,678	9,5	84	69,18	0,69	5,5	26	72,93	0,81	1
Mirage IV	164	56	0,486	62	79	53,16	0,51	20,5	19,03	59,07	0,58	3,5
M-50	348	68	0,422	83,5	165	68,35	0,43	39	50	64,39	0,47	10
B -1b	593	63	0,933	86	273,18	63,63	0,93	42,5	80,1	64,62	0,97	10,5
Transport aircrafts												
An- 72	455	88.	1,1152	10	210,05	86,19	1,15	3	59,03	88,07	1,09	0
An -124	1031	76	1,075	39	472	76	1,07	18	142,43	81,35	1,17	5
An- 225	1111	75	1,080	59	573	76	1,08	31	168	76,49	1,09	8
Boeing747	969	70	0,921	48	443	69,42	0,92	24,5	134	72,35	0,96	8,5
Boeing757	524	79	0,803	24	248	79,1	0,81	15	68	85,14	0,79	3
Boeing737	471	80	1,053	26	223,10	78,83	1,06	14	62,01	84,69	1,13	1,5
Boeing767	655	71	0,857	39	306,01	71,82	0,86	21	89	73,57	0,86	6,5
Airbus340	848	73	1,048	47	395,13	74,01	1,04	22,5	114,1	73,15	1,01	6
Airbus300	634	77	0,840	7	292,06	80,32	0,83	5,5	87,05	82,28	0,86	3,5
Airbus310	622	75	0,968	12	284,25	73,56	0,93	6	82	80,79	0,95	2

3   Examination of the recognition accuracy at resolution 0.07; 0.15 and 0.5 m.

Radioholographic images of 30 types of aircrafts are received at resolution 0.07; 0.15 and 0.5 m [3]. A classification of the aircrafts is accomplished, in class military and class transport, by two classifiers based on probability neural network (PNN) and fuzzy logic respectively, which work parallel. A graphics user interface is created for easier examination. The results are given in Table 2. For convenience by (1) is marked correct recognition and by (0) is marked wrong classification. The way of work of Fuzzy classifier and PNN classifier are shown in Fig. 2 and Fig. 3 respectively. The results shows that the Fuzzy classifier gives better results - the accuracy vary from 95% to 100% for military aircrafts and from 90% to 100% for transport aircrafts. In some cases is possible the classifier to be unable to take decision. The accuracy of

Table 2: Accuracy of recognition at resolution 0.07; 0.15 and 0.5 m

Accuracy of recognition						
Resolution(m)						
Types of aircrafts	0,07		0,15		0,5	
	PNN classifier	Fuzzy classifier	PNN classifier	Fuzzy classifier	PNN classifier	Fuzzy classifier
Military aircrafts						
Mig- 29	**0**	1	**0**	1	**0**	1
Mirag2000	1	1	1	1	1	1
Su- 27	1	1	1	1	1	1
Su- 35	1	1	1	1	1	1
Su- 37	1	1	1	1	1	1
F-15	**0**	1	1	1	1	1
Mig -25	1	1	1	1	1	1
Mig MF1	1	1	1	1	1	1
Rafale	1	1	1	1	1	1
Yak- 38	**0**	1	**0**	1	1	1
Jaguar	1	1	1	1	**0**	1
Su- 34	1	1	1	1	1	1
B-58	1	1	1	1	1	1
Su- 24	1	1	1	1	1	1
Tu- 22	1	1	1	1	1	1
Tornado	1	1	1	1	1	1
Buccaneer	1	1	1	1	**0**	1
Mirage IV	1	1	1	1	1	1
M-50	1	1	1	1	1	1
B -1b	1	**not recognized**	1	1	1	**0**
Transport aircrafts						
An- 72	1	1	1	1	1	**0**
An -124	1	1	1	1	1	1
An- 225	1	1	1	1	1	1
Boeing747	**0**	1	**0**	1	1	1
Boeing757	1	1	1	1	1	1
Boeing737	1	1	1	1	1	1
Boeing767	1	1	1	1	1	1
Airbus340	1	1	1	1	1	1
Airbus300	1	1	1	1	1	1
Airbus310	1	1	1	1	1	1

PNN classifier for military aircrafts vary from 85% to 90% and for transport aircrafts is from 90% to 100%.

Logical rules for taking decisions can be lead in the two classifiers in parallel work. The following abbreviations are accepted:military - M, transport - T

If Fuzzy- not recognized and PNN- M, then M If Fuzzy- M and PNN- T, then M IF Fuzzy- T and PNN- M, then M If Fuzzy- T and PNN- T, then T If Fuzzy- M and PNN- M, then M

In this case the accuracy of the system for military aircrafts reaches 100% no mater of the resolution, and for the transport aircrafts at resolution 0.07 and 0.15 is 100%, and at resolution 0.5 is 90%.
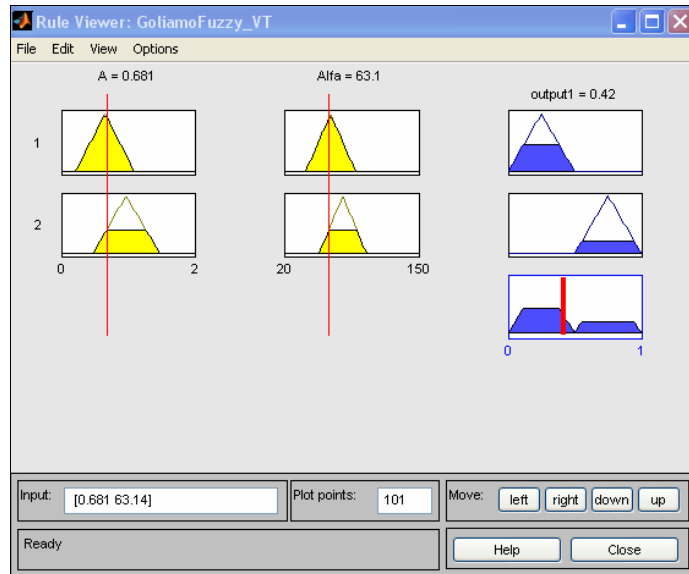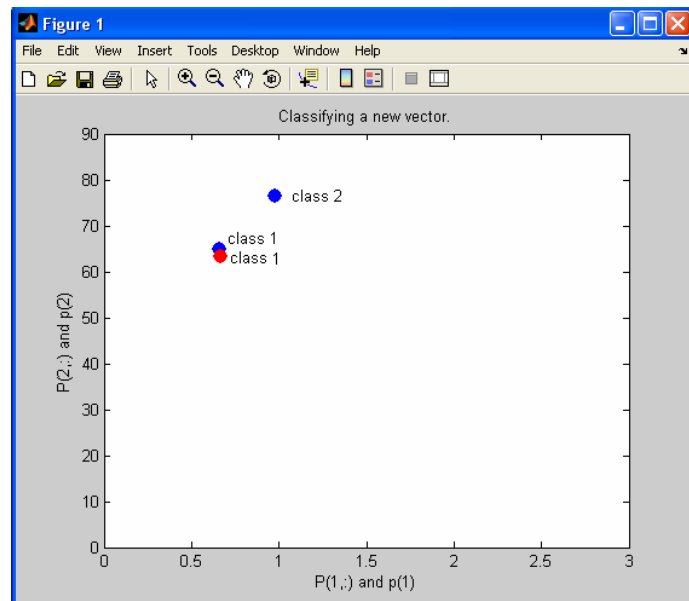
Figure 2: Way of work of Fuzzy classifier

Figure 3: Way of work of PNN classifier

4   Conclusions

At a corresponding device realization, the system can be applied at radiolocational stations working in centimeter and millimeter frequency area (according to NATO standards X and K) which can provide a resolution from 0.07 to 0.5m. The high accuracy of recognition gives the opportunity to analyse and observe objects which got in exceptional situation (technical failure, terroristic act and s.o.) or entering of a unknown object in a air corridor. The system can manage full day automatic monitoring without human intervention. The appliance of two different methods for classification increase the protection of the system against cyberattacks.

References

[1] Chua,L., T. Roska, Cellular Neural Networks and Visual Computing: Foundations and Applications, 2005

[2] Kostova, M.,V. Djurov. Classificators of Radioholographic Images of Aeroplanes with Application of Probability Neural Network and Fuzzy Logic, ADVANCED TOPICS ON NEURAL NETWORKS, Proceedings of the 9TH WSEAS International Conferenses on NEURAL .NETWORKS (NN' 08), Sofia, Bulgaria, May, 2-4, 2008, ISBN: 978-960-6766-56-5, ISSN: 1790-5109, pp.78-83[4.2]

[3] Lazarov, A.D., Minchev Ch.N., Spektral 2-D Image Reconstruction in ISAR with linear frequency modulated Signals , 20th Digital Avionics Systems Conference Daytona Beach, USA,Fl, Oct.13-18 , 2001

[4] Slavowa A., V. Mladenov, Cellular Neural Networks: Theory and Applications , Nova Science Publishers, Inc. New York, 2004

Scalable PCG Solution Algorithms for μFEM Elasticity Systems

Nikola Kosturski

In this study we analyze the performance of a parallel PCG solver for micro finite element method (μFEM) elasticity systems on the IBM Blue Gene/P massively parallel computer. The test problems arise from the μFE simulation of human bones. The voxel representation of the bone micro structure is obtained from a high resolution computer tomography (CT) image. Linear elasticity models at micro and macro levels are applied. A numerical homogenization scheme is applied to upscale the elastic properties of the reference volume element (RVE). The RVE of the studied trabecular bone tissue has a strongly heterogeneous micro structure composed of solid and fluid phases (see Fig. 1). Nonconforming Crouzeix–Raviart (C.–R.) elements are used to

32 × 32 × 32   64 × 64 × 64   128 × 128 × 128
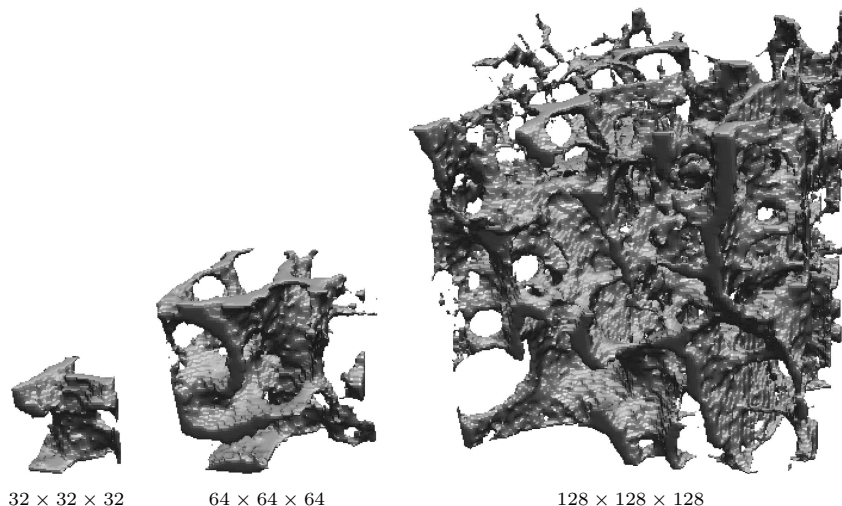
Figure 1: Structure of the solid phase of the bone specimens.

discretize the resulting linear elasticity problems. The solver for the arising large-scale linear systems is based on a parallel PCG [1] method. The preconditioner is the parallel algebraic multigrid implementation BoomerAMG [8]. The presented parallel tests well illustrate the scope of efficient applications.

In a number of earlier articles dealing with μFE simulation of bone structures (see, e.g., [6, 7]), the contribution of the fluid phase is neglected. In this study the fluid phase, located in the pores of the solid skeleton, is considered as an almost incompressible linear elastic material (see for some additional details, e.g., [5]). The bulk modulus K_f is fixed and the related Poisson ration ν_f is approaching the incompressible limit of 0.5. The elasticity modulus E_s and the Poisson ratio ν_s of the solid phase

93

as well as the bulk modulus of the fluid phase K_f are taken from [2]. Let us note that the C.–R. elements (after a proper modification of the related bilinear form) provide a locking-free approximation of almost incompressible elasticity problems.

Let us briefly describe the applied numerical upscaling scheme. A more detailed description along with a description of the locking-free FEM discretization can be found in [4].
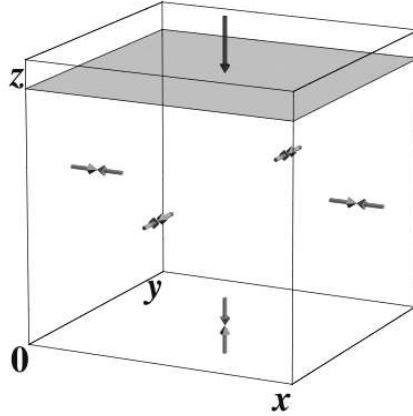
Figure 2: Boundary conditions of the RVE problem.

We study the implementation of a numerical homogenization scheme of Dirichlet boundary conditions type [3]. The RVE boundary value problem (BVP) has zero normal displacements on five of the faces of the cube and a (small) nonzero constant normal displacement on the sixth face (see Fig. 2).

By symmetry arguments, it simply follows that the homogenized stress and strain tensors have zero shear components, and therefore the following relation between the homogenized normal stress and strain components holds

$$
\begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{bmatrix} = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1 & \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} \\ \frac{\nu}{1-\nu} & 1 & \frac{\nu}{1-\nu} \\ \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{bmatrix}.
$$

Let us now consider the case, where the nonzero constant normal displacement (in z direction) is applied on the top face of the RVE cube. Then $\varepsilon_x = \varepsilon_y = 0$ and therefore

$$
\sigma_x = \sigma_y = \frac{E\nu}{(1+\nu)(1-2\nu)}\varepsilon_z, \quad \text{and} \quad \sigma_z = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)}\varepsilon_z.
$$

From these relations we can directly calculate the homogenized elasticity coefficients as follows

$$\nu = \frac{1}{1+p}, \quad E = \frac{(1+\nu)(1-2\nu)}{1-\nu}r,$$

where

$$p = \frac{\sigma_z}{\sigma_x} = \frac{\sigma_z}{\sigma_y}, \quad r = \frac{\sigma_z}{\varepsilon_z}.$$

Similar relations hold when the nonzero displacements are alternatively applied in the x and y directions.

Here, we focus our attention on the parallel performance of the solver for the RVE BVP. The presented results are based on three different test cases of RVE. The voxel representations (with sizes $32 \times 32 \times 32$, $64 \times 64 \times 64$, and $128 \times 128 \times 128$) of the related bone micro structures are extracted from a high resolution CT image, taken from [9]. The elastic properties of the bone materials are taken from [2]. The elasticity modulus of the solid phase is $E_s = 14.7\,\text{GPa}$ and the related Poisson ratio is $\nu_s = 0.325$. The fluid phase is considered as an almost incompressible material with bulk modulus $K_f = 2.3\,\text{GPa}$. In the presented parallel tests, the Poisson ratio is ν_f varied between 0.4 and 0.499999, and respectively $E_f = 3K_f(1-2\nu_f)$.

The studied methods and algorithms are targeted to the case of μFEM numerical homogenization, which leads to large-scale linear systems. The solver is based on the PCG method, where BoomerAMG is the preconditioner. A relative PCG stopping criterion in the form

$$\mathbf{r}_k^T C^{-1} \mathbf{r}_k \leq \varepsilon^2 \mathbf{r}_0^T C^{-1} \mathbf{r}_0,$$

where $\mathbf{r}_k$ stands for the residual on the k-th step of the PCG method, is used. The selected coarsening algorithm for the BoomerAMG preconditioner is *Falgout-CLJP* with a *strength threshold* of 0.02. *Modified classical interpolation* is applied. The selected relaxation method is *hybrid symmetric Gauss-Seidel or SSOR* in lexicographical ordering on each processor. In order to fit in the available operating memory three levels of aggressive coarsening are used and the maximum number of elements per row for the interpolation is restricted to six.

The presented parallel tests are performed on the IBM Blue Gene/P machine at the Bulgarian Supercomputing Center (see `http://www.scc.acad.bg/`). This supercomputer consists of 2048 PowerPC 450 based compute nodes, each with four cores running at 850 MHz and 2 GB RAM. It is equipped with a torus network for the point to point communications capable of 5.1 GB/s and a tree network for global communications with a bandwidth of 1.7 GB/s. Our code is compiled using the IBM XL C++ compiler with the following options: "`-O5 -qstrict`".

The parallel times and the weak scaling of the linear solver for $\nu = 0.4$, 0.499, and 0.499999 are presented in Tables 1, 2, and 3 respectively. The number of processors is denoted with p. N_{it} stands for the number of performed PCG iterations. The setup time for the BoomerAMG preconditioner T_{setup} is given separately from the solution time T_{solve}. T stands for the total time including both the setup and the solution stage. The *weak scaling* of the solver is measured on the smallest sample ($32 \times 32 \times 32$) with two consecutive uniform mesh refinement steps, and on the next

95

sample ($64 \times 64 \times 64$) with one refinement. After that, the parallel times for the largest sample are presented.

Table 1: Parallel times and weak scaling for $\nu = 0.4$, $\varepsilon = 10^{-3}$.

Bone sample	p	DOF	N_{it}	T_{setup}	T_{solve}	T	Weak scaling
$32 \times 32 \times 32$	1	608 256	8	32	59	91	
	8	4 792 320	9	45	67	112	82 %
	64	38 043 648	10	72	77	149	61 %
$64 \times 64 \times 64$	8	4 792 320	9	47	68	115	
	64	38 043 648	9	72	71	143	80 %
$128 \times 128 \times 128$	64	38 043 648	9	75	71	146	

Table 2: Parallel times and weak scaling for $\nu = 0.499$, $\varepsilon = 10^{-3}$.

Bone sample	p	DOF	N_{it}	T_{setup}	T_{solve}	T	Weak scaling
$32 \times 32 \times 32$	1	608 256	13	32	95	127	
	8	4 792 320	12	45	90	135	94 %
	64	38 043 648	17	79	129	209	61 %
$64 \times 64 \times 64$	8	4 792 320	17	50	127	178	
	64	38 043 648	19	77	144	220	81 %
$128 \times 128 \times 128$	64	38 043 648	13	83	104	187	

Table 3: Parallel times and weak scaling for $\nu = 0.499999$, $\varepsilon = 10^{-3}$.

Bone sample	p	DOF	N_{it}	T_{setup}	T_{solve}	T	Weak scaling
$32 \times 32 \times 32$	1	608 256	9	32	69	101	
	8	4 792 320	10	45	78	123	83 %
	64	38 043 648	11	92	91	183	55 %
$64 \times 64 \times 64$	8	4 792 320	16	50	121	172	
	64	38 043 648	14	90	113	202	85 %
$128 \times 128 \times 128$	64	38 043 648	11	93	92	185	

The observed numbers of iterations strongly depend on the Poisson ratio of the fluid phase. The worst case is for $\nu = 0.499$. Less iterations are needed in last case, where the Poisson ratio is much closer to the incompressibility limit. The parallel times depend both on the size of the problems and on the structure of the considered bone samples. However, similar parallel times for both setup and solution steps are

observed for the problems with equal numbers of degrees of freedom (DOF) when the numbers of the performed PCG iterations are equal. The results indicate that the BoomerAMG setup stage has a worse scalability than the PCG solution stage.

The homogenized elastic material coefficients for the three test specimens are presented next in Table 4. The coefficients (E_{fluid}, ν_{fluid}) correspond to the case when the fluid phase of the bone micro structure is considered as an almost incompressible linear elastic material. If the fluid phase is not taken into account (a solid skeleton with empty pores is considered), the resulting homogenized material properties are (E_{empty}, ν_{empty}). A significant difference between the two sets of results is observed. The percentages of the volume of the solid phase for the samples are also included in the table.

Table 4: Homogenized elastic properties of the bone samples.

	$32 \times 32 \times 32$	$64 \times 64 \times 64$	$128 \times 128 \times 128$
Solid phase	26 %	19 %	18 %
E_{empty}	1.10 GPa	0.58 GPa	0.70 GPa
ν_{empty}	0.141	0.158	0.142
E_{fluid}	1.86 GPa	1.05 GPa	1.23 GPa
ν_{fluid}	0.396	0.436	0.425

The presented parallel tests on the BlueGene/P massively parallel computer demonstrate the ability to simulate, with reasonable efficiency, bones with strongly heterogeneous structure at the level of detail provided by the micro CT resolution.

The presented numerical homogenization results clearly demonstrate the significant contribution of the fluid phase to the homogenized material properties.

Acknowledgements

References

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1996.

[2] S. Cowin, *Bone poroelasticity*, J. Biomechanics 32 (1999) 217–238.

[3] R. Kohut, *The determination of Young modulus E and Poisson number ν of geocomposite material using mathematical modelling*, private communication, 2008.

[4] N. Kosturski, S. Margenov, *Numerical Homogenization of Bone Microstructure*, in: Large Scale Scientific Computations 2009, to appear as a special volume of Lecture Notes in Computer Science, 2009.

[5] J. Kraus, S. Margenov, *Robust Algebraic Multilevel Methods and Algorithms*, Radon Series on Computational and Applied Mathematics, 5, de Gruyter, 2009

[6] S. Margenov, Y. Vutov, *Preconditioning of Voxel FEM Elliptic Systems*, TASK Quaterly 11 (2007) 117–128.

[7] A. Wirth, C. Flaig, T. Mueller, R. Muller, P. Arbenz, G. van Lenthe, *Fast smooth-surface micro-finite element analysis of large-scale bone models*, Journal of Biomechanics 41 (1) (2008) 100.

[8] Lawrence Livermore National Laboratory, *Scalable Linear Solvers Project*, `http://www.llnl.gov/CASC/linear_solvers/`.

[9] Vertebral Body Data Set ESA29-99-L3, `http://bone3d.zib.de/data/2005/ESA29-99-L3`.

Challenges for the Iterative Solution of Phase-Separation Flow Problems

Maya Neytcheva, Petia Boyanova

1 Introduction

We are interested in the numerical solution of problems, modelled by the so-called Cahn-Hilliard equation.

The basic Cahn-Hilliard equation (after John W. Cahn and John E. Hilliard) is a partial differential equation which describes the process of phase separation, by which the two components of a binary fluid spontaneously separate and form domains pure in each component.

The problem reads as follows

$$
\begin{aligned}
\frac{\partial C}{\partial t} - \varepsilon(\Delta(\Psi'(C) - \Delta C)) &= 0 \quad \text{in } \Omega_T \equiv \Omega \times (0, T) \\
\frac{\partial C}{\partial \mathbf{n}} = \frac{\partial}{\partial \mathbf{n}}(\Psi'(C) - \Delta C) &= 0 \qquad \text{on } \partial\Omega_T \\
C &= C_0 \qquad \text{in } \Omega \times (0)
\end{aligned}
\tag{1}
$$

where C is the concentration of the two different phases, Ψ is the so-called double-well potential with two global minima at ± 1, e.g., $\Psi(C) = (C+1)^2(C-1)^2$, ε is a positive but small parameter and defines the *interaction length*. The boundary conditions imply that the mixture can not pass through walls. Further, $\frac{\partial C}{\partial \mathbf{n}} = 0$ means that the total free energy decreases with time. The function $\eta = \Psi'(C) - \Delta C$ is the so-called *chemical potential.*

As is seen from (1), Cahn-Hilliard (C-H) equation is a fourth-order nonlinear parabolic equation, which turns out to be very stiff and hard to solve numerically.

In related works, various formulations of the Cahn-Hilliard equation have been considered, describing different physical phenomena, in conjunction with phase separation. Formulation (1), referred to as the 'solid model', is among the simplest, in which the diffusion is the only transport process. The parameter ε determines the slope of the surfaces, formed by the concentration, and the difficulty in solving (1) increases with decreasing ε. In the discrete case, special care has to be taken on relating the space discretization to ε in order to resolve the layers well enough.

A more advanced 'fluid model' is the one described below,

$$
\frac{\partial C}{\partial t} + (\mathbf{u} \cdot \nabla)C - \nabla \cdot (k\nabla\eta) = 0,
\tag{2}
$$

where k is the so-called mobility. In (2), $\mathbf{u}$ denotes the velocity of a hydrodynamic flow, the presence of which means that material can be transported both by diffusion

and flow field. The velocity is computed by solving the Navier-Stokes equations for fluid flow, coupled with C-H equation,

$$
\begin{aligned}
\rho \frac{\partial \mathbf{u}}{\partial t} + (\rho \mathbf{u} \cdot \nabla)\mathbf{u} &= -\nabla p + \nabla \cdot (\mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)) - \eta \nabla C + \mathbf{F} \\
\nabla \cdot \mathbf{u} &= 0
\end{aligned}
\tag{3}
$$

with $\rho(C)$ abd $\mu(C)$ being the fluid density and viscosity and $\mathbf{F}$ - body force. Solving (2) and (3) as a coupled system is a substantial computational challenge and therefore the equations are usually solved in an alternating fashion until convergence is achieved. The area of research where Cahn-Hilliard equations are applied, is progressively extending. For example, it has recently been found that in addition to the above well-known models a new model of phase separation is required to describe the phase-separation behaviour of a dynamically asymmetric mixture, which is composed of fast and slow components. Such 'dynamic asymmetry' can be induced by either the large size difference or the difference in transition temperature between the components of a mixture. The former often exists in so-called complex fluids, such as polymer solutions, micellar solutions, colloidal suspensions, emulsions and protein solutions. This new type of phase separation is referred to as 'viscoelastic phase separation' since viscoelastic effects play a dominant role.
In the sequel we consider only the models (1) and (2), and the solution of the algebraic systems arising from their discretization by the Finite Element method (FEM) in space and a suitable (stable) discretization in time.

2 Discretization and matrix formulation

To avoid discretizations of the higher order derivatives, C-H equation is decomposed into a coupled system of two second-order PDEs. Then, the 'fluid mode' reads as

$$
\begin{aligned}
\eta + \nabla^2 C - Cn^2 \, \Psi'(C) &= 0, \\
\frac{\partial C}{\partial t} + (\mathbf{u} \cdot \nabla)C - \frac{1}{Pe}\nabla \cdot (k\nabla \eta) &= 0.
\end{aligned}
\tag{4}
$$

where Cn and Pe are Cahn and Pecle numbers appearing after the problem is made dimensionless. We recall, that C is a continuous scalar variable that describes the diffusive interface profile and has a constant value in each phase, ± 1.
Consider first a discretization in space using the finite element method (FEM). Here we omit the details since the space discretization is straight forward. Then, the semidiscrete problem takes the form

$$
\begin{aligned}
M\eta_h &+ \left(KC_h - (C_h^3 - C_h)\right) &= 0, \\
\tfrac{1}{Pe}K\eta_h &+ \quad M\frac{\partial C_h}{\partial t} + VC_h &= 0,
\end{aligned}
\tag{5}
$$

where M, K, V are the mass, stiffness and discrete convection matrices. Within this work we use standard piece-wise linear FEM basis functions. More detailed analysis of that space discretization can be found in [5].

To discretize in time we use the fully implicit Euler method and therefore the fully discrete system reads as

$$
\begin{array}{rcl}
M\eta_h^{k+1} & - & KC_h^{k+1} - ((C_h^{k+1})^3 - C_h^{k+1}) = 0 \\
\frac{\delta_t}{Pe} K\eta_h^{k+1} & + & MC_h^{k+1} + \delta_t \, VC_h^{k+1} = MC_h^k
\end{array}
\tag{6}
$$

The so-arising system of algebraic equations is nonlinear due to the chemical potential function and is solved by Newton's method. The linear systems with the Jacobian, which have to be handled at each nonlinear iteration, are solved using an iterative method. Both the system matrix and the corresponding Jacobian have a two-by-two block structure,

$$
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},
\tag{7}
$$

which is utilized when constructing a suitable preconditioner.

3   Numerical solution methods

The real-life applications of interest, which are modeled via C-H, lead to large scale linear systems. Therefore, due to their lower computer resource demands, iterative techniques become the methods of choice. For achieving numerical efficiency (fewer iterations to converge), iterative solvers require computationally efficient and robust preconditioners. The construction of such preconditioners is our main interest.

There is a vast amount of theoretical results and experience on how to construct efficient preconditioners for matrices in the form (7) The best known preconditioners utilize the available matrix structure, i.e.,

$$
M = \begin{bmatrix} B_{11} & 0 \\ A_{21} & S \end{bmatrix} \begin{bmatrix} I_1 & B_{11}^{-1} A_{21} \\ 0 & I_2 \end{bmatrix} \quad \text{or} \quad M = \begin{bmatrix} B_{11} & 0 \\ A_{21} & S \end{bmatrix},
\tag{8}
$$

where $S \sim S_A \equiv = A_{22} - A_{21} A_{11}^{-1} A_{12}$, $B_{11} \sim A_{11}$ (or $B_{11}^{-1} \sim A_{11}^{-1}$). For details on preconditioners for matrices of two-by-two block form and their properties, we refer to [3] and [4], for example.

The existing theory states that there arises the need for an accurate approximation of the Schur complement S_A. The task to approximate the Schur complement of a matrix in a two-by-two block form is studied in various frameworks, such as hierarchical bases and nested meshes, algebraically constructed levels, systems of PDEs. The framework is also related to (A)MG methods.

Clearly, to approximate S_A, and in particular, by a sparse matrix, is not an easy task, since for large problems S_A is still a large matrix. Furthermore, even if A is sparse, S_A might be dense.

In this work we precondition A by

$$
M = \begin{bmatrix} [A_{11}] & 0 \\ A_{21} & S \end{bmatrix} \begin{bmatrix} I_1 & [A_{11}]^{-1} A_{12} \\ 0 & I_2 \end{bmatrix},
\tag{9}
$$

where $S \sim S_A$. The notation $[A_{11}]$ is used to denote that an inner (cheap) iterative method is used to solve systems with the pivot block.

To approximate the Schur complement we apply the so-called 'element-by-element' (EBE) technology. Briefly, the EBE idea is as follows. We notice, that when discretizing in space, the arising element stiffness matrix is in a block two-by-two form, $A_k = \begin{bmatrix} A_{11,k} & A_{12,k} \\ A_{21,k} & A_{22,k} \end{bmatrix}$. A_k is of small size and is easy to manipulate. We compute exactly $S_k = A_{22,k} - A_{21,k} A_{11,k}^{-1} A_{12,k}$ and then assemble those in the usual FEM way, $S = \sum_{k=1}^{m} S_k$, where m is the number of finite elements in the discretization mesh. In this way we construct a sparse approximation of the Schur complement in a cheap way. The EBE technology replaces the exact Schur complement

$$ S_A = A_{22} - \left(\sum_k R_k^T A_{21,k} R_k \right) \left(\sum_l R_l^T A_{11,l} R_l \right)^{-1} \left(\sum_m R_m^T A_{12,m} R_m \right) $$

by $S = A_{22} - \sum_k R_k^T \left(A_{21,k} A_{11,k}^{-1} A_{12,k} \right) R_k$, where R_m are the Boolean maps from local to global numbering of the degrees of freedom.

The question how good is the so-constructed approximation is relevant. It is shown that the approximation is very good for symmetric and positive definite matrices. For general nonsymmetric matrices, the study is not completed and there are both very promising results as well as cases where the quality of the approximation is not high enough to ensure optimal order preconditioned method. For details, we refer to [2], [6] and the references therein.

4 Numerical experiments

Below, we illustrate the behaviour of the solution of C-H problem and the performance of an iterative solution method. The preconditioner is as in (9). All experiments are performed in Matlab. The iterative solver is the Generalized Conjugate Gradient (GCG) method (cf. [1]). In Figure 1 we show the solution of problem (1) for $\varepsilon = 0.01$ on a regular mesh of size 64×64, time increment 10^{-5} and 50 time-steps. The space discretization parameter h is chosen in such a way that there are about six discretization points in the layer regions. Table 1 illustrates the quality of the EBE approximation of the Schur complement for problem (4). Here $Ca = 300, Pe = 0.1$ As we can see, the iteration counts are relatively few but nearly double with increasing the problem size four times. For smaller problem sizes, some eigenvalue information for $S^{-1}S_A$ is included, showing some increase in both real and imaginary parts of the eigenvalues.
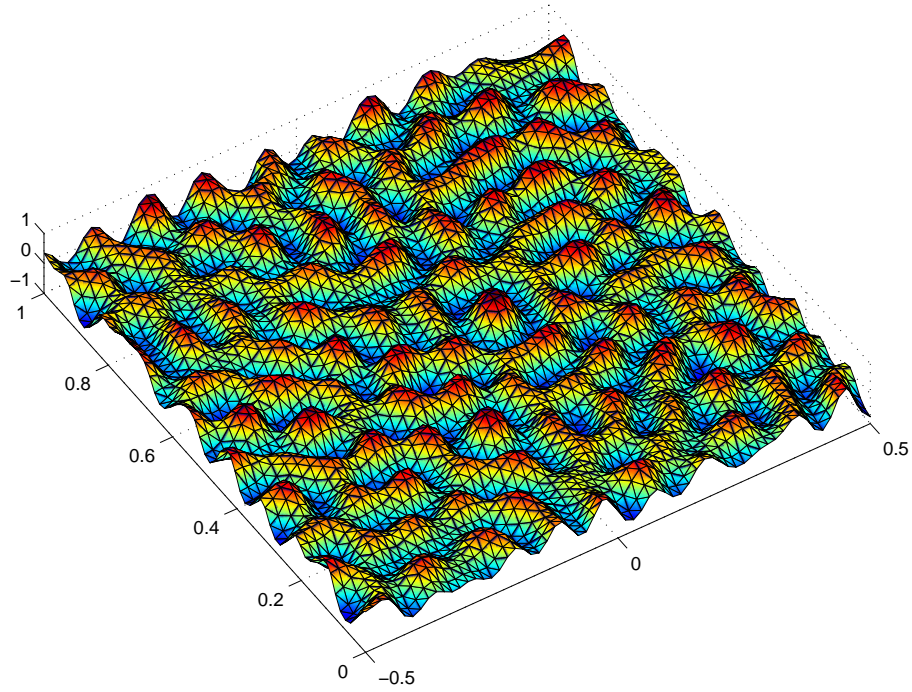
Figure 1:

Size	Iter	$eig(S^{-1}S_A)$		
		$min(Real)$	$max(Real)$	$max(Imag)$
1122/ 561	11	1	3.0642	0.2728
4290/ 2145	23	1	10.5635	1.3129
16770/ 8385	49	-	-	-
66306/16770	88	-	-	-

Table 1:

103

5 Conclusions and outlook

From Table 1 we see that the element-by-element Schur complement approximation shows reasonable performance, however it is not fully robust with respect to the space discretization parameter. Therefore, the technique how to approximate the Schur complement should be improved in a future work. One idea how to improve the quality of the Schur complement approximation is to use coarse grid correction. The second direction of research is to use nonconforming FEM to discretize the C-H equation. Then, for the Croizeux-Raviart element, the diagonal blocks in (7) become diagonal and the corresponding Schur complement can be explicitly computed. However, nonconforming FEM will lead to increase in the number of degrees of freedom (roughly three times in 2D). A thorough analysis would be necessary to compare the increase of computational complexity due to discretization with the computational gains due to a better matrix structure.

References

[1] O. Axelsson, *Iterative Solution Methods*, Oxford University Press, 1994.

[2] O. Axelsson, R. Blaheta, M. Neytcheva. Preconditioning for boundary value problems using elementwise Schur complements. *SIAM Journal on Matrix Analysis and Applications*, 31 (2009), 767-789.

[3] O. Axelsson, M. Neytcheva, Preconditioning methods for linear systems arising in constrained optimization problems, *Numerical Linear Algebra with Applications*, 10 (2003), 3-31.

[4] M. Benzi, G. H. Golub, and J. Liesen. Numerical Solution of Saddle Point Problems, *Acta Numerica*, 14 (2005), pp. 1-137.

[5] C. Elliott, D. French, F. Milner, A second order splitting method for the Cahn-Hilliard equation, *Numerische Matematik*, 54 (1989), 575-590.

[6] M. Neytcheva On element-by-element Schur complement approximations. Submitted to *Linear Algebra and Its Applications*, 2009.

Upscaling of Deformable Porous Media with Applications to Bone Modelling

Peter Popov

1   Introduction

In this work we consider upscaling the deformable porous microstructure of trabecular bones. The work is in progress, with the eventual goal being performing coarse-scale simulations of coupled fluid flow and mechanical deformations of bone tissue. Currently, a number of upscaling approaches exist, aimed at resolving the coarse scale mechanical response of bones [6, 9]. The fluid phase is not explicitly accounted for, eventhough some works modify the effective elastic coefficient in order to account for the fluid. In contrast, we consider Biot's diphasic model of poroelasticity [2, 3], in which a macroscopic point is simultaneously occupied by fluid and solid. We explicitly compute the appropriate macroscopic coefficients from the porous microstructure.

2   The physical problem at the fine scale

The physical processes under consideration span two length-scales. On the macroscopic level, one has fluid, diffusing through a porous solid. At the microscale the solid has a complex pore geometry and interacts with a Stokes flow. We assume good scale separation, with the usual small parameter ε being the ratio of the fine to the coarse length scales. We denote the fine scale domain by Ω_ε^0, which contains two subdomains - a fluid part $\mathcal{F}_\varepsilon^0$ and solid part $\mathcal{S}_\varepsilon^0$. The superscript 0 indicates the reference, or undeformed configuration of the body. The interface between the solid and fluid domains is denoted by $\Gamma_\varepsilon^0 = \partial \mathcal{F}_\varepsilon^0 \cap \partial \mathcal{S}_\varepsilon^0$. The physics is described by the strongly coupled, stationary fluid structure interaction problem [4]: *Find the interface Γ_ε, velocity $\mathbf{v}_\varepsilon$, pressure p_ε and displacements $\mathbf{u}_\varepsilon$ such that:*

$$\Gamma_\varepsilon = \left\{ \mathbf{X} + \mathbf{u}_\varepsilon(\mathbf{X}) |\ \forall\ \mathbf{X} \in \Gamma_\varepsilon^0 \right\}, \tag{1}$$

the Stokes and Elasticity equations are satisfied:

$$-\nabla p_\varepsilon + \mu \Delta \mathbf{v}_\varepsilon + \mathbf{f} = 0, \qquad \nabla \cdot \mathbf{v}_\varepsilon = 0 \qquad \qquad in\ \mathcal{F}_\varepsilon \tag{2}$$

$$-\nabla \cdot \mathbf{S}(\mathbf{E}_\varepsilon) = \mathbf{f} \qquad \qquad in\ \mathcal{S}_\varepsilon^0 \tag{3}$$

with the interface condition

$$\det\left(\mathbf{I} + \nabla \mathbf{u}_\varepsilon\right)\left(-p_\varepsilon \mathbf{I} + 2\mu \mathbf{D}_\varepsilon\right)\left(\mathbf{I} + \nabla \mathbf{u}_\varepsilon\right)^{-T} \mathbf{n}_0 = \mathbf{S}(\mathbf{E}_\varepsilon)\mathbf{n}_0 \qquad \qquad on\ \Gamma_\varepsilon^0. \tag{4}$$

Here $\mathbf{X}$ is the material and $\mathbf{x}$ the spatial coordinate, μ is the fluid viscosity, $\mathbf{f}$ is the body force. Further, $\mathbf{S}$ is the first Piola-Kirchhoff stress tensor in the solid. It

depends, linearly or nonlinearly, on the strain $\mathbf{E}(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u}(\mathbf{X}) + \nabla\mathbf{u}(\mathbf{X})^T\right)$. Recall, that given a Cauchy stress $\boldsymbol{\sigma}$ one has $\mathbf{S} = \det(\mathbf{I} + \nabla\mathbf{u}_\varepsilon)\boldsymbol{\sigma}\left(\mathbf{I} + \nabla\mathbf{u}_\varepsilon\right)^{-T}$. $\mathbf{D}(\mathbf{v}) = \frac{1}{2}\left(\nabla\mathbf{v}(\mathbf{x}) + \nabla\mathbf{v}(\mathbf{x})^T\right)$ is the rate of stretching tensor, and $\mathbf{n}_0$ is the reference normal to the interface. Note, that the interface condition (4) introduces a nonlinearity into the problem, regardless of the constitutive form for $\mathbf{S}$.

3 Upscaled Biot model

Recall that in classical linear poroelasticity it is assumed that the solid is governed by the linearized constitutive relationship $\mathbf{S} = \mathcal{L} : \mathbf{E}$, where $\mathcal{L}$ is the elasticity tensor. The form of the homogenized equations depends on the magnitude of the displacements of the microscopic interface Γ_ε^0. If one assumes infinitesimal displacements in the sense that there is no difference between Γ_ε^0 and Γ_ε, and in addition, a quasi-steady regime, the media is described by the classical Biot's law (c.f. e.g. [2, 5]):

$$\nabla \cdot \left(\mathcal{L}^* : \mathbf{e}\left(\mathbf{u}_0\right) - \mathbf{A}^* p_0\right) = \mathbf{0}, \tag{5a}$$

$$\nabla \cdot \left(\mathbf{K}^* \nabla p_0\right) = \left(\nabla\phi_f\right) \cdot \frac{\partial \mathbf{u}_0}{\partial t} + \mathbf{A}^* : \mathbf{e}\left(\frac{\partial \mathbf{u}_0}{\partial t}\right) + \beta^* \frac{\partial p_0}{\partial t}. \tag{5b}$$

Here p_0 and $\mathbf{u}_0$ are the macroscopic pressure and displacements. In the above equations, the acoustic effects in both phases are disregarded. These equations form the phenomenological theory of [2]. Later, they were re-derived via asymptotic homogenization (c.f. e.g. [1, 8]). They involve three different sets of cell problems needed to determine the effective properties $\mathcal{L}^*$, $\mathbf{A}^*$, $\mathbf{K}^*$ and β^*.

To obtain the cell problems a Representative Element of Volume (REV) Y is defined. Since we have two domains, we denote with $Y_{\mathcal{F}}$ the fluid part and with Y_{S} the solid part of the REV, that is, $Y = Y_{\mathcal{F}} \cup Y_{\mathrm{S}}$. The initial interface is denoted by Γ_0^Y.

The first set of cell problems defines $\mathcal{L}^*$ and is (c.f., e.g. [3, 5]):

$$\nabla \cdot \left(\mathcal{L} : \mathbf{e}\left(\boldsymbol{\varphi}^{mn}\right)\right) = -\nabla \cdot \left(\mathcal{L} : \delta_{mn}\right) \qquad \text{in } Y_{\mathrm{S}}, \tag{6a}$$

$$\left(\mathcal{L} : \mathbf{e}\left(\boldsymbol{\varphi}^{mn}\right)\right)\mathbf{n}_0 = -\left(\mathcal{L} : \delta_{mn}\right)\mathbf{n}_0 \qquad \text{on } \Gamma_0^Y, \tag{6b}$$

with $\boldsymbol{\varphi}^{mn}$ being Y periodic and with zero mean. Note, that due to symmetry, the cell problems can be reduced to just six. The cell solutions $\boldsymbol{\varphi}^{mn}$ then determine effective elastic properties $\mathcal{L}^*$ as follows:

$$\mathcal{L}^*_{ijkl} = \langle\mathcal{L}_{ijpq}\left(\delta_{pk}\delta_{ql} + e_{pq}\left(\boldsymbol{\varphi}_{mn}\right)\right)\rangle_Y. \tag{7}$$

Here $\langle\cdot\rangle_Y$ is the Y- averaging operator. Note that the effective tensor $\mathcal{L}^*$ is simply the homogenized elasticity coefficient in the absence of any fluid.

Next, the Biot coefficient $\mathbf{A}^*$ is determined from the single cell problem

$$\nabla \cdot \left(\mathcal{L} : \mathbf{e}\left(\boldsymbol{\eta}\right)\right) = 0 \qquad \text{in } Y_{\mathrm{S}}, \tag{8a}$$

$$\left(\mathcal{L} : \mathbf{e}\left(\boldsymbol{\eta}\right)\right)\mathbf{n}_0 = \mathbf{n}_0 \qquad \text{on } \Gamma_0^Y, \tag{8b}$$
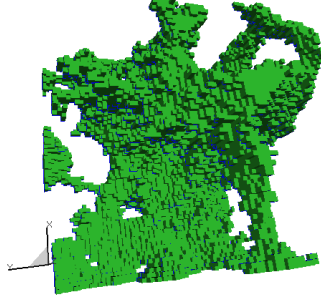
106

Figure 1: CAT scan of trabecular bone tissue, largest REV (64^3).

with $\boldsymbol{\eta}$ being Y periodic and with zero mean. The solution yields $\mathbf{A}^*$ as follows:

$$A_{ij}^* = \phi\delta_{ij} + \langle \mathcal{L}_{ijpq}e_{pq}\left(\boldsymbol{\eta}\right)\rangle_Y. \tag{9}$$

The Biot coefficient takes into account the contribution of the fluid pressure p_0 into the momentum equation (6a), as well as the pore volume change term $\mathbf{A}^* : \mathbf{e}\left(\frac{\partial \mathbf{u_0}}{\partial t}\right)$ in the balance of mass (6b), due to macroscopic deformations $\mathbf{u}_0$. The pore volume change in (6b) due to p_0 is captured by the β^* coefficient, defined as:

$$\beta^* = \langle e_{ii}\left(\boldsymbol{\eta}\right)\rangle_Y. \tag{10}$$

The permeability is determined via the final set of cell problems:

$$\Delta\mathbf{w}_i + \nabla q_i = -\mathbf{e}_i, \qquad \nabla \cdot \mathbf{w}_i = 0 \qquad \text{in } Y_{\mathcal{F}}, \tag{11a}$$

$$\mathbf{w}_i = \mathbf{0} \qquad \text{on } \Gamma_0^Y, \tag{11b}$$

with the permeability $\mathbf{K}^*$ being defined as

$$K_{ij}^* = \langle \mathbf{w}_i \cdot \mathbf{e}_j\rangle_Y. \tag{12}$$

4   Numerical Results

In this section we present a numerical example of homogenizing a real porous microstructure of a human thigh bone. A series of REVs were considered, with 16^3, 32^3 and 64^3 voxels, each being a subset of the next larger size (Figure 1). The voxelized geometry, obtained via CAT scan, were used in a number of studies, related to upscaling the elastic properties only [6, 9]. The goal here was to obtain both the elastic ($\mathcal{L}^*$) and Biot's coefficients ($\mathbf{A}^*$, β^*), so that simulations based on Biot's poroelastic model could be performed.

The voxelized geometry was converted to a tetrahedral mesh. Due to the large number of elements in a directly meshed voxelized geometry we also attempted to generate a coarser mesh by smoothing the original geometry. This was done by first starting
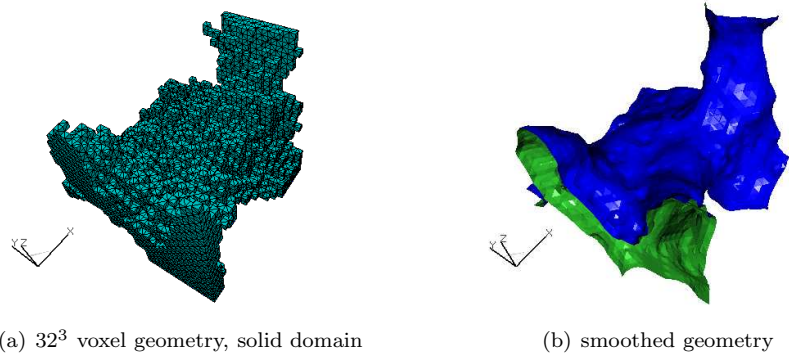
107

(a) 32^3 voxel geometry, solid domain   (b) smoothed geometry

Figure 2: Smoothing of the voxelized Fluid-solid interface

Table 1: Elastic properties for several REV sizes

REV Voxels^3	Solid Phase	E_{11}^*	E_{22}^* GPa	E_{33}^*	ν_{12}^*	ν_{21}^*	ν_{13}^*	ν_{31}^*	ν_{23}^*	ν_{32}^*
16^3	0.47%	5.72	5.69	4.09	0.32	0.32	0.31	0.22	0.31	0.22
32^3	0.26%	2.26	1.72	1.46	0.27	0.21	0.20	0.13	0.30	0.26
64^3	0.19%	1.08	0.90	1.10	0.23	0.19	0.17	0.17	0.19	0.23

with the original tetrahedral mesh of the entire REV, fluid and solid voxels being marked with two distinct markers. Secondly, all elements are removed except those bordering the Fluid-solid interface. Next, a standard Laplace equation is solved in the remaining voxels, with a Dirichlet boundary condition of 0 at the boundary of the fluid voxels which is not boundary of the REV and a Dirichlet boundary condition of 1 at the boundary of the solid voxels which is not boundary of the REV. The remainder of the boundary had natural boundary conditions. The Fluid-Solid interface is then reconstructed as particular level surface of the solution to this Laplace problem. The resulting level surface (Figure 2(b)) retained the topology of the original interface and was much smoother than the voxel-based interface. Unfortunately, we were not able to then generate a mesh from this surface. For that reason, all the computations were performed on voxel-derived meshes (Figure 2(a)).

The homogenization was conducted for linear isotropic elastic material with $E = 14.7 GPA$ and $\nu = 0.35$. The solution of the elastic cell problem (6), for the 32^3 REV is shown in Figure 3. As the geometry is not periodic, we used linear boundary conditions, instead of periodic. A summary of the results for several REV sizes is given in Table 1. Shown are the Young modulii in the three coordinate directions (E_{11}, E_{22} and E_{33}) as well as the Poisson ratios. From the table it is seen that the REV sizes were not sufficiently large to obtain converging properties. The anisotropy in the coefficients is however decreasing with increasing REV size. This implies that a factor of 2 or 4 increase in REV size will be sufficient for the volume to be truly
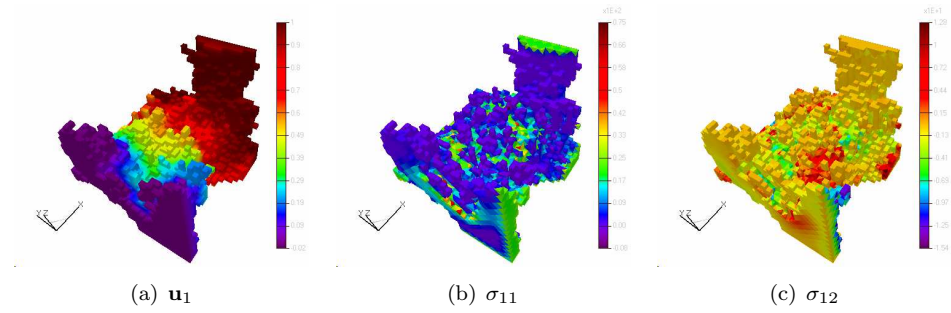
108

(a) $\mathbf{u}_1$ (b) σ_{11} (c) σ_{12}

Figure 3: First cell problem (out of six) for computing elastic modulii

Table 2: Fluid-Solid coupling coefficients

REV	A_{11}^*	A_{22}^*	A_{33}^*	β^*
16^3	0.79	0.79	0.89	0.021
32^3	0.46	0.48	0.49	0.014
64^3	0.37	0.38	0.37	0.011

representative. This will be addressed in future works either by meshing the smoothed interface or by optimizing the existing solver.

The results for the Biot coefficient $\mathbf{A}^*$ were also obtained for the same geometries by solving the cell problem (8). The tensor is diagonal with the three entries given in Table 2. The cell solution for the 32^3 case is shown in Figure 4. It is seen that as the REV is increased, the tensor becomes almost isotropic.

5   Conclusions

We have successfully computed the macroscopic coefficients for Biot's poroelasticity equations for a couple of REVs representative of the microstructure of human thigh
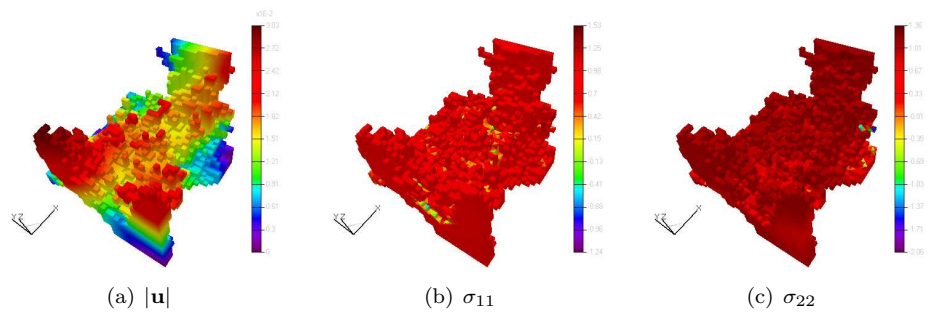

(a) $|\mathbf{u}|$ (b) σ_{11} (c) σ_{22}

Figure 4: Cell problem for computing fluid-solid coupling term

109

bone tissue. We obtained anisotropic elastic and biot tensors. In the future, we plan to first find a suitable way to mesh the smoothed FSI interface, so that bigger REVs can be handled. Secondly, we plan to solve representative macroscopic poroelastic problems and compare the results with the various proposed schemes for upscaling the bone microstrocure to purely elastic mascroscopic problems. If it is shown that fluid flow plays an important role, we further plan to compare the classical macroscopic model of Biot with schemes particularly designed for large pore-level deformations [7].

References

[1] J.-L. Auriault and E. Sanchez-Palencia. Étude du comportement macroscopique d'un milieu poreux saturé déformable. *Journal de Mécanique*, 16:575–603, 1977.

[2] M.A. Biot. General theory of three dimensional consolidation. *J. Appl. Phys.*, 12:155–164, 1941.

[3] U. Hornung. *Homogenization and porous media*. Interdisciplinary Applied Mathematics. 6.New York, Springer, 1997.

[4] O. Iliev, A. Mikelic, and P Popov. On upscaling certain flows in deformable porous media. *SIAM Multiscale Modeling and Simulation*, 7:93–123, 2008.

[5] Cheo K. Lee and Chiang C. Mei. Re-examination of the equations of poroelasticity. *Int. J. Eng. Sci.*, 35:329–352, 1997.

[6] S. Margenov and Y. Vutov. Preconditioning of voxel fem elliptic systems. *TASK, Quaterly*, 11:117–128, 2007.

[7] P. Popov, Y. Efendiev, and Y. Gorb. Multiscale modeling and simulation of fluid flows in highly deformable porous media. In I. Lirkov, S. Margenov, and J. Wasniewski, editors, *Large-Scale Scientific Computing*, volume 5910 of *Lecture notes in computer science*. Springer, 2010.

[8] E. Sanchez-Palencia. *Non-Homogeneous Media and Vibration Theory*, volume 127 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, 1980.

[9] A. Wirth, C. Flaig, T. Mueller, R. Mueller, P. Arbenz, and G. van Lenthe. Fast smooth-surface micro-finite element analysis of large-scale bone models. *Journal of Biomechanics*, 41:S100, 2008.

A Hybrid Multiscale Multigrid Approach by Incorporating Analytical Results

Rosangela F. Sviercoski, Svetozar Margenov

Abstract

In this paper, convergence results for a multigrid solver for a linear elliptic equation is presented, which uses the continuous homogenized problem to define coarse-grid operators. The differences from previous approaches are that the upscaled tensor is given by an analytical form presented in [11], which generally is a full tensor, and it is applicable to media not necessarily periodic. The numerical results indicate the robustness of the algorithm by comparing average convergence rates with the standards arithmetic and harmonic averages.

1 Introduction

Complex multiscale problems of flows in porous media are routinely encountered as part of research and development activities in a number of engineering, environmental and biomedical fields. A major challenge in computational sciences is the development of efficient methods for predicting the mechanical and flow behavior of systems with complex hierarchical structures. The presence of multiple length-scales in such media leads to numerical discretizations with very large number of degrees of freedom $(10^7 - 10^9)$. One option for solving these systems is by using iterative methods, especially hierarchical ones like multigrid, which often shows to have optimal complexity. However, in many interesting cases, particularly when the heterogeneous media is described by step functions with high contrast coefficient, optimal performance of multigrid is still a challenging problem. This may happen due to, for example, the mesh anisotropy, which is caused by elements having very large aspect ratio, which typically appears as a factor in the condition number of the stiffness matrix, which easily generates a highly ill-conditioned problem [7]. Another related problem is that the smallest eigenvalues in absolute value do not correspond to very smooth eigenfunctions. It is therefore, not easy to represent them on coarser grids [3]. Algorithms for dealing with these problems already exist, such as [1, 3, 4, 6, 8] among others, but yet there has been no definite answer. For example, in [3, 4, 8], the algorithms are defined by specific prolongation and restriction operators for periodic oscillations of the media. Because of this particular case, some have closed forms showing convergence criteria as a function of the period ε and discretization size δ.

In this paper, we also present convergence results for a multigrid solver for a linear elliptic equation (1), by using the continuous homogenized problem to define coarse-grid operators. Unlike previous works, there is no constrain regarding periodicity of the media, and the upscaled tensor is given by an analytical form presented in [11], which may be a full tensor. The results are still preliminaries and no explicit formalism has been yet developed. In the next section, the main problem and its

approach is presented and, in the last section, convergence results using two types of media. The first one is a randomly generated described by a polynomial type and, the second one is defined by step function with various degree of contrasts.

2 Setting of the Problem

Without loss of generality, we consider the two-dimensional boundary value problem (BVP) for $u(x)$ on $\Omega = [0,1]^2$ as:

$$\begin{cases} -\nabla \cdot (\mathbf{K}(x)\nabla u(x)) = f(x) & x \in \Omega \\ u(x) = g(x) & x \in \partial\Omega \end{cases} \tag{1}$$

where $\mathbf{K}(x) = k(x)\mathbf{I} \in M^{2\times 2}$ is positive definite, and may be called a permeability tensor, with $k(x) > 0$ being a scalar function and $\mathbf{I} \in M^{2\times 2}$ the identity matrix. The existence of an upscaled coefficient $\overline{\mathbf{K}} \in M^{2\times 2}$, positive definite, derived from periodic homogenization, and applicable to non-periodic media is based on the mathematical concept of G-convergence [5].

The homogenized coefficient $\overline{\mathbf{K}}$, replaced in (1), lead to the upscaled or homogenized solution, $\overline{u}(x)$, which converges to $u(x)$ in the sense of $H-$convergence [5, 10]. In this case, $\overline{\mathbf{K}}$ (2) is generally obtained by solving numerically the periodic cell-problem for $w_i(x)$ $i = 1, 2$ at each partition of Ω, χ_j, and it is obtained by [5]:

$$\overline{\mathbf{K}} = \frac{1}{|\chi_j|} \int_{\chi_j} \mathbf{K}(x) \begin{pmatrix} 1 + \frac{\partial w_1}{\partial x_1} & \frac{\partial w_1}{\partial x_2} \\ \frac{\partial w_2}{\partial x_1} & 1 + \frac{\partial w_2}{\partial x_2} \end{pmatrix} dx \tag{2}$$

In this paper, an analytical approximation to $\overline{\mathbf{K}}$ from [11] is used, which is given by:

$$\mathbf{K}^\star = \begin{pmatrix} 0.5\left(K_{11}^a + K_{11}^g\right) & -b\sin(2\theta) \\ -b\sin(2\theta) & 0.5\left(K_{22}^a + K_{22}^g\right) \end{pmatrix} \tag{3}$$

where K_{ii}^a and K_{ii}^g are the arithmetic and geometric averages of the well known Cardwell and Parsons bounds (CPB) [2]. The angle θ and the coefficient b, are derived from the rotation of a diagonal matrix and the center of mass of the cell. The coefficient is then given by:

$$b = \begin{cases} \frac{K_{ii}^\star - K_{jj}^\star}{2\cos(2\theta)} & if K_{ii} \neq Kjj \\ \frac{\tilde{K}_{ii} + \hat{K}_{jj}}{4} - \frac{\tilde{K}_{ii} + \tilde{K}_{jj}}{4} & if K_{ii} = K_{jj} \end{cases} \tag{4}$$

where $\tilde{K}_{ii}$ and $\hat{K}_{ii}$ are the lower and upper bounds of (CPB), respectively. As presented in [11], $\mathbf{K}^\star$ approximates the $\overline{\mathbf{K}}$ which is derived by solving $w_i(x)$'s with periodic boundary condition, even when $\mathbf{K}(x)$ is not periodic.

2.1 Analytic Homogenized Coarse Grid Operator

The sequence of grids: $\Omega_0 \subset \Omega_1 \subset ... \subset \Omega_h \subset ... \subset \Omega_k = \Omega$ is defined and the corresponding sequence of equations here is discretized by using bilinear rectangular

finite elements. The approximation $\mathbf{K}^\star$ leads to a sequence of linear operators from the coarse to fine level $L_0^\star U_0^\star = f_0, ..., L_h^\star U_h^\star = f_h, L_k U_k = f_k$, where a coarse-grid operator at a given level h, is:

$$(L_h^\star)_{ij} = \int_\Omega \nabla \varphi_j^h \mathbf{K}^\star \nabla \psi_i^h dx \tag{5}$$

where $\{\varphi_j^h\}_{j=1}^{n_h}$ and $\{\psi_j^h\}_{j=1}^{n_h}$ are test functions at that level. Application of a two-level multigrid to, say $L_k U_k = f_k$ at the nth iteration, is usually done by [3]:

1. Presmoothing. Compute an approximation to U_k^n, i.e. $U_k^{n\eta}$ by applying η steps of a iteration to the system at the $k-$level, with initial value U_k^n on the $k-$grid level. This can be formalized as: $U_k^{n\eta} = S^\eta(U_k^n, L_k, f_k)$.

2. Coarse grid correction. Define a coarse grid, $k-1$ level, and its operator $L_{k-1}^\star$, then: (2.1) - Restrict the residual to the $k-1$ grid, by solving $r_{k-1} = I_k^{k-1}(f_k - L_k U_k^{n+\frac{1}{2}})$. (2.2) - Solve for the corrector e_{k-1}, with $L_{k-1}^\star e_{k-1} = r_{k-1}$. (2.3) - Compute the approximation $\overline{U}$ by interpolating the correction back to the grid k, to have: $\overline{U} = U_k^{n\eta} + I_{k-1}^k e_{k-1}$.

3. Postsmoothing. Use $\overline{U}$ as the initial value to give: $U_k^{n+1} = S^{\eta_{pos}}(\overline{U}, L_k, f_k)$.

The resulting two-level multigrid method is then:

$$G = S^{\eta_{pos}}(I - I_k^{k-1}(L_{k-1}^\star)^{-1}I_{k-1}^k L_k)S^{\eta_{pre}} \tag{6}$$

This procedure can be repeated until the coarsest level is reached, where the correction is resolved exactly, corresponding to 1 V-cycle. In this paper, the intergrid operators were defined by using the standard full weight restriction and bilinear interpolation, like early works by [1, 3]. The smoothing, S in this paper, was done by applying Gauss-Seidel procedure η times. At each level, the corresponding homogenized operator is obtained by first dividing the entire domain into many cells. At each cell, $\mathbf{K}^\star$ is determined and $L^\star$ maintains the form of the homogenized operator but with still varying coefficient from one cell to another. The next section, the results from applying the above algorithm will be presented.

3   Numerical Results

The goal here is to show that the algorithm derived using $\mathbf{K}^\star$, which is generally given as an anisotropic and full tensor, can produce equal or better results than using the standard diagonal tensors given by either the arithmetic or the harmonic averages, $\mathbf{K}^a$ and $\mathbf{K}^h$, respectively. The BVP is given by:

$$\begin{cases} -\nabla \cdot (\mathbf{K}(x)\nabla u(x)) = f(x) & x \in \Omega \\ u(0, x_2) = 10, \ u(1, x_2) = 1 & x \in \partial\Omega \\ \frac{\partial u}{\partial \eta}(x_1, 0) = \frac{\partial u}{\partial \eta}(x_1, 1) = 0 & x \in \partial\Omega \end{cases} \tag{7}$$

The media $\mathbf{K}(x)$ are shown in Figure 1 taken from [11]. The medium on the right is described by step function, whereas the random medium on the left is a realization of a log-normal homogeneous stochastic process with expected value $E[Y(x,\omega)] = -1$ on $\Omega = [0,1]^2$. $Y(x,\omega)$ is a Gaussian process with covariance kernel given by $R(x,y) = \sigma^2 \exp\left(-\frac{|x_1-x_2|^2}{2L_1^2} - \frac{|y_1-y_2|^2}{2L_2^2}\right)$, with $\sigma = 4$ and $L_1 = 0.1$ is the correlation length in $x-$direction and, $L_2 = 0.2$ is the correlation in the $y-$direction. The results are shown on Tables 1-3, where the average convergence rate is shown for each case of $\mathbf{K}(x)$. The average convergence rate, ρ_a, is obtained as the average of the residual over the total number of V-cycles, starting from the 2nd up to the 10th V-cycle (in those tables). In another words, $\rho_a = 1 - [10^{(\log_{10}(residual(10)/residual(2))/8)}]$. In all of them, the fine-scale problem has a grid size of 63×63, followed by coarse-grids with 31×31 (30×30 upscaled tensors), 15×15 (14×14 upscaled tensors), 7×7 (6×6 upscaled tensors) and 3×3 (2×2 upscaled tensors), respectively. For the Fig. 1 (right), the mesh resolves the coefficient jumps, i.e., the value of $k(x)$ is unique within each rectangular element. All examples use an initial guess of $u = 0$ and solve the coarsest grid problem directly.
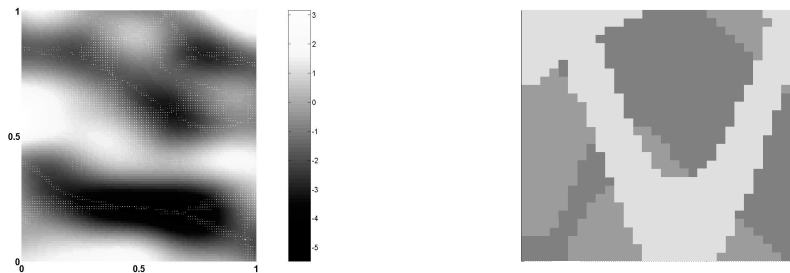

Figure 1: Media used on the simulations presented on Tables 1-3.

$\eta_{pre} = \eta_{pos}$	3	5	10	15	20	25
$\mathbf{K}^\star$	0.4364	0.6387	0.6302	0.6700	0.7041	0.7300
$\mathbf{K}^a$	0.4336	0.5930	0.5823	0.6305	0.6723	0.7037
$\mathbf{K}^h$	0.4316	0.7578	0.8124	0.8327	0.8336	0.8395

Table 1: Average convergence rates as a function of number of smoothing steps η, after 10 V-cycle iterations. Here $\mathbf{K}(x)$ is of polynomial type, Fig. 1(left). In this case, $\mathbf{K}^\star$ performed better than $\mathbf{K}^a$, but worse than $\mathbf{K}^h$.

In general, the operator $L^\star$ derived from $\mathbf{K}^\star$ showed to be a robust operator. This is because neither one of the others, that are commonly used [1, 3], showed to be the best on all the cases. Indeed, by looking at Table 1, the harmonic average showed better results than the others, but this was not consistent throughout the experiment, and it even failed to converge for the high contrast coefficient Table 3.

114

$\eta_{pre} = \eta_{pos}$	2	5	10	15	20	25
$\mathbf{K}^{\star}$	0.0257	0.5677	0.6871	0.7350	0.7687	0.7946
$\mathbf{K}^{a}$	0.1376	0.5720	0.6680	0.7172	0.7521	0.7896
$\mathbf{K}^{h}$	...	0.4879	0.6764	0.7340	0.7641	0.7829

Table 2: Average convergence rates as a function of number of smoothing steps η, after 10 V-cycle iterations, for $\mathbf{K}(x)$ as step function on Fig. 1 (right), where on the dark gray $\mathbf{K}(x) = 0.1$, on the gray $\mathbf{K}(x) = 0.5$ and on the light gray $\mathbf{K}(x) = 1$. In this case, $\mathbf{K}^{\star}$ showed to be a better option than the others. The sign ... indicates that it did not converge.

$\eta_{pre} = \eta_{pos}$	20	25	30	35	40	45	50
$\mathbf{K}^{\star}$	0.1230	0.3200	0.4293	0.4877	0.5303	0.5563	0.5846
$\mathbf{K}^{a}$	0.3955	0.5938	0.6642	0.7022	0.7261	0.7435	0.7576
$\mathbf{K}^{h}$	...	...	...	...	...	...	...

Table 3: Average convergence rates as a function of number of smoothing steps η, after 10 V-cycle iterations. $\mathbf{K}(x)$ is described as step function on Fig. 1 (right), where the dark gray $\mathbf{K}(x) = 0.1$, gray $\mathbf{K}(x) = 1$ and light gray $\mathbf{K}(x) = 10$. In this case, using $\mathbf{K}^{\star}$ demonstrated being a better option than the harmonic average, $\mathbf{K}^{h}$, which failed to converge, indicated by the sign ... above.

3.1 Conclusion

It is known that the efficiency of a multigrid method is tightly coupled to both the coarse-grid operatorï¿½s approximation of the fine-grid and the ability of the inter-grid transfer operators to approximate the interaction between the various scales. In this paper a coarse grid operator, $L^{\star}$, derived from an analytical approximation to the homogenized coefficient, $\mathbf{K}^{\star}$, showed to be a robust operator, in the sense that it converged, even when others failed to do so. As mentioned before, these are preliminary results suggesting that further improvements should be pursued. In this sense, the next step will be to derive more specific interpolation and restriction operators, instead of the standard ones used here. This may include the use of analytical approximation of the cell problems presented in [9, 10], in light of what has been presented, for example, in [8, 7].

References

[1] Alcouffe, R. E., Brandt, A., Dendy, J. E., Painter, J. W. *The Multi-grid Method for the Diffusion Equation with Strongly Discontinuous Coefficient*, SIAM J. Sci. Stat. Comput. (2), 4, pp. 430-454 (1981).

[2] Cardwell, W. T., Parsons, R.L. *Average Permeabilities of Heterogeneous Oils Sands.* Trans. Am. Inst. Mining. Met. Pet. Eng. pp 34-42 (1944).

[3] Enquist, B., Luo, E. *Convergence of a Multigrid Method Elliptic Equations with Highly Oscillatory Coefficients*, SIAM J. N. Anal. 34 (6), pp. 2254-2273 (1997).

[4] Fish, J., Belsky, V. *Multi-grid Method for Periodic Heterogenous Media - Part II: Multiscale Modeling and Quality Control in Multidimensional Case*, Comput. Meth. Appl. Mech. Eng. 126, pp. 17-38, (1995).

[5] Jikov, V. V., Kozlov, S.M., Oleinik, O.A. *Homogenization of Differential Operators and Integral Functionals.* Springer Verlag (1994).

[6] Knapek, S. *Matrix-dependent Multigrid-homogenization for Diffusion Problems*, SIAM J. Sci. 20 (2), pp.515-533, (1999).

[7] Kraus J., Margenov, S. *Robust Algebraic Multilevel Methods and Algorithms* Radon Series on Comput. and Appl. Mathematics (5), De Gruyter, Berlim (2009).

[8] Neuss N., Jaeger, W., Wittum G., *Homogenization and Multigrid*, Computing, 66 (1), pp. 1-26 (2001).

[9] Sviercoski, R. F., Winter, C. L., Warrick, A.W. *Analytical Approximation for the Generalized Laplace's Equation with Step Function Coefficient.* SIAM Journal of Applied Mathematics vol. 68 (5) - pp. 1268-1281, (2008).

[10] Sviercoski, R. F. , Travis, B. J., Hyman, J. M. *Analytical Effective Coefficient and First-Order Approximation for Linear Flow through Block Permeability Inclusions.* Comp. Math. Appl., 55, pp. 2118-2133, (2008).

[11] Sviercoski, R. F. *An Analytical Effective Tensor for Flows through Generalized Composites.* Submitted (2009).

Approximation Properties of Coarse Spaces by Algebraic Multigrid

Panayot S. Vassilevski

Abstract

This note provides an application of the "window"-based spectral AMG method (cf. [FVZ05] or [Va08]) for proving energy error estimates desired in providing coarse (upscaled) discretization of fairly general classes of PDEs. Computationally more efficient versions of the original window-based spectral AMG, as well as a new method utilizing local element matrices, are outlined summarizing the results from a forthcoming report.

1   The strong approximation property

We are given a s.p.d. $n \times n$ sparse matrix A and let $P : \mathbf{R}^{n_c} \mapsto \mathbf{R}^n$, $n_c < n$, be a given (rectangular) interpolation matrix.
We are interested in the following *strong approximation property*:
For any fine-grid vector $\mathbf{u} \in \mathbf{R}^n$ there is a coarse interpolant $P\mathbf{u}_c$ such that

$$\|A\|\|\mathbf{u} - P\mathbf{u}_c\|_A^2 \leq C_A \|A\mathbf{u}\|^2. \tag{1}$$

If the problem of our main interest

$$A\mathbf{u} = \mathbf{f},$$

comes from a finite element discretization of a PDE on a domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3), then $\mathbf{f} = (f_i)$ comes from a given r.h.s. function $f(\mathbf{x}) \in L_2(\Omega)$, where the entries f_i are computed as the following integral moments

$$f_i = (f, \varphi_i) \equiv \int\limits_\Omega f(\mathbf{x})\varphi_i \, d\mathbf{x}.$$

Above, φ_i runs over a basis of the fine–grid finite element space V_h associated with a triangulation of Ω with characteristic fine-grid mesh size h. For a nodal (Lagrangian) basis, the index "i" runs over the set of fine degrees of freedom $x_i \in \mathcal{N}_h$. The unknown $\mathbf{u}$ stands for the coefficient vector of the finite element, Galerkin, approximation u_h to the solution of the underlined PDE posed variationally, i.e., $u_h \in V_h$ solves the discretized PDE in a variational form associated with a given bilinear form $a(.,\ .)$, stated as follows

$$a(u_h,\ \varphi) = (f,\ \varphi) \text{ for all } \varphi \in V_h.$$

As an example, we consider a second order self–adjoint elliptic bilinear form $a(u,\ \varphi) = \int\limits_\Omega k(\mathbf{x})\ \nabla u \cdot \nabla \varphi \, d\mathbf{x}$ for $u,\ \varphi \in H_0^1(\Omega)$ and a given positive coefficient function $k =$

117

$k(\mathbf{x})$, $\mathbf{x} \in \Omega$, the given polygonal/polyhedral domain in $\mathbf{R}^d$, $d = 2$ or 3. Using a standard piecewise linear conforming finite element space V_h on a quasiuniform triangulation $\mathcal{T}_h$, as it is well-known, the stiffness matrix $A = (a(\varphi_j, \varphi_i))$ computed from a nodal Lagrangian basis $\{\varphi_i\}_{\mathbf{x}_i \in \mathcal{N}_h}$ of V_h satisfies

$$\|A\| \simeq h^{d-2}. \tag{2}$$

The equivalence constants above generally depend on the variation $\dfrac{\max\limits_{\mathbf{x} \in \Omega} k(\mathbf{x})}{\min\limits_{\mathbf{x} \in \Omega} k(\mathbf{x})}$ but are mesh-independent.

Assume, that we have come up with a coarse space $V_H \subset V_h$ such that the coefficient vectors of functions in V_H viewed as elements of V_h can be represented as the range of an interpolation mapping P. We can define respective coarse basis functions by forming $P\mathbf{e}_{i_c}$ for each coarse coordinate vector $\mathbf{e}_{i_c} \in \mathbf{R}^{n_c}$ that has a single nonzero entry at the i_cth position. Then, consider the fine–grid function $\phi_{i_c}^{(H)}$ that has coefficient vector the i_cth column of P, i.e., equal to $P\mathbf{e}_{i_c}$. The set of functions $\{\phi_{i_c}^{(H)}\}$ forms the coarse basis of interest. The parameter H stands for characteristic size of the support of the coarse basis functions.

The above matrix–vector strong approximation property (1) admits the following finite element function form:

$$\|A\| \, a(u_h - u_H, \ u_h - u_H) \le C_A \sum_{\mathbf{x}_i \in \mathcal{N}_h} f_i^2 = C_A \sum_{\mathbf{x}_i \in \mathcal{N}_h} \left(\int_\Omega f(\mathbf{x}) \varphi_i \, d\mathbf{x} \right)^2.$$

Using Cauchy-Schwarz inequality, we have

$$\sum_{\mathbf{x}_i \in \mathcal{N}_h} \left(\int_\Omega f(\mathbf{x}) \varphi_i \, d\mathbf{x} \right)^2 \le \sum_{\mathbf{x}_i \in \mathcal{N}_h} \int_{\text{support } (\varphi_i)} f^2(\mathbf{x}) \, d\mathbf{x} \int_\Omega \varphi_i^2 \, d\mathbf{x}.$$

For a fairly general class of basis functions (including piecewise linears) on a quasiuniform mesh, we have

$$\int_\Omega \varphi_i^2 \, d\mathbf{x} \simeq |\text{support } (\varphi_i)| \simeq h^d.$$

Due to the bounded overlap of the supports of the finite element basis functions, we also have

$$\sum_{\mathbf{x}_i \in \mathcal{N}_h} \int_{\text{support } (\varphi_i)} f^2(\mathbf{x}) \, d\mathbf{x} \le \kappa \, \|f\|_0^2.$$

Thus, we arrive at the energy error estimate of our main interest (using (2))

$$a(u_h - u_H, \ u_h - u_H) \le C_A \kappa \, \frac{h^d}{\|A\|} \|f\|_0^2 \simeq C_A \kappa \, h^2 \, \|f\|_0^2.$$

In practice, we typically have $C_A = \mathcal{O}\left(\left(\frac{H}{h}\right)^2\right)$ with a constant in the $\mathcal{O}$ symbol, independent of the two mesh sizes (h and H); see, e.g., Corollary 1 later on. Thus, we get the following final upscaling energy error estimate:

$$a(u_h - u_H, \ u_h - u_H) \le c_A \ H^2 \|f\|_0^2. \tag{3}$$

In the remaining sections, we summarize a few AMG methods that provide *strong approximation property.*

2   Efficient window-based spectral AMG methods

In [FVZ05] (see also [Va08]), the following AMG method was proposed that exhibits strong approximation property. The original version tends to lead to relatively large coarse spaces so that the resulting two (and multi)–level methods have unacceptably high complexities. In the present section, we propose several approaches in the attempt to reduce the complexity of the original method.

Given an overlapping partition $\{w\}$ of the set of indices $i = 1, 2, \ldots, n$, we extract the rows of a given $n \times n$ matrix A with indices from any given set (called window) w. The respective rectangular matrix is denoted by A_w. By proper reordering, A_w can be written as follows

$$A_w = [A_{ww}, \ A_{w,\chi}].$$

Here, A_{ww} is the principal submatrix of A (row and column indices from w) and $A_{w,\chi}$ is the submatrix of A with columns outside w (and row indices from w).

We are interested, for a proper nonnegative diagonal matrix D_w, in the normal matrices $A_w^T D_w A_w$. The diagonal matrices D_w provide a partition of unity, i.e., if I_w stands for extension by zero outside the set w, then $\sum_w I_w D_w I_w^T = I$. This property ensures that

$$\sum_w \mathbf{v}^T A_w^T D_w A_w \mathbf{v} = \sum_w \mathbf{v}^T A^T I_w D_w I_w^T A \mathbf{v} = \|A\mathbf{v}\|^2, \qquad A_w = I_w^T A. \tag{4}$$

The method in question uses the symmetric semi-definite Schur complements S_w defined as follows:

$$\mathbf{v}_w^T S_w \mathbf{v}_w = \inf_{\mathbf{v}_\chi} \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}^T A_w^T D_w A_w \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}. \tag{5}$$

The original method utilizes the eigenvectors of the semidefinite Schur complements S_w,

$$S_w \mathbf{p}_k = \lambda_k \ \mathbf{p}_k, \ k = 1, \ldots, n_w. \tag{6}$$

For efficiency reason, for a given tentative interpolation matrix $\widetilde{P}$, we use in (6) instead the modified Schur complements

$$\mathbf{v}_w^T S_w \mathbf{v}_w = \inf_{\mathbf{v}_\chi \in I_\chi^T \mathrm{Range}(\widetilde{P})} \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}^T A_w^T D_w A_w \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}. \tag{7}$$

In what follows, we denote the exact window Schur complement with S_w^*.

We first form local interpolation matrices P_w by putting together the first $n_w^c \geq 1$ eigenvectors (in the lower part of the spectrum of S_w), i.e.,

$$P_w = \left[\mathbf{p}_1, \ldots, \mathbf{p}_{n_w^c} \right]. \tag{8}$$

The corresponding eigenvalues (ordered in an increasing order) are such that $\lambda_k \leq \mathrm{tol}\, \lambda_{\max}(S_w)$ for $k \leq n_w^c$ and

$$\lambda_k(S_w) > \mathrm{tol}\, \|S_w\| = \mathrm{tol}\, \lambda_{\max}(S_w) \ \ \text{for } k > n_w^c. \tag{9}$$

Here, we have the freedom to choose the pre-selected tolerance "tol" (a number between zero and one) that may also vary with w.

The eigenvectors $\{\mathbf{p}_k\}_{k=1}^{n_w}$ are orthogonal and assumed normalized.

The global P is computed based on another partition of unity set of nonnegative $n_w \times n_w$ diagonal matrices $\{Q_w\}$ that satisfy

$$I = \sum_w I_w Q_w I_w^T.$$

Then P is defined as follows

$$P = \sum_w I_w Q_w \left[0, \ P_w, \ 0 \right] = \sum_w I_w Q_w P_w (I_w^c)^T. \tag{10}$$

Here, I_w^c maps the local indices of the eigenvectors coming from the window w to their global indices expanding the result with zeros elsewhere. Thus we have defined a process that from a tentative $\widetilde{P}$ produces another one P. This can be iterated several times (by possibly changing the parameters such as $\{w\}$ and tol). In the next theorem, we formulate conditions ensuring that P admits a strong approximation property.

Theorem 1 *Consider the iterated window spectral AMG interpolation matrix P constructed on the basis of the modified window Schur complements using a $\widetilde{P}$ that satisfies the following estimate*

$$\sum_w \| D_w^{\frac{1}{2}} A_{w,\,\chi} (\mathbf{v}_\chi - I_\chi^T \widetilde{P} \mathbf{w}_c) \|^2 \leq \mu\, \|A\mathbf{v}\|^2. \tag{11}$$

That is, $\widetilde{P}$ is such that for any $\mathbf{v}$, when restricted to a complementary set χ, there is a coarse vector $\mathbf{w}_c$ (depending on $\mathbf{v}$ and the set χ) such that for a fixed number $\mu > 0$ (11) holds. Then, if we choose $\mathrm{tol} = \frac{1}{\delta} \leq 1$ in the two-level spectral decomposition defining the local P_w so that (see (9))

$$\|S_w\| \leq \delta\, \lambda_{m_w+1}(S_w),$$

and if we also assume the quasiuniformity of the windows, i.e., the estimate

$$\beta\, \|A\|^2 \leq \|S_w^*\|, \tag{12}$$

120

then, the following main strong approximation property holds for P

$$\|A\|^2 \|\mathbf{v} - P\mathbf{v}_c\|^2 \leq \eta \, \|A\mathbf{v}\|^2. \tag{13}$$

Here, $\eta = \frac{\delta}{\beta}\,(1 + \sqrt{\mu})^2$, where μ is from (11).

3 A new "window"-based spectral AMG method for finite element matrices

Here, we present a modified version of the method applied to finite element matrices A. The difference is in the eigenproblems that we use. Also, it utilizes a special partition of unity matrices. A main additional assumption is that the window sets are covered exactly by fine–grid elements and that we have access to the respective fine–grid element matrices so that we can assemble the semi–definite local matrices further denoted by Λ_w. Therefore, we have the identity $A = \sum_w I_w \Lambda_w I_w^T$.

We solve eigenproblems associated with the pair of matrices Λ_w and S_w, where S_w is the exact window-based Schur complement (as introduced before). The eigenproblems read (compare with (6)):

$$S_w \mathbf{p}_k = \lambda_k \, \Lambda_w \, \mathbf{p}_k, \ k = 1, \ldots, \, n_w, \tag{14}$$

where the eigenvalues are numbered in an increasing order and the eigenvectors are Λ_w–normalized.

Since, the matrices Λ_w can also be only semi–definite, to have real eigenvalues the nullspace of Λ_w should be contained in the nullspace of S_w, which is the case for finite elliptic matrices (Laplacian–like as well as elasticity).

It is clear that we can choose the eigenvectors $\mathbf{p}_k$ be orthogonal to the nullspace of Λ_w (and Λ_w–orthogonal to each other). Let the columns of $P_w^{(0)}$ span the nullspace of Λ_w. Then, we have $\mathbf{p}_k^T P_w^{(0)} = 0$.

Based on a preselected tolerance tol $\in [0, 1)$, we choose n_w^c such that $\lambda_k > \text{tol} \, \lambda_{n_w}$ for $k > n_w^c$. The local interpolation matrices are defined similarly as before (cf., (8)), now augmented with the nullspace, i.e.,

$$P_w = \left[P_w^{(0)}, \, \mathbf{p}_1, \, \ldots, \, \mathbf{p}_{n_w^c} \right]. \tag{15}$$

To define the global one, we use special diagonal matrices $\{Q_w\}_w$ with nonnegative entries that provide partition of unity, i.e., we have

$$I = \sum_w I_w Q_w I_w^T.$$

Each Q_w has entries on its diagonal $q_{w,\,i}$, $i \in w$, defined as follows:

$$q_{w,\,i} = \frac{\|\Lambda_w\|}{\sum\limits_{w':\,i\in w'} \|\Lambda_{w'}\|}. \tag{16}$$

At the end we formulate our main result.

Theorem 2 *Let A be a given finite element s.p.d. matrix. Consider a given set of windows $\{w\}$ where each window w is exactly covered by fine–grid elements. Assume also that the local finite element matrices Λ_w corresponding to the sets w are available. The nullspace (if nonempty) of the local matrices Λ_w is assumed known (explicitly computed). That is, let the nullspace of Λ_w be represented by the range of an explicitly available local matrix $P_w^{(0)}$. Assume that this nullspace is contained in the nullspace of the window Schur complement S_w (defined in (5)). The global interpolation matrix P is defined as in (10) based on the local interpolation matrices (15) and the weights $q_{w,\,i}$ (entries of the diagonal partition of unity matrices Q_w) are defined in (16). Then, the following global strong approximation property holds*

$$\|\mathbf{v} - P\mathbf{v}_c\|_A^2 \leq \kappa \, \max_w \, Cond^+(\Lambda_w) \, \max_w \, \left(\frac{1}{tol \, \lambda_{\max}(\Lambda_w^+ S_w)} \right) \, \|A\mathbf{v}\|^2.$$

Above, $\kappa \geq 1$ depends on the overlap of $\{w\}$, $tol \in (0,1]$ (in general depending on w) is the tolerance used to define the portion of the eigenvectors $\mathbf{p}_k$ in the lower part of the spectrum computed in (14) used to define P_w, $Cond^+(\Lambda_w) = \frac{\|\Lambda_w\|}{\lambda_{\min}^+(\Lambda_w)}$ is the effective condition number of Λ_w computed in a subspace orthogonal to the nullspace of Λ_w. Finally, $\lambda_{\max}(\Lambda_w^+ S_w) = \max_k \lambda_k$ where λ_k are from (14).

Corollary 1 *For finite element s.p.d. matrices A coming from second order elliptic problems, the constructed finite element modification of the window-based spectral AMG method, the following strong approximation property holds*

$$\|\mathbf{v} - P\mathbf{v}_c\|_A^2 \leq C \, \kappa \, \left(\frac{H}{h} \right)^2 \, \max_w \, \frac{1}{tol \, \|A_{w_0, w_0}\|} \, \|A\mathbf{v}\|^2.$$

References

[FVZ05] R. Falgout, P. S. Vassilevski, and L. T. Zikatanov, "*On two–grid convergence estimates,*" Numerical Linear Algebra with Applications, **12**(2005), pp. 471–494.

[Va08] Panayot S. Vassilevski, "MULTILEVEL BLOCK FACTORIZATION PRECONDITIONERS, Matrix-based Analysis and Algorithms for Solving Finite Element Equations," Springer, New York, 2008. 514 p.

The Role of Information Pattern in Approximation of Control Systems

Vladimir M. Veliov

We consider a control system

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(0) = x^0 \in \mathbf{R}^n, \qquad t \in [0, T], \tag{1}$$

where $u \in \mathcal{U} \subset L_1([0, T] \mapsto U)$, U is a convex compact subset of $\mathbf{R}^r$. For $u \in \mathcal{U}$ denote by $x[u]$ the solution of (1) that corresponds to u (assuming existence and uniqueness). Control theory and the set-membership estimation theory raise two main approximation problems related to (1): (i) approximate the set of trajectories, $\mathcal{X} = \{x[u] : u \in \mathcal{U}\}$, of (1); (ii) approximate the reachable set, $R = \{x[u](T) : u \in \mathcal{U}\}$, of (1).

Since the set of admissible controls $\mathcal{U}$ contains rather irregular functions[1] it is natural to split the approximation problems of (1) into two parts:

(P1) Replace the set of admissible controls $\mathcal{U}$ by a finitely parameterized subset $\mathcal{V}_N$ consisting only of functions u for which (1) can be discretized efficiently;
(P2) Apply a discretization scheme for solving (1) for $u \in \mathcal{V}_N$.

The requirement that $\mathcal{V}_N$ is a finitely parameterizable set (say, with a "degree of freedom" proportional to N) is needed to make the approximation "computable". Moreover, for each $u \in \mathcal{V}_N$ equation (1) should be well discretizable by single step methods, that is, the restrictions of the functions from $\mathcal{V}_N$ to each interval $[t_k, t_{k+1}]$ are sufficiently regular (for example, polynomial functions of a fixed degree and with uniformly bounded coefficients).[2] Then the error analysis of the discretization can be carried out in the usual way as for differential equations. Therefore we focus on the error analysis for the problem (P1), where one needs to estimate the *uniform error*

$$H_C(\mathcal{X}, \mathcal{X}_N) = \sup_{u \in \mathcal{U}} \inf_{v \in \mathcal{V}_N} \|x[v] - x[u]\|_{C[0,T]}$$

and the *terminal error*

$$H(R, R_N) = \sup_{u \in \mathcal{U}} \inf_{v \in \mathcal{V}_N} |x[v](T) - x[u](T)|,$$

where $\mathcal{X}_N$ and R_N are the set of trajectories and the reachable set corresponding to the set $\mathcal{V}_N$ of admissible controls.

---

[1] We mention that the reachable set R is usually not generated by "nice" controls (differentiable, Lipschitz, continuous). Even more, control functions of unbounded variation or non-integrable in Riemann sense may generate points of R that are not reachable by other controls, as in Fuller's phenomenon or as in [10]. This creates the main difficulty of approximating (1) by discrete schemes.

[2] Of course, there is a trade-off in choosing $\mathcal{V}_N$: the larger is $\mathcal{V}_N$, the better is the approximation to $\mathcal{X}$ and R by controls from $\mathcal{V}_N$; on the other hand, the lower is the accuracy of discretization.

Let us assume that the mapping $u \longrightarrow x[u]$ is continuous in L_1 and $\mathcal{V}_N$ is compact in the same space, hence the infimum in v is achieved. (These assumptions are normally fulfilled, as it will be the case in all considerations below.) Then there exists a mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_N$ such that

$$\sup_{u \in \mathcal{U}} \|x[\pi_N(u)] - x[u]\|_{C[0,T]} = H_C(\mathcal{X}, \mathcal{X}_N), \tag{2}$$

or

$$\sup_{u \in \mathcal{U}} |x[\pi_N(u)](T) - x[u](T)| = H(R, R_N), \tag{3}$$

respectively (the mapping π_N needs not be the same in the two equalities). This formulation has an advantage: one can study the information pattern of the mapping π_N that provides the best approximations in (2) or (3), or at least some approximation with a given order of accuracy with respect to N. Namely, we can distinguish the following cases of a mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_N$:

Definition 1 *(i) The mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_N$ is called "local" if for every $k = 0, \ldots, N-1$, and for every $u', u'' \in \mathcal{U}$ with $u'(t) = u''(t)$ on $[t_k, t_{k+1}]$ it holds that $\pi_N(u')(t) = \pi_N(u'')(t)$ on $[t_k, t_{k+1}]$;*
(ii) The mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_N$ is called "non-anticipative" if for every $k = 1, \ldots, N$, and for every $u', u'' \in \mathcal{U}$ with $u'(t) = u''(t)$ on $[0, t_k]$ it holds that $\pi_N(u')(t) = \pi_N(u'')(t)$ on $[0, t_k]$;
(iii) The mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_N$ is called "anticipative" if it is not non-anticipative.

We shall see that for the same approximating set of inputs $\mathcal{V}_N$ it may happen that a certain order of approximation can be achieved by anticipative (non-anticipative) *approximating mappings* π_N but cannot be achieved by non-anticipative (resp. local) mappings π_N. That is, the information pattern of the approximating mapping π_N may play a role for the order of the accuracy.

It is to be stressed that in different problems related to the control system (1) one may need to restrict the choice of the approximation mapping to a prescribed information pattern: local or non-anticipative. This is the case, for example, if one has to simulate a real system modeled by (1) only knowing the current, or the past information about the input u. For other problems, say for an optimal open-loop control problem one can freely employ anticipative approximation mappings to pass directly to mathematical programming.

The concept of information pattern in the approximation theory for control systems (or systems with "deterministic" uncertainties) opens a new filed of research in this area. The aim of this note is to revisit some known results in the area from this perspective.

First we recall a few often used approximation sets $\mathcal{V}_N$. For any natural number N denote $h = T/N$, $t_i = ih$, $i = 0, \ldots, N$.

Example 1.

$$\mathcal{V}_N = \mathcal{V}_{N,0} := \{u \in \mathcal{U} :\ u(t) \text{ is constant on each } (t_{i-1}, t_i))\}.$$

Example 2.

$$\mathcal{V}_N = \mathcal{V}_{N,1} := \{u \in \mathcal{U} :\ u(t) \text{ is p.-w. constant with at most 1 jump in each } (t_{i-1}, t_i)\}.$$

Example 3. Denote by U^{extr} the set of all extreme points of U and define

$$\mathcal{V}_N = \mathcal{V}_{N,0}^{\text{extr}} := \{u : [0,T] \mapsto U^{\text{extr}} :\ u(t) \text{ is constant on each } (t_{i-1}, t_i))\}.$$

Clearly $\mathcal{V}_{N,0}^{\text{extr}} \subset \mathcal{V}_{N,0} \subset \mathcal{V}_{N,1}$.

Below we reformulate some known results in a way which exhibits the information pattern of the approximation mapping implicitly involved in the proofs, and comment the validity of the results if a more stringent information pattern has to be used.

One commonly used approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,0}$ is defined as

$$\pi_N(u)(t) = \frac{1}{h} \int_{t_{k-1}}^{t_k} u(s)\, \mathrm{d}s \text{ for } t \in (t_{k-1}, t_k). \tag{4}$$

Obviously it is local and even more, it is independent of the specific form of the equation (1).

Let us consider first a liner control system, where $f(x, u) = Ax + Bu$ in (1). From a general result in [3] it follows that there exists a *local* approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,0}$ (namely, defined by (4)) such that

$$\|x[\pi_N(u)] - x[u]\|_{C[0,1]} \le ch \qquad \forall\, u \in \mathcal{U}. \tag{5}$$

In the same time the results in [13, 1] imply that there exists an *anticipative* approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,0}$ (which is not explicitly defined in these papers) such that

$$|x[\pi_N(u)](T) - x[u](T)| \le ch^2 \qquad \forall\, u \in \mathcal{U}. \tag{6}$$

We mention that the result holds for an arbitrary convex and compact set U, therefore it applies also to the "pathological" examples in [10] mentioned in footnote 2. A second order approximation as above cannot be achieved by using local approximation mappings.

An important extension is proved in [9]: there exists an *anticipative* approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,0}$ that ensures simultaneously (5) and (6). This non-trivial result opens the door to error estimates for non-linear systems by local linearization. Of course, in view of the previous paragraph, this simultaneous approximation cannot be achieved in the class $\mathcal{V}_{N,0}$ by local approximation mappings.

We recall also two results, again for a linear system as above, for the class $\mathcal{V}_{N,1}$ of approximating controls, restricted to polyhedral sets U. From the results in [5] it can be deduced that there exists a *local* approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,1}$ for which an estimation such as in 6 holds. On the other hand, the results in [1] show that there exists an *anticipative* $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,1}$ such that

$$|x[\pi_N(u)](T) - x[u](T)| \le ch^3 \qquad \forall\, u \in \mathcal{U}.$$

Third order accuracy cannot be achieved in the class $\mathcal{V}_{N,1}$ by local approximation mappings.

Now we turn to the non-linear case, assuming that the right-hand side in (1) has the form

$$f(x,u) = g_0(x) + \sum_{i=1}^{m} g_i(x)u_i, \qquad u = (u_1, \ldots, u_m) \in U.$$

We assume that U is convex and compact and that g_i are sufficiently smooth. There are only a few results obtaining higher than first order approximations in the nonlinear case.[3] The first is that in [12], where a second order approximation of $\mathcal{X}$ is proved (or an approximation of order $3/2$ for a more general form of f than the above one) assuming, however, that $f(x,U)$ is uniformly strongly convex, which is rather strong for many applications. The implicitly involved approximation mapping π_N is *local*. Another group of results concern the case of commutative affine systems, i.e. such that the Lie brackets $[g_i, g_j]$ are all zero for $i, j \ge 1$. A rather general indirect (variational) estimation of $H(R, R_N)$ in the class $\mathcal{V}_{N,0}$ is obtained in [14]. It allows to obtain a second order estimation of $H(R, R_N)$ provided that R and R_N have the so called *external ball property*. This property is present if R and R_N are convex, for example. The approximation mappings π_N implicitly involve are *anticipative*. The drawback is that the external ball property does not hold, in general, and is difficult for verification.

A more advanced result for commutative systems is proved in a paper in preparation by the author and M. Krastanov. Namely, there exists a constant C such that for every N there exists an (*anticipative*) approximation mapping $\pi_N : \mathcal{U} \mapsto \mathcal{V}_{N,0}$ such that

$$\|x[\pi_N(u)] - x[u]\|_{C[0,T]} \quad \le \quad Ch, \tag{7}$$

$$|x[\pi_N(u)](T) - x[u](T)| \quad \le \quad Ch^{1.5}. \tag{8}$$

It is an open question if the estimation is sharp, but in any case the second estimation cannot be achieved by using local approximation mappings only. Whether the result

---

[3]Higher than first order approximations to optimal control problems are known. However, most of these results are based on a priori assumption that the optimal control is sufficiently regular (i.e. is Lipschitz with derivative having bounded variation), see e.g. [4, 7]. The results recalled in the present paper are applicable in the optimal control context without such assumptions.

holds for a non-local but non-anticipative approximation mappings is another open question.

Higher than first order approximations are particularly difficult for non-commutative systems. An estimation like in (8) for class of non-commutative bilinear systems is proved in [9]. The class approximating controls is again $\mathcal{V}_{N,0}$ and the employed approximation mapping π_N is *anticipative*. A different result for non-commutative systems without drift term ($g_0 = 0$) and with $U = [0,1] \times [0,1] \subset \mathbf{R}^2$ is presented in [8]. The error estimate in terms of the trajectories is like in (7) but with h^2 in the right-hand side (that is, of second order) and it is achieved by *local* approximation mappings. However, the approximating control set $\mathcal{V}_N$ is much reacher even than $\mathcal{V}_{N,1}$, which creates a substantial problem in the numerical realization. The main contribution in this paper is, in fact, how to cope with this problem.

The last issue we briefly address is that of approximations using the class $\mathcal{V}_{N,0}^{\mathrm{extr}}$ of controls. This issue is of substantial importance for numerical treatment of optimal control problems for switching systems, see e.g. [11]. The following estimation is proved, essentially, in [2, 6]: for the approximating class of controls $\mathcal{V}_{N,0}^{\mathrm{extr}}$,

$$H_C(\mathfrak{X}, \mathfrak{X}_N) \leq Ch^{1/2}. \tag{9}$$

This estimation is proved for the general system (1) under Lipschitz continuity of f. In [15] the author of the present paper conjectured that a first order estimation holds in (9) and proved this in several particular classes of systems. The paper [9] also contains a small contribution in this direction. A substantial progress in proving the conjecture is done in [11], where however, U is assumed to be a polyhedral set. In all the above contributions the (implicitly or explicitly) involved approximation mapping π_N is *non-local* and *non-anticipative*. Also, it is quite clear that local approximation mappings cannot provide even (9). In general, the problem of first order approximation is still open.

References

[1] B.D. Doitchinov and V.M. Veliov. Parametrisations of integrals of set-valued mappings and applications. *J. Math. Anal. and Appl.*, **179**(2):483–499, 1993.

[2] T. Donchev. Approximation of lower semicontinuous differential inclusions. *Numer. Funct. Anal. and Optimiz.*, **22**(1&2):55–67, 2001.

[3] A. Dontchev and E. Farkhi. Error estimates for discretized differential inclusion. *Computing*, **41**(4):349–358, 1989.

[4] A.L. Dontchev, W.W. Hager, and V.M. Veliov. Second-order Runge-Kutta approximations in control constrained optimal control, *SIAM J. Numerical Anal.*, **38**(1):202–226, 2000.

[5] R. Ferretti. High-order approximations of linear control systems via Runge-Kutta schemes. *Computing*, **58**(4):351–364, 1997.

[6] G. Grammel. Towards fully discretized differential inclusions. *Set-Valued Analysis*, **11**(1):1–8, 2003.

[7] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, **87**:247–282, 2000.

[8] M. Krastanov and V.M. Veliov. High-order approximations to multi-input non-commutative control systems. To appear in *Large-Scale Scientific Computations*, Springer, Berlin, Heidelberg, 2010.

[9] A. Pietrus and V.M. Veliov. On the Discretization of Switched Linear Systems. *Systems&Control Letters*, **58**:395–399, 2009.

[10] D. Silin. On the Variation and Riemann Integrability of Optimal Control of Linear Systems. *Dokl. Akad. Nauk SSSR*, **257**:548–550, 1981.

[11] S. Sager. Reformulations and Algorithms for the Optimization of Switching Decisions in Nonlinear Optimal Control. *Journal of Process Control*, to appear. (URL http://sager1.eu/sebastian/downloads/Sager2009b.pdf)

[12] V.M. Veliov. Second order discrete approximations to strongly convex differential inclusions. *Systems & Control Letters*, **13**:263–269, 1989.

[13] V.M. Veliov. Second order discrete approximations to linear differential inclusions. *SIAM J. Numer. Anal.*, **29**(2):439–451, 1992.

[14] V.M. Veliov. On the time-discretization of control systems. *SIAM J. Control Optim.*, **35**(5):1470–1486, 1997.

[15] V. Veliov. Relaxation of Euler-Type Discrete-Time Control System. *Research Report No 273*, ORCOS, TU-Wien, 2003. (http://www.eos.tuwien.ac.at/OR/research/reports/RR273.pdf).