Southern Federal University
I.I. Vorovich Institute of Mathematics, Mechanics,
and Computer Science

Proceedings of

Eighth China-Russia Conference

NUMERICAL ALGEBRA WITH APPLICATIONS

Editors: Zhong-Zhi Bai, Galina V. Muratova

24–27 June 2019
Rostov-on-Don, RUSSIA

Rostov-on-Don – Taganrog
Southern Federal University Press
2019

UDC 512.761.5(075.8)
BBC 22.132я73

Editors: Zhong-Zhi Bai, Galina V. Muratova

NUMERICAL ALGEBRA WITH APPLICATIONS
Proceedings of Eighth China-Russia Conference, 24–27 June 2019, Rostov-on-Don, RUSSIA / Editors: Zhong-Zhi Bai, Galina V. Muratova ; Southern Federal University. – Rostov-on-Don ; Taganrog : Southern Federal University Press, 2019. – 144 p.
ISBN 978-5-9275-3155-4

The Eighth China-Russia Conference on Numerical Algebra with Applications (CRC-NAA'19) follows the traditions of its seven predecessors and provides a forum for researches to present papers on recent advances in the overall field of numerical algebra \& computational methods and their applications. It is organized by I.I. Vorovich Institute of Mathematics, Mechanics, and Computer Science of Southern Federal University. The topics of CRC-NAA'19 include, but are not limited to: mathematical modeling, applications of numerical methods and algorithms to solve problems of mathematical modeling, linear and nonlinear equations systems, preconditionining techniques, parallel computing.

UDC 512.761.5(075.8)
BBC 22.132я73

ISBN 978-5-9275-3155-4

CONTENTS

Plenary Lectures 6

EQUILIBRIUM AND STABILITY OF THE INHOMOGENEOUS NONLINEARLY ELASTIC BODIES
Karyakin M., Obrezkov L., Pustovalova O. 7

COMPUTING EIGENPAIRS OF HERMITIAN MATRICES IN PERFECT KRYLOV SUBSPACES
Bai Z.-Z., Miao C.-Q. 12

MULTIGRID METHOD FOR BIOLOGICAL AND MEDICAL PROBLEMS
Muratova G. V., Bavin V. V. 15

ASYMPTOTIC APPROXIMATIONS FOR ONE RADIATIVE-CONDUCTIVE HEAT TRANSFER PROBLEM
Amosov A.A., Krymov N.E. 20

COMPUTER SIMULATION OF MIGRATION OF LIQUID CYLINDRICAL INCLUSIONS IN A CRYSTAL IN SOME CASES OF INTERFACIAL ENERGY ANISOTROPY
Garmashov S.I., Karpenko A.S. 25

NUMERICAL ANALYSIS OF COSYMMETRY VIOLATION IN FILTRATION CONVECTION PROBLEM
Govorukhin V.N. 30

A P-VERSION TWO LEVEL SPLINE METHOD FOR 2D NAVIER-STOKES EQUATIONS
Han D.-F. 34

MINIMUM NORM PARTIAL QUADRATIC EIGENVALUE ASSIGNMENT FOR VIBRATING STRUCTURES USING RECEPTANCE METHOD
Liu H., He B.-X. , Chen X.-P. 39

THE REGION OF TURING INSTABILITY IN SCHNAKENBERG SYSTEM
Lysenko S.A., Revina S.V. 44

THE GENERALIZED HSS METHOD WITH A FLEXIBLE SHIFT-PARAMETER
Meng G.-Y., Wen R.-P. 49

COMPUTATIONAL EXPERIMENTS BASED ON REDUCED MODEL OF A LONG AND SHALLOW STREAM
Nadolin K.A., Zhilyaev I.V. 54

NUMERICAL STUDY OF EFFECTIVE MODULI OF POROUS PIEZOCOMPOSITES WITH METALLIZED INTERFACES IN UNIT CELL

WITH CUBIC OR SPHERICAL PORES
Nasedkin A.V., Nassar M.E. 60

FOURIER METHOD FOR SOLVING THE CAUCHY PROBLEM FOR AN ELLIPTIC EQUATION.
Sorokin S.B. 65

CONVECTION IN A POROUS MEDIUM: COSYMMETRY AND ITS CONSERVATION THROUGH A FINITE-DIFFERENCE APPROXIMATION
Tsybulin V.G. 70

PMHSS PRECONDITIONERS FOR STOKES CONTROL OPTIMIZATION PROBLEMS
Wang Z.-Q., Cao S.-M. 71

A THREE-TERM ACCELERATED TECHNIQUE FOR HSS-TYPE METHODS
Wen R.-P., Jiang W. 76

A FAST NULLSPACE METHOD BASED ON MATRIX EXPONENTIAL FOR THE UNSTEADY STOKES EQUATIONS
Yang X. 81

ON THE QUADRATIC EQUATIONS AND MATRIX FUNCTION
Zhang X. 86

Invited talks 91

BIPARTITE GRAPHS IN THE PROBLEM OF NFA MINIMIZATION AND ALGORITHMS FOR CALCULATING THEIR NUMBER
Abramyan M.E. 92

ABOUT SOME NEW APPROACHES TO TEXT SENTIMENT ANALYSIS AND TOXICITY ANALYSIS USING DEEP LEARNING METHODS
Abramyan M.E., Litovchenko D.E. 97

ABOUT SOME FEATURES OF THE HEURISTIC INPLEMENTATION OF BRANCH AND BOUND METHOD FOR THE PROBLEM OF NFA MINIMIZATION
Abramyan M.E., Melnikov B.F. 102

FOURIER ANALYSIS OF MULTIGRID METHOD WITH HSS METHODS AS SMOOTHERS
Andreeva E.M., Muratova G.V. 107

THE SECOND ORDER DIFFERENCE SCHEME OF THE CONVECTION-DIFFUSION PROBLEM WITH THE ROBIN'S BOUNDARY CONDITIONS
Chikina L.G., Shabas I.N., Chikin A.L. 112

POSITIVE DEFINITENESS AND M-MATRIX CONDITIONS OF A CONVECTION-DIFFUSION DIFFERENCE OPERATOR WITH BOUNDARY CONDITIONS OF THE THIRD KIND
Chikina L.G., Shabas I.N., Martynova T.S. 116

A REGILARIZATION METHOD ON THE BASE AUGMENTED SYSTEM FOR IDENTIFICATION PROBLEM FOR INTENSITY OF ATMOSPHERIC POLLUTION SOURCE
Chubatov A.A., Karmazin V.N. 121

PHYSICAL AND MATHEMATCAL MULTIFUNCTIONAL MODELS, SPECIALIZED NUMERICAL ALGORITHMS AND SOFTWARE FOR MODELING OF PROPAGATION OF CONTAMINATION, TAKING INTO ACCOUNT THREE-DIMENSIONAL URBAN OR INDUSTRIAL BUILDING
Dzama D.V., Sorokovilova O.S., Asfandiyarov D.G. 124

INVESTIGATION OF THE STABILITY OF TWO-DIMENSIONAL FLOWS CLOSE TO THE SHEAR
Kirichenko O.V., Revina S.V. 130

SOME COMPUTATIONAL ALGEBRAIC PROBLEMS, REDUCIBLE TO THE KNAPSACK PROBLEM, AND THEIR SOLUTION USING LUNAR ARITHMETIC
Melnikov B.F., Korabelshchikova S.Yu. 135

NUMERICAL STUDY OF THE SKEW-SYMMETRIC PRECONDITIONERS
Pichugina O.A. 139

Plenary Lectures

EQUILIBRIUM AND STABILITY OF THE INHOMOGENEOUS NONLINEARLY ELASTIC BODIES[1]

Karyakin M.*,****, Obrezkov L.*******, Pustovalova O.***

* *Southern Federal University, Rostov-on-Don, Russia*

** *Southern Mathematical Institute – the Affiliate of VSC RAS, Vladikavkaz, Russia*

*** *Lappeenranta Technological University, Lappeenranta, Finland*

The purpose of this work is to study the effect of the inhomogeneity of the material properties of a nonlinearly elastic body on its equilibrium and stability under various types of loading, taking into account large strains.

Large Bending Strains of Inhomogeneous Panel

It is well known that pure bending is one of the basic deformations that is actively used for experimental determination and refinement of model parameters for both elastic and nonlinearly elastic behavior of media. The appearance of new highly elastic structural materials, as well as a significant increase in attention to the tasks of modeling biological substances and their artificial substitutes, capable in particular of withstanding large strains, causes a new wave of interest to the theoretical problems and experimental techniques that allow constructing correct and verifiable mathematical models of continuous media. Another reason that makes it necessary to address classical problems is the intensive study, modeling and numerical analysis of the processes of creating and testing the performance of composite materials, which are a matrix with relatively rigid fibers possessing predetermined properties [1]. As it was shown in [2] the calculation of the strength of such microstructures insistently requires considering the possibility of destruction of the composite due to loss of stability.

A detailed analysis of various aspects of the equilibrium of bent homogeneous nonlinearly elastic panel was presented in [3]. It was shown that for the most common models of nonlinear elastic behavior of the panel material there is a maximum on the loading diagram which indicates a possible loss of stability at bending. In [4]–[7] a series of specific stability problems for large bending strains with the use of the bifurcation analysis for a homogeneous panel were solved.

In this paper, a similar approach is applied to a more complex model: its material parameters are functions of the vertical coordinate – the position of the point along the thickness of the panel. We study the influence of this inhomogeneity to the loading diagram or, more precisely, the position of its maximum

[1] Supported by Program for Fundamental Research of the RAS Presidium N 1 on the Strategic Directions of the Science Development “Fundamental Problems of Mathematical Modeling” (project 114072870112)

point that can be treated (with some limitations) as an onset of the stability loss.

In order to the traditional semi-inverse scheme be applicable to the inhomogeneous bar without significant changes we restrict considerations by the case when the material of the bar is inhomogeneous in the vertical direction only. We assume also that only shear modulus μ is variable. This assumption is not general but it allows comparing the results with [3], where for all models the same value of the Poisson ratio equal to $1/4$ was used.

It should be noted that both the process of derivation of ordinary differential equations of boundary value problem and their numerical analysis is very laborious, especially for the complex multiparameter expression of the strain energy function. The fast and reliable solution of such problems is feasible with the help of the system of computer automation of the semi-inverse method of the nonlinear theory of elasticity, which is presented in [8]. The realized numerical scheme is based upon the well-known shooting method. The specific feature of BVP arising in finite elastostatics is substantial nonlinearity of the boundary conditions. That means that one extra step should be introduced into the classical scheme: after choosing the value of shooting parameter, say function value at the initial point, we need to solve some scalar nonlinear equation to find the derivative of this function at this point. This step requires additional accuracy to choose one of generally speaking multiple roots of the nonlinear equation.

The effect of inhomogeneity on the feature of a nonlinearly elastic material bending diagram found in previous works – the maximum point followed by the falling section – was investigated. It has been established, in particular, that if the heterogeneity of the material is very large, i.e. the values of the elastic modulus in the upper and lower layers differ by more than ten times, the effect of inhomogeneity consists in moving the maximum point to the region of larger strains.

The effect of inhomogeneity of the material of the panel on the relationship between the position of the maximum point and the point of buckling on the bending diagram was studied. It was established, in particular, that for a panel with a softer lower edge, the bifurcation points are located to the left of the maximum points, i.e. the panel will lose stability on the ascending part of the diagram.

It seems that the obtained results can be used to develop and elaborate new experimental methods for identifying the parameters of mathematical models that describe the mechanical properties of materials and structures that are experiencing large strains.

Cylinder with Inner Stresses Due to a Disclination

Another example of inhomogeneity can be related with the initial stresses that exist in an unloaded body. As an example of such situation the problem on the equilibrium and stability of a cylinder with wedge disclination was considered.

The concept of disclinations, which arose in the study of certain properties of liquid crystals, later found application in the description of various biological objects, such as protein polymers, wood, nematoid structures of human skin et al. A new wave of interest to the nonlinear theory of elastic disclinations in recent years is associated with the active use of disclination models in the description of nanostructures of various kinds [9].

The study of disclinations in the framework of the nonlinear theory of elasticity was initiated in [10], a number of propositions and ideas of which were later refined and developed in the works of L.M. Zubov and his followers [11]-[13].

The use of models and methods of the nonlinear theory of elasticity allows not only to analyze the stress-strain state of a body with a defect but also to investigate questions of the stability of the constructed solutions. The traditional scheme of stability studies [14, 15] is based on the linearization of three-dimensional nonlinear boundary problems in the neighborhood of the constructed solution and the study of the possibility of the existence of a nontrivial solution of these linear problems depending upon parameters, i.e. certain characteristics of the deformed state or external influences. This work can be viewed as a continuation of the studies begun in [12] and related to the study of the equilibrium and stability of a cylinder containing a wedge disclination, and a comparative analysis of the use in this connection more general models of nonlinear-elastic behavior of compressible media.

The simplified version of the Blatz and Ko model and the five-constant Murnaghan model were used for the simulation. The main method to construct the equilibrium state was the semi-inverse method of the nonlinear theory of elasticity, the stability study was carried out within the framework of the static bifurcation approach.

It should be noted that the choice of the model did not show a qualitative effect on the stress-strain state. If the parameters of the models were chosen so that in the linear approximation they corresponded to the same material, the difference in the distribution of normal stresses was no more than 10 percent.

More important difference was found when we compare results for close-to-solid cylinder with that of similar to the thick-walled tube. In the first case the stability region is described by curves with different mode numbers that is rather untypical. In addition, one can see the condensing of bifurcation curves at one of the boundaries of the stability region. Increasing the mode number n, starting from a certain value (about fifty), practically does not change the critical pressure. This fact agrees with the results obtained earlier in [12] for the harmonic material model. This, in particular, means that for such body it is impossible to predict the form of the stability loss based on the analysis of linearized equations. For a thinner cylinder, the $n = 2$ mode is always preferred; the corresponding critical pressure is minimal. Another feature of thick cylinders is the existence of a zone where the loss of stability can occur only due to the presence of disclination without the application of external pressure. Such an area may be absent in the thinner cylinders.

A comparative analysis of the models used at various values of the parameters showed, in particular, that thick-walled cylinders are characterized by condensing of bifurcation curves, which can also occur in cases where there is no disclination, but its presence makes the condensation areas much more extensive. The presence of disclination can be both a stabilizing factor and vice versa – significantly reduce the value of the critical pressure.

Different types of the inhomogeneity of the cylinder material can also have a noticeable influence upon the stability. In particular, they can significantly change the region of stability of the unloaded cylinder having wedge disclination of the given strength.

Bibliography

1. *Levy A.J. , Shukla A., Xie M.* Bending and buckling of a class of nonlinear fiber composite rods // J. Mech. Phys. Solids, 2006, vol. 54 (5), pp. 1064–1092.

2. *Karamanos S.A.* Bending instabilities of elastic tubes // Int. J. Solids Struct., 2002, vol. 39 (8), pp. 2059-2085.

3. *Karyakin M., Kalashnikov V., Shubchinskaya N.* The specific features of the pure bending of the elastic panel undergoing large strains // Int. J. Eng. Sci., 2014, vol. 80, pp. 90–105.

4. *Triantafyllidis N.* Bifurcation phenomena in pure bending // J. Mech. Phys. Solids, 1980, vol. 28(3-4), pp. 221–245.

5. *Haughton D.M.* Flexure and compression of incompressible elastic plates // Int. J. Eng. Sci., 1999, vol. 37 (13), pp. 1693–1708.

6. *Coman C.D. , Destrade M.* Asymptotic results for bifurcations in pure bending of rubber blocks // Q. J. Mechanics Appl. Math., 2008, vol. 61 (3), pp. 395–414.

7. *Destrade M., Gilchrist M.D., Murphy J.G.* Onset of nonlinearity in the elastic bending of blocks // J. Appl. Mech., 2010, vol. 77 (6).

8. *Gavrilyachenko T.V. , Karyakin M.I., Sukhov D.Yu.* Designing of the interface for nonlinear boundary value problem solver using Maple // Proceedings of the International Conference on Computational Sciences and its Applications. Los Alamitos-Washington-Tokyo: ICCSA, 2008, pp. 284–291.

9. *Romanov A. E.* Mechanics and physics of disclinations in solids // European Journal of Mechanics A. Solids, 2003, 22 pp. 727–741

10. *de Wit R.* Linear theory of static disclinations // Fundamental Aspects of Dislocation Theory, Nat. Bur. Stand. (US), 1970, pp. 651–673.

11. *Zubov L. M.* Nonlinear Theory of Dislocations and Disclinations in Elastic Bodies, Springer –Verlag, Berlin–Heidelberg–New York–Tokyo, 1997.

12. *Zelenin A. A., Zubov L. M.* Stability and postcritical behavior of an elastic cylinder with disclination // Mechanics of solids, 1989, 24, pp. 97–103.

13. *Karyakin M. I., Zubov L. M.* Theory of Isolated and Continuously Distributed Disclinations and Dislocations in Micropolar Media // Advanced Structured Materials, vol 7. Mechanics of Generalized Continua, Springer–Verlag, Berlin–Heidelberg, 2011, pp. 275–290.

14. *Zubov L. M. , Sheidakov D. N.* Instability of a hollow elastic cylinder under tension, torsion, and inflation // Journal of Applied Mechanics, Transactions ASME, 2008, 75, pp. 0110021–0110026.

15. *Obrezkov L.* Equilibrium and stability of nonlinearly elastic cylinder made of Blatz-Ko material // Engineering Transactions, 2016, 4, pp. 457–463.

COMPUTING EIGENPAIRS OF HERMITIAN MATRICES IN PERFECT KRYLOV SUBSPACES[1]

Bai Z.-Z., Miao C.-Q.

State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, P.R. China

Consider the large and sparse standard Hermitian eigenvalue problem

$$Ax = \lambda x, \quad \text{with} \quad \|x\| = 1, \tag{1}$$

where $A \in \mathbb{C}^{n \times n}$ is a Hermitian matrix, $\lambda \in \mathbb{R}$ is an eigenvalue of the matrix A, and $x \in \mathbb{C}^n$ is the corresponding eigenvector; see [14]. Here and in the sequel, we use $\| \cdot \|$ to denote the Euclidean norm of either a vector or a matrix.

We are interested in computing the smallest eigenpair, that is, the eigenvalue of the smallest magnitude and the corresponding eigenvector, of the Hermitian eigenvalue problem (1) by iteration methods based on projections onto Krylov subspaces [7, 17]. This class of iteration methods can be essentially categorized into the standard and the rational Krylov subspace methods, with both of them first building orthogonal bases for the corresponding Krylov subspaces by using the Lanczos process; then projecting the Hermitian eigenvalue problem (1) onto the orthogonal subspaces, obtaining Hermitian eigenvalue problems of triangular matrices with much smaller sizes; and, finally, extracting the desired Ritz pairs for the Hermitian eigenvalue problem (1) through computing the eigenpairs of the Hermitian triangular matrices; see, e.g., [14, 15, 16].

If the extreme eigenvalues of the Hermitian eigenvalue problem (1) are well separated from the interior eigenvalues, then they can be accurately and effectively computed by the standard Krylov subspace methods. Otherwise, we can transform the problem by employing the shift-and-invert technique, making the extreme eigenvalues of the transformed problem to have good separations, so that the Krylov subspace methods can work efficiently on the transformed Hermitian eigenvalue problem. On the other hand, the rational Krylov subspace methods can accurately and effectively compute the extreme eigenpairs that may be poorly separated from the interior eigenpairs, but these methods are too expensive to be used to compute the extreme eigenpairs of good separations. Of course, these two classes of Krylov subspace methods may encounter stability problem in finite precision arithmetic, because the orthogonality among the computed vectors in the orthogonal basis are easily lost during the Lanczos process when the Krylov subspace is gradually enlarged. In actual applications, in order to

[1] Supported by The National Natural Science Foundation (No. 11671393 and No. 11911530082), P.R. China.

further accelerate the convergence rates of these Krylov subspace methods, appropriate preconditioning strategy is often indispensible. For more details, we refer to [12, 5, 6, 17] and the references therein.

In this paper, in order to integrate the advantages and avoid the disadvantages of the standard and the rational Krylov subspace methods, by technically unionizing these two kinds of subspaces we introduce a new subspace, called the perfect Krylov subspace. Then we present a class of perfect Krylov subspace methods for computing the smallest eigenpair of the Hermitian eigenvalue problem (1). In theory, we demonstrate that this method has local, semilocal and global convergence properties, with both quotient convergence factor and quotient convergence order being dependent on the order of the perfect Krylov subspace; and by experiments, we show that this method outperforms the standard and the rational Krylov subspace methods in terms of iteration counts and computing times, as well as in terms of absolute error with respect to the eigenvalue and the eigenvector.

Bibliography

1. Bai Z.-Z. and Miao C.-Q., On local quadratic convergence of inexact simplified Jacobi-Davidson method, *Linear Algebra Appl.* 2017. Vol. 520. P. 215–241.

2. Bai Z.-Z. and Miao C.-Q., On local quadratic convergence of inexact simplified Jacobi-Davidson method for interior eigenpairs of Hermitian eigenproblems, *Appl. Math. Lett.* 2017. Vol. 72. P. 23–28.

3. Berns-Müller J., Graham I.G. and Spence A., Inexact inverse iteration for symmetric matrices, *Linear Algebra Appl.* 2006. Vol. 416. P. 389–413.

4. Druskin V. and Knizhnerman L., Extended Krylov subspaces: approximation of the matrix square root and related functions, *SIAM J. Matrix Anal. Appl.* 1998. V. 19. P. 755–771.

5. Freitag M.A. and Spence A., Convergence of inexact inverse iteration with application to preconditioned iterative solves, *BIT* 2007. V. 47, P. 27–44.

6. Freitag M.A. and Spence A., A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems, *IMA J. Numer. Anal.* 2008. V. 28. P. 522–551.

7. Golub G.H. and Van Loan C.F., Matrix Computations, Third Edition, *The Johns Hopkins University Press*, Baltimore, 1996.

8. Golub G.H. and Ye Q., An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems, *SIAM J. Sci. Copmut.* 2002. V. 24. P. 312–334.

9. Jagels C. and Reichel L., The extended Krylov subspace method and orthogonal Laurent polynomials, *Linear Algebra Appl.* 2009. V. 431. P. 441–458.

10. Jagels C. and Reichel L., Recursion relations for the extended Krylov subspace method, *Linear Algebra Appl.* 2011. V.434. P. 1716–1732.

11. Knizhnerman L. and Simoncini V., A new investigation of the extended Krylov subspace method for matrix function evaluations, *Numer. Linear Algebra Appl.* 2010. V. 17. P. 615–638.

12. Morgan R.B. and Scott D.S., Preconditioning the Lanczos algorithm for sparse symmetric eigenvalue problems, *SIAM J. Sci. Comput.* 1993. V. 14. P. 585–593.

13. Ortega J.M. and Rheinboldt W.C., Iterative Solution of Nonlinear Equations in Several Variables, *SIAM*, Philadelphia, PA, 2000.

14. Parlett B.N., The Symmetric Eigenvalue Problem, *SIAM*, Philadelphia, PA, 1998.

15. Ruhe A., Rational Krylov sequence methods for eigenvalue computation, *Linear Algebra Appl.* 1984. V. 58. P. 391–405.

16. Ruhe A., Rational Krylov: A practical algorithm for large sparse nonsymmetric matrix pencils, *SIAM J. Sci. Comput.* 1998. V. 19. P. 1535–1551.

17. Saad Y., Numerical Methods for Large Eigenvalue Problems, Second Edition, *SIAM*, Philadelphia, PA, 2011.

18. Saad Y., On the rates of convergence of the Lanczos and the block-Lanczos methods, *SIAM J. Numer. Anal.* 1980. V. 17. P. 687–706.

19. Shank S.D. and Simoncini V., Krylov Subspace methods for large-scale constrained Sylvester equations, *SIAM J. Matrix Anal. Appl.* 2013. V. 34. P. 1448–1463.

20. Sleijpen G.L.G. and Van der Vorst H.A., A Jacobi-Davidson iteration method for linear eigenvalue problems, *SIAM J. Matrix Anal. Appl.* 1996. V. 17. P. 401–425.

MULTIGRID METHOD FOR BIOLOGICAL AND MEDICAL PROBLEMS[1]

Muratova G. V., Bavin V. V.

Vorovich Institute of Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, Russia

Multigrid methods and its modifications for solving biological and medical problems are considered in this paper. We present short review of some approaches for modeling these problems. As an example we research the cable equation. Cable theory models the propagation of a pulse along the nerve fibers, composed of equivalent electrical circuits, using a parabolic differential equation in partial derivatives. We present a neuron model based on cable equation. To implement the model an algebraic multigrid method is used. Some numerical results are presented.

I Introduction

Multigrid belongs to effective iterative methods for solving large scale linear algebraic equation systems arising from the discretization of partial differential equations. Multigrid approach is based on using a different level grid sequence that allows to resolve conflicts between fast converge high frequency and slowly converge low frequency components of the error to achieve high efficiency. Due to the structure of the method, it can be adapted for different types of modeling problems.

MGM is actively used in the implementation of the models in biology and medicine. Efficient numerical simulation of processes occurring in models of biology and medcine requires the solution of large scale linear systems of equations, obtained after discretization of differential equations, which can be solve by multigrid method.

II Models in biology and medicine

Mathematical modeling of both normal physiological and pathological processes is currently one of the most relevant areas in scientific research.[1] The important field of research in medcine is the problems of hemodynamic, the functioning of the respiratory and digestive organs. [2] The Navier-Stokes equations are the base of the models describing these processes and many others.

To solve the system of linear algebraic equations obtained after approximation multigrid method can be used. Multigrid method, thanks to its structure, allows

[1] This research was supported by the Russian Foundation for Basic Research under Grants N 19-51-53013 GFEN-a

to signicantly increase the efficiency of the basic iterative method, combining the usual iterative process with a technique called coarse-grid correction. In addition, the multigrid method makes it possible to increase the efficiency by adapting its components to the problem in question. [3]

Multigrid method is suitable for modeling at the cellular level. [4, 5] Modeling the drug diffusion through the human skin is consider in [4]. A good approximation for the cells of the stratum corneum (the corneocytes) can be achieved by tetrakaidecahedral elements that allow for a realistic dense packing, similar to what can be observed in reality. The resulting mesh features mostly isotropic elements in the higher levels. Starting with a coarse grid, a distributed multigrid hierarchy is built through intertwined parallel rearrangements and redistribution. [4]

The mesh-based discretisation of the above models leads to a large system of equations. Multigrid preconditioned BiCGStab solver is used on the lower anisotropic levels. Those levels only span a small part of the available processes. This solver serves as a base solver for a multigrid preconditioned BiCGStab solver with a highly scalable Jacobi smoother on higher levels. This multigrid solver spans the whole range of available processes. [4]

Employing geometric multigrid methods and carefully designed refinement and distribution strategies, demonstrate the applicability and efficiency of simulation approaches in massively parallel computations. [4]

III Neural models

The modeling neural activity of the brain is one of the most important directions of research. Neuroscience is among the biological subdisciplines where the use of mathematical techniques are most established and recognized.

There are many different approaches for simulating brain activity.

While every approach has its advantages and limitations, such as computational cost, integrated and methods-spanning simulation approaches, depending on the network size could establish new ways to investigate the brain. In [5] it's presented a hybrid simulation approach, that makes use of reduced 1D-models using e.g., the NEURON simulatorwhich couples to fully resolved models for simulating cellular and sub-cellular dynamics, including the detailed three-dimensional morphology of neurons and organelles. The calcium model used in the presented simulations is based on a cytosolic diffusion equation with boundary conditions that include the plasma membrane calcium exchange mechanisms.To solve the linear system of equations, a geometric multigrid solver was used as a pre-conditioner along with BiCGstab as a basic solver for the linear part. [6]

One of the major problems for researchers in modeling networks is the problem of choosing the scale of the simulation. On the one hand the detailed models, describing signal propagation processes inside the neuron, are needed to account for the complex computational capabilities of a neuron as computing element of network. On the other hand large scaled networks, providing the opportunity to study the functional interaction of different neurons are required. Models based on cable theory have good detailing due to the segmentation of neural ramification and sufficient computational simplicity, which demonstrates the optimum approximation to the choice of scale.

IV Cable equation

For the description of propagation of a nerve pulse cable equation is used, which allows to represent nerve cell in the form of a section of cable placed inside a suitable environment and having insulation, which plays the role of the cell membrane. After the difference approximation of the differential equations of a model that includes a set of neurons, a large non-symmetric sparse system of linear algebraic equations is obtained. It can have the discontinuous coefficients. So it has a low convergence, when we use standard relaxation methods.

Cable equation has the following form (1):

$$\frac{d^2V}{dx^2} \cdot \frac{1}{r_l} = \frac{V}{r_m} + c_m \cdot \frac{dV}{dt}, \tag{1}$$

where r_l, r_m – axial and membrane resistance, respectively, c_m – capacity, membrane potential – V.

V Numerical experiments

Consider the problem of modeling the action potential propagation along the nerve fibers at the site of a neural network.

The initial conditions – $V(X,0) = V_r, X$ – the computational domain.

Boundary conditions:

$\frac{dV}{dx}(X,0) = 0$ when X is the end of nerve fibers,

$\frac{d^2V}{dx^2} \cdot \frac{1}{r_l} = \frac{V}{r_m} + c_m \cdot \frac{dV}{dt} + g(t) \cdot (V_k - V)$ when X is the synapse, where g – conductivity, r_m – membrane resistance, r_l – longitudinal resistance, c_m – membrane capacitance, V_k – reverse potential.

After discretization of the differential equation we obtain (2):

$$\frac{1}{r_l} \cdot \left(\frac{V_{i-1}^{k+1} - 2 \cdot V_i^{k+1} + V_{i+1}^{k+1}}{\Delta x^2}\right) = \frac{V_i^{k+1}}{r_m} + c_m \cdot \frac{V_i^{k+1} - V_i^k}{\Delta t}. \tag{2}$$

To solve such systems of linear equations, an algebraic multigrid approach with the RS, PMIS algorithm for coarsening is used. RS(Ruge-Stuben) is a traditional coarsening approach. The RS algorithm is based on two heuristic criteria that achieve optimal convergence and minimal computational cost.Therefore, the first criterion is strictly observed, and the second one is guidance. PMIS (parallel changes independent set), the algorithm of coarsening, is based on the same principles as the RS algorithm except that a heuristic criterion is not strictly observed. Unlike the RS coarsening, PMIS is not sequential. Analysis of the solution of cable equation shows that the use of multigrid algorithm, gives a significant advantage on grids in higher size.

Table 1. AMG algorithms for network size > 900000 sections.

AMM type	complexity	iterations	tsetup	tsolve	t
RS	1.91	86	24.8	266.3	91.1
PMIS(GPU)	1.77	92	65.6	86.3	51.9

Table 2. AMG algorithms for network size > 10000 sections.

AMM type	complexity	iterations	tsetup	tsolve	t
RS	1.82	45	4.6	14.6	9.2
PMIS(GPU)	1.54	35	12.2	3.2	5.4

Table 3. AMG algorithms for network size > 1000 sections.

AMM type	complexity	iterations	tsetup	tsolve	t
RS	1.53	32	4.1	8.1	2.2
PMIS(GPU)	1.52	15	9.4	5.6	5.0

VI Conclusion

The short review of the models in biology and medicine is presened. Multigrid methods for soving linear algebraic equation systems applied to the problem of simulating the propagation of action potential along nerve fibers are discussed. Considered methods provide acceptable convergence rate, and variation to achieve the desired error value does not exceed the permissible limit.Numerical experiments using a variety of methods for solving such equations, for different sizes of neural networks: RS, PMIS algorithm are presented. The efficiency of the algebraic multigrid with PMIS algorithm for solving the cable equation is demonstrated.

Bibliography

1. *I. B. Petrov* Mathematical modeling in medicine and biology based on continuum mechanics models. MIPT, 2009, no. 1, 1, P. 5-16.

2. *A.Danilov, A.Lozovskiy, M.Olshanskii, Yu.Vassilevski* A finite element method for the Navier-Stokes equations in moving domain with application to hemodynamics of the left ventricle. Russian J. Numer. Anal. Math. Modelling, V.32, No.4, 2017. P. 225-236.

3. *Chiachi Chiu* A multigrid method for pattern formation problems in biology. Differential Integral Equations 16, no. 2, 2003. P. 201-220.

4. *S. Reiter, A. Vogel, G. Wittum* Large scale simulations of continuum models using parallel geometric multigrid methods. NIC Series 49, 2018. P. 377-384.

5. *Jennifer Young, Sorin Mitran* Numerical Model of Cellular Blebbing: A Volume-Conserving, Fluid-Structure Interaction Model of the Entire Cell. J Biomech, 43(2), 2011. P. 210-220.

6. *Stephan Grein* 1D-3D hybrid modeling – from multi-compartment models to full resolution models in space and time. Front. Neuroinform, 2014.

ASYMPTOTIC APPROXIMATIONS FOR ONE RADIATIVE-CONDUCTIVE HEAT TRANSFER PROBLEM[1]

Amosov A.A., Krymov N.E.

National Research University Moscow Power Engineering Institute, Moscow, Russia

In applications, it is of great importance to study the heat transfer process in periodic media containing vacuum interlayers or cavities through which the heat transfer is realized by radiation. A numerical solution of such problems requires considerable computational efforts and becomes, in fact, impossible for a large number of heat transferring elements, especially in the case of two-dimensional and three-dimensional structures. Therefore, it is important to construct effective approximation methods.

This article continues a series of papers [1]-[9] devoted to the construction and substantiation of special discrete, semi-discrete and asymptotic approximations of radiative-conductive heat exchange problems in periodic systems of heat-conducting elements separated by a vacuum.

In this paper, we consider a stationary problem of radiative-conductive heat transfer in a periodic system consisting of n^2 absolutely black heat-conducting rods of circular cross section with diameter $\varepsilon = 1/n$ packed in a square box $\Omega = (0,1)^2$ with boundary Γ (Fig. 1). For each rod, we assign a disc G_{ij} of radius $\varepsilon/2$ with the center at the point $x_{ij} = (\varepsilon(i-1/2), \varepsilon(j-1/2))$, $1 \le i \le n$, $1 \le j \le n$.

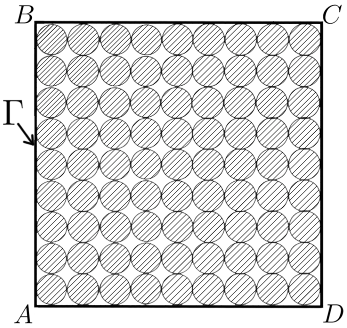


Figure 1. System of rods.

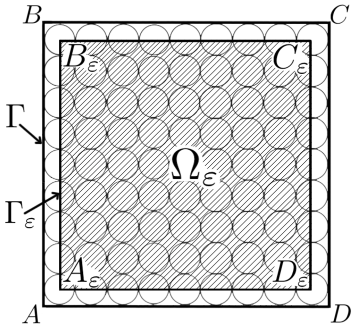


Figure 2. Ω_ε, Γ_ε and γ_ε.

The stationary process of radiative-conductive heat exchange in the system of rods $G = \bigcup_{i,j} G_{ij}$ is described by the following boundary value problem:

$$-div(\lambda \nabla u) = f, \quad x \in G, \tag{1}$$

$$\lambda \frac{\partial u}{\partial n} + h(u) = \int\limits_{\partial G} h(u(\xi))\varphi(\xi, x)\, d\sigma(\xi) + \int\limits_{\Gamma} h(u_\Gamma(\xi))\varphi(\xi, x)\, d\sigma(\xi), \quad x \in \partial G. \tag{2}$$

[1] The work supported by the Russian Science Foundation (grant 19-01-00033).

The sought function is the absolute temperature $u(x) = u(x_1, x_2)$. Here λ is the coefficient of thermal conductivity, f is the density of thermal sources; $h(u) = \sigma_0|u|^3u$, where $\sigma_0 > 0$ is the Stefan-Boltzmann constant; u_Γ is the temperature on Γ; $n(x)$ is the outward normal to ∂G for $x \in \partial G$ and the normal to Γ for $x \in \Gamma$; $d\sigma(x)$ is a natural measure on $\partial G \cup \Gamma$; φ is a visual factor:

$$\varphi(\xi, x) = \begin{cases} \dfrac{\cos(n(\xi),\, x-\xi)\cos(n(x),\, \xi - x)}{2|x-\xi|}, & \text{if } [x,\xi] \cap G = \varnothing \\ 0, & \text{if } [x,\xi] \cap G \neq \varnothing \end{cases}.$$

Integrating equation (1) over G_{ij}, taking into account condition (2) and assuming that the value of the temperature u is approximately equal to a constant u_{ij} on G_{ij}, we come to the discrete problem for values $h(u_{ij})$. This problem can be considered as a difference approximation of the following non-standard boundary value problem (the first asymptotic approximation):

$$-\varepsilon\Delta h(v) = \frac{\pi}{4}f, \quad x \in \Omega_\varepsilon, \tag{3}$$

$$\varepsilon D_n h(v) - \varepsilon^2\frac{\pi-2}{4}D_s^2 h(v) + h(v) = h(u_\Gamma(x_\Gamma)) + \varepsilon\frac{\pi}{8}f_\Gamma, \quad x \in \Gamma_\varepsilon, \tag{4}$$

$$\varepsilon\widehat{D}_n h(v) + h(v) = \widehat{h}_\Gamma + \varepsilon\frac{\pi}{16}\widehat{f}_\Gamma, \quad x \in \gamma_\varepsilon. \tag{5}$$

Here $\Omega_\varepsilon = (\varepsilon/2, 1-\varepsilon/2)^2$, Γ_ε is the boundary of Ω_ε and $\gamma_\varepsilon = \{A_\varepsilon, B_\varepsilon, C_\varepsilon, D_\varepsilon\}$ is the set of its corner points (Fig. 2); D_n and D_s are derivatives with respect to external normal and tangent to Γ_ε, $\widehat{D}_n|_{x=A_\varepsilon} = -\dfrac{1}{2}\left(\dfrac{\partial}{\partial x_1} + \dfrac{\partial}{\partial x_2}\right)$, $\widehat{D}_n|_{x=B_\varepsilon} = \dfrac{1}{2}\left(-\dfrac{\partial}{\partial x_1} + \dfrac{\partial}{\partial x_2}\right)$, $\widehat{D}_n|_{x=C_\varepsilon} = \dfrac{1}{2}\left(\dfrac{\partial}{\partial x_1} + \dfrac{\partial}{\partial x_2}\right)$, $\widehat{D}_n|_{x=D_\varepsilon} = \dfrac{1}{2}\left(\dfrac{\partial}{\partial x_1} - \dfrac{\partial}{\partial x_2}\right)$. Besides, $x_\Gamma \in \Gamma$ is the closest point to $x \in \Gamma_\varepsilon$ and f_Γ denotes the average value of the function f over the segment $[x, x_\Gamma]$. In addition, $\widehat{h}_\Gamma|_{x=A_\varepsilon}$ is the average of $h(u_\Gamma)$ over $\Gamma \cap \{|x| < \varepsilon/2\}$ and $\widehat{f}_\Gamma$ is the the mean value of f over the square with the side length $\varepsilon/2$ and the left lower vertex at the point A. Values $\widehat{h}_\Gamma$ and $\widehat{f}_\Gamma$ in points B_ε, C_ε, D_ε are calculated in an analogical way.

We consider the solution v of problem (3)-(5) as an approximation to the solution u of problem (1), (2). Note that this problem is linear with respect to $h(v)$. It is easy to see that problem (3)-(5) does not contain any information about the value of the thermal conductivity λ. As we will see below, it is permissible for well-conducting materials and leads to large errors for materials with a small value of λ.

Taking into account the local structure of the solution of problem (1), (2) we come to a more complex nonlinear boundary value problem – the second asymptotic approximation:

$$-\varepsilon\Delta H(v)=\frac{\pi}{4}f,\quad x\in\Omega_\varepsilon,\tag{6}$$

$$\varepsilon D_nH(v)-\varepsilon^2\frac{\pi-2}{4}D_s^2H_*(v)+H_\Gamma(v)=H_\Gamma(u_\Gamma(x_\Gamma))+\varepsilon\frac{\pi}{8}f_\Gamma,\quad x\in\Gamma_\varepsilon,\tag{7}$$

$$\varepsilon\widehat{D}_nH_*(v)+H_\Gamma(v)=\widehat{H}_\Gamma+\varepsilon\frac{\pi}{16}\widehat{f}_\Gamma,\quad x\in\gamma_\varepsilon.\tag{8}$$

Here

$$H(v)=\int\limits_0^v\frac{h'(t)dt}{1+(1-\alpha)\dfrac{\varepsilon}{\lambda}h'(t)},\quad H_*(v)=\int\limits_0^v\frac{h'(t)}{\left(1+\dfrac{\varepsilon}{2\lambda}h'(t)\right)^2}dt,$$

$$H_\Gamma(v)=\int\limits_0^v\frac{1+(1+\alpha)\dfrac{\varepsilon}{2\lambda}h'(t)}{\left(1+\dfrac{\varepsilon}{2\lambda}h'(t)\right)^2}h'(t)dt,\quad \alpha\approx0.178.$$

Value $\widehat{H}_\Gamma|_{x=A_\varepsilon}$ is the average of $H(u_\Gamma)$ over $\Gamma\cap\{|x|<\varepsilon/2\}$. Values $\widehat{H}_\Gamma$ in points B_ε, C_ε, D_ε are calculated in an analogical way.

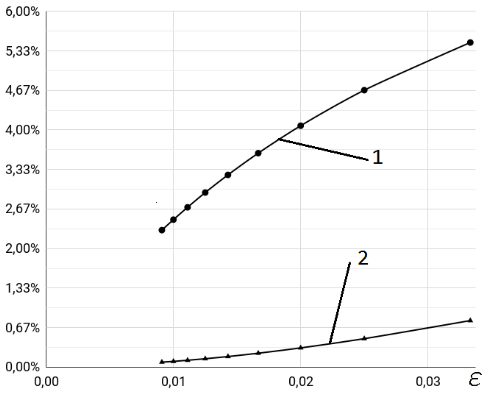
We consider the solution v of problem (6)-(8) as an approximation to the solution u of problem (1), (2).

Below we present the results of some computational experiments that allow us to draw preliminary conclusions about the quality of the proposed approximations. Consider a system of rods packed in a square box Ω with a side equal to 1 m. The temperature u_Γ of its left and right boundaries continuously varies between $300K$ and $1000K$ with two symmetrically located maxima. At the upper and lower boundaries the temperature is constant: $u_\Gamma=300K$. Besides, $f=0$, that is, there are no heat sources or sinks.

To solve boundary-value problems (3)-(5) and (6)-(8) we used finite-difference methods with very small mesh steps (much less than ε). The obtained approximations were compared with an "exact" solution of the problem (1), (2), obtained by the finite-difference method using a sufficiently detailed radial grid on each disc G_{ij}. To estimate the relative errors of approximate solutions we use the following value:

$$\delta v=\frac{\sqrt{\sum_{i,j}\left(u(x_{ij})-v(x_{i,j})\right)^2}}{\sqrt{\sum_{i,j}u(x_{i,j})^2}}.$$

Figure 3 shows graphs of relative errors as functions of ε when the value of thermal conductivity $\lambda=5\,W/(m\cdot K)$. Figure 4 shows graphs of relative errors as functions of λ when the value of rods diameter $\varepsilon=0.02$. It can be seen that the second asymptotic approximation turns out to be much more accurate than the first.



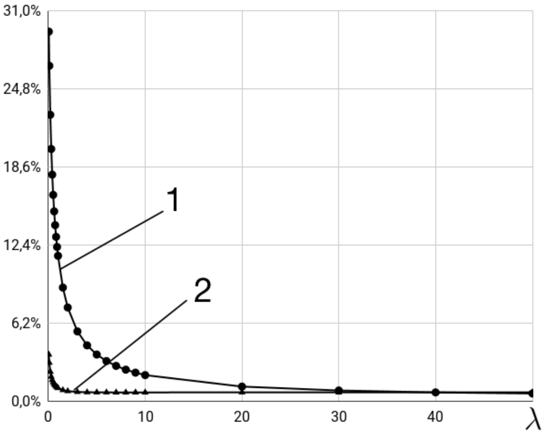


Figure 3. Dependence of relative errors on ε at $\lambda = 5W/(m \cdot K)$;

Figure 4. Dependence of relative errors on λ at $\varepsilon = 0.02\,m$;

1 – the first asymptotic approximation, 2 – the second asymptotic approximation

The distribution of the obtained temperature values is shown on Figures 5 and 6 using thermograms, where the transition from the maximum temperature to the minimum corresponds to transition on the gray scale from white to black.

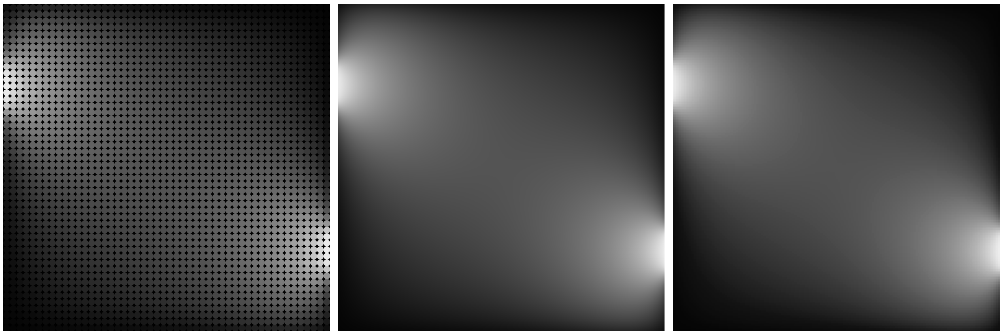


Figure 5. Exact solution, the first and the second asymptotic approximations in the case $\lambda = 200W/(m \cdot K)$

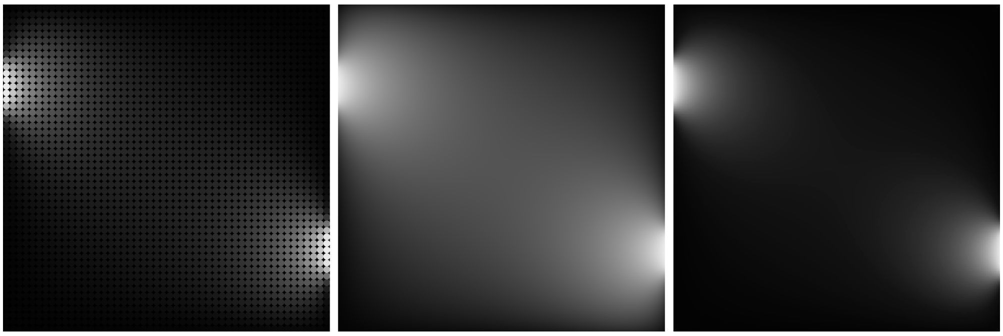


Figure 6. Exact solution, the first and the second asymptotic approximations in the case $\lambda = 0.05W/(m \cdot K)$

In the case of the large value of $\lambda = 200\,W/(m \cdot K)$ both asymptotic approximations give good approximations to the exact solution, while for the small value of $\lambda = 0.05\,W/(m \cdot K)$ the first approximation turns out to be practically unsuitable.

Bibliography

1. *Amosov A.A.* Semidiscrete and asymptotic approximations to a solution to the heat transfer problem in a system of heat shields under radiation [in Russian]// Modern Problems of Mathematical Simulation. Rostov-na-Donu. 2007. P. 21–36.

2. *Amosov A.A., Gulin V.V.* Semidiscrete and asymptotic approximations in the heat transfer problem in a system of heat shields under radiation [in Russian] // Vestnik MEI. 2008. N 6. P. 5–15.

3. *Amosov A.A.* Semidiscrete and asymptotic approximations for the nonstationary radiative-conductive heat transfer problem in a periodic system of grey heat shields // J. Math. Sci. (United States). 2011. Vol. 176. N 3. P. 361–408.

4. *Kremkova A.A.* Semidiscrete and asymptotic approximations for the radiative-conductive heat transfer problem in the two-dimensional periodic structure [in Russian] // Vestnik MEI. 2012. N 6. P. 151–161.

5. *Amosov A.A., Kremkova A.A.* An estimate of the error of semi-discrete approximate method for solving the radiative-conductive heat transfer problem in the two-dimensional periodic structure [in Russian] // Vestnik MEI. 2013. N 6. P. 22–36.

6. *Amosov A.A., Maslov D.A.* Two stationary radiative-conductive heat transfer problems in a system of two-dimensional plates // J. Math. Sci. (United States). 2015. Vol. 210. N 5. P. 3–14.

7. *Amosov A.A., Maslov D.A.* Semidiscrete approximations for the stationary radiative-conductive heat transfer problem in the two-dimensional system of plates // Russ. J. Numer. Anal. Math. Modelling. 2016. Vol. 31. N 1. P. 1–17.

8. *Amosov A.A.* Asymptotic approximations for the stationary radiative-conductive heat transfer problem in the two-dimensional system of plates // Russ. J. Numer. Anal. Math. Modelling. 2017. Vol. 32. N 3. P. 173–185.

9. *Amosov A.A., Krymov N.E.* Approximations for the stationary problem of radiative-conductive heat exchange in a system of rods of circular cross section [in Russian] // Vestnik MEI. 2017. N 5. P. 94–100.

COMPUTER SIMULATION OF MIGRATION OF LIQUID CYLINDRICAL INCLUSIONS IN A CRYSTAL IN SOME CASES OF INTERFACIAL ENERGY ANISOTROPY[1]

Garmashov S.I., Karpenko A.S.

Southern Federal University, Rostov-on-Don, Russia

One of the methods for studying the processes of crystallization and dissolution can be based on the phenomenon of migration of liquid inclusions in a non-uniformly heated crystal [1] – [7]. In order to extract information from experimental data on the migration of such inclusions [6, 7], it is important to have an adequate mathematical model of this process. In the general case, the liquid inclusion in a crystal is a three-dimensional object, the shape of which can change during its migration through the crystal. Therefore, in order to describe such process, it is necessary to solve three-dimensional non-stationary equations of mass and heat transfer in a volume with moving boundaries. Besides, obtaining experimental data on the shapes and velocities of liquid inclusions in the form of drops is a rather complicated task, especially if the crystal is opaque in the visible region of the spectrum and the process temperature is much higher than the room temperature. Therefore, it makes a sence to study the steady-state migration of the liquid inclusions with cylindrical shapes, when it is possible to analyze the shape of not the entire inclusion, but only its cross-section.

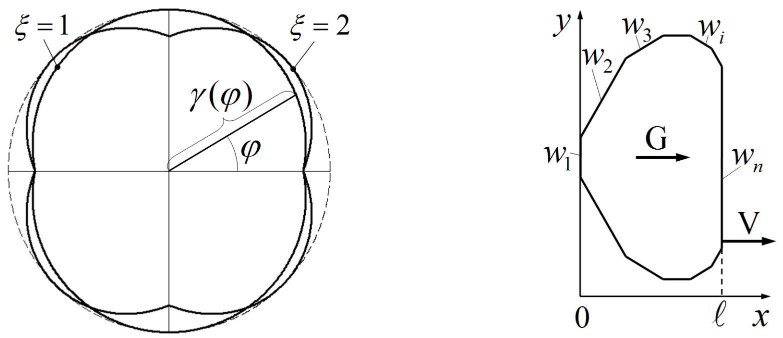
The steady-state velocity and cross-sectional shape of a liquid cylindrical inclusion migrating in a stationary uniform field of the temperature gradient, as was shown in [2, 3], depend on the specific interfacial energy and interface kinetics, as well as their anisotropies. The interfacial energy anisotropy (IEA) is known to determine the shape of a crystal (or an inclusion in a crystal) in equilibrium (i.e., in the absence of a temperature gradient). When the crystal is not uniformly heated, the equilibrium conditions at the solid-liquid interfaces are violated, and the inclusion shape deviates from the equilibrium one due to the occurrence of supersaturation (undersaturation) of the crystal solution at the flat (atomically-smooth) parts of the inclusion boundary. As was shown in [1] for the case of liquid inclusions in the form of flat interlayers, the interface kinetics most strongly limits the mass transfer in thin interlayers. For this reason, the dependence of the velocity (V) of the interlayer on its thickness (l) increases monotonously from 0 (at $l = 0$) and asymptotically approaches (at $l \longrightarrow \infty$) some maximum - so called the diffusive velocity corresponding to the absence of interface limitations. In the papers devoted to the migration of cylindrical inclusions, the $V(l)$ dependence was not analyzed in details, and, therefore, it is interesting to calculate and analyze this dependence not only with varying interface kinetics but the degree of the IEA.

[1] The work supported by the Russian Science Foundation (grant 19-01-00033).

Generally, the IEA is described by a complicated dependence of the specific interfacial energy on the orientation of the crystal facets. In this paper, we consider the case when the IEA can be approximated by a function $\gamma(\varphi)$ (Fig. 1a) in the form:

$$\gamma(\varphi) = \gamma_{min} + (\gamma_{max} - \gamma_{min})|\sin(\xi\varphi)|, 0 \leq \varphi < 2\pi, \tag{1}$$

where γ is the specific interfacial energy of the crystal facet with the orientation given by the angle φ; γ_{min}, γ_{max} are the minimal and maximal values of $\gamma(\varphi)$; $\xi = 1$ or 2. The facets corresponding to the sharp minima of $\gamma(\varphi)$ at $\varphi = n\pi/\xi$ $(n \in Z)$ are considered as the singular facets.



(a) (b)

Figure 1. Schematics of (a) the IEA types used in the present model (Eq.(1)) and (b) the approximated cross-sectional shape of the cylindrical inclusion migrating with the velocity V in a crystal under the action of the temperature gradient **G**

The value of $\xi = 1$ can correspond to the case of the migration of a cylindrical inclusion in the direction normal to the close-packed crystal planes (e.g., the inclusion migration in a crystal of Si in the <111> direction [3]). The value of $\xi = 2$ is characteristic of the inclusions with 4 singular facets (e.g., the migration of brine inclusions in a crystal of KCl in the <100> direction [2]). We assume that the temperature gradient is perpendicular to the singular facets, the mechanism of the interface processes corresponds to the two-dimensional nucleation [1], and there are no temperature oscillations.

In order to calculate the cross-sectional shape of the cylindrical inclusion at the given velocity of its migration, we use the conclusion from [3] that the distribution of actual concentrations in the liquid phase in the steady state should be flat and inclined in the direction of the inclusion migration. Based on both this conclusion and the flow balance conditions in the inclusion and at its interfaces, the desired shape can be calculated by means of the numerical integration, as described in [3, 7], or by the method of facets proposed in [2] and used in [4].

The method of facets is based on the approximation of the curvilinear interfaces by a set of flat sections (facets) (Fig.1b). It is preferable in that it allows us to calculate the cross-sectional shape of the inclusion simply for arbitrary anisotropies of the interfacial energy and interface kinetics, but the rate of its convergence, as will be shown below, is relatively low. Nevertheless, we use the method of facets for solving the problem considered in the present work because we are interested to analyze the influence of the various types of the IEA on the velocities and cross-sectional shapes of the cylindrical inclusions. Briefly, the algorithm for constructing the cross-sectional shape for the cylindrical inclusion is the following.

First of all, we set the inclusion velocity V in the range from 0 to the diffusive velocity, and then, using this value, calculate both the slope of the flat distribution $(C(x,y))$ of actual concentrations of the crystal solution over the inclusion cross-section and the supersaturation (undersaturation) at the singular facets according to a given mechanism of the interface processes. If after this we set arbitrarily the size (w_1) of the first (singular) facet (Fig. 1b) and take into account the decrease of the liquidus at it because of the capillary effects [2, 3, 5], then the distribution $C(x,y)$ for the given velocity V can be completely determined.

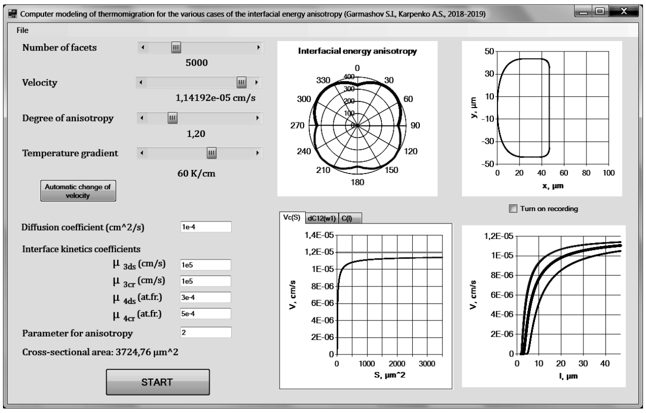
Since the supersaturations at the curvilinear (atomically-rough) parts of the inclusion boundary are negligibly small, the distribution of the equilibrium concentrations along these parts coincides with the earlier calculated distribution $C(x,y)$ of the actual concentrations. Hence, the sizes $(w_2, w_3$, and so on) of the next facets approximating the curvilinear interface (Fig. 1b) must be chosen so that the equilibrium concentrations at them exactly coincide with the actual concentrations for these facets. As a result, the entire shape of the inclusion can be constructed. However, the size (w_n) of the last (singular) facet (Fig. 1b) might be such that the flow balance conditions at it will not be satisfied. Therefore, the size (w_1) of the first facet should be changed so that the flow balance conditions would not be violated at the last facet. The cross-section shape constructed in this way (Fig. 1b) will correspond to the given migration velocity V. It should be noted that one value of the inclusion velocity V can correspond to several shapes of the cross-sections with different areas, all other conditions being equal.

According to the above algorithm, a computer program was developed (Fig. 2). The program allows us to calculate the velocity and shape of the inclusion as functions of the area and thickness of the cross-section for varying type and degree of the IEA and degree of difficulty of the interface processes.

Fig. 3 illustrates the rate of convergence of the method of facets in the cases when the cross-sectional shape is slightly and significantly different from the equilibrium one. In both cases, the order of the convergence rate is about 1. We used 5000 facets for approximating the cross-sectional shape.

With the developed program, the dependences of the velocities of cylindrical inclusions on the thicknesses of their cross-sections were calculated (Fig. 4) for the IEA types corresponding to 2 and 4 singular facets, as well as for the case of migration of the flat interlayers [1], with other conditions being equal.

It follows from the obtained results that the velocities of the cylindrical inclusions are higher than the ones of the flat interlayers with the same thicknesses, all other conditions being equal, but approaching the rates of the latter with increasing the degree of the IEA $(\gamma_{max}/\gamma_{min})$. The velocities for the inclusions



Computer modeling of thermomigration for the various cases of the interfacial energy anisotropy (Garmashov S.I., Karpenko A.S., 2018-2019)

Number of facets 5000

Velocity 1,14192e-05 cm/s

Degree of anisotropy 1,20

Temperature gradient 60 K/cm

Diffusion coefficient (cm^2/s) 1e-4

Interface kinetics coefficients

μ_{3ds} (cm/s) 1e5

μ_{3cr} (cm/s) 1e5

μ_{4ds} (at.fr.) 3e-4

μ_{4cr} (at.fr.) 5e-4

Parameter for anisotropy 2

Cross-sectional area: 3724,76 μm^2

☐ Turn on recording

Figure 2. The computer program developed for calculating the velocities and cross-sectional shapes of migrating cylindrical inclusions by the method of facets

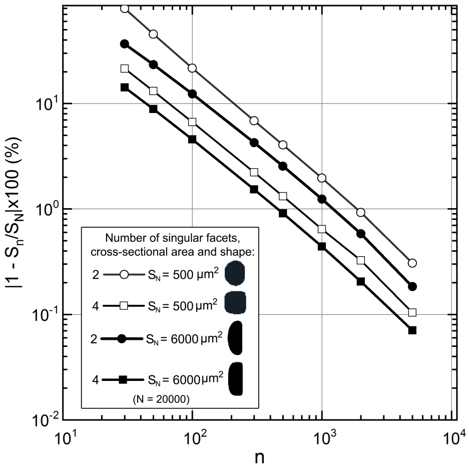


Figure 3. The relative errors of calculating the cross-section area S_n as functions of the number of facets (n) approximating the inclusion shape with 2 and 4 singular facets in the cases of relatively low and high degrees of its deformation

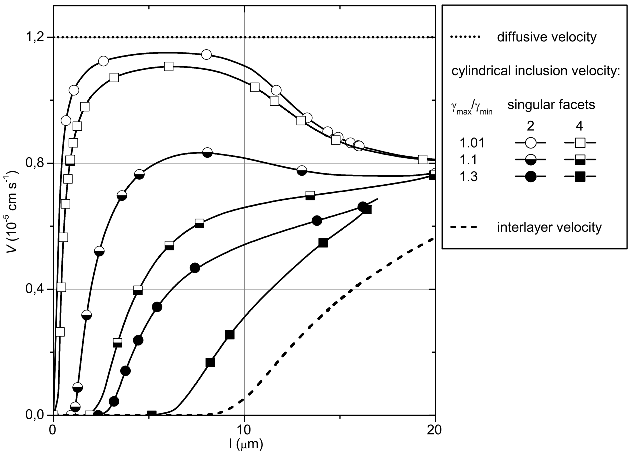


Figure 4. The calculated velocities of the cylindrical inclusions (at the different types and degrees of the IEA) and flat interlayers as functions of their thicknesses

with 4 singular facets are lower than for the ones with 2 singular facets under the same degree of the IEA. Interestingly, in contrast to the case of the flat interlayers, for that the $V(l)$ dependences are monotonic, the similar dependences in the case of the cylindrical inclusions can take the form of curves with a maximum (at the small degrees of the IEA).

Bibliography

1. *Tiller W.A.* Migration of a liquid zone through a solid: Part I. // J. Appl. Phys. 1963. Vol. 34. P. 2757–2762.

2. *Cline H.E., Anthony T.R.* Nonequilibrium morphology of liquid inclusions migrating in solids // J. Appl. Phys. 1977. Vol. 48. P. 5096–5104.

3. *Garmashov S.I., Gershanov V.Yu.* Velocity and cross-section shape of liquid cylindrical inclusions migrating normally to close-packed planes of a non-uniformly heated crystal under stationary thermal conditions // J. Cryst. Growth. 2009. Vol. 311, N 2. P. 413–419.

4. *Garmashov S.I., Surnin V.I.* A computer model of steady-state cross-sectional shape of liquid cylindrical inclusion migrating through a crystal with account of thermal gradient direction and anisotropy of interfacial energy and interface kinetics // Abstracts of Lecturers and Young Scientists of Second China-Russia Conference on Numerical Algebra with Applications (CRC-NAA'13), June 25-29, 2013, Rostov-on-Don, Russia. Rostov-on-Don: Southern Federal University Publishing, 2013. P. 74–76.

5. *Gershanov V.Yu., Garmashov S.I.* Inverse Gibbs–Thomson effect // Tech. Phys. 2015. Vol. 60, N 1. P. 61–65.

6. *Garmashov S.I., Protsenko V.V.* Techniques of calculation of interfacial energy anisotropy from experimental data on cross-sectional shapes of cylindrical inclusions migrating in a crystal // Communication on Applied Mathematics and Computation. 2018. Vol. 32, N 2. P. 190–201.

7. *Garmashov S.I.* On a technique of studying the interface kinetics and anisotropy of specific interfacial energy based on experiments on migration of liquid cylindrical inclusions in a crystal under stationary thermal conditions // Crystallography Reports. 2018. Vol. 63, N 5. P. 844–848.

NUMERICAL ANALYSIS OF COSYMMETRY VIOLATION IN FILTRATION CONVECTION PROBLEM[1]

Govorukhin V.N.

Southern Federal University, Rostov-on-Don, Russia

The property and theory of cosymmetry were introduced to explain a phenomenon of existence of one-parameter family of steady-state flows in the planar Darcy convection problem. The branching of one-parameter set of equilibria (closed curve of equilibria in phase space) is a typical bifurcation in cosymmetric dynamical system [1, 2]. It implies the existence, at fixed physical parameters, an infinite number of substantially different steady-state regimes.

An analytic study of the one-parameter family of steady-state regimes in cosymmetric problems is possible for parameters values near bifurcation of its appearance [1, 2], and numerical approaches can be used for other situations. Development of special numerical methods to studying cosymmetric dynamical systems started in [3, 7] and continued in [4, 5, 6, 7, 8, 9].

If cosymmetry is violated the continuous families of steady-state regimes break down or disappear [11] and as result can happen unusual bifurcations. By V.I. Yudovich was proposed the selective function for studying these destructions. This function indicates which points of the equilibrium curve conserves under small violation of cosymmetry. It has been shown that two scenarios are possible for stable family of equilibria: the disintegration of the family on a finite number of equilibria or appearance of slow periodic motions. However, what might be happening with partially stable families (consisting of stable and unstable arcs) under cosymmetry breaking perturbations remain largely open. This paper is devoted to numerical analysis of such bifurcation phenomena and outlines the numerical approaches for investigation both cosymmetric and close to cosymmetric tasks.

The problem of filtration convection in the Darcy-Boussinesq approximation in a region $\Omega \in R^2$ and in the presence of internal heat sources is reduced to the following system of differential equations

$$\begin{aligned} \Delta\psi &= \theta_x, && (1)\\ \theta_t + \psi_y\theta_x - \psi_x\theta_y &= \Delta\theta + \lambda\psi_x + \delta\, f(x,y). && (2) \end{aligned}$$

Here x and y are Cartesian coordinates, t is time, $\psi(t,x,y)$ is a stream function and $\theta(t,x,y)$ is a deviation of the temperature from a linear vertical equilibrium profile. λ is an analogue of the Rayleigh number, $f(x,y)$ is a function which a result of external actions, for example, intensity distribution function of the internal heat sources or infiltration and δ is a parameter.

[1] Supported by RFBR Grant 18-01-00453

We consider a rectangular domain $\Omega = [0,a] \times [0,b]$. On the boundary of Ω, the following conditions are assumed to hold:

$$[\theta]_{\partial\Omega} = [\psi]_{\partial\Omega} = 0. \tag{3}$$

The initial condition for system has the form

$$\theta(0,x,y) = \theta_0(x,y) \tag{4}$$

where $\theta_0(x,y)$ is a function defined in Ω.

For any fixed value of t we can express ψ from (1) in terms of θ by solving the Dirichlet problem for the Poisson equation.

For $\delta = 0$, problem (1)–(4) has cosymmetry which is given by the function $L\theta = \psi$. This means that the following equality holds:

$$\int_\Omega (\triangle\theta - \psi_y\theta_x + \psi_x\theta_y + \lambda\psi_x)\psi\, dxdy = 0. \tag{5}$$

In cosymmetric case for each transition of λ through the value $\lambda_{m,n} = 4\pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)$, $n,n = 1,2,\dots$ corresponds the bifurcation of the onset of a one-parameter family of equilibria (steady state regimes), see [1, 2].

When $\delta \neq 0$, the cosymmetry of the problem is violated. Corresponding selection equation, see [11] has the form

$$\psi_y\theta_x - \psi_x\theta_y = \Delta\theta + \lambda\psi_x, \; \int_\Omega f(x,y)\,\psi\, dxdy = 0 \tag{6}$$

here ψ can be expressed in terms of θ from (2).

The numerical analysis of a *cosymmetric and near-cosymmetric systems* leads to *computational problems* due to their specific properties. First of all numerical methods must *conserve the cosymmetry of the problem.* It includes preserving the cosymmetry by discretizations of partial differential equations. The violation of cosymmetry properties by discretizations may lead to the spurious behaviour such as destruction of the family of equilibria. The other problem is a strongly degeneration of equilibria in cosymmetric case and impossibility to use standard methods for their study. The algorithm of *continuation of the family of equilibria* along the hidden parameter should be used for this purpose. The *analysis of special selective function* must applied for studying near-cosymmetric systems and destruction of families of equilibria. This selective function indicates which points of the equilibrium curve preserves under violation of cosymmetry. All mentioned stages of numerical research as applied to convection in porous medium problems are considered in this paper.

In this work, we use the Galerkin method for analysis ot the problem (1)–(4). Functions ψ and θ are approximated by the series

$$\psi = \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{i,j}(t)\phi_{i,j}(x,y), \quad \theta = \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{i,j}(t)\phi_{i,j}(x,y), \\ \phi_{i,j}(x,y) = \tfrac{2}{\sqrt{ab}}\sin\left(i\tfrac{\pi x}{a}\right)\sin\left(j\tfrac{\pi y}{b}\right). \tag{7}$$

Substitution of (7) into (1) and projection operations lead to a system of ODEs of order $N = n^2$. Similarly, we obtain a approximating system for the selective function (6). It is easy to prove that the obtained approximation preserves the cosymmetry of the original problem.

The system of ODE's for $\delta = 0$ is cosymmetric and their cosymmetry is defined by approximation of ψ from (2). Only those steady-state solutions on the family $\widehat{\theta}(s)$ survives for small perturbations violating the cosymmetry for which the value of (6) approximation are equal zero. Here s is a parameter on the curve of equilibria. Thus, for the analysis of family of equilibria destruction we need to be able to calculate their values for any s.

The continuation method is based on the cosymmetric implicit function theorem, see [10]. The first version of this method was provided in detail in [7]. Here is presented the algorithm for continuation of the equilibria curve along a hidden parameter simultaneously with the analysis of its destruction under cosymmetry violating perturbations. As F and D we denote approximations of steady problem (1)-(4) for $\delta = 0$ and of selection function (6) respectively. By $\widehat{\theta}(s)$ we denote equilibrium point on the family corrsponding to value s of the curve parametrization.

The method can be presented as the following algorithm:

1. Find a point $\widehat{\theta}_0 = \widehat{\theta}(s_k)$, $k = 0$ on the curve of equilibria by modified Newton method. Choice the direction of continuation along the curve. Calculate the kernel of Jacobi matrix of F at the point $\widehat{\theta}(s_k)$. Selection function $D(s)$ calculation in $\widehat{\theta}(s_0)$.

2. Do one step of Runge-Kutta method for special Cauchy problem with start point s_k, $\widehat{\theta}(s_k)$. On each stage of RK-method a eigenvector ϕ_0 corresponding to zero eigenvalue (RHS of special Cauchy problem) must be calculated.

3. Checking of accuracy of predicted point of the curve of equilibria. We use following criteria for accuracy control: $\left\| F\,\widehat{\theta}(s_{k+1}) \right\| < \epsilon$, where $F\,\theta$ is a right hand side of cosymmetric ODE system and ϵ is accuracy. If the accuracy of equilibria calculations is not satisfactory then correct step of Runge-Kutta method $h = h/2$ and goto to item 2 else go to next item.

4. Correct the point $\widehat{\theta}(s_{k+1})$ using iterations of Newton method. The result of this step is a next point $\widehat{\theta}(s_{k+1})$ of the curve of equilibria.

5. Selection function calculation in $\widehat{\theta}(s_{k+1})$. If $D\,\widehat{\theta}(s_k) \cdot D\,\widehat{\theta}(s_{k+1}) < 0$ then finding s_* corresponding to zero of selective function. This equilibrium is survives under non-cosymmetric perturbation. Using of dense output algorithm provides to obtain $\widehat{\theta}(s)$ for any $s \in [s_k, s_{k+1}]$ with accuracy without additional calculations.

6. Check of conditions of the end of computation (step too small, number of steps, curve of equilibria is closed. etc.). If at least one of conditions is true then computation stops. Otherwise transition to item two of algorithm.

The described algorithm was applied to analyze a number of problems of filtration convection with various functions $f(x, y)$. The results will be presented in a talk.

Bibliography

1. *Yudovich V. I.* Cosymmetry, degeneration of solutions of operator equations, and onset of filtration convection // Mat. Zametki 49. 1991. P.142–148.

2. *Yudovich V. I.* Secondary cycle of equilibria in a system with cosymmetry, its creation by bifurcation and impossibility of symmetric treatment of it // Chaos. N5. 1995. P.402–411.

3. V. N. Govorukhin, Computer experiments with cosymmetric models, Z. Angew. Math. Mech. 76 (1996) 559–562.

4. *Govorukhin V. N.* Numerical investigation of loss of stability by secondary steady-state regimes in the Darcy plane convection problem // Dokl. Ross. Akad. Nauk. V.363. 1998. P. 772–774.

5. *Govorukhin V. N.* Analysis of families of secondary steady-state regimes in the problem of plane flow through a porous medium in a rectangular vessel, Fluid Dynamics 34 (1999) 652–659.

6. *Karasösen B., Tsybulin V. G.* Finite-difference approximations and cosymmetry conservation in filtrational convection problem // Phys. Lett. A. N262. 1999. P.321–329.

7. *Govorukhin V. N.* Calculation of one-parameter families of stationary regimes in a cosymmetric case and analysis of plane filtrational convection problem // Continuation Methods in Fluid Dynamics. Vol. 74 of Notes on Numerical Fluid Mechanics. Vieweg. 2000, P. 133–144.

8. *Govorukhin V.N., Shevchenko I.V.* Numerical investigation of the second transition in the problem of plane convective flow through a porous medium // Fluid Dynamics. V.38. 2003. P.760–771.

9. *Govorukhin V. N.* On the action of internal heat sources on convective motion in a porous medium heated from below // J. Appl. Mech. Tech. Phy. N5. 2014. P. 225–233.

10. *Yudovich V. I.* Implicit function theorem for cosymmetric equations // Mat. Zametky. V.60. 1996. P. 313–317.

11. *Yudovich V. I.* Bifurcations under perturbations violating cosymmetry // Doklady Physics V.49 (9). 2004. P. 522–526.

A P-VERSION TWO LEVEL SPLINE METHOD FOR 2D NAVIER-STOKES EQUATIONS[1]

Han D.-F.

Department of Mathematics, Hangzhou Normal University, Hangzhou 310006, China

In two dimensional case, the weak form of the stream function formulation for the Navier-Stokes equations is standard [1], that is, to find a solution $\psi \in H^3(\Omega) \bigcap H_0^2(\Omega)$ satisfying the boundary conditions, which should be specified later, such that

$$\nu a(\psi, \phi) + b(\psi; \psi, \phi) = \langle f, \phi \rangle, \qquad \forall \phi \in H_0^2(\Omega), \tag{1}$$

where ν is viscosity, and the bilinear functional $a(\cdot, \cdot)$ and the trilinear functional $b(\cdot; \cdot, \cdot)$ are defined by

$$a(\psi, \phi) = \int_\Omega \Delta\psi \Delta\phi dxdy, \qquad b(\xi; \psi, \phi) = \int_\Omega \Delta\xi \left(\frac{\partial\psi}{\partial x}\frac{\partial\phi}{\partial y} - \frac{\partial\psi}{\partial y}\frac{\partial\phi}{\partial x} \right) dxdy,$$

where function $f \in L^2(\Omega)$ in the right hand side of (1), and $\langle \cdot, \cdot \rangle$ denotes the inner product of $L^2(\Omega)$.

First we introduce some notations for the ease of describing the spline methods. Denote $\triangle$ as the triangulation of the computational domain Ω, and

$$\mathcal{S}_h^d = \{ s \in C^1(\Omega) : s|_t \in \mathbf{P}_d, \forall t \in \triangle \}.$$

is a spline space of the degree d with C^1-smoothness on $\triangle$, and the subscription h means the mesh size of $\triangle$. We know that $\mathcal{S}_h^d$ is a finite dimensional function space which is dense in $C^1(\Omega)$. The full approximation power order could be achieved following the multivariate spline theory. For the ease of reference, we presented one result in the following lemma.

Lemma 1. Assuming $d \geq 5$ and $\psi \in H^{d+1}(\Omega)$. $\psi_d \in \mathcal{S}_h^d$ is the spline approximation for ψ, then

$$|\psi - \psi_d|_2 \leq C \inf_{\omega \in \mathcal{S}_h^d} |\psi - \omega|_2 \leq C h^{d-1}.$$

In the current research, we only consider the case of Dirichlet boundary condition, which is $(\psi, \frac{\partial\psi}{\partial \mathbf{n}})\big|_{\partial\Omega} = \mathbf{g} := (g_1, g_2)$. Other types of boundary condition could be treated as the same manner as in the finite element method. Let $\mathbf{n}$ be the normal direction of the boundary $\partial\Omega$, and then denote

$$\mathcal{S}_{h,\mathbf{g}}^d = \{ s \in \mathcal{S}_h^d : s|_{\partial\Omega} = g_1, \frac{\partial s}{\partial \mathbf{n}}|_{\partial\Omega} = g_2 \}$$

[1] Supported by the National Foundation of Natural Science with Grant No.11471092, and the Natural Science Foundation of Zhejiang Province under grant LZ13A010003.

as the spline space satisfying the Dirichlet boundary condition above. In the homogeneous case, it is written as $\mathcal{S}^d_{h,\mathbf{0}}$.

Then the spline methods for the model problem (1) can be stated as follows: To find a solution $\psi_d \in \mathcal{S}^d_{h,\mathbf{g}}$ such that

$$\nu a(\psi_d, \phi_d) + b(\psi_d; \psi_d, \phi_d) = \langle f, \phi_d \rangle, \qquad \forall \phi_d \in \mathcal{S}^d_{h,\mathbf{0}}(\Omega), \tag{2}$$

where the bilinear form $a(\cdot,\cdot)$, the trilinear form $b(\cdot;\cdot,\cdot)$ and the inner product $\langle \cdot,\cdot \rangle$ are defined as before. The advantage of the spline method is that we do not need to construct the explicit basis functions of S^d_h with different degree d. In this sense, we are intended to regard it as an alternative approach for the finite element methods. The dual problem corresponding to (2) is necessary to derive the error with H^1-norm: To find a function $w_f \in H^2_0(\Omega) \bigcap H^3(\Omega)$ such that

$$\nu a(\phi, w_f) + q(\zeta; \phi, w_f) = \langle \phi, f \rangle, \quad \forall \phi \in H^2_0(\Omega), \tag{3}$$

which is necessary to derive the error with H^1-norm, where $\zeta = \psi$ or ψ_d and

$$q(\xi; \psi, \phi) = b(\xi; \psi, \phi) + b(\psi; \xi, \phi).$$

In the spline method, any spline $\psi_d \in \mathcal{S}^d_{h,\mathbf{g}}$ is written into its piecewise B-form on any triangle element t, that is

$$\psi_d = \sum_{i+j+k=d} c_{i,j,k} B_{i,j,k}(x,y), \qquad \forall t \in \triangle, \tag{4}$$

where $B_{i,j,k}(x,y)$ is the triangular Bernstein polynomial defined on any triangle. Denote $\left(c_{i,j,k}\right)_{i+j+k=d}$ as the B$\acute{e}$zier coefficients of the spline ψ_d. It is geometrically simple to formulate the global C^1-smoothness constrains with the B$\acute{e}$zier coefficients.

We can use the algorithm to elevate the degree of a polynomial from d to $d+1$. As for arbitrary $D > d$, by repeatedly executing the algorithm $D-d$ times.

Lemma 2. (Degree elevation algorithm)

Let ψ_d be a polynomial of degree d defined on a triangle t in the form of (4), then ψ_d can be elevated to a polynomial of degree $d+1$ with the form $\psi_d = \sum\limits_{i+j+k=d+1} \tilde{c}_{ijk} B_{ijk}(x,y)$, where the new B$\acute{e}$zier coefficients are calculated by

$$\tilde{c}_{ijk} = (\frac{i}{d+1} c_{i-1,j,k} + \frac{j}{d+1} c_{i,j-1,k} + \frac{k}{d+1} c_{i,j,k-1}), \tag{5}$$

for $i+j+k = d+1$. The coefficients are assumed to be zero with negative subscripts.

Then the new p-version of the two level method for the stream function formulation of Navier-Stokes equations can be concluded as the following algorithm, which gives high order spline approximate solutions ψ_D and ψ_D^*.

Algorithm. (The p-version two level spline method)

Step 1. Find $\psi_d \in \mathcal{S}_{h,\mathbf{0}}^d$ by solving (2).
Step 2. Do interpolation for ψ_d using the degree elevation algorithm (5) .
Step 3. Find $\psi_D \in \mathcal{S}_{h,\mathbf{0}}^D$, which solves

$$\nu a(\psi_D, \phi) + q(\psi_d; \psi_D, \phi) = b(\psi_d; \psi_d, \phi) + \langle f, \phi\rangle, \qquad \forall \phi \in \mathcal{S}_{h,\mathbf{0}}^D. \qquad (6)$$

Step 4. Find a better solution $\psi_D^* \in \mathcal{S}_{h,\mathbf{0}}^D$, which solves

$$\nu a(\psi_D^*, \phi)+q(\psi_d; \psi_D^*, \phi) = q(\psi_d; \psi_D, \phi)-b(\psi_D; \psi_D, \phi)+\langle f, \phi\rangle, \quad \forall \phi \in \mathcal{S}_{h,\mathbf{0}}^D. \quad (7)$$

It is worth to remark that Step 4 of the above algorithm is only effective when the high order D is much larger than the low order d, for e.g., $d = 5$ and $D = 11$. It is interested when D is fairly large and much lower d could be used to reduce the computational costs. A precise description for the restrictions between the values of d and D in Step 3 and Step 4 are given respectively by the theorems later.

It is then straightforward to derive our first error estimations of the high order spline solution ψ_D and ψ_D^* in the Algorithm.

Theorem 1. Suppose $\psi \in H^{D+1}(\Omega)$ is the exact solution of the equation (1) and ψ_D is obtained by the Algorithm. If $d < D \leq 2d-1$ $(d \geq 5)$, then we have

$$|\psi - \psi_D|_2 \leq C(h^{D-1} + |\ln h|^{1/2}h^{2d-1}) \leq Ch^{D-1}.$$

Namely, ψ_D has the optimal asymptotic accuracy in energy norm $|\cdot|_2$ with respect to h.

Theorem 2. Let $\psi \in H^{D+1}(\Omega)$ is the exact solution of the equation of (1) and ψ_D^* is obtained by the Algorithm. If $d < D \leq 3d-2 (D > d \geq 5)$, then

$$|\psi - \psi_D^*|_2 \leq Ch^{D-1},$$

i.e, ψ_D^* has the optimal asymptotic accuracy in energy norm $|\cdot|_2$ with respect to h.

Numerical calculations with the proposed two level spline methods are implemented, and then applied to some benchmark problems in the literature. All the numerical calculations here are performaned on a laptop equipped with Intel CPU T7500 and 2.00GB memory.

Example. Consider the 2D Navier-Stokes equation [4] on unique square $\Omega = (0,1)^2$. This example is only used to validate the numerical solver, more practical boundary condition is considered in the third example. In this sense, we choose a proper right-hand term f, such that the exact solution being

$$\psi(x,y) = x^2(x-1)^2y^2(1-y)^2. \tag{8}$$

For this test problem, all requirements of the theory concerning the geometry of the domain and the smoothness of the data are satisfied. Meshes with three different sizes are used in our calculation, which corresponding to meshes with size $h = 1/8, 1/14$ and $1/16$ respectively. They are $\mathcal{S}^5(\triangle_{1/4})$ and $\mathcal{S}^6(\triangle_{1/4})$ and $\mathcal{S}^7(\triangle_{1/4})$ and $\mathcal{S}^8(\triangle_{1/4})$. The cpu-time, number of Newton's iterations(Iter), Degree Of Freedoms (Dof), L^2-error,H^1-error and H^2-error of the stream function for the one-level method with $\nu = 1/1000$ for different spline spaces are tabulated in Table 1. On the low degree spline space level, all nonlinear problems are solved by executing the Newton's iteration repeatedly until both the norm of the difference in successive iterates and the norm of the residual are within a fixed tolerance ϵ. Because of its small size, this nonlinear system takes very little time to solve compared with the larger linear systems. in this calculation, $\epsilon = 5e-10$. The CPU-time, the L^2 errors, the H^1 errors and H^2 errors for the two level method are presented in Table 2. Compare with those in Table 1, the

Table 1. One level method for Example with ($\nu = 1/1000, \triangle_{1/4}$)

Spline space	CPU-time	Dof	Iter	L^2 error	H^1 error	H^2 error
$\mathcal{S}^5(\triangle_{1/4})$	5.65s	672	3	1.39e-06	3.36e-05	1.10e-03
$\mathcal{S}^6(\triangle_{1/4})$	6.97s	896	3	6.67e-08	2.14e-06	8.02e-05
$\mathcal{S}^7(\triangle_{1/4})$	11.58s	1152	5	1.76e-09	6.71e-08	3.28e-06
$\mathcal{S}^8(\triangle_{1/4})$	24.01s	1440	10	1.42e-11	1.64e-10	2.43e-09

Table 2. Example: two level method($\nu = 1/1000$)

	two level	CPU-time	L^2 error	H^1 error	H^2 error
$\lvert\psi_D - \psi\rvert$	$\mathcal{S}^5, \mathcal{S}^7$	5.03s	1.88e-09	6.79e-08	3.28e-06
	$\mathcal{S}^5, \mathcal{S}^8$	6.30s	7.68e-10	1.01e-08	2.40e-07
	$\mathcal{S}^6, \mathcal{S}^8$	7.63s	2.02e-11	1.84e-10	2.53e-09
$\lvert\psi_D^* - \psi\rvert$	$\mathcal{S}^5, \mathcal{S}^7$	6.02s	1.74e-09	6.71e-08	3.28e-06
	$\mathcal{S}^5, \mathcal{S}^8$	7.31s	1.74e-11	1.53e-10	2.27e-09
	$\mathcal{S}^6, \mathcal{S}^8$	11.65s	3.06e-11	2.02e-10	2.23e-09

CPU-time for the two level algorithm is much smaller than the corresponding CPU-time for the one-level method. For example, in $S^8(\triangle_{1/4})$, we save about 65%. We anticipate the savings to increase if the mesh is refined simultaneously.

The advantage of the new two level framework is that high order spline solutions could be obtained by simply solving one or two linearized problems in the high order spline spaces. Furthermore, the modified Newton step is also effective however optional for the whole calculations. In total, the two level spline methods in this paper saves a fairly amount of computational costs compared with traditional nonlinear iterative method. Convergence analysis and numerical results are coincide and show the efficiency of the two level schemes. Based on this efficient solver for the Navier-Stokes equations, applications in the artery graft design problems are under investigation.

Bibliography

1. M. D. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice and Algorithms.* Academic Press , London, 1989.

2. M. J. Lai and P. Wenston. Bivariate splines for fluid flows. *Computers & Fluids*, 33(5):1047–1073, 2004.

3. X. L. Hu and D. F. Han. Quadrature-free spline method for two dimensional Navier-Stokes equations. *Appl. Math. J. Chinese Univ.*, 23(1):31–42, 2008.

4. F. Fairag. Numerical computations of viscous incompressible flow problems using a two-level finite element method. *SIAM J. Sci. Comput.*, 24:1919–1929, 2002.

5. F. Fairag. Two-level finite element method for the stream function formulation of the Navier-Stokes equations. *Comput. Math. Appl.*, 36:117–127, 1998.

6. X. P. Shao, D. F. Han and X. L. Hu. A p-version two level spline method for semi-linear elliptic equations. *J. Comput. Math.*, 30(5):544–554, 2012.

7. X. L. Hu, D. F. Han and M. J. Lai. Bivariate splines of various degrees for numerical solution of partial differential equations. *SIAM J. Sci. Comput.*, 29:1338–1354, 2007.

MINIMUM NORM PARTIAL QUADRATIC EIGENVALUE ASSIGNMENT FOR VIBRATING STRUCTURES USING RECEPTANCE METHOD[1]

Liu H.[*], **He B.-X.**[*], **Chen X.-P.**[**]

[*] *Nanjing University of Aeronautics and Astronautics, Nanjing, China*

[**] *Taizhou University, Taizhou , China*

1. Introduction

Consider the following second-order system by multi-input control

$$\mathbf{M}\ddot{\mathbf{x}}\left(t\right)+\mathbf{C}\dot{\mathbf{x}}\left(t\right)+\mathbf{K}\mathbf{x}\left(t\right)=\mathbf{B}\mathbf{u}\left(t\right), \tag{1}$$

where $\mathbf{M},\mathbf{C},\mathbf{K}\in\mathbb{R}^{n\times n}$ are system matrices, $\mathbf{B}\in\mathbb{R}^{n\times m}$ is the full column rank control matrix and $\mathbf{u}\left(t\right)\in\mathbb{R}^{m}$ is the control vector. The associated open-loop pencil is given by $\mathbf{P}\left(\lambda\right)=\lambda^{2}\mathbf{M}+\lambda\mathbf{C}+\mathbf{K}$.

In general, active control using velocity and displacement state feedback can be used to assign the eigenvalues. Hence, we consider the control vector $\mathbf{u}\left(t\right)$ taking the following form

$$\mathbf{u}\left(t\right)=\mathbf{F}^{\top}\dot{\mathbf{x}}\left(t\right)+\mathbf{G}^{\top}\mathbf{x}\left(t\right), \tag{2}$$

where $\mathbf{F},\mathbf{G}\in\mathbb{R}^{n\times m}$ are state feedback matrices. Then the closed-loop system corresponding to (1) is

$$\mathbf{M}\ddot{\mathbf{x}}\left(t\right)+\left(\mathbf{C}-\mathbf{B}\mathbf{F}^{\top}\right)\dot{\mathbf{x}}\left(t\right)+\left(\mathbf{K}-\mathbf{B}\mathbf{G}^{\top}\right)\mathbf{x}\left(t\right)=\mathbf{0}. \tag{3}$$

Mathematically, the partial quadratic eigenvalue assignment problem is to find the matrices $\mathbf{F},\mathbf{G}\in\mathbb{R}^{n\times m}$ such that a few eigenvalues of the closed-loop pencil

$$\mathbf{P}_{c}\left(\lambda\right)=\lambda^{2}\mathbf{M}+\lambda\left(\mathbf{C}-\mathbf{B}\mathbf{F}^{\top}\right)+\left(\mathbf{K}-\mathbf{B}\mathbf{G}^{\top}\right) \tag{4}$$

are altered as required and the resting eigenpairs remain unchanged, i.e., possessing the no spill-over property.

In this talk, we consider the following problem.

Problem MNPQEAP. Given the open-loop eigenvalues $\{\lambda_k\}_{k=1}^{2n}$ and the corresponding eigenvector set $\{\mathbf{v}_k\}_{k=1}^{2n}$ and a self-conjugate set $\{\mu_k\}_{k=1}^{p}$, find the state feedback matrices $\mathbf{F},\mathbf{G}\in\mathbb{R}^{n\times m}$, where the Frobenius norms of $\mathbf{F},\mathbf{G}$ are minimized, such that the closed-loop pencil (4) has the desired eigenvalues $\{\mu_k\}_{k=1}^{p}$, and the eigenpairs $\{\lambda_k,\mathbf{v}_k\}_{k=p+1}^{2n}$.

[1] Supported by NSFC under grant 11401305, 11571171

2. An Iterative Method for Solving the MNPQEAP

The receptance matrix is

$$\mathbf{H}(s) = \left(s^2\mathbf{M} + s\mathbf{C} + \mathbf{K}\right)^{-1}.$$

Denote

$$\mathbf{w}_k = \mathbf{H}(\mu_k)\left(\mathbf{b}_1\left(\mu_k\mathbf{f}_1^\top + \mathbf{g}_1^\top\right) + \ldots + \mathbf{b}_m\left(\mu_k\mathbf{f}_m^\top + \mathbf{g}_m^\top\right)\right)\mathbf{w}_k. \tag{5}$$

$$\mathbf{r}_{\mu_k,j} = \mathbf{H}(\mu_k)\,\mathbf{b}_j, \tag{6}$$

and

$$\alpha_{\mu_k,j} = \left(\mu_k\mathbf{f}_j^\top + \mathbf{g}_j^\top\right)\mathbf{w}_k, k = 1,\ldots,p, j = 1,\ldots,m. \tag{7}$$

$$\mathbf{W}_k = \begin{bmatrix} \mu_k\mathbf{w}_k^\top & 0 & \ldots & 0 & \mathbf{w}_k^\top & 0 & \ldots & 0 \\ 0 & \mu_k\mathbf{w}_k^\top & \ldots & 0 & 0 & \mathbf{w}_k^\top & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_k\mathbf{w}_k^\top & 0 & 0 & \ldots & \mathbf{w}_k^\top \end{bmatrix},$$

$$\mathbf{V}_l = \begin{bmatrix} \lambda_l\mathbf{v}_l^\top & 0 & \ldots & 0 & \mathbf{v}_l^\top & 0 & \ldots & 0 \\ 0 & \lambda_l\mathbf{v}_l^\top & \ldots & 0 & 0 & \mathbf{v}_l^\top & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_l\mathbf{v}_l^\top & 0 & 0 & \ldots & \mathbf{v}_l^\top \end{bmatrix},$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_m \\ \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_m \end{bmatrix}, \alpha_k = \begin{bmatrix} \alpha_{\mu_k,1} \\ \vdots \\ \alpha_{\mu_k,m} \end{bmatrix}.$$

Then Problem MNPQEAP can be transformed into the following linear systems

$$\mathbf{A}\mathbf{y} = \begin{bmatrix} \alpha \\ \mathbf{0} \end{bmatrix}, \tag{8}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{W}_k \\ \mathbf{V}_l \end{bmatrix} \in \mathbb{C}^{2mn\times 2mn}, k = 1,\ldots,p, l = p+1,\ldots,2n,$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \in \mathbb{C}^{mp}.$$

Obviously, the solutions to the partial quadratic eigenvalue assignment problem are not unique when $m > 1$, hence, we consider the MNPQEAP for multi-input control system (1) via solving the following problem.

$$\min J := \tfrac{1}{2}\left(\|\mathbf{F}\|_F^2 + \|\mathbf{G}\|_F^2\right), \tag{9}$$

where $\|\cdot\|_F$ means the Frobenuis matrix norm.

Denote

$$\mathbf{Y} = [\mathbf{F}, \mathbf{G}],$$

then the objective function is given by $J := \frac{1}{2}\left(\|\mathbf{Y}\|_F^2\right)$.

Let

$$\mathbf{y} = \begin{bmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_m \\ \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_m \end{bmatrix} = \operatorname{vec}(\mathbf{Y}). \tag{10}$$

Considering the linear systems (8), we can rewrite $\mathbf{W}_k = \mathbf{B}_k^\top \otimes \mathbf{A}_k, k = 1, \ldots, p$ and $\mathbf{V}_l = \mathbf{B}_l^\top \otimes \mathbf{A}_l, l = p+1, \ldots, 2n$ by using Kronecker product, where

$$\mathbf{A}_k = \mathbf{w}_k^\top \in \mathbb{C}^{1\times n}, \mathbf{B}_k = \begin{bmatrix} \mu_k & & \\ & \ddots & \\ & & \mu_k \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{C}^{2m\times m}, k = 1, \ldots, p,$$

and

$$\mathbf{A}_l = \mathbf{v}_l^\top \in \mathbb{C}^{1\times n}, \mathbf{B}_l = \begin{bmatrix} \lambda_l & & \\ & \ddots & \\ & & \lambda_l \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{C}^{2m\times m}, l = p+1, \ldots, 2n.$$

Then the coefficient matrix $\mathbf{A}$ of (8) can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{B}_1^\top \otimes \mathbf{A}_1 \\ \vdots \\ \mathbf{B}_{2n}^\top \otimes \mathbf{A}_{2n} \end{bmatrix},$$

therefore the linear systems (8) is equivalent to

$$\begin{bmatrix} \mathbf{B}_1^\top \otimes A_1 \\ \vdots \\ \mathbf{B}_{2n}^\top \otimes \mathbf{A}_{2n} \end{bmatrix} \operatorname{vec}(\mathbf{Y}) = \begin{bmatrix} \operatorname{vec}(\mathbf{D}_1) \\ \vdots \\ \operatorname{vec}(\mathbf{D}_{2n}) \end{bmatrix}, \tag{11}$$

where

$$\mathbf{D}_k = [\alpha_{\mu_k,1}, \ldots, \alpha_{\mu_k,m}] \in \mathbb{C}^{1\times m}, k = 1, \ldots, p,$$

$$\mathbf{D}_l = [0, \ldots, 0] \in \mathbb{C}^{1\times m}, l = p+1, \ldots, 2n.$$

Considering that $\operatorname{vec}(\mathbf{D}_i) = \left(\mathbf{B}_i^\top \otimes \mathbf{A}_i\right) \operatorname{vec}(\mathbf{Y}) = \operatorname{vec}(\mathbf{A}_i\mathbf{Y}\mathbf{B}_i), i = 1, \ldots, 2n$, then we can transform (8) into the following matrix equations

$$\begin{cases} \mathbf{A}_1\mathbf{Y}\mathbf{B}_1 = \mathbf{D}_1, \\ \vdots \\ \mathbf{A}_p\mathbf{Y}\mathbf{B}_p = \mathbf{D}_p, \\ \mathbf{A}_{p+1}\mathbf{Y}\mathbf{B}_{p+1} = \mathbf{D}_{p+1}, \\ \vdots \\ \mathbf{A}_{2n}\mathbf{Y}\mathbf{B}_{2n} = \mathbf{D}_{2n}. \end{cases} \tag{12}$$

Therefore the solutions of the MNPQEAP are equivalent to the minimum Frobenius norm solutions of the matrix equations (12). The MCG algorithm for solving (12) can be listed as follows.

Algorithm 1

```
1. Compute  $\mathbf{Z}_1 = \mathbf{P}_1 = \sum_{k=1}^{2n} \mathbf{A}_k^H (\mathbf{D}_k - \mathbf{A}_k\mathbf{Y}_1\mathbf{B}_k)\mathbf{B}_k^H$;
2. Compute  $\mathbf{R}_1 = \begin{bmatrix} \mathbf{D}_1 - \mathbf{A}_1\mathbf{Y}_1\mathbf{B}_1 \\ \vdots \\ \mathbf{D}_{2n} - \mathbf{A}_{2n}\mathbf{Y}_1\mathbf{B}_{2n} \end{bmatrix}$;
3. Set  $i = 1$;
4. Compute  $\mathbf{Y}_{i+1} = \mathbf{Y}_i + \frac{\|\mathbf{R}_i\|_F^2}{\|\mathbf{P}_i\|_F^2}\mathbf{P}_i$;
5. Compute  $\mathbf{Z}_{i+1} = \sum_{k=1}^{2n} \mathbf{A}_k^H (\mathbf{D}_k - \mathbf{A}_k\mathbf{Y}_{i+1}\mathbf{B}_k)\mathbf{B}_k^H$;
6. Compute  $\mathbf{P}_{i+1} = \mathbf{Z}_{i+1} - \frac{\operatorname{tr}\left(\mathbf{Z}_{i+1}^H\right)}{\|\mathbf{P}_i\|_F^2}\mathbf{P}_i$;
7. Compute  $\mathbf{R}_{i+1} = \begin{bmatrix} \mathbf{D}_1 - \mathbf{A}_1\mathbf{Y}_{i+1}\mathbf{B}_1 \\ \vdots \\ \mathbf{D}_{2n} - \mathbf{A}_{2n}\mathbf{Y}_{i+1}\mathbf{B}_{2n} \end{bmatrix}$;
8. If $\mathbf{R}_{i+1} = \mathbf{0}$, then stop; Otherwise, set $i = i + 1$, go to step 4.
```

Bibliography

1. *Liu H., He B.X., Chen X.P.* Minimum norm partial quadratic eigenvalue assignment for vibrating structures using receptance method // Mechanical Systems and Signal Processing. 2019. Vol. 123. P. 131–142.

2. *Bai Z.J., Lu M., Wan Q.Y.* Minimum norm partial quadratic eigenvalue assignment for vibrating structures using receptances and system matrices// Mechanical Systems and Signal Processing. 2018. Vol. 112. P. 265–279.

3. *Ram Y.M., Mottershead J.E.* Receptance method in active vibration control// AIAA Journal. 2007. Vol. 45. P. 562–567.

THE REGION OF TURING INSTABILITY IN SCHNAKENBERG SYSTEM

Lysenko S.A., Revina S.V.

Southern Federal University, Rostov-on-Don, Russia

The self-organization has inspired scientific study in different fields such as physics, chemistry, biology and ecology. In 1952 Alan Turing published a paper [1] where he proposed a reaction-diffusion model for pattern formation. Turing described diffusion-driven instability. A diffusion-driven instability, or Turing instability, occurs when a steady state of reactions-diffusion system, stable in absence of diffusion, becomes unstable when diffusion is present.

Most of the study of Turing patterns is done on chemical reaction-diffusion systems. Schnakenberg developed his kinematic reaction model [2] as a simple chemical model exibiting limit-cycle behavior. He showed that such a model would need to involve three reactions

$$2U + V \to 3U, \qquad B \to V, \qquad U \rightleftarrows A, \tag{1}$$

Suppose that $u(t)$, $v(t)$ are the concentrations of chemicals U and V respectively, depending on time t; a and b are constant concentrations of chemicals A and B. Hence the system of ordinary differential equations corresponding the reaction (1) takes the form

$$\frac{du}{dt} = u^2v - u + a, \quad \frac{dv}{dt} = -u^2v + b, \tag{2}$$

This is a Schnakenberg system in absence of diffusion.

Next one can introduce a spatial dependence of concentration and take in account the movement of chemicals by adding diffusion. Suppose that concentrations $u = u(x,t)$, $v = v(x,t)$ depend not only on time t but on spatial variable x, too. Let x belong to bounded domain $\Omega \subset R^m$ as $m = 1, 2, 3$. We suppose, that for $m = 2; 3$ the boundary is sufficiently smooth $\partial\Omega \in C^2$, or Ω is rectangle. Let $\Delta = \frac{\partial^2}{\partial x_1^2} + ... + \frac{\partial^2}{\partial x_m^2}$ be Laplace operator. The resulting partial differential system is given by

$$u_t = D_1\Delta u + f(u,v), \quad v_t = D_2\Delta v + g(u,v), \tag{3}$$

here

$$f(u,v) = u^2v - u + a \quad g(u,v) = -u^2v + b. \tag{4}$$

Introducing the change of variables $x_i \to \sqrt{D_1}x_i; \quad i = 1, 2, ..m$ and new notation of diffusion coefficient $d = \frac{D_1}{D_2}$ we rewrite system (3) as

$$u_t = \Delta u + f(u,v), \quad v_t = d\Delta v + g(u,v). \tag{5}$$

Suppose that the flow of matter through the boundary of the region is zero. This means that Neumann boundary conditions are satisfied.

$$\frac{\partial u}{\partial n}|_{\partial\Omega} = \frac{\partial v}{\partial n}|_{\partial\Omega} = 0. \tag{6}$$

Here n is outer normal to boundary.

We call the system (5) subject to boundary conditions (6) with reaction terms f and g describing (4) Schnakenberg system with diffusion or diffusive Schnakenberg system.

Suppose that some boundary conditions are satisfied for the system (5)-(6). If initial conditions don't depend on x, then a solution doesn't depend on x, too. In this sense R^2 is invariant subspace of the system (5)-(6) for every diffusion coefficient.

It is known [3] that the equilibrium of the system (2) has a form

$$(u_0, v_0) = (a+b, \quad \frac{b}{(a+b)^2}). \tag{7}$$

Following conditions are fulfilled

$$a+b > 0, \quad b > 0. \tag{8}$$

Using linearization method one can investigate the stability of the steady state (u_0, v_0). Let J be the Jacobi matrix of ODE system (2) at point (u_0, v_0):

$$J = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix} |_{(u_0,v_0)}. \tag{9}$$

Here

$$\begin{aligned} f_u &= \tfrac{b-a}{a+b}, \quad & f_v &= (a+b)^2, \\ g_u &= -\tfrac{2b}{a+b}, \quad & g_v &= -(a+b)^2. \end{aligned} \tag{10}$$

Then linearized system (2) has a form

$$\frac{dy}{dt} = Jy, \quad y \in R^2. \tag{11}$$

Now let consider diffusive Schnakenberg system. We linearize (5) near equilibrium (u_0, v_0):

$$\tilde{u}_t = \triangle\tilde{u} + f_u\tilde{u} + f_v\tilde{v}, \quad \tilde{v}_t = d\triangle\tilde{v} + g_u\tilde{u} + g_v\tilde{v}, \tag{12}$$

here f_u, f_v, g_u, g_v have been defined at (10). It is known that the spectrum of a linear operator defined by the lefthand side of the system (12) is discrete.

The equilibrium (u_0, v_0) of a system with diffusion (4) – (6) is called Turing unstable if two conditions are met. First, the eigenvalues of the non-diffusion system (11) linearized in the neighborhood of the equilibrium state lie strictly in the left half-plane of the complex plane. Secondly, there is an eigenvalue of a linearized system with diffusion (12), which lies in the right half-plane.

The system (4) — (6) contains the reaction parameters a, b that satisfy the conditions (8) and the diffusion coefficient d. For practical applications, it is important to be able to find the Turing instability region, as well as the critical value of the diffusion parameter.

Area on the parameter plane, containing those parameters for which the Turing instability occurs and the diffusion coefficient is fixed is called the Turing instability region.

Now let the parameters a and b be fixed, and the diffusion coefficient d change. Then the eigenvalues of the linearized system (12) can be considered as functions of the parameter d. They change their position on the complex plane when the parameter d changes. It is easy to show (this will be shown below) that for d less than a certain value, all the eigenvalues of the system (12) lie strictly in the left half-plane of the complex plane.

We will be interested in the critical case when the eigenvalues intersect the imaginary axis. In the case of general position, this is possible either when the imaginary axis is traversed by a pair of purely imaginary eigenvalues (then there is an oscillatory loss of stability), or when the eigenvalue passes through zero (which corresponds to a monotonous loss of stability). It is known that Turing instability refers to monotonous loss of stability [3].

The critical value of the parameter d is the value of d_c for which the spectrum of the linearized problem (12) lies strictly in the left half-plane of the complex plane, except for the eigenvalue $\lambda(d_c) = 0$, and the intersection of the imaginary axis occurs transversally:

$$\lambda'|_{d=d_c} \neq 0, \tag{13}$$

the prime means differentiation by the parameter d.

The main purpose of this work is to find the region of Turing instability in the parameter plane $a; b$, as well as the critical value of the diffusion coefficient. In [4] diffusive Schnakenberg system with time delay is considered. Critical conditions for Turing instability are derived in the plane of diffusion coefficients. Based on Murray's results [3], we propose an original approach of finding the Turing instability region on the plane of reaction parameters.

Find the trace and the determinant of the matrix J and write out the conditions under which the eigenvalues of the system without diffusion (11) lie strictly in the left half-plane of the complex plane:

$$TrJ \equiv \frac{b - a - (a+b)^3}{a+b} < 0, \quad DetJ \equiv (a+b)^2 > 0. \tag{14}$$

From (8) it follows that $Det(J) > 0$ is always. Consequently, the conditions (14) for the diffusionless approximation take the form:

$$b - a < (a+b)^3. \tag{15}$$

We introduce new variables:

$$Y = b - a; \quad X = a + b. \tag{16}$$

The convenience of this replacement is quite obvious. Note that this change of variables is proposed in paper [5]. We have introduced it independently of this work.

Then taking into account (8) the condition (15) takes the form:

$$X > 0; \quad Y < X^3. \tag{17}$$

The convenience of replacing (16) is fairly obvious. This replacement is also used in [4]. We have introduced it independently of this work.

Taking into account the diffusionless approximation (17), the necessary conditions for Turing instability in the variables (X, Y) take the form:

$$X > 0; \quad Y < X^3; \quad Y \geq \frac{1}{d}X^3 + \frac{2}{\sqrt{d}}X^2. \tag{18}$$

Note that (18) implies a constraint on Y (and the parameters a and b)

$$Y = b - a > 0, \tag{19}$$

and also on the diffusion coefficient d

$$d > 1. \tag{20}$$

Let ψ_k be eigenfunction and μ_k be eigenvalue of Laplace operator subject to Neumann boundary conditions, $k = 0, 1, 2...$

$$\Delta\psi_k + \mu_k\psi_k = 0, \quad x \in \Omega, \quad \frac{\partial\psi_k}{\partial n}|_{\partial\Omega} = 0. \tag{21}$$

Let $h(\mu)$ be a polinomial

$$h(\mu_k) \equiv d\mu_k^2 + (X^2 - d \cdot \frac{Y}{X})\mu_k + X^2. \tag{22}$$

The curve $Y = X^3$ on the plane (X, Y), corresponding to the zeroing of the trace of the matrix J, is called the curve of the zero trace.

The curve $Y_0 = Y_0(X)$ on the (X, Y) plane, corresponding to the zeroing of the discriminant of the polynomial $h(\mu)$, will be called the discriminant curve:

$$Y_0 = \frac{1}{d}X^3 + \frac{2}{\sqrt{d}}X^2. \tag{23}$$

Thus, in the half-plane $X > 0$, the region of necessary conditions for Turing instability is bounded by a zero trace curve and a discriminant curve. Using the change of variables (16), we can similarly determine the zero trace curve and the discriminant curve in the plane of the initial parameters of the system (a, b). These definitions have not been encountered in the literature and are introduced here for ease of visualization of the Turing instability region.

Further, we can obtain inequalities describing the Turing instability domains for fixed values of a, b, d, l:

1. The conditions of Turing instability are not satisfied and cannot be satisfied when d is changed:

$$Y \geq X^3 \quad \bigcup \quad Y \leq \mu_1 X. \tag{24}$$

2. The conditions of Turing instability are not satisfied, but can be satisfied when d is changed:

$$Y < X^3 \quad \bigcap \quad Y < \frac{1}{d}X^3 + \frac{2}{\sqrt{d}}X^2. \tag{25}$$

3. The necessary, but not sufficient, Turing instability conditions are satisfied:

$$Y < X^3 \quad \bigcap \quad Y \geq \frac{1}{d}X^3 + \frac{2}{\sqrt{d}}X^2 \quad \bigcap \quad Y < \min_k \left(\frac{\mu_k + 1}{\mu_k \cdot d} \cdot X^3 + \mu_k X\right). \tag{26}$$

4. The sufficient conditions for Turing instability are satisfied:

$$Y < X^3 \quad \bigcap \quad Y \geq \min_k \left(\frac{\mu_k + 1}{\mu_k \cdot d} \cdot X^3 + \mu_k X\right). \tag{27}$$

To visualize the Turing instability region, one of the authors (S.A. Lysenko) has developed a software package.

Bibliography

1. *Turing A.M.* The chemical basis of morphogenesis // Philosofical Transactions of the Royal Society of London. Series B, Biological Sciences, 1952, vol. 237, No 641, pp. 37–72.

2. *Schnakenberg J.* Simple chemical reaction systems with limit cycle behaviour // Journal of Theoretical Biology, 1979, vol. 81, No 3, pp. 389–400.

3. *Murray J. D.* Mathematical biology II: Spatial models and biomedical applications. —- Berlin, Heidelberg: Springer-Verlag, 1993.

4. *Jiang W., Wang H., Cao X.* Turing instability and Turing-Hopf bifurcation in diffusive Schnakenberg systems with gene expression time delay // arXiv: 1803.00164v1[math.DS] 1 Mar 2018.

5. *Liu P., Shi J., Wang Y., Feng X.* Bifurcation analysis of reaction-diffusion Schnakenberg model // Journal of Mathematical Chemistry, 2013. Vol. 51, pp. 2001-2019.

THE GENERALIZED HSS METHOD WITH A FLEXIBLE SHIFT-PARAMETER[1]

Meng G.-Y.*, Wen R.-P.**

* *Xinzhou Teachers University, Xinzhou, Shanxi, China*

** *Taiyuan Normal University, Taiyuan, Shanxi, China*

To solve a large sparse non-Hermitian and positive definite system of linear equations

$$Ax = b. \tag{1}$$

Bai, Golub and Ng first proposed the efficient Hermitian and skew-Hermitian splitting (HSS) iteration method [4] with a fixed parameter in 2003.

$$\begin{cases} (\alpha I + H)x_{k+\frac{1}{2}} = (\alpha I - S)x_k + b, \\ (\alpha I + S)x_{k+1} = (\alpha I - H)x_{k+\frac{1}{2}} + b, \end{cases} \tag{2}$$

where I is the identity matrix and α is a fixed shift-parameter.

Note that the HSS iteration (2) may also be considered as a splitting iteration induced from the splitting of the matrix A as follows,

$$A = M(\alpha) - N(\alpha),$$

where

$$M(\alpha) = \frac{1}{2\alpha}(\alpha I + H)(\alpha I + S) \quad \text{and} \quad N(\alpha) = \frac{1}{2\alpha}(\alpha I - H)(\alpha I - S). \tag{3}$$

It was proved that the HSS iteration method converges unconditionally to the unique solution of the linear systems (1). The optimal shift-parameter is estimated as

$$\alpha_{\text{opt}} = \arg\min_{\alpha} \left\{ \max_{\lambda_{\min} \le \lambda \le \lambda_{\max}} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| \right\} = \sqrt{\lambda_{\min}\lambda_{\max}},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and the maximum eigenvalues of the matrix H, respectively.

Because of its outstanding performance and elegant mathematical properties, the HSS iteration method obtains widespread attention; see [1, 2, 6, 8] and the references therein. It is noticed that the parameter α plays an important role in these HSS class iteration methods. We can see the extremal eigenvalues and determinants of some matrices are required and which may greatly decrease the computing efficiency of the HSS iteration methods. In order to compute the

[1] Supported by the National Natural Science Foundation of China (No.11371275),the National Natural Science Foundation of Shanxi Province, China (No.2014011019-3)

optimal parameter, Bai, Golub and Li [3] used the positive real roots of the equation

$$(\alpha^2+q^2)^2(\alpha^2-\lambda_{\max}^2)(\alpha^2-\lambda_{\min}^2)=(\alpha^2-q^2)^2(\alpha^2-\lambda_{\min}\lambda_{\max}),$$

or

$$(\alpha^2+q^2)^2(\lambda_{\max}^2-\alpha^2)(\alpha^2-\lambda_{\min}^2)=(\alpha^2-q^2)^2(\alpha^2-\lambda_{\min}\lambda_{\max})$$

to estimate the optimal shift-parameter $\alpha_{opt}>0$ satisfying

$$\rho(M(\alpha_{\text{opt}}))=\min\{\rho(M(\alpha))|\alpha>0\}.$$

While Huang [7] chosen the optimal parameters using a cubic polynomial equation, which comes from the minimization of the F-norm. Chen [5] used Euclidean norm to estimate the optimal parameter for the HSS iteration method. Wen [10] presented the Quasi-Chebyshev accelerated iteration methods(QCA), which utilize the optimization models to determine the optimal parameters in the each iteration, as follows,

Let $A=M-N, \overline{x}_{k+1}=M^{-1}Nx_k+M^{-1}b$. Solve the following system of linear equations

$$x_{k+1}=\omega_{k+1}(\overline{x}_{k+1}-x_{k-1})+x_{k-1},$$

where ω_{k+1} is the solution of the following optimization problems:

- when A is a symmetric positive definite matrix, set $x=\omega(\overline{x}_{k+1}-x_{k-1})+x_{k-1}$,
$$\min_{\omega}\frac{1}{2}x^TAx-x^Tb;$$
- when A is not symmetric positive definite, set $r=Ax-b$,
$$\min_{\omega} r^T(\alpha I+H)^{-2}r.$$

Pearcy [9] has designed the device of changing iteration shift-parameters on the half-step

$$\begin{cases} x_{k+\frac{1}{2}}=-\left(H+\alpha_{k+\frac{1}{2}}D\right)^{-1}\left[(V-\alpha_{k+\frac{1}{2}}D)x_k-b\right], \\ x_{k+1}=-(V+\alpha_{k+1}D)^{-1}\left[(H-\alpha_{k+1}D)x_{k+\frac{1}{2}}-b\right], \end{cases}$$

where D is a positive definite nomalizing matrix, H, V are positive definite with $A=H+V$, $\alpha_{\frac{1}{2}}\geq\alpha_1\geq\alpha_{1+\frac{1}{2}}\geq\cdots\geq\alpha_{t-\frac{1}{2}}\geq\alpha_t$, and $\alpha_k=\alpha_{k(\operatorname{mod} t)}$ for $k>t$, but the convergence of this alternating direction iteration (ADI) method depended on the positive definite of the splitting matrices H and V.

Motivated by the QCA method and ADI method, shift-parameters $\alpha_k, k=1,2,\cdots$ are constructed by the minimization of residuals.

Method 1(MWZ$_1$−HSS)

Step 1. Compute $r_k = b - Ax_k$.

Step 2. Solve the following system of linear equations:

$$\begin{cases} (\alpha_{k+1}I + H)x_{k+\frac{1}{2}} = (\alpha_{k+1}I - S)x_k + b, \\ (\alpha_{k+1}I + S)x_{k+1} = (\alpha_{k+1}I - H)x_{k+\frac{1}{2}} + b, \end{cases} \tag{4}$$

where α_{k+1} is the solution of the following optimization problem

$$\min_{\alpha} r_{k+1}^{*} \left(\alpha I - H\right)^{-2} r_{k+1}, \tag{5}$$

here, $r_{k+1} = N(\alpha)M(\alpha)^{-1}r_k$.

Step 3. If $\|r_{k+1}\|_2 \leq \varepsilon$, stop; otherwise, $k \Leftarrow k+1$ and go to Step 1.

By the optimization model (5), it can be deduced

$$\frac{d}{d\alpha}\left(\|(\alpha I - H)^{-1}r_{k+1}\|_2^2\right) = -2r_k^{*}(\alpha I + H)^{-3}r_k. \tag{6}$$

It is worthy to note that the solution of the minimization model (5) is equivalent to compute the root

$$\frac{d}{d\alpha}\left(\|(\alpha I - H)^{-1}r_{k+1}\|_2^2\right) = 0.$$

However, the computational formula is far away from actual applications, since the computational cost of the matrix $(\alpha I + H)^{-3}$ is expensive. Instead, we approximate the root of

$$f(\alpha) = \|(\alpha I - H)^{-1}r_{k+1}\|_2^2 = 0$$

by the Newton method. An alternative procedure might be to approximate the value of

$$f(\alpha) = \|(\alpha I + H)^{-1}r_k\|_2^2 = 0$$

by Lemma. Hence, Method 1 could be rewritten into a practical form.

Method 2 (MWZ$_2$−HSS)

Step 1. Compute $r_k = b - Ax_k$.

Step 2. Solve the systems of linear equations:

$$\begin{cases} (\alpha_{k+1}I + H)x_{k+\frac{1}{2}} = (\alpha_{k+1}I - S)x_k + b, \\ (\alpha_{k+1}I + S)x_{k+1} = (\alpha_{k+1}I - H)x_{k+\frac{1}{2}} + b, \end{cases}$$

where α_{k+1} is the root of the equation

$$f(\alpha) = \|(\alpha I + H)^{-1}r_k\|_2^2 = 0.$$

Step 3. If $\|r_{k+1}\|_2 \leq \varepsilon$, stop; otherwise, $k \Leftarrow k+1$ and go to Step 1.

Example Consider the two-dimensional convection-diffusion equation

$$-(u_{xx}+u_{yy})+\beta(u_x+u_y)=g(x,y),$$

on the unit square $(0,1)\times(0,1)$ and subject to Dirichlet-type boundary condition.

Table 1. Iteration steps and CPU times (m=32).

Method		β					
		50	100	500	1000	5000	10000
BGN-HSS	IT	45	46	56	74	155	216
	CPU	0.20	0.22	0.25	0.32	0.67	0.94
H-HSS	IT	66	35	50	70	163	252
	CPU	0.29	0.15	0.22	0.31	0.71	1.10
MWZ$_1$-HSS	IT	32	41	58	70	113	115
	CPU	0.81	1.03	1.44	1.73	2.80	2.84
MWZ$_2$-HSS	IT	38	43	56	62	83	92
	CPU	0.19	0.21	0.27	0.30	0.41	0.45

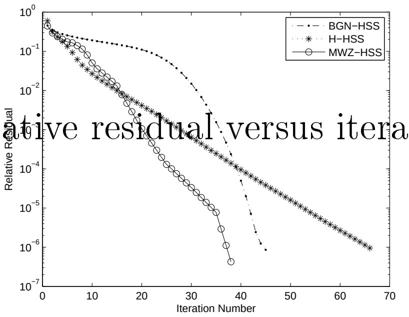


Figure 1. Relative residual versus iteration number with $\beta = 50$.

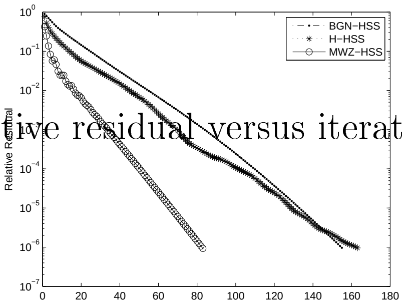


Figure 2. Relative residual versus iteration number with $\beta = 5000$.

From Table 1, it can be seen that for any different β, the numbers of iteration steps of MWZ$_1$-HSS and MWZ$_2$-HSS methods are less than that of BGN-HSS and H-HSS methods at almost the same CPU times. The case $\beta = 50$ and $\beta = 5000$ of these observations can be further illustrated by the iteration pictures plotted in Figure 1 and Figure 2. Clearly, the convergence of MWZ$_2$-HSS method is better than that of BGN-HSS and H-HSS methods.

Bibliography

1. *Bai Z.-Z.* Block alternating splitting implicit iteration methods for saddle-point problems from time-harmonic eddy current models // Numer. Linear Algebra Appl. 2012. Vol. 19. P. 914–936.

2. *Bai Z.-Z., Golub G. H.* Accelerated Hermitian and skew-Hermitian splitting iteration methods for saddle-point problems // SIAM J. Numer. Anal. 2007. Vol. 27. P. 1–23.

3. *Bai Z.-Z., Golub G. H., Li C.-K.* Optimal parameter in Hermitian and skew-Hermitian splitting method for certain two-by-two block matrices // SIAM J. Sci. Comput. 2006. Vol. 28. P. 583–603.

4. *Bai Z.-Z., Golub G. H., Ng M. K.* Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems // SIAM J. Matrix Anal. Appl. 2003. Vol. 24. P. 603–626.

5. *Chen F.* On choices of iteration parameter in HSS method // Appl. Math. Comput. 2015. Vol. 271. P. 832–837.

6. *Guo X.-X., Wang S.* Modified HSS iteration methods for a class of non-Hermitian positive-definite linear systems // Appl. Math. Comput. 2012. Vol. 218. P. 10122–10128.

7. *Huang Y.-M.* A practical formula for computing optimal parameters in the HSS iteration methods // J. Comput. Appl. Math. 2014. Vol. 255. P. 142–149.

8. *Li W., Liu Y.-P., Peng X.-F.* The generalized HSS method for solving singular linear systems // J. Comput. Appl. Math. 2012. Vol. 236. P. 2338–2353.

9. *Pearcy C.* On convergence of alternating direction procedures // Numer. Math. 1962. Vol. 4. P. 172–176.

10. *Wen R.-P.,Wang C.-L., Meng G.-Y.* Quasi-Chebyshev accelerated iteration methods based on optimization for linear systems // Comput. Math. Appl. 2013. Vol. 66. P. 934–942.

COMPUTATIONAL EXPERIMENTS BASED ON REDUCED MODEL OF A LONG AND SHALLOW STREAM[1]

Nadolin K.A.*, Zhilyaev I.V.**

* *Southern Federal University, Rostov-on-Don, Russia*

** *Southern Scientific Center of Russian Academy of Sciences, Rostov-on-Don, Russia*

The main goal of the study is a validation of a simplified 3D mathematical model for passive admixture spreading in shallow flows. The tested model is oriented to the hydrological and ecological problems, and it can be applied to natural streams like rivers and channels. The earlier proposed model of the elongated, shallow and weakly curved stream [5] takes into account the structure of a stream-bed for evaluation of flow velocity in every point of domain. This is a model advantage, which allows calculation of the admixture spreading in a channel with varying width and depth more accurately than by using in-depth averaged models. For example, we can observe the opposite flow in a near-surface zone, which may be caused e.g. by the wind. The results of numerical experiments show that this reduced 3D model adequately describes the admixture spreading processes in natural streams with acceptable accuracy.

Introduction

Mathematical models of various types are used for evaluation of the hydrological characteristics of streams and for simulation of the admixture spreading [2, 3, 8]. The most accurate are three-dimensional models, which are based on full equations of turbulent flow. However, the high accuracy of these simulations cannot be obtained in practice because the data of the real hydrological measurements are not precise enough and no initial and boundary conditions for 3D partial differential equations are available. In addition, the complexity and computational costs of numerical experiments with 3D mathematical models are increased due to the geometry of the model domain, which is extremely elongated along the flow direction. Natural water flows have significant difference in size of they length, width, and depth. The ratio between the average depth and width for the typical lowland river varies from 1:10 to 1:200.

The main aim of this work is to validate the simplified mathematical model for spreading process in natural streams.

In [5], the simplified equations for channel flow hydrodynamics and mass transfer is proposed. The hydrodynamical part of this reduced 3D mathematical model was studied in [6].

[1] This work was supported by the Vladimir Potanin Foundation, project ID GK190000844.

This article focuses on testing the model by comparing the data of hydrological experiment, published in [1] and the numerical results obtained on the base of the model. The computer simulations were performed by finite-element software COMSOL©[7].

I Problem Statement

Let us consider a relatively slow stream in a non-deformable rigid bed $z = h(x, y)$. The channel flow is shallow, elongated, and weakly curved. In a mathematical sense, the *'shallow and elongated'* assumption means that the stream bed geometry has the ratio $D : W : L \approx \epsilon \ll 1$. Here D is the average depth, W is the average width and L is the length of the section of the stream under consideration; $\epsilon \ll 1$ – a value that was used in [5] as a small parameter. Also, *'weakly curved'* means that $\partial h \partial y \sim \epsilon$ and $\partial h \partial x \sim \epsilon^2$.

Let us introduce Cartesian coordinates such that the plane (xy) is located on the flow surface and z-axis is directed toward the bottom. We assume that the x-axis is directed along the flow, and the y-axis is perpendicular to x and directed from the left to the right bank. The origin lies in the inlet section at equal distances from the banks. The equations of the 3D reduced mathematical model for the passive admixture transport in *shallow, elongated and weakly curved stream* in dimensionless variables can be written as

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} + w\frac{\partial c}{\partial z} = \frac{\partial}{\partial z}\left(d\frac{\partial c}{\partial z}\right) - \lambda c, \tag{1}$$

$$c\Big|_{t=0} = c^0, \qquad \frac{\partial c}{\partial x}\bigg|_{x=0} = \pi_0, \qquad \frac{\partial c}{\partial z}\bigg|_{z=h} = \frac{\partial c}{\partial z}\bigg|_{z=\xi} = 0, \tag{2}$$

$$\frac{\partial}{\partial z}\left(\nu\frac{\partial u}{\partial z}\right) = -\mathit{Re}GI, \qquad u\Big|_{z=h} = 0, \qquad \frac{\partial u}{\partial z}\bigg|_{z=\xi} = 0, \tag{3}$$

$$\frac{\partial p}{\partial z} = G, \qquad p\Big|_{z=\xi} = 0, \tag{4}$$

$$\frac{\partial}{\partial z}\left(\nu\frac{\partial v}{\partial z}\right) = \mathit{Re}\frac{\partial p}{\partial y}, \qquad v\Big|_{z=h} = 0, \qquad \frac{\partial v}{\partial z}\bigg|_{z=\xi} = 0, \tag{5}$$

$$\frac{\partial w}{\partial z} = -\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right), \qquad w\Big|_{z=h} = 0, \tag{6}$$

$$\frac{\partial \xi}{\partial t} + u\Big|_{z=\xi}\frac{\partial \xi}{\partial x} + v\Big|_{z=\xi}\frac{\partial \xi}{\partial y} - w\Big|_{z=\xi} = 0. \tag{7}$$

Here c is the concentration of admixture; u, v and w are the components of a velocity vector along the longitudinal (x), transversal (y) and vertical (z) directions, respectively. The known function $h(x, y)$ describes the shape of the stream-bed and the unknown function $\xi(t, x, y)$ describes a slightly deformable free surface of the flow. The known functions $c^0(x, y, z)$ and $\pi_0(t, y, z)$ set the initial distribution of concentration and its inflow through the inlet, respectively.

Equations (1)-(7) contain a set of parameters: d – the dimensionless coefficient of the turbulent diffusion in z- direction; λ – the decay factor for the admixture; ν – the normalized viscosity, which allows taking into account changes in the viscosity of the turbulent flow in accordance with the Boussinesq turbulence hypothesis; Re is the Reynolds number; G is the gravity parameter and I is the slope of the flow.

For more details about derivation of the equations (1)-(6) see [5].

II Solution of the Hydrodynamics System

Equations (1),(2) form the concentration part, and equations (3)-(7) form the hydrodynamics part of the model for the shallow and elongated stream. These subsystems are consistent according to the precision of the approximation [5].

The hydrodynamic part does not depend on the concentration part and its solution can be explicitly written as

$$p = G(z - \xi), \qquad u = ReGI\left(J_2 - \xi J_1\right), \tag{8}$$

$$v = ReG\tfrac{\partial \xi}{\partial y}\left(J_2 - \xi J_1\right), \tag{9}$$

$$w = ReG\left(I\tfrac{\partial}{\partial x}\left(J_4 - \xi J_3\right) + \tfrac{\partial}{\partial y}\left(\left(J_4 - \xi J_3\right)\tfrac{\partial \xi}{\partial y}\right)\right). \tag{10}$$

Here we introduced the notations

$$J_1 = \int\limits_z^{h(x,y)} \frac{d\tau}{\nu}, \quad J_2 = \int\limits_z^{h(x,y)} \frac{\tau d\tau}{\nu}, \quad J_3 = \int\limits_z^{h(x,y)} J_1 d\tau, \quad J_4 = \int\limits_z^{h(x,y)} J_2 d\tau. \tag{11}$$

The pressure and velocity components in (8) are expressed in terms of the free surface function ξ, which is determined from the kinematic boundary condition (7).

Combination of (8) and (7) allows performing the kinematic boundary condition (7) in the following form

$$\begin{aligned}\tfrac{\partial \xi}{\partial t} = ReG\Big[I\left(\tfrac{\partial}{\partial x}\left(J_4 - \xi J_3\right) - \left(J_2 - \xi J_1\right)\tfrac{\partial \xi}{\partial x}\right) + \\ + \left(J_4 - \xi J_3\right)\tfrac{\partial^2 \xi}{\partial y^2} + \tfrac{\partial \xi}{\partial y}\left(\tfrac{\partial}{\partial y}\left(J_4 - \xi J_3\right) - \left(J_2 - \xi J_1\right)\tfrac{\partial \xi}{\partial y}\right)\Big].\end{aligned} \tag{12}$$

where functions (11) and their derivatives are calculated within $z = \xi$ (i.e. on the free surface).

Equation (12) was solved with the finite-element software COMSOL©[7]. For detailed description of these numerical experiments, see [6].

III Testing the Model

To verify the proposed model, we used the data which was published in [1], where the transfer of an admixture in the Severn River was studied. The section of river under study flows through the territory of Wales (Great Britain) between the settlements of Llanidloes and Caersws. In article [1], the observations of British hydrologists, who studied the distribution of tracer – coloring matter, were published. The diffusion coefficient of this substance is 10^{-6} cm/s.

The goal of that experiment was to collect and publish data of diffusive transfer of admixture for testing mathematical models proposed by various authors. The concentration of admixture was monitored in a section of a river about 14 kilometers long by 6 observation stations located downstream. British authors describe in detail the geometry of the river-bed and the flow velocity in the considered section of the river, as well as other hydrological characteristics of the water stream obtained as a result of measurements that lasted more than 10 hours.

The width of the channel on the considered area was measured at 86 points and varies from 13 to 48 meters with an average value of 20 meters. The depth of the flow was measured in each of the 86 sections with an interval of 1 meter. (The average depth was 0.6 meters.)

Considering that the average distance between measuring stations is 2 kilometers, the approximate value of the parameter ϵ is 0.01, which satisfies the requirements of the mathematical model (1)-(7).

Thus, the British authors provided data required for the mathematical modeling and performing computational experiments to calculate the mass transfer of passive admixture in a natural water flow using proposed reduced 3D mathematical model (1)-(7).

Figure 1 shows the reconstructed flow region. The reconstruction of the river-bed geometry was made on the base of data presented in [1] for the section of the river Severn between stations A and F.

Figure 2 shows the values of concentration at different times. The solid line corresponds to the concentration of a substance calculated using a reduced 3D model of a long, shallow and slightly curved flow (1)-(7). Circles on the graph depict the results of measurements of the concentration of a substance at times when the admixture flows through the cross-sections of the stream near the observational stations.

Conclusions

The simulation of the passive admixture spreading in channel flows based on complete 3D hydrodynamic and mass transfer equations system is very computationally expensive. Therefore, mathematical models, which give a simplified but adequate description of the process, could be implemented. Such models

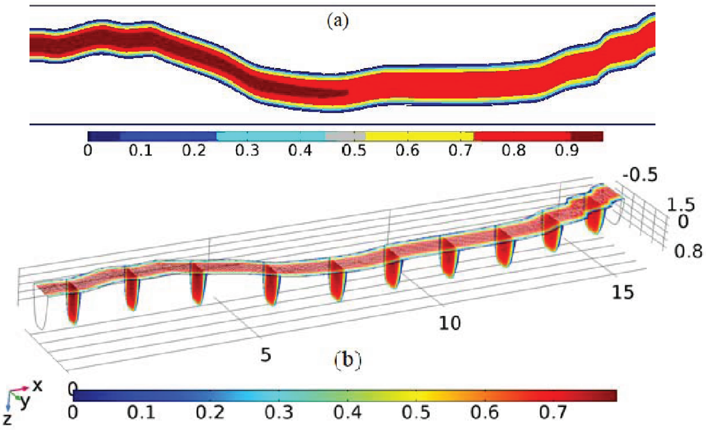


Figure 1. Reconstructed stream-bed function $h(x, y)$ and the velocity field: (a) – the horizontal plane view (depth difference is colored according to presented scale); (b) – a set of segments with cross-sections and colored velocity field.

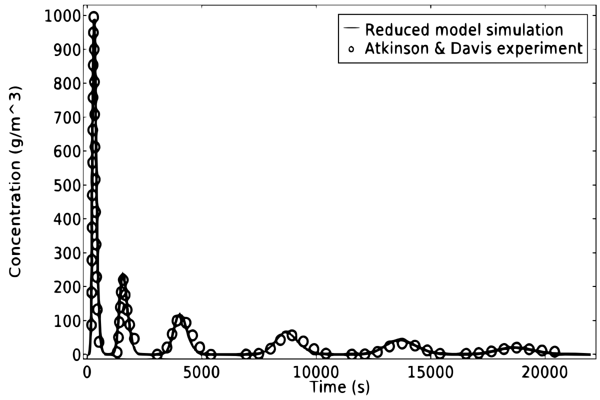


Figure 2. The concentration at times when admixture passes each of the six measurement stations (A-F).

should consider the key features of natural streams. The equations of a shallow, elongated and weakly curved stream (1)-(7) describe the flow dynamics as a three-dimensional, however, they are much simpler than the full 3D equations.

The proposed mathematical model of a long shallow and weakly curved flow can be applied only for simulation of slow flows, which can be described by steady-state model equations.

The results of the numerical simulation that are given in this article show that the proposed reduced 3D model of a long shallow flow adequately describes its hydrodynamics and mass transfer of the passive admixture. It can be used to simulate the spreading of pollutants in such streams.

The research is partly supported by a grant from the Vladimir Potanin Charity Fund. The results of the study will be used in the Master Program "Mathematical Modeling, Numerical Methods and Software".

Bibliography

1. Davis P., Atkinson T., Wigley T.: Longitudinal dispersion in natural channels: 1. Experimental results from the River Severn, U.K. Hydrol. Earth Syst. Sci. **4**:3, 345Ц353 (1999)

2. Knight, D.W.: River hydraulics — a view from midstream. J. Hydr. Res. **51**:1, 2–18 (2013)

3. Luk, G.K.Y., Lau, Y.L. Watt, W.E.: Two-dimensional mixing in rivers with unsteady pollutant source. J. Env. Eng. **116**, 125–143 (1990)

4. Monin, A.S., Yaglom, A.M.: Statistical fluid mechanics. Cambridge, MIT Press (1979).

5. Nadolin, K.A.: An approach to simulating passive mass transport in channel flows. Mat. Model. **21**:2, 4–28 (2009)

6. Nadolin, K.A., Zhilyaev, I.V.: A Reduced 3D Hydrodynamic Model of a Shallow, Long, and Weakly Curved Stream, Water Res. **44**:2, 237–245 (2017)

7. Pryor R.W.: Multiphysics Modeling Using COMSOL: A First Principle Approach. // Jones & Bartlett Publisers, Sudbury, MA (2011)

8. Stansby, P.K.: Coastal hydrodynamics — present and future. J. Hydr. Res. **51**:4, 341-350 (2013)

NUMERICAL STUDY OF EFFECTIVE MODULI OF POROUS PIEZOCOMPOSITES WITH METALLIZED INTERFACES IN UNIT CELL WITH CUBIC OR SPHERICAL PORES[1]

Nasedkin A.V.[*], **Nassar M.E.**[*,**]

[*] *Southern Federal University, Rostov-on-Don, Russia*

[**] *Menoufia University, Menouf, Egypt*

This paper deals with the modeling problem of finding the effective characteristics of microporous piezoceramic materials with infinitely thin metallization of pore surfaces [1] based on the effective moduli method of mechanics of composites and the finite element method [2, 3]. In contrast to [2, 3], one cell of a porous material in the form of a cube with one pore located in the center is considered. The influence of the pore surface metallization and of the pore shape on the values of the effective moduli for porous piezoceramic PZT-4 is investigated numerically.

I Homogenization Problem

Let Ω be a unit cubic cell of piezoelectric material with one pore of cubic or spherical form; a is the cubic cell side; $\Omega = \Omega^m \cup \Omega^p$; Ω^m is the part of Ω with main piezoelectric material or matrix; Ω^p is the pore; $\Gamma = \partial\Omega$ is the external boundary of the cell; $\Gamma^p = \partial\Omega^p$ is the boundary of the pore; $\mathbf{n}$ is the unit normal vector external with respect to the volume of the main piezoelectric material Ω^m.

We assume that in the Cartesian coordinate system $Ox_1x_2x_3$ the unit cell Ω occupies the region $|x_k| \leq a/2$, $k = 1, 2, 3$. Then, in the case of a cubic pore with side b $(b < a)$, the domain Ω^p will be defined by the inequalities $|x_k| \leq b/2$, $k = 1, 2, 3$, and in the case of a spherical pore with radius $R < a/2$, the domain Ω^p is given by the inequality $|\mathbf{x}| \leq R$.

In accordance with the effective moduli method, we will consider the system of differential equations of the static theory of electroelasticity (piezoelectricity) in the volume Ω

$$\mathbf{L}^*(\nabla) \cdot \mathbf{T} = 0, \quad \nabla \cdot \mathbf{D} = 0, \quad \mathbf{T} = \mathbf{c}^E \cdot \mathbf{S} - \mathbf{e}^* \cdot \mathbf{E}, \quad \mathbf{D} = \mathbf{e} \cdot \mathbf{S} + \boldsymbol{\epsilon}^S \cdot \mathbf{E}, \quad (1)$$

$$\mathbf{S} = \mathbf{L}(\nabla) \cdot \mathbf{u}, \quad \mathbf{E} = -\nabla\varphi, \quad \mathbf{L}^*(\nabla) = \begin{bmatrix} \partial_1 & 0 & 0 & 0 & \partial_3 & \partial_2 \\ 0 & \partial_2 & 0 & \partial_3 & 0 & \partial_1 \\ 0 & 0 & \partial_3 & \partial_2 & \partial_1 & 0 \end{bmatrix}, \quad (2)$$

where $\mathbf{T} = \{\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{23}, \sigma_{13}, \sigma_{12}\}$ is the array of stress components σ_{ij}; $\mathbf{S} = \{\varepsilon_{11},\ \varepsilon_{22},\ \varepsilon_{33},\ 2\varepsilon_{23},\ 2\varepsilon_{13},\ 2\varepsilon_{12}\}$ is the array of strain components ε_{ij}; $\mathbf{D}$

[1] Supported by grant 16-01-00785 of the Russian Foundation for Basic Research.

is the electric flux density vector; $\mathbf{E}$ is the electric field vector; $\mathbf{u} = \mathbf{u}(\mathbf{x})$ is the vector function of displacements; $\varphi = \varphi(\mathbf{x})$ is the scalar function of electric potential; $\mathbf{c}^E$ is the 6×6 matrix of elastic stiffness moduli $c^E_{\alpha\beta}$, measured at constant electric field (E); $\mathbf{e}$ is the 3×6 matrix of piezomoduli $e_{i\alpha}$; $\boldsymbol{\epsilon}^S$ is the 3×3 matrix of dielectric permittivities ϵ^S_{ij}, measured at constant strains (S); $(...)^*$ is the operation of transposition for matrices and vectors; $\mathbf{c}^E = \mathbf{c}^{E\,m}$, $\mathbf{e} = \mathbf{e}^m$, $\boldsymbol{\epsilon}^S = \boldsymbol{\epsilon}^{S\,m}$ for $\mathbf{x} \in \Omega^m$; $\mathbf{c}^E = \mathbf{c}^{E\,p}$, $\mathbf{e} = \mathbf{e}^p$, $\boldsymbol{\epsilon}^S = \boldsymbol{\epsilon}^{S\,p}$ for $\mathbf{x} \in \Omega^p$.

At the external boundary Γ of the volume Ω, we will take the conventional boundary conditions of the effective moduli method

$$\mathbf{u} = \mathbf{L}^*(\mathbf{x}) \cdot \mathbf{S}_0, \quad \varphi = -\mathbf{x} \cdot \mathbf{E}_0, \quad \mathbf{x} \in \Gamma, \tag{3}$$

where $\mathbf{S}_0$ and $\mathbf{E}_0$ are the arrays with constant values of size 6 and 3, respectively.

For ordinary porous piezoceramics at the interfaces of pores and matrix the following boundary condition should be satisfied

$$\mathbf{n} \cdot \mathbf{D} = 0, \quad \mathbf{x} \in \Gamma^p. \tag{4}$$

However, in the problem with infinitely thin metallizaton of pore surface, the boundaries Γ^p should be considered as equipotential surface, and so instead of (4) it is necessary to use the conditions

$$\varphi = \Phi^p, \quad \mathbf{x} \in \Gamma^p, \quad \int_{\Gamma^p} \mathbf{n} \cdot \mathbf{D}\, d\Gamma = 0, \tag{5}$$

where Φ^p is an unknown value of potential on the surface Γ^p.

Hence, the transition from the problem with ordinary porosity to the problem with the pore surface metallization formally consists in the change of the boundary condition (4) for nonelectrodized surface to the boundary conditions (5). Note that the boundary conditions (5) on Γ^p are usually used for the free electrode surfaces.

In the case of porous piezoceramics of $6mm$ class, in order to find its ten independent effective constants ($c^{E\,\mathrm{eff}}_{11}$, $c^{E\,\mathrm{eff}}_{12}$, $c^{E\,\mathrm{eff}}_{13}$, $c^{E\,\mathrm{eff}}_{33}$, $c^{E\,\mathrm{eff}}_{44}$, e^{eff}_{31}, e^{eff}_{33}, e^{eff}_{15}, $\epsilon^{S\,\mathrm{eff}}_{11}$, $\epsilon^{S\,\mathrm{eff}}_{33}$), it is enough to solve five static problems (1)–(4) or (1)–(3), (5) with various values of $\mathbf{S}_0$ and $\mathbf{E}_0$, having set one of the component $S_{0\beta}$, E_{0k} ($\beta = 1, 2, ..., 6$; $k = 1, 2, 3$) in the boundary conditions (3) not equal to zero ($\langle(...)\rangle = 1/|\Omega| \int_\Omega (...)\, d\Omega$):

– problem I

$$S_\beta = \varepsilon_0 \delta_{1\beta},\ \mathbf{E}_0 = 0\ \Rightarrow\ c^{E\,\mathrm{eff}}_{1j} = \langle\sigma_{jj}\rangle/\varepsilon_0,\ j = 1, 2, 3,\ e^{\mathrm{eff}}_{31} = \langle D_3\rangle/\varepsilon_0\,, \tag{6}$$

– problem II

$$S_\beta = \varepsilon_0 \delta_{3\beta},\ \mathbf{E}_0 = 0\ \Rightarrow\ c^{E\,\mathrm{eff}}_{j3} = \langle\sigma_{jj}\rangle/\varepsilon_0,\ j = 1, 2, 3,\ e^{\mathrm{eff}}_{33} = \langle D_3\rangle/\varepsilon_0\,, \tag{7}$$

– problem III

$$S_\beta = \varepsilon_0 \delta_{4\beta},\ \mathbf{E}_0 = 0\ \Rightarrow\ c^{E\,\mathrm{eff}}_{44} = \langle\sigma_{23}\rangle/\varepsilon_0,\ e^{\mathrm{eff}}_{15} = \langle D_2\rangle/\varepsilon_0\,, \tag{8}$$

– problem IV

$$\mathbf{S}_0 = 0,\ E_k = E_0\delta_{1k}\ \Rightarrow\ e_{15}^{\mathrm{eff}} = -\langle\sigma_{13}\rangle/E_0,\ \epsilon_{11}^{S\,\mathrm{eff}} = \langle D_1\rangle/E_0\,, \tag{9}$$

– problem V

$$\mathbf{S}_0 = 0,\ E_k = E_0\delta_{3k}\ \Rightarrow\ e_{3j}^{\mathrm{eff}} = -\langle\sigma_{jj}\rangle/E_0,\ j = 1,3,\ \epsilon_{33}^{S\,\mathrm{eff}} = \langle D_3\rangle/E_0\,. \tag{10}$$

II Finite Element Results and Discussion

The boundary problems (1)–(4) or (1)–(3), (5) with (6)–(10) were solved numerically in the ANSYS finite element package. The 10-node tetrahedral elements SOLID227 were used with option of piezoelectric analysis. The finite element mesh was created in ANSYS with a limit on the maximum edge length of elements equal to $\tilde{a}/8$, where $\tilde{a} = 1$ is the dimensionless cell edge length. Finite elements inherit the material properties of the main piezoelectric material and the pore associated with the volumes Ω^m and Ω^p.

The calculations of the effective moduli were performed for porous piezoceramics PZT-4 with the following values of material constants of the material matrix: $c_{11}^{E\,m} = 13.9\cdot 10^{10}$, $c_{12}^{E\,m} = 7.78\cdot 10^{10}$, $c_{13}^{E\,m} = 7.74\cdot 10^{10}$, $c_{33}^{E\,m} = 11.5\cdot 10^{10}$, $c_{44}^{E\,m} = 2.56\cdot 10^{10}$ (N/m^2); $e_{33}^{m} = 15.1$, $e_{31}^{m} = -5.2$, $e_{15}^{m} = 12.7$ (C/m^2); $\epsilon_{11}^{S\,m} = 730\varepsilon_0$, $\epsilon_{33}^{S\,m} = 635\varepsilon_0$, $\varepsilon_0 = 8.85\cdot 10^{-12}$(F/m) is the dielectric permittivity of vacuum. For the pores, we have set negligible values of the elastic moduli $c_{\alpha\beta}^{E\,p} = \kappa c_{\alpha\beta}^{E\,m}$, piezomoduli $e_{i\alpha}^{p} = \kappa$ (C/m^2), $\kappa = 10^{-10}$, and $\epsilon_{ii}^{S\,p} = \varepsilon_0$.

Some computational results are demonstrated in Figures 1, 2, 3, where $r(c_{33}^{E}) = c_{33}^{E\,\mathrm{eff}}/c_{33}^{E\,m}$ are the values of the effective moduli $c_{33}^{E\,\mathrm{eff}}$, related to the corresponding values of the moduli $c_{33}^{E\,m}$ of dense ceramic and so on. The full curves denote relative values for the composite material with full pore surface metallization, and the dotted curves denotes relative values for the material without pore surface metallization. The curves with squares denote the results for the volume Ω with cube pore, and the curves with circles denote the results for the volume with spherical pore.

As it can be seen from Fig. 1, the relative values of the effective stiffnesses $r(c_{33}^{E})$ decrease with the increase of porosity $p = |\Omega^p|/|\Omega|$, both for ordinary porous material and for porous material with pore surface metallization, and this decrease is slightly stronger for the cases of pore metallization and for the cubic pore. The relative values of the effective moduli of dielectric permittivity $r(\epsilon_{33}^{S})$ for ordinary porous piezoceramics decrease with the increase of porosity, on the contrary, the effective moduli of dielectric permittivity for porous piezoceramics with metallized pore surfaces increase with the increase of porosity. Here, the corresponding values of $r(\epsilon_{33}^{S})$ for the case of cubic pore are less compared to the case of the spherical pore.

The greater interest is the behavior of piezomoduli (Fig. 2, 3). Indeed, for ordinary porous piezoceramics the piezomoduli e_{3j}^{eff} decrease with the increase of

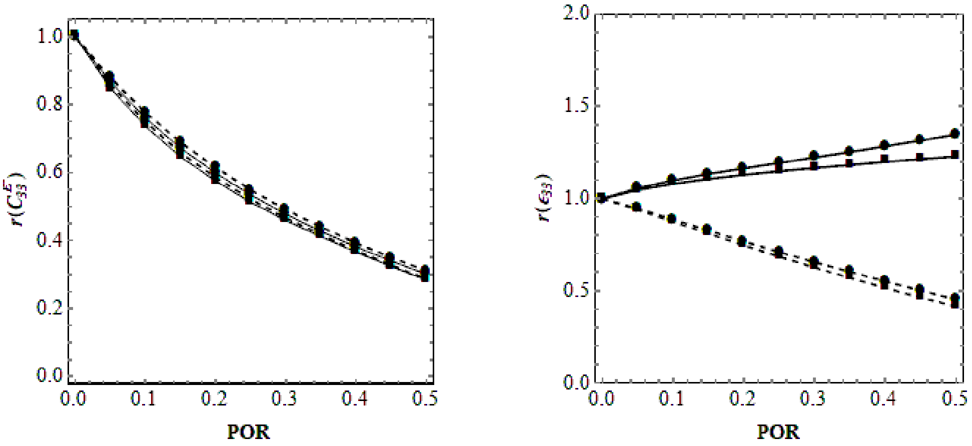


Figure 1. Dependencies of effective elastic stiffness and dielectric permittivity on porosity

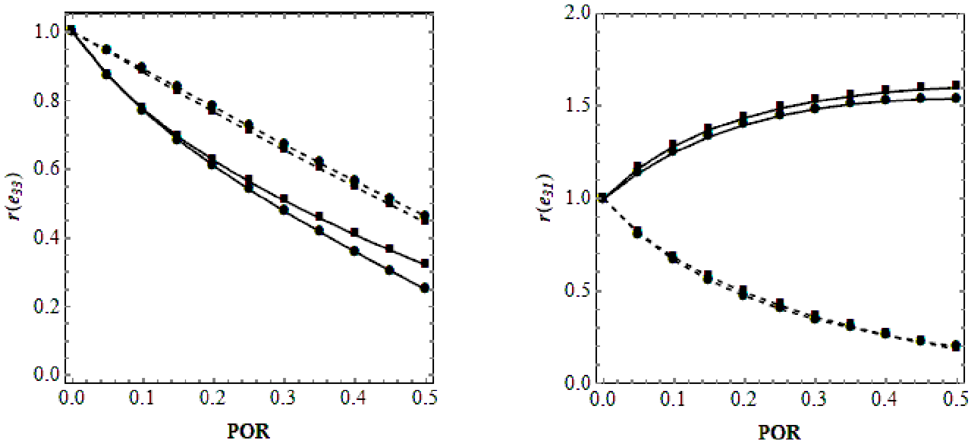


Figure 2. Dependencies of the effective piezomoduli e_{3j} on porosity

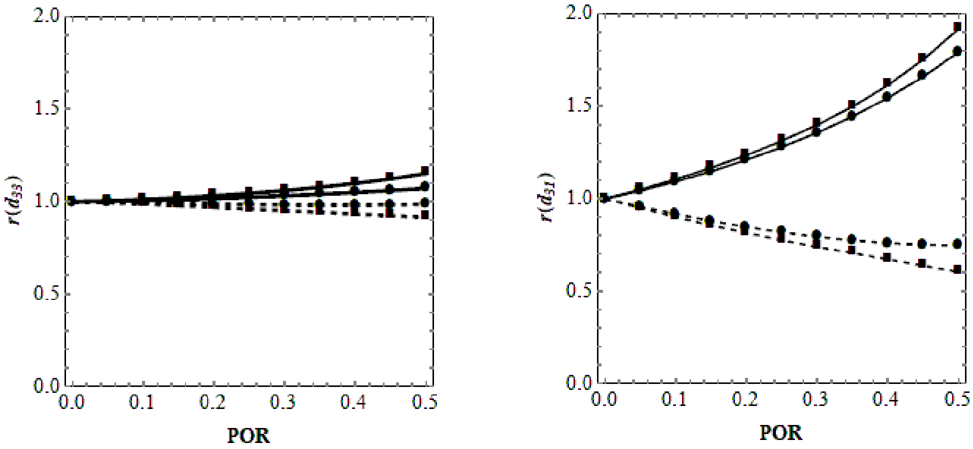


Figure 3. Dependencies of the effective piezomoduli d_{3j} on porosity

porosity. However, for piezoceramics with metallized pore surface the piezomodulus e_{33}^{eff} also decreases with increase of p, and it decreases faster than for the non-metallized pore surface, whereas the piezomodulus e_{31}^{eff} increases with the increase of porosity.

For the piezomodulus d_{33}^{eff} of an ordinary porous piezoceramics, its unusual property of weak dependence on porosity is well known, however, the values of piezomodulus $|d_{31}^{\mathrm{eff}}|$ decrease with the increase of p. As can be seen from Fig. 3, for porous piezoceramics with metallized pore surfaces the values of piezomoduli d_{33}^{eff} and $|d_{31}^{\mathrm{eff}}|$ increase with the increase of porosity, moreover, piezomodulus $|d_{31}^{\mathrm{eff}}|$ grows faster than d_{33}^{eff}. For example, for $p = 0.5$ the effective piezomodulus $|d_{31}^{\mathrm{eff}}|$ is almost two times greater than the similar value for dense piezoceramics.

Interestingly, in contrast to elastic stiffnesses and dielectric constants, the amplitude values of piezomoduli for metallized cubic pores are slightly larger than for metallized spherical pores. However, for ordinary porous ceramics, these values turn out to be larger for spherical pores, except for the values of the piezomodulus e_{33}^{eff}. These differences are not very significant and require further analysis.

Thus, summarizing the above, we can conclude that the shape of the pores has certain effects on the effective moduli of the porous material, though it is not so extensive. Therefore, for more complex models of representative volumes, it is quite possible to consider cubic pores instead of more physical spherical pore shapes, as was accepted in [2, 3]. Further studies of the influence of pore shape for this problem could be executed to take into account the mechanical properties of metallized pore surfaces, the local alloying pore surfaces and the heterogeneously polarized porous piezoceramic materials similarly to [4].

Bibliography

1. *Rybyanets A.N., Shvetsov I.A., Lugovaya M.A., Petrova E.I., Shvetsova N.A.* Nanoparticles transport using polymeric nano- and microgranules: novel approach for advanced material design and medical applications // J. Nano- Electron. Phys. 2018. Vol. 10. N 2. P. 02005–1-02005-6.

2. *Nasedkin A.V., Nasedkina A.A., Rybyanets A.N.* Mathematical modeling and computer design of piezoceramic materials with random arrangement of micropores and metallized pore surfaces / In: Proc. 2016 Int. Conf. on "Physics, Mechanics of New Materials and Their Applications". Eds. I.A. Parinov, S.-H. Chang, M.A. Jani. New York, Nova Science Publishers, 2017, pp. 385–392.

3. *Nasedkin A., Nasedkina A., Rybyanets A.* Finite element simulation of effective properties of microporous piezoceramic material with metallized pore surfaces // Ferroelectrics. 2017. Vol. 508. P. 100–107.

4. *Nasedkin A.V., Nasedkina A.A., Rybyanets A.N.* Numerical analysis of effective properties of heterogeneously polarized porous piezoceramic materials with local alloying pore surfaces // Material Physics and Mechanics. 2018. Vol. 40. N 1. P. 12–21.

FOURIER METHOD FOR SOLVING THE CAUCHY PROBLEM FOR AN ELLIPTIC EQUATION.

Sorokin S.B.[*][**]

[*] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS*

[**] *Novosibirsk State University, Novosibirsk, Russia*

One widely used approach to solving the Cauchy problem for an elliptic equation is to reduce it to the inverse problem. As a rule, an iterative procedure is used to solve the latter. In the work an economical direct method for the numerical solution of the inverse problem in rectangular form is described. The idea is based on the expansion of the desired solution with respect to a basis consisting of eigenfunctions discrete inverse problem operator.

The presented algorithm can be applied for elliptic operator with variable coefficients (of course of a special type). In this case, it is impossible to obtain an analytical solution of the continuation problem (the well-known direct algorithms for solving the Cauchy problem for the Laplace equation are based precisely on this). Therefore, an economical algorithm that allows one to obtain a solution to a discrete problem for the number of arithmetic operations of order N, where N is the number of equations in a difference problem, can be useful.

Formulation of the problem

In the domain $\Omega = \{(x_1, x_2) \in R^2 : \ x_1 \in (a, b), \ x_2 \in (c, d)\}$ consider the initial-boundary value problem (the continuation problem)

$$\begin{aligned}
&-\frac{\partial}{\partial x_1}(k_1(x_1)\frac{\partial u}{\partial x_1}) - \frac{\partial}{\partial x_2}(k_2(x_1)\frac{\partial u}{\partial x_2}) + k_3(x_1)u = 0, \quad (x_1, x_2) \in \Omega,\\
&-k_1(a)\frac{\partial u}{\partial x_1}(a, x_2) = 0, \quad x_2 \in [c, d], \quad u(a, x_2) = f(x_2), \quad x_2 \in [c, d],\\
&-k_2(x_1)\frac{\partial u}{\partial x_2}(x_1, c) = 0, \quad k_2(x_1)\frac{\partial u}{\partial x_2}(x_1, d) = 0, \quad x_1 \in [a, b].
\end{aligned} \tag{1}$$

It is necessary to find the function $u(x_1, x_2)$ in the domain Ω according to the data of $f(x_2)$.

Following [1],[2], we reduce the continuation problem (1) to the inverse problem:

Determine the function $q(x_2)$ from the relations

$$\begin{aligned}
&-\frac{\partial}{\partial x_1}(k_1(x_1)\frac{\partial u}{\partial x_1}) - \frac{\partial}{\partial x_2}(k_2(x_1)\frac{\partial u}{\partial x_2}) + k_3(x_1)u = 0, \quad (x_1, x_2) \in \Omega,\\
&-k_1(a)\frac{\partial u}{\partial x_1}(a, x_2) = 0, \quad x_2 \in [c, d], \quad u(b, x_2) = q(x_2), \quad x_2 \in [c, d],\\
&-k_2(x_1)\frac{\partial u}{\partial x_2}(x_1, c) = 0, \quad k_2(x_1)\frac{\partial u}{\partial x_2}(x_1, d) = 0, \quad x_1 \in [a, b]
\end{aligned} \tag{2}$$

using additional information

$$u(a,x_2)=f(x_2),\quad x_2\in(c,d). \tag{3}$$

We will solve the inverse problem (2), (3) at a discrete level.

Discretization of the inverse problem

Construct uniform grids:

$$\bar\omega_1=\{x_{1,i}=x_{1,i-1}+h_1,\ 1\le i\le N_1+1,\ x_{1,0}=a,\ x_{1,N_1+1}=b,\ h_1=\frac{b-a}{N_1+1}\},$$
$$\bar\omega_2=\{x_{2,j}=x_{2,j-1}+h_2,\ 1\le j\le N_2+1,\ x_{2,0}=c,\ x_{2,N_2+1}=d,\ h_2=\frac{d-c}{N_2+1}\},$$
$$\bar\omega=\{\ x_{i,j}=(x_{1,i},x_{2,j})\in\bar\Omega,\quad x_{1,i}\in\bar\omega_1,\quad x_{2,j}\in\bar\omega_2\}.$$

Scalar products in the spaces of grid functions defined on these grids and on the grid $\bar\omega=\bar\omega_1\bigcup\bar\omega_2$, set as follows (at the right end of the grid $\bar\omega_1$, the functions take a zero value)

$$(u^h,v^h)_{\bar\omega_1}=0.5h_1u^h(x_{1,0})v^h(x_{1,0})+\sum_{i=1}^{N_1}u^h(x_{1,i})v^h(x_{1,i})h_1,$$

$$(u^h,v^h)_{\bar\omega_2}=0.5h_2u^h(x_{2,0})v^h(x_{2,0})+\sum_{j=1}^{N_2}u^h(x_{2,j})v^h(x_{2,j})h_2+$$
$$+0.5h_2u^h(x_{2,N_2+1})v^h(x_{2,N_2+1}),$$
$$(u^h,v^h)_{\bar\omega}=(1,(u^h,v^h)_{\bar\omega_2})_{\bar\omega_1}.$$

Discrete analogue of the inverse problem (2),(3) we formulate as follows:

Among all the tasks

$$\Lambda_1u^h(x_{1,i},x_{2,j})+k_2(x_{1,i})\Lambda_2u^h(x_{1,i},x_{2,j})=0,\quad i=\overline{0,N_1},\ j=\overline{0,N_2+1} \tag{4}$$

$$u^h(x_{1,N_1+1},x_{2,j})=q^h(x_{2,j}),\quad j=\overline{0,N_2+1} \tag{5}$$

differing from each other by the grid function $q^h(x_{2,j})$, it is necessary to indicate such a problem (such a grid function $q^h(x_{2,j})$) for its solution $u^h(x_{1,i},x_{2,j})$ satisfies equality

$$u^h(x_{1,0},x_{2,j})=f(x_{2,j}),\quad x_{2,j}\in[c,d]. \tag{6}$$

Here the action of the operators Λ_1 and Λ_2 are determined by the equations [3],[4]:

$$\Lambda_1y^{(1)}(x_{1,i})=\begin{cases}-\dfrac{2}{h_1}k_1(x_{1,0}+\dfrac{h_1}{2})y^{(1)}_{x_1}(x_{1,0}),\quad i=0,\\ -(k_1(x_{1,i}-\dfrac{h_1}{2})y^{(1)}_{\bar x_1})_{x_1}(x_{1,i})+k_3(x_{1,i})y^{(1)}(x_{1,i}),\quad i=\overline{1,N_1},\end{cases}$$

$$\Lambda_2 y^{(2)}(x_{2,j}) = \begin{cases} -\dfrac{2}{h_2} y^{(2)}_{x_2}(x_{2,0}), & j = 0, \\ -y^{(2)}_{\overline{x}_2 x_2}(x_{2,j}), & j = \overline{1, N_2}, \\ \dfrac{2}{h_2} y^{(2)}_{\overline{x}_2}(x_{2,N_2+1}), & j = N_2 + 1. \end{cases}$$

Also, as described, for example, in [2] we can, using task (4)-(5), enter operator

$$A_h \,:\, q^h(x_{2,j}) \;\to\; u^h(x_{1,0}, x_{2,j}),$$

where $u^h(x_{1,i}, x_{2,j})$ direct problem solution (4)-(5).

Then the inverse problem (4)-(6) can be written in the following form:

$$A_h\, q^h(x_{2,j}) \;=\; f(x_{2,j}). \tag{7}$$

Solution of the spectral problem for the operator A_h

Take the eigenfunction $\mu^{(2)}_k$ of the spectral problem

$$\Lambda_2 \mu^{(2)}(x_{2,j}) = \lambda^{(2)} \mu^{(2)}(x_{2,j}), \quad j = \overline{0, N_2 + 1}. \tag{8}$$

The eigenvalues of this task are written explicitly [4].

We act on $\mu^{(2)}_k$ with the operator A_h. In accordance with the definition, to calculate $A_h \mu^{(2)}_k$ we need to solve the problem

$$\Lambda_1 w_k(x_{1,i}, x_{2,j}) + k_2(x_{1,i}) \Lambda_2 w_k(x_{1,i}, x_{2,j}) = 0, \quad i = \overline{0, N_1},\ j = \overline{0, N_2 + 1}, \tag{9}$$

$$w_k(x_{1,N_1+1}, x_{2,j}) = \mu^{(2)}_k(x_{2,j}), \quad j = \overline{0, N_2 + 1}. \tag{10}$$

After that, it is required to calculate the trace of the obtained solution on the left border of the computational domain. This will be the desired value $A_h \mu^{(2)}_k$.

Dirichlet boundary conditions (10) problems (9)-(10) we take into account on the right side. As a result, we get the task

$$\Lambda_1 w_k(x_{1,i}, x_{2,j}) + \Lambda_2 w_k(x_{1,i}, x_{2,j}) = f_k(x_{1,i}, x_{2,j}), \quad i = \overline{0, N_1},\ j = \overline{0, N_2 + 1}. \tag{11}$$

$$w_k(x_{1,N_1+1}, x_{2,j}) = 0, \quad j = \overline{0, N_2 + 1}, \tag{12}$$

where the right side is given by the formula

$$f_k(x_{1,i}, x_{2,j}) = \begin{cases} 0, \quad i = \overline{0, N_1 - 1},\ j = \overline{0, N_2 + 1}, \\ k_1(x_{1,N_1} + \dfrac{h_1}{2}) \dfrac{\mu^{(2)}_k(x_{2,j})}{h_1^2}, \quad i = N_1,\ j = \overline{0, N_2 + 1}. \end{cases} \tag{13}$$

We will solve the problem (11)-(13) by the single-row decomposition method [3].

Considering the grid functions $w_k(x_{1,i},x_{2,j})$ and $f_k(x_{1,i},x_{2,j})$ with fixed i as the grid functions of the argument j, decompose them in basis $\mu^{(2)}_m$, $\quad m = 0,1,...,N_2+1$:

$$\begin{gathered} w_k(x_{1,i},x_{2,j}){=}\sum_{m=0}^{N_2+1} w_{k,m}(i)\ \mu^{(2)}_m(x_{2,j}),\ w_{k,m}(i){=}(w_k(x_{1,i},x_{2,j}),\mu^{(2)}_m(x_{2,j}))_{\bar\omega_2}, \\ f_k(x_{1,i},x_{2,j}) = \sum_{m=0}^{N_2+1} f_{k,m}(i)\ \mu^{(2)}_m(x_{2,j}),\ f_{k,m}(i) = (f_k(x_{1,i},x_{2,j}),\mu^{(2)}_m(x_{2,j}))_{\bar\omega_2}. \end{gathered} \tag{14}$$

Substituting these expansions into the equation (11) and taking into account the equalities $\Lambda_2\mu^{(2)}(x_{2,j}) = \lambda^{(2)}\mu^{(2)}(x_{2,j}), \quad j = \overline{0, N_2+1}$, we have

$$[\sum_{m=0}^{N_2+1} \Lambda_1 w_{k,m}(i) + \lambda^{(2)}_m k_2(x_{1,i})w_{k,m}(i)]\mu^{(2)}_m(x_{2,j}) = \sum_{m=0}^{N_2+1} f_{k,m}(i)\mu^{(2)}_m(x_{2,j}).$$

Using linear independence of eigenfunctions, we obtain for each number $m = 0,1,...,N_2+1$ linear equation system

$$\Lambda_1 w_{k,m}(i) + \lambda^{(2)}_m k_2(x_{1,i})w_{k,m}(i) = f_{k,m}(i), \quad i = \overline{0, N_1}. \tag{15}$$

In accordance with the scalar products introduced and the form $f_k(x_{1,i},x_{2,j})$ (see (13)) for each fixed number k we have

$$f_{k,m}(i){=}(f_k(x_{1,i},x_{2,j}),\mu^{(2)}_m(x_{2,j}))_{\bar\omega_2}{=}\begin{cases} 0, \quad i=\overline{0,N_1-1},\ m=\overline{0,N_2+1}, \\ k_1(x_{1,N_1}{+}\frac{h_1}{2})(\frac{\mu^{(2)}_k}{h_1^2},\mu^{(2)}_m)_{\bar\omega_2}{=}k_1(x_{1,N_1}{+}\frac{h_1}{2})\frac{1}{h_1^2}\delta_{km}, \\ i=N_1,\ m=\overline{0,N_2+1}. \end{cases} \tag{16}$$

Therefore, for a fixed k, all systems (15) (designed to determine the Fourier coefficients $w_{k,m}(i), \quad i = \overline{0,N_1};\ m = \overline{0,N_2+1}$) for numbers $m \neq k$ will be uniform. Therefore, their solutions are identically zero: $w_{k,m}(i) = 0, i = \overline{0,N_1}, \forall \quad m \neq k$.

The only non-uniform system from (15) will be the system with the number $m = k$. Having in mind (16), this system is written as follows:

$$\begin{aligned} &\Lambda_1 w_{k,k}(i) + \lambda^{(2)}_k k_2(x_{1,i})w_{k,k}(i) = 0, \quad i = \overline{0,N_1-1}, \\ &\Lambda_1 w_{k,k}(N_1) + \lambda^{(2)}_k k_2(x_{1,N_1})w_{k,k}(N_1) = k_1(x_{1,N_1}+\tfrac{h_1}{2})\frac{1}{h_1^2}. \end{aligned} \tag{17}$$

In summary, the solution (14) of the problem (11)-(13) (with the number k) recorded in the form $w_k(x_{1,i},x_{2,j}) = w_{k,k}(i)\mu^{(2)}_k(x_{2,j})$. The trace of the solution obtained on the left boundary of the computational domain $x = x_{1,0}$ is equal to $w_{k,k}(0)\mu_k^{(2)}(x_{2,j})$. Which means

$$A_h\mu_k^{(2)}(x_{2,j}) = w_{k,k}(0)\mu_k^{(2)}(x_{2,j}).$$

So we have established that the eigenfunctions of $\mu_k^{(2)}(x_{2,j})$ from the spectral problem (8) are the eigenfunctions of the operator A_h, eigenvalues of the corresponding $w_{k,k}(0)$ are calculated from (17).

Inverse problem solving $A_h\, q^h(x_{2,j}) \;=\; f(x_{2,j})$

We will seek a solution in the form

$$q^h(x_{2,j}) = \sum_{k=0}^{N_2+1} \alpha_k\mu_k^{(2)}(x_{2,j}). \tag{18}$$

We expand the additional information (3) according to the basis $\mu_k^{(2)}(x_{2,j})$, $k = \overline{0, N_2+1}$, consisting of eigenfunctions of the operator $A_h \;:\; f(x_{2,j}) \;=\; \sum\limits_{k=0}^{N_2+1} \beta_k\mu_k^{(2)}(x_{2,j})$. Substitute the last two expansions into the equation $A_h\, q^h(x_{2,j}) \;=\; f(x_{2,j}):$

$$\sum_{k=0}^{N_2+1} \alpha_k w_{k,k}(0)\mu_k^{(2)}(x_{2,j}) = \sum_{k=0}^{N_2+1} \beta_k\mu_k^{(2)}(x_{2,j}).$$

It follows that the coefficients α_k are equal $\alpha_k = \dfrac{\beta_k}{w_{k,k}(0)}$. Substituting them into (18), we obtain the solution of the inverse problem

$$q^h(x_{2,j}) = \sum_{k=0}^{N_2+1} \frac{\beta_k}{w_{k,k}(0)}\mu_k^{(2)}(x_{2,j}).$$

Bibliography

1. *Kabanikhin S.I., Karchevsky A.L.* Method for solving the Cauchy Problem for an Elliptic Equation, Journal of Inverse and Ill-posed Problems, 3. 1 (1995), 21-46.

2. *Kabanikhin S.I* Inverse and ill-posed problems, Novosibirsk: Siberian Scientific Publishing House, 2009.

3. *Samarskii A.A.,Nikolaev E.S.* Methods for solving grid equations. M.: Science. 1978. 529 p.

4. *Samarskii A.A., Andreev V.B.* Difference methods for elliptic equations, M.: Nauka. 1979. 152c.

CONVECTION IN A POROUS MEDIUM: COSYMMETRY AND ITS CONSERVATION THROUGH A FINITE-DIFFERENCE APPROXIMATION[1]

Tsybulin V.G.

Southern Federal University, Rostov-on-Don, Russia

The onset of convective flows in a porous rectangle occupied by a heat-conducting fluid heated from below is analyzed. Darcy–Boussinesq model in the case both anisotropic medium and fluid is formulated. It is shown that there are combinations of physical parameters for which the system has a nontrivial cosymmetry and a one-parameter family of stationary convective regimes branches off from the mechanical equilibrium. For the two-dimensional convection equations in a porous medium, finite-difference approximations preserving the cosymmetry of the original system are developed. Numerical results demonstrate the appearance of a family of convective regimes and its disintstructione when the approximations do not inherit the cosymmetry property [1, 2].

Bibliography

1. *Abdelhafez M.A., Tsybulin V.G.* Numerical Simulation of Convective Motions in an Anisotropic Porous Medium and Conservation of Cosymmetry // Computational Mathematics and Mathematical Physics, 2017, Vol. 57, N 10, P. 1706–1719.

2. *Abdelhafez M.A., Tsybulin V.G.* Anisotropic Problem of Darcy Convection: Family of Steady Flows and Its Disintegration during the Destruction of Cosymmetry // Fluid Dynamics, 2018, Vol. 53, N 6, P. 738–749.

[1] Supported by Southern Federal University

PMHSS PRECONDITIONERS FOR STOKES CONTROL OPTIMIZATION PROBLEMS[1]

Wang Z.-Q., Cao S.-M.

Shanghai Jiao Tong University, Shanghai, P.R. China

Flow control has been widely used in petroleum, chemical, and aeronautical engineering, and becomes a very active research area. It is obvious that developing efficient numerical methods for flow control is one of the keys to its successful applications. In this work, we consider the numerical solutions of Stokes control problems, to make a preparation for more complicated fluid dynamic problems, such as Navier-Stokes control problems. The present study focuses on the solution of the multiple saddle point problems generated by the discretize-then-optimize approach. The aim is constructing the iterative solvers which are independent of not only the mesh size of finite element discretization, but also the regularization parameter of the optimization. Some successful solvers and preconditioners for Stokes control optimization problems have been generated from different perspectives. Some studies design the preconditioners according to properties of saddle point matrices, especially the approximation of Schur complement matrices. In [1], a parameter-robust block diagonal preconditioner is derived based on the nonstandard norm argument. In [2], the block diagonal and block triangular preconditioners are generated based on a commutator argument [3]. Instead, [4] consider the achieved KKT system as a special structured block two-by-two linear system

$$\mathbf{A}\mathbf{x} \equiv \begin{bmatrix} \mathbf{W} & -\mathbf{T} \\ \mathbf{T} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} \equiv \mathbf{g}, \tag{1}$$

and design a preconditioner based on [5] and the references therein. In [6], authors present the preconditioned modified Hermitian/Skew-Hermitian splitting (PMHSS) iteration method and the corresponding preconditioner for Poisson control optimization problems. The PMHSS preconditioner has nearly the same workload as the preconditioner in [5]. But it could be performed on the short-term recurrence iteration method, for instance, MINRES, Chebyshev and so on, see the references [7, 8]. The convergence of the PMHSS iteration method is studied when $\mathbf{W}, \mathbf{T} \in \mathbb{R}^{n\times n}$ are real, symmetric and positive semidefinite matrices with at least one of them, e.g., $\mathbf{W}$, being positive definite. In the present paper, we analyze the convergence of PMHSS iteration method when $\mathbf{W}$ and $\mathbf{T}$ are positive semidefinite matrix and saddle point matrix, respectively. The PMHSS preconditioner is studied in terms of the eigenvalues distribution and the computing complexity. As the preconditioner in [4], PMHSS preconditioner needs the solutions of two saddle point linear systems in every iteration. The saddle point problems could be solved by inner iterations, such as flexible GMRES,

[1] This research is supported by National Natural Science Foundation of China (11371022)

parameterized and preconditioned Uzawa iterations and so on. Alternatively, to save the computing cost and avoid the decision of inner tolerance, we suggest a modified PMHSS preconditioner and analyze the eigenvalues distribution of the corresponding preconditioned matrix.

I PMHSS iteration and preconditioner for optimality systems

We consider the Stokes control problem:

$$\min_{u,f} \frac{1}{2}\|y-y_d\|_2^2 + \frac{1}{2}\beta\|u\|_2^2$$

$$\begin{aligned} \text{subject to} -\nabla^2 y + \nabla p &= u \quad \text{in} \quad \Omega, \\ \nabla \cdot y &= 0 \quad \text{in} \quad \Omega, \\ y &= g_D \quad \text{on} \quad \partial\Omega, \end{aligned} \tag{2}$$

where Ω is a domain in $\mathbb{R}^2$ or $\mathbb{R}^3$, $\partial\Omega$ is the boundary of Ω. The desired state function y_d and the boundary value function g_D are given. As the Stokes equation is self-adjoint, the discretize-then-optimize and optimize-then-discretize processes are mathematically equivalent and lead to the same solution. Here, we study the algebraic systems which are obtained by discretize-then-optimize approach. The rectangular Taylor-Hood finite element method is performed as it is inf-sup stable. Specifically, the velocity y and control u are approximated by the combinations of Q_2-basis functions $\{\phi_j\}$, $j = 1, \cdots, n_v$, while the pressure p is approximated by the combinations of Q_1-basis functions $\{\psi_j\}$, $k = 1, \cdots, n_p$. Then, the first order necessary optimality condition of the discretized optimization problem yields the following linear system

$$\mathcal{A}^R \begin{bmatrix} \mathbf{y} \\ \mathbf{p} \\ \mathbf{l} \\ \mathbf{m} \end{bmatrix} \equiv \begin{bmatrix} M & 0 & -F^T & -B^T \\ 0 & 0 & -B & 0 \\ F & B^T & M & 0 \\ B & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{p} \\ \mathbf{l} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{3}$$

where

$$F = \sqrt{\beta}\int_\Omega \nabla\phi_i : \nabla\phi_j, \quad M = \int_\Omega \phi_i\phi_j, \quad B = -\sqrt{\beta}\int_\Omega \psi_k \cdot \nabla\phi_j,$$

$$b = \int_\Omega y_d\phi_i - \sum_{j=n_y+1}^{n_y+n_\partial} \int_\Omega \nabla\phi_i : \nabla\phi_j,$$

$$f_i = -\sqrt{\beta}\sum_{j=n_y+1}^{n_y+n_\partial} y_j \int_\Omega \nabla\phi_i : \nabla\phi_j, \quad g_i = \sqrt{\beta}\sum_{j=n_y+1}^{n_y+n_\partial} y_j \int_\Omega \psi_i\nabla\cdot\phi_j,$$

$$f = [f_i], \quad g = [g_i],$$

$\mathbf{l}$ and $\mathbf{m}$ are scaled Lagrange multipliers, corresponding to $\mathbf{y}$ and $\mathbf{p}$, respectively. The matrices $M \in \mathbb{R}^{n_\nu \times n_\nu}$ and $F \in \mathbb{R}^{n_\nu \times n_\nu}$ referred to mass matrix and scaled stiffness matrix, are symmetric positive definite. The matrix $B \in \mathbb{R}^{n_p \times n_\nu}$ is full row rank. We refer to [9, 2, 4] for details of the finite element discretization. The four-by-four block matrix $\mathcal{A}^R$ in (3) can be partitioned into

$$\mathcal{A}^R := \begin{bmatrix} \mathbf{W} & -\mathbf{T} \\ \mathbf{T} & \mathbf{W} \end{bmatrix}, \tag{4}$$

where

$$\mathbf{W} = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{T} = \begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix} \tag{5}$$

are symmetric positive semidefinite matrix and nonsingular saddle point matrix, respectively. Recently, Bai et al. define a preconditioned modified Hermitian/Skew-Hermitian splitting (**PMHSS**) iteration method [6] when $\mathbf{W}$ and $\mathbf{T}$ are symmetric positive semi-definite on the basis of the matrix splitting

$$\mathcal{A}^R = \mathbf{F} - \mathbf{G},$$

where

$$\mathbf{F} := \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{W}+\mathbf{T} & 0 \\ 0 & \mathbf{W}+\mathbf{T} \end{bmatrix}, \quad \text{and} \quad \mathbf{G} := \begin{bmatrix} \mathbf{T} & -\mathbf{W} \\ \mathbf{W} & \mathbf{T} \end{bmatrix}. \tag{6}$$

The PMHSS iteration scheme for solving the two-by-two block system can be written as

$$\mathbf{F}\mathbf{x}^{(k+1)} = \mathbf{G}\mathbf{x}^{(k)} + \mathbf{g}.$$

The matrix $\mathbf{F}$ is refer to as PMHSS preconditioner as well. The PMHSS iteration method and preconditioner achieve very good effect due to the clustering eigenvalues distribution and the normalization of the eigenvectors, when $\mathbf{W}$ and $\mathbf{T}$ are symmetric positive definite. In the present study, we explore the performance of PMHSS iteration method and preconditioner working on the linear system in (3).

Lemma I.1 *Let $\mathbf{W}$ and $\mathbf{T}$ be the matrices in (5), where M and F are symmetric positive definite and B is full row rank. Then the eigenvalues of $\mathbf{T}^{-1}\mathbf{W}$ are all nonnegative.*

Lemma I.2 *Let τ and x be the eigenvalues and corresponding eigenvectors of $\mathbf{T}^{-1}\mathbf{W}$. Define $E = (\mathbf{W}+\mathbf{T})^{-1}(\mathbf{W}-\mathbf{T})$. Then, the eigenvalues of E are $\mu = \dfrac{\tau-1}{\tau+1}$, with x the corresponding eigenvectors.*

As a result, we have $\mu = \dfrac{\tau-1}{\tau+1}$, which leads to $-1 \leq \mu < 1$. Moreover, $\mu = -1$ are the eigenvalues of multiplicity n_p at least.

Theorem I.1 *Let μ_j , $j = 1, \ldots\cdot, n_v + n_p$ be the eigenvalues of E. Then the eigenvalues of $\mathbf{F}^{-1}\mathcal{A}^R$ are*

$$\lambda_j^{\pm} = \frac{1}{2}(1 \pm \mathrm{i}\mu_j).$$

According to Theorem I.1, the spectral of preconditioned coefficient matrix $\mathbf{F}^{-1}\mathcal{A}^R$ are located in the unitary segment between $\frac{1}{2}(1+\mathrm{i})$ and $\frac{1}{2}(1-\mathrm{i})$. The eigenvalues of PMHSS iterative matrix

$$\mathbf{L} = \mathbf{F}^{-1}\mathbf{G} = I - \mathbf{F}^{-1}\mathcal{A}^R$$

are $1-\lambda_j^{\pm} = \frac{1}{2}(1\mp\mathrm{i}\mu_j)$. PMHSS iteration method is convergent since the spectral radius of iterative matrix is no more than $\frac{1}{\sqrt{2}}$.

When PMHSS preconditioning is performed on Krylov subspace methods, the generalized residual equation $\mathbf{Fr} = \mathbf{z}$ should be solved in every iteration. The main workload is solving two linear saddle point problems with

$$\mathbf{W} + \mathbf{T} = \begin{bmatrix} F+M & B^T \\ B & 0 \end{bmatrix} \tag{7}$$

in every iteration. The saddle point equations should be solved by inner iteration methods, we refer to the discussion in [4].

We will propose a new preconditioner to avoid solving saddle point equations in this talk.

Bibliography

1. *Zulehner W.* Nonstandard Norms and Robust Estimates for Saddle Point Problems // SIAM Journal on Matrix Analysis and Applications. 2011. Vol. 32. N 2. P. 536–560.

2. *Pearson J. W.* On the development of parameter-robust preconditioners and commutator arguments for solving Stokes control problems //Electronic Transactions on Numerical Analysis. 2015. Vol. 44. P. 53-725.

3. *Cahouet J, Chabard J.P.* Some fast 3D finite element solvers for the generalized Stokes problem//International Journal for Numerical Methods in Fluids. 1998. Vol. 8. N 8. P. 869-895.

4. *Axelsson O., Farouq S., Neytcheva M.*Comparison of preconditioned Krylov subspace iteration methods for PDE-constrained optimization problems : Stokes control// Numerical Algorithms. 2017. Vol. 74. N 1. P. 19–37.

5. *Axelsson O., Neytcheva M., Ahmad B.*A comparison of iterative methods to solve complex valued linear algebraic systems // Numerical Algorithms. 2014. Vol. 66. N 4. P. 1017–1398.

6. *Bai Z.-Z., Benzi M., Chen F., Wang Z.-Q.* Preconditioned MHSS iteration methods for a class of block two-by-two linear systems with applications to distributed control problems// IMA Journal of Numerical Analysis. 2013. Vol. 33. N 1. P. 343-369.

7. *Bai Z.-Z.* On preconditioned iteration methods for complex linear systems//Journal of Engineering Mathematics. 2015. Vol. 93. N 1. P. 41-60.

8. *Wang Z.-Q.* On a Chebyshev accelerated splitting iteration method with application to two-by-two block linear systems // Numerical Linear Algebra with Applications. 2018. Vol. 25. e2172.

9. *Elman H. C. , Silvester D. J., Wathen A. J.* Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics. Oxford University Press, USA, 2014.

A THREE-TERM ACCELERATED TECHNIQUE FOR HSS-TYPE METHODS[1]

Wen R.-P., Jiang W.

Key Laboratory for Engineering and Computational Science, Shanxi Provincial Department of Education, Taiyuan Normal University, Jinzhong 030619, Shanxi Province, P. R. China

We consider the large sparse non-Hermitian and positive definite system of the form

$$Ax = b, \qquad A \in \mathbb{C}^{n\times n} \text{ nonsingular}, \qquad x, b \in \mathbb{C}^n, \tag{1}$$

based on the *Hermitian and skew-Hermitian* (HS) splitting

$$A = A_H + A_S$$

of the coefficient matrix A with $A_H = \frac{1}{2}(A+A^*)$, $A_S = \frac{1}{2}(A-A^*)$, where A^* is the conjugate transpose of A, Bai, Golub and Ng[1] have proposed the Hermitian and skew-Hermitian Splitting (HSS) iteration method in 2003. Our interest for the HSS iteration method is how to find a way to accelerate its convergence rate and/or to modify the iteration method to construct an extrapolated iteration so as to improve its efficiency. We present a three-term acceleration (TTA) strategy apply to some HSS-type iteration methods, say HSS, PHSS, GHSS or GPHSS iteration methods, for solving a non-Hermitian positive definite system of linear equations in this study. Taking the GHSS method as an example we introduce the detail of the TTA strategy for some HSS-type methods based on the optimization technique. The acceleration factor ω is obtained by the optimization technique. We study the convergence theory of new methods and also discuss the convergence rate. Finally, some numerical results indicate that the TTA strategy is efficient in computations.

Here is a brief review for some related modifications of HSS iteration method, and then a three-term acceleration scheme for them is proposed.

Method 1 (*The HSS Method*) Given an initial guess x_0, for $k = 0, 1, 2, \cdots$, until $\{x_k\}$ converges, compute

$$\begin{cases} (\alpha I + A_H)x_{k+\frac{1}{2}} = (\alpha I - A_S)x_k + b, \\ (\alpha I + A_S)x_{k+1} = (\alpha I - A_H)x_{k+\frac{1}{2}} + b, \end{cases} \tag{2}$$

where I is the identity matrix and α is a given positive constant.

Note that the HSS iteration scheme (2) may also be considered as a splitting iteration method induced from the splitting of the matrix A as follows,

$$A = M - N,$$

[1] Supported by NSF of China (11371275), NSF of Shanxi Province (201601D011004), and the SYYJSKC-1803.

where

$$M = \frac{1}{2\alpha}(\alpha I + H)(\alpha I + S) \quad \text{and} \quad N = \frac{1}{2\alpha}(\alpha I - H)(\alpha I - S). \tag{3}$$

The preconditioned HSS (PHSS) method based on the above particular matrix splitting. Especially, we might consider that the method is designed for solving another preconditioned linear system $\bar{A}\bar{x} = \bar{b}$ with $\bar{A} = R^{-*}AR^{-1}, \bar{x} = Rx$ and $\bar{b} = R^{-*}b$. Where $R \in \mathbb{C}^{n\times n}$ is a prescribed nonsingular matrix, and $R^{-*} = (R^{-1})^*$. We usually take a Hermitian positive definite matrix $P = R^*R$. Thus, the PHSS method[2] was defined as follows:

Method 2 (*The PHSS Method*) Given an initial guess x_0, for $k = 0, 1, 2, \cdots$, until $\{x_k\}$ converges, compute

$$\left\{ \begin{array}{l} (\alpha P + A_H)x_{k+\frac{1}{2}} = (\alpha P - A_S)x_k + b, \\ (\alpha P + A_S)x_{k+1} = (\alpha P - A_H)x_{k+\frac{1}{2}} + b, \end{array} \right. \tag{4}$$

where α is a given positive constant, and P is a Hermitian positive definite matrix (in particular, it reduces to the HSS method with $P = I$).

Let

$$M = \frac{1}{2\alpha}(\alpha P + A_H)(\alpha P + A_S), \quad N = \frac{1}{2\alpha}(\alpha P - A_H)(\alpha P - A_S), \tag{5}$$

the splitting $A = M - N$ induces the PHSS iteration method.

Split A_H into the sum of two Hermitian positive semidefinite matrices: $A_H = A_G + A_K$, where A_K is of simple form (e.g., diagonal) and to associate A_K to the skew-Hermitian portion A_S of A so that $A = A_G + (A_S + A_K)$. The generalized HSS (GHSS) scheme[3] is obtained as follows:

Method 3 (*The GHSS Method*) Given an initial guess x_0, for $k = 0, 1, 2, \cdots$, until $\{x_k\}$ converges, compute

$$\left\{ \begin{array}{l} (\alpha I + A_G)x_{k+\frac{1}{2}} = (\alpha I - A_S - A_K)x_k + b, \\ (\alpha I + A_S + A_K)x_{k+1} = (\alpha I - A_G)x_{k+\frac{1}{2}} + b, \end{array} \right. \tag{6}$$

where α is a given nonnegative constant.

When either A_G or A_K is positive definite, the resulting scheme, which reduces to the original HSS method when $A_K = 0$, is shown to be convergent for all $\alpha > 0$.

Let

$$M = \frac{1}{2\alpha}(\alpha I + A_G)(\alpha I + A_S + A_K), \quad N = \frac{1}{2\alpha}(\alpha I - A_G)(\alpha I - A_S - A_K), \tag{7}$$

the splitting $A = M - N$ induces to the GHSS iteration method.

With different properties of the matrices H and S, it is natural to produce different effects on the parameter in (4). Due to this consideration, the GPHSS method[4] was proposed in the following:

Method 4 (*The GPHSS Method*) Given an initial guess x_0, for $k = 0, 1, 2, \cdots$, until $\{x_k\}$ converges, compute

$$\begin{cases} (\alpha P + A_H)x_{k+\frac{1}{2}} = (\alpha P - A_S)x_k + b, \\ (\beta P + A_S)x_{k+1} = (\beta P - A_H)x_{k+\frac{1}{2}} + b, \end{cases} \tag{8}$$

where α is a given nonnegative constant, β is a given positive constant and P is a Hermitian positive definite matrix as the above (in particular, when we choose $\alpha = \beta$, it reduces to the PHSS method, it reduces to the GHSS method when without preconditioner, also it reduces to the HSS method with $\alpha = \beta$ and without preconditioner).

Let

$$M = \frac{1}{\alpha+\beta}(\alpha P + A_H)(\beta P + A_S), \quad N = \frac{1}{\alpha+\beta}(\beta P - A_H)(\alpha P - A_S), \tag{9}$$

the splitting $A = M - N$ induces the GHSS iteration method.

We describe the three-term accelerate to **Methods** 1-4 above.

Method 5 (*A three-term acceleration (TTA) to some HSS-type iteration methods*) Given an initial point x_0, a precision $\epsilon > 0$. For $k = 0, 1, 2, \cdots$ until $\{x_k\}$ converges, do

Step 1. Solve the system of linear equations as follows:

$$\bar{x}_{k+1} = M^{-1}Nx_k + M^{-1}b.$$

Let

$$x_{k+1} = \omega_{k+1}\bar{x}_{k+1} + (1 - \omega_{k+1})x_{k-1}, \tag{10}$$

where

$$x_0 \in \mathbb{R}^n, \quad x_1 = M^{-1}Nx_0 + M^{-1}b,$$

ω_{k+1} is the solution of the following optimization problem.

(a) (TTA-HSS) For **Method 1**, let $z = Ax - b,\ x = \omega(\bar{x}_{k+1} - x_{k-1}) + x_{k-1}$,

$$\min_{\omega} z^T(\alpha I + A_H)^{-2}z. \tag{11}$$

(b) (TTA-GHSS) For **Method 3**, let $z = Ax - b,\ x = \omega(\bar{x}_{k+1} - x_{k-1}) + x_{k-1}$,

$$\min_{\omega} z^T(\alpha I + A_G)^{-2}z. \tag{12}$$

(c) (TTA-GPHSS) For **Method 2** and **Method 4**, let $z = Ax-b,\ x = \omega(\bar{x}_{k+1} - x_{k-1}) + x_{k-1}$,

$$\min_{\omega} z^T(\alpha P + A_H)^{-2}z. \tag{13}$$

Step 2. If $\|z_{k+1}\|_2 < \epsilon$, stop; Otherwise, $k \Leftarrow k + 1$ and go to Step 1.

To avoid the tedious computation, however, we can determine the approximations of the acceleration factor ω by using optimization method (for example, the Newton method). Evidenly, we will find the optimal point in the following hyperplane.

$$\mathcal{H} = \{y | y = \omega(Qy_k + M^{-1}b - y_{k-1}) + y_{k-1},\ \omega \in \mathbb{R}\}.$$

Next, we provide the convergence theory of **Method 5** corresponding to (11)-(13), respectively.

Lemma 1 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (3). Then

$$\|(\alpha I + A_H)^{-1} N M^{-1} (\alpha I + A_H)\| < 1. \tag{14}$$

Furthermore, if A is a normal matrix, then

$$\|(\alpha I + A_H)^{-1} N M^{-1} (\alpha I + A_H)\| = \rho(M^{-1}N). \tag{15}$$

Theorem 1 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (3). ω_k is solved by the quadratic programming (11). Let $z = Ax - b$. If $\langle(\alpha I + A_H)^{-1}\bar{z}_{k+1}, (\alpha I + A_H)^{-1}(\bar{z}_{k+1} - z_{k+1})\rangle \geq \varphi$. Then $\{x_k\}$ generated by (10) of **Method 5** converges to the unique solution of the system of linear equations (1). Further, the convergence rate satisfies

$$r \geq -(\ln \|(\alpha I + A_H)^{-1} N M^{-1} (\alpha I + A_H)\| + \ln(\sin\varphi)). \tag{16}$$

If A is a normal matrix, then

$$r \geq -(\ln \rho(M^{-1}N) + \ln(\sin\varphi)).$$

Lemma 2 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (7). Then

$$\|(\alpha I + A_G)^{-1} N M^{-1} (\alpha I + A_G)\| < 1.$$

Furthermore, if A_K is diagonal, then

$$\|(\alpha I + A_G)^{-1} N M^{-1} (\alpha I + A_G)\| = \rho(M^{-1}N).$$

Theorem 2 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (7). ω_k is determined by the quadratic programming (12). Let $z = Ax - b$. If $\langle(\alpha I + A_G)^{-1}\bar{z}_{k+1}, (\alpha I + A_G)^{-1}(\bar{z}_{k+1} - z_{k+1})\rangle \geq \varphi$. Then $\{x_k\}$ generated by (10) of **Method 5** converges to the unique solution of the system of linear equations (1). Further, the convergence rate satisfies

$$r \geq -(\ln \|(\alpha I + A_G)^{-1} N M^{-1} (\alpha I + A_G)\| + \ln(\sin\varphi)). \tag{17}$$

If A_K is diagonal, then

$$r \geq -(\ln \rho(M^{-1}N) + \ln(\sin \varphi)). \tag{18}$$

For the other case of **Method 5**, say (13), we can analogously give the following convergence theorem.

Lemma 3 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (5) or (9). Then

$$\|(\alpha P + A_H)^{-1} N M^{-1} (\alpha P + A_H)\| < 1.$$

Theorem 3 Let $A = M - N$ be a splitting of the non-Hermitian positive definite matrix A. Assume that M, N are given by (5) or (9). ω_k is obtained by the quadratic programming (13). Let $z = Ax - b$. If $\langle (\alpha P + A_H)^{-1} \bar{z}_{k+1}, (\alpha P + A_H)^{-1} (\bar{z}_{k+1} - z_{k+1}) \rangle \geq \varphi$. Then $\{x_k\}$ generated by (10) of **Method 5** converges to the unique solution of the system of linear equations (1). Further, the convergence rate satisfies

$$r \geq -(\ln \|(\alpha P + A_H)^{-1} N M^{-1} (\alpha P + A_H)\| + \ln(\sin \varphi)).$$

For HSS iteration method, the three-term accelerated iteration method has more effective when HSS iteration method is not effective or the optimal parameter α_{opt} is difficultly found. On the other hand, as we know, when the parameter α is larger, the conjugate gradient method to solve the former half-step linear system of (2) becomes stable and speeds up its convergence, it is because the condition number of $\alpha I + H$ becomes smaller. Hence, the three-term accelerated iteration method is effective and valuable.

Bibliography

1. *Bai Z.-Z., Golub G.H., Ng M.K.* Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems// SIAM J. Matrix Anal. Appl. 2003. Vol.–24. N 3. P. 603–626.

2. *Bai Z.-Z., Golub G.H., Pan J.-Y.* Preconditioned Hermitian and skew-Hermitian splitting methods for non-Hermitian positive semidefinite linear systems // Numer. Math. 2004. Vol. 98. N 1. P. 1–32.

3. *Benzi M.* A Generalization of the Hermitian and skew-Hermitian splitting iteration // SIAM J. Matrix Anal. Appl. 2009. Vol.–31. N 2. P. 360–374.

4. *Yang A.-L., An J., Wu Y.-J.* A generalized preconditioned HSS method for non-Hermitian positive definite linear systems // Appl. Math. Comput. 2010. Vol. 216. N 3. P. 1715–1722.

A FAST NULLSPACE METHOD BASED ON MATRIX EXPONENTIAL FOR THE UNSTEADY STOKES EQUATIONS[1]

Yang X.*

** Nanjing University of Aeronautics and Astronautics, Nanjing, China*

We consider the numerical solution of the unsteady Stokes equations modeling “low-speed” incompressible viscous flow as follows

$$\left\{ \begin{array}{rcl} \frac{\partial \vec{u}}{\partial t} - \nu \nabla^2 \vec{u} + \nabla p & = & \vec{f} \ \text{ in } \Omega \times [0, \mathrm{t}], \\ \nabla \cdot \vec{u} & = & 0 \ \text{ in } \Omega \times [0, \mathrm{t}], \end{array} \right. \tag{1}$$

where $\vec{u}$ is the velocity of fluid. p is the pressure of fluid. $\vec{f}$ is a given external force. $\Omega \subset \mathbb{R}^d$ $(d = 2,\ 3)$ is an open bounded domain. $\nu > 0$ is the kinematic viscosity. A boundary value problem is adding conditions to the system (1) on boundary $\partial\Omega = \partial\Omega_{\mathrm{D}} \cup \partial\Omega_{\mathrm{N}}$ as

$$\vec{u} = \vec{v} \text{ on } \partial\Omega_{\mathrm{D}}, \quad \nu \frac{\partial \vec{u}}{\partial \vec{n}} - \vec{n} p = 0 \text{ on } \partial\Omega_{\mathrm{N}},$$

where $\vec{n}$ is the outward-pointing normal to the boundary.

Semi-discretization in space of system (1) leads to a system of differential-algebraic equations (DAEs), i.e., the unsteady discrete Stokes equations

$$\left(\breve{\mathcal{B}} \frac{\mathrm{d}}{\mathrm{d}t} + \breve{\mathcal{A}} \right) \breve{\mathbf{x}} := \left\{ \begin{pmatrix} H & 0 \\ 0 & 0 \end{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} + \begin{pmatrix} \breve{A} & \breve{B}^T \\ \breve{B} & 0 \end{pmatrix} \right\} \begin{pmatrix} \breve{\mathbf{u}} \\ \breve{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \breve{\mathbf{f}} \\ \breve{\mathbf{g}} \end{pmatrix} := \breve{\mathbf{b}}, \tag{2}$$

where H and $\breve{A} \in \mathbb{R}^{n \times n}$, symmetric positive definite, represent velocity mass matrix and discrete diffusion, respectively. $\breve{B}^T \in \mathbb{R}^{n \times m}$ and $\breve{B} \in \mathbb{R}^{m \times n}$, full column rank and full row rank, represent discrete gradient and negative discrete divergence, respectively. $\breve{\mathbf{u}} \in \mathbb{R}^n$ and $\breve{\mathbf{p}} \in \mathbb{R}^m$ are the discrete velocity and pressure. $\breve{\mathbf{f}} \in \mathbb{R}^n$ and $\breve{\mathbf{g}} \in \mathbb{R}^m$ are constant forcing and boundary terms. An initial value problem is adding initial data $\breve{\mathbf{x}}(0) = \breve{\mathbf{x}}_0 \in \mathbb{R}^{n+m}$ to the system (2).

The unsteady discrete Stokes equations (2) is algebraically equivalent to a simple form through a block-diagonal scaling, say,

$$\mathcal{S} \ = \ \begin{pmatrix} L & 0 \\ 0 & I \end{pmatrix}$$

[1] Supported by the National Natural Science Foundation (No. 11101213, No. 11401305, No. 11571171 and No. 11701409), P.R. China.

where $H = LL^T$ is the Cholesky factorization of the matrix H. The above mentioned simple form results from

$$\mathcal{S}^{-1}\left(\breve{\mathcal{B}}\frac{\mathrm{d}}{\mathrm{d}t}+\breve{\mathcal{A}}\right)\mathcal{S}^{-T}\mathcal{S}^{T}\breve{\mathbf{x}} = \mathcal{S}^{-1}\breve{\mathbf{b}},$$

or equivalently,

$$\left\{\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\frac{\mathrm{d}}{\mathrm{d}t} + \begin{pmatrix} L^{-1}\breve{A}L^{-T} & L^{-1}\breve{B}^T \\ \breve{B}L^{-T} & 0 \end{pmatrix}\right\}\begin{pmatrix} L^T\breve{\mathbf{u}} \\ \breve{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} L^{-1}\breve{\mathbf{f}} \\ \breve{\mathbf{g}} \end{pmatrix}.$$

For the simplicity of notations, we simply rewrite the above form as

$$\left(\mathcal{B}\frac{\mathrm{d}}{\mathrm{d}t}+\mathcal{A}\right)\mathbf{x} := \left\{\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\frac{\mathrm{d}}{\mathrm{d}t} + \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}\right\}\begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} := \mathbf{b}, \quad (3)$$

where $A = L^{-1}\breve{A}L^{-T}$, $B = \breve{B}L^{-T}$, $\mathbf{u} = L^T\breve{\mathbf{u}}$, $\mathbf{p} = \breve{\mathbf{p}}$, $\mathbf{f} = L^T\breve{\mathbf{f}}$, $\mathbf{g} = \breve{\mathbf{g}}$. In the sequel, we consider the numerical solution to the unsteady discrete Stokes equations (3).

The most frequently used methods for unsteady discrete Stokes equations (3) are time-stepping methods including Runge-Kutta methods [1] and linear multistep methods [1]. The basic idea of time-stepping methods is to adopt temporal discretization to (3) on a prescribed time-level-sequence, then the evaluation of solution $\mathbf{x}$ is needed on each time level. Therefore, a small time-step-size always leads to a large number of evaluations of solution $\mathbf{x}$, thus making the workload of time-stepping methods increasing intensively. In order to overcome the disadvantages of time-stepping methods with small time-step-size, we employ a class of nullspace methods whose workload are independent of the length of the time interval.

An idea to solve the unsteady discrete Stokes equations (3) is to eliminate the algebraic constraints and solve a reduced system of ordinary differential equations (ODEs). Specifically, we suppose that $Z \in \mathbb{R}^{n\times(n-m)}$ is a matrix with orthonormal columns spanning the nullspace [2, 3, 4, 5, 6] of the matrix B, i.e., $Z^TZ = I$ and $BZ = 0$. Thus, the columns of B^T together with the columns of Z become a basis of $\mathbb{R}^n$, say, $\mathbb{R}^n = \mathrm{span}\{B^T, Z\}$. Then the first block $\mathbf{u}$ of the solution $\mathbf{x}$ to (3) can be written as

$$\mathbf{u} \;=\; \widehat{\mathbf{u}}_B + \widehat{\mathbf{u}}_Z \quad (4)$$

where $\widehat{\mathbf{u}}_B$ is the component of $\mathbf{u}$ in $\mathrm{span}\{B^T\}$, specifically,

$$\widehat{\mathbf{u}}_B \;=\; B^T(BB^T)^{-1}\mathbf{g}, \quad (5)$$

and $\widehat{\mathbf{u}}_Z$ is the component of $\mathbf{u}$ in $\mathrm{span}\{Z\}$, specifically,

$$\widehat{\mathbf{u}}_Z \;=\; e^{-tP_ZA}P_Z\left(\mathbf{u}_0 - P_{AZ}\widetilde{\mathbf{f}}\right) + P_{AZ}\widetilde{\mathbf{f}}, \quad (6)$$

where $\widetilde{\mathbf{f}} = \mathbf{f} - AB^T(BB^T)^{-1}\mathbf{g}$, and $P_{AZ} = Z(Z^TAZ)^{-1}Z^T$ is called the scaled projection matrix (when $A = I$, P_{AZ} is reduced to the orthogonal projection matrix P_Z). The expressions of the projection matrices P_{AZ} and P_Z need to involve the nullspace basis Z explicitly, which is impractical. In actual implementation, we have an idea to avoid the explicit access to the nullspace basis Z. In addition, the second block $\mathbf{p}$ of the solution $\mathbf{x}$ to (3) is given by

$$\mathbf{p} \;=\; (BB^T)^{-1}(B\mathbf{f} - BA\mathbf{u}). \tag{7}$$

The formulas (4-7) lead to a nullspace method for the unsteady discrete Stokes equations (3) as follows.

Algorithm .1 The nullspace method:

1. Solve $BB^T\mathbf{u}_B = \mathbf{g}$ to get $\mathbf{u}_B$, and compute $\widehat{\mathbf{u}}_B = B^T\mathbf{u}_B$ which is the component of $\mathbf{u}$ in span$\{B^T\}$;
2. Compute the scaled projection term $P_{AZ}\widetilde{\mathbf{f}}$;
3. Compute the matrix exponential vector product $e^{-tP_ZA}P_Z\left(\mathbf{u}_0 - P_{AZ}\widetilde{\mathbf{f}}\right)$;
4. Compute the component $\widehat{\mathbf{u}}_Z$ of $\mathbf{u}$ in span$\{Z\}$ by (6);
5. Compute $\mathbf{u} = \widehat{\mathbf{u}}_B + \widehat{\mathbf{u}}_Z$;
6. Solve $BB^T\mathbf{p} = B\mathbf{f} - BA\mathbf{u}$ to get $\mathbf{p}$.

The main workload of the Algorithm .1 is the matrix exponential vector product $e^{-tP_ZA}P_Z\mathbf{y}$ with $\mathbf{y} = \left(\mathbf{u}_0 - P_{AZ}\widetilde{\mathbf{f}}\right)$, which can be reformulated to be an integral in complex plane, i.e.,

$$e^{-tP_ZA}P_Z\,\mathbf{y} \;=\; \frac{1}{2\pi\imath}\int_{\Gamma} e^s F(s)\,\mathrm{d}s = I_\Gamma$$

with $F(s) = (sI + tP_ZA)^{-1}P_Z\,\mathbf{y}$. Moreover, I_Γ can be approximated by a linear combination of the values of the integrand evaluated at a small number of complex numbers $\{s_k\}_{k=1}^N$ [7, 8, 9, 10, 11, 12],

$$I_\Gamma \;\approx\; -\sum_{k=1}^{N} w_k F(s_k). \tag{8}$$

Each evaluation of the integrand leads to solve a linear system,

$$(s_kI + tP_ZA)\;F(s_k) \;=\; P_Z\mathbf{y}, \quad k = 1, \ldots, N. \tag{9}$$

We remark that if the sizes of matrices A and B are quite small, it is acceptable to solve the shifted linear systems (9) by direct linear solvers, e.g., Gaussian elimination. Otherwise, iterative linear solvers would be considered, e.g., Krylov subspace methods [13]. In this work, we consider generalized minimum residual method (GMRES) for the linear systems (9). The convergence of GMRES for the linear systems (9) is established, which provide an upper bound on the convergence rate of the GMRES iterates.

The numerical tests show that the nullspace method, i.e., Algorithm .1, for the unsteady Stokes equations (3) is much faster than the frequently used time-stepping methods, i.e., BDF(p), since the nullspace method solves a much smaller number of linear systems.

Bibliography

1. U.M. Ascher, L.R. Petzold, Computer Methods for Ordinary Differential Equations and Differential–Algebraic Equations, SIAM, Philadelphia, 1998.

2. J.C. Dunn, Second–order multiplier update calculations for optimal control probelms and related large scale nonlinear programs, SIAM J. Optim. 3(1993) pp. 489–502.

3. D. James, Implicitly nullspace iterative methods for constrained least squares problems, SIAM J. Matrix Anal. Appl. 13(1992) pp. 962–978.

4. Ph. L. Toint, D. Tuyttens, On large–scale nonlinear network optimizations, Math. Program. Ser. B 48(1990) pp. 125–159.

5. T.F. Coleman, A. Verma, A Preconditioned Conjugate Gradient Approch to Linear Equality Constrained Minimization, Technical Report, Department of Computer Sciences, Cornell University, Ithaca, NY, 1998.

6. N.I.M. Gould, M.E. Hribar, J. Nocedal, On the solution of equality constrained quadratic programming problems arising in optimizaiton, SIAM J. Sci. Comput. 23(2001) pp. 1376–1395.

7. G. Meinardus, Approximation of Functions: Theory and Numerical Mehtods, Springer, Berlin, 1967.

8. W.J. Cody, G. Meinardus, R.S. Varga, Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems, J. Approximation Theory 2(1969), pp. 50–65.

9. L.N. Trefethen, M.H. Gutknecht, The Carathéodory–Fejér method for real rational approximation, SIAM J. Numer. Anal. 20(1983) pp. 420–436.

10. A.P. Magnus, Asymptotics and super asymptotics of best rational approximation error norms for the exponential function (the '1/9' problem) by the Carathéodory–Fejér method, in Nonlinear Methods and Rational Approximation, II, A. Cuyt et al., eds., pp. 173–185, Kluwer, Dordrecht, 1994.

11. M. López–Fernández, C. Palencia, A. Schädle, A spectral order method for inverting sectorial Laplace transforms, SIAM J. Numer. Anal. 44(2006) pp. 1332–1350.

12. J.A.C. Weideman, L.N. Trefethen, The exponentially convergent trapezoidal rule, SIAM Review 56(2014) pp. 385–458.

13. Y. Saad, Iterative Mehtods for Sparse Linear Systems (second ed.), SIAM, Philadelphia, 2003.

ON THE QUADRATIC EQUATIONS AND MATRIX FUNCTION[1]

Zhang X.

Department of Mathematics Science, Guizhou Normal University, Guiyang 550025, P.R. China

Abstract: The octonions $\mathbb{O}$ is an 8-dimensional algebra over the real numbers $\mathbb{R}$ with basis $\{1 := e_0, e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$. The multiplication law over $\mathbb{O}$ is non-associative and non-commutative. For an example:

$$(e_1(e_4+1))e_7 = e_2 + e_6, \qquad e_1((e_4+1)e_7) = -e_2 + e_6.$$

Solving linear equations over non-associative rings $\mathbb{O}$ is not a duplicate work over communitive or associative rings. For an example: In complex field,

$$x^2 + bx + c = (x + \tfrac{b}{2})^2 + c - \tfrac{b^2}{4} \tag{1}$$

always holds, thus $x^2 + bx + c = 0$ can convert to $y^2 = 1$ where $y = \pm\frac{(x+\frac{b}{2})}{\sqrt{(\frac{b^2}{4}-c)}}$. However, the equality is invalid in octonion algebra. We give the explicit expressions of octonionic quadratic equation:

$$x^2 + bx + c = 0,$$

$$x^2 + xb + c = 0,$$

$$x^2 + ax + xb + c = 0.$$

The preparation work is as follows:

First,we provide the representation of an octonion number. For any

$$a = a_0 + a_1e_1 + \cdots + a_7e_7 \in \mathbb{O},$$

it is correspondent to a vector in $\mathbb{R}^8$,

$$\phi(a) := \begin{pmatrix} a_0, & a_1, & a_2, & a_3, & a_4, & a_4, & a_6, & a_7 \end{pmatrix} \in \mathbb{R}^8.$$

The real part and the imaginary part are correspondent to

$$\phi_r(a) := \begin{pmatrix} a_0, & 0, & 0, & 0, & 0, & 0, & 0, & 0 \end{pmatrix} \in \mathbb{R}^8,$$
$$\phi_i(a) := \begin{pmatrix} 0, & a_1, & a_2, & a_3, & a_4, & a_5, & a_6, & a_7 \end{pmatrix} \in \mathbb{R}^8.$$

Then the caculation over $\mathbb{O}$ can be converted to $\mathbb{R}$.

Secondly, we cite other important work. Let B, E and D be real numbers such that

[1] Supported by national fund 11401125

(i) $E \geq 0$, and
(ii) $B < 0$ implies $B^2 < 4E$.
Then the real system

$$\begin{aligned} &T^3 + (B - 2N)T + D = 0 \\ &N^2 - (B + T^2)N + E = 0 \end{aligned}$$

has at most two solutions (T, N) satisfying $T \in \mathbb{R}$ and $N \geq 0$ as follows.
(a) If $D = 0, B^2 \geq 4E$, then $T = 0$, $N = (B \pm \sqrt{B}^2 - 4E)/2$.
(b) If $D = 0, B^2 < 4E$, then $T = \pm\sqrt{2\sqrt{E} - B}$, $N = \sqrt{E}$.
(c) If $D \neq 0$ and z is the unique positive root of the real polynomial $z^3 + 2Bz^2 + (B^2 - 4E)z - D^2$, then $T = \pm\sqrt{z}$, $N = (T^3 + BT + D)/2T$.
Thirdly, to solve the equations,we obtain the alternativity $(a^{-1}ba)^2 = a^{-1}b^2a$.

Finally, we get the following main results. The solutions of the octonionic quadratic equation $x^2 + bx + c = 0$ can be obtained by following formulas:

Case 1. If $b, c \in \mathbb{R}$ and $b^2 < 4c$, then

$$x = \frac{1}{2}(-b + q^{-1}\sqrt{4c - b^2}e_1 q),$$

where q is any nonzero octionion.

Case 2. If $b, c \in \mathbb{R}$ and $b^2 \geq 4c$, then

$$x = \frac{1}{2}(-b \pm \sqrt{b^2 - 4c}).$$

Case 3. If $b \in \mathbb{R}$ and $c \notin \mathbb{R}$, then

$$x = -\frac{b}{2} \pm \frac{r}{2} \mp \sum_{i=1}^{7} \frac{c_i e_i}{r},$$

where $c = c_0 + \sum\limits_{i=1}^{7} c_i e_i$ and $r = \sqrt{\frac{1}{2}\left(b^2 - 4c_0 + \sqrt{(b^2 - 4c_0)^2 + 16\sum\limits_{i=1}^{7} c_i^2}\right)}$.

Case 4. If $b \notin \mathbb{R}$, then

$$x = \frac{-b_0}{2} - (b' + T)^{-1}(c' - N),$$

where $b_0 = \mathrm{Re}(b)$, $b' = \mathrm{Im}(b)$, $c' = c - \frac{b_0}{2}(b - \frac{b_0}{2})$, and (T, N) is chosen as follows: let

$$B = |b'|^2 + 2(c')_0, \quad D = 2(\bar{b}'c')_0, \quad E = |c'|^2. \tag{2}$$

Then

(i) $T = 0$, $N = (B \pm \sqrt{B^2 - 4E})/2$ provided that $D = 0$, $B^2 \geq 4E$,

(ii) $T = \pm\sqrt{2\sqrt{E} - B}$, $N = \sqrt{E}$ provided that $D = 0$, $B^2 < 4E$,

(iii) $T = \pm\sqrt{z}$, $N = \frac{T^3+BT+D}{2T}$ and z is the unique positive root of the real polynomial $z^3 + 2Bz^2 + (B^2 - 4E)z - D^2$.

Note that not all octonionic quadratic equations can be solved by using above conclusions. For example, this method is invalid for the equation $xax+bx+c = 0$ and the following equations: for $0 \neq a$,

$$\begin{aligned} x^2a + bx + c = 0 &\quad \text{is equivalent to} \quad x^2 + (mx)n + k = 0, \\ ax^2 + bx + c = 0 &\quad \text{is equivalent to} \quad x^2 + m(nx) + k = 0, \\ x^2a + xb + c = 0 &\quad \text{is equivalent to} \quad x^2 + (xn)m + k = 0, \\ ax^2 + xb + c = 0 &\quad \text{is equivalent to} \quad x^2 + m(xn) + k = 0, \end{aligned}$$

The non-linear problems are also interesting in matrix equations. We focus on some quadratic matrix equations because it can solve some special matrices. Consider the relations between some **linear matrix function** and the **nonlinear matrix function**:

$$\Phi_1 = \{Y = A_r^- | AYA = A, YAY = Y\}, \qquad \Phi_2 = \{B^- | BXB = B\}.$$

The general solution to $BXB = B$ can be expressed as

$$X = X_0 - L_B V_1 - V_2 R_B,$$

where X_0 is a special solution of $BXB = B$, V_1 and V_2 are arbitrary matrices. The expression of reflexive g-inverses of A can be expressed as

$$A_r^- = (A^\dagger - L_A V_3)A(A^\dagger - V_4 R_A),$$

where V_3 and V_4 are arbitrary matrices. In this talk, we characterize the distribution of the solutions to some matrix equation. And discuss the relations between the sets

$$\Phi_3 = \{Y = A_r^- | AYA = A, YAY = Y\}, \qquad \Phi_4 = \{X | BXC = D\}.$$

The nontrivial generalization is also obtained. We get the relations between Φ_3 and

$$\Phi_n = \left\{ \sum_{i=1}^{n} B_i^- \middle| B_i B_i^- B_i = B_i \right\}.$$

The skill is simple and practical: Two matrices A and B with the same size are equal if and only if $r(A - B) = 0$. Two sets Φ_1 and Φ_2 have a common element if and only if

$$\min_{A\in\Phi_1, B\in\Phi_2} r\,(A - B) = 0.$$

$\Phi_1 \subseteq \Phi_2$ if and only if

$$\max_{A\in\Phi_1}\min_{B\in\Phi_2} r\,(A-B)=0.$$

We establish the ranks of matrix function

$$f(X_1,X_2,X_3,X_4)=D_1-A_1X_1-X_2B_1-(A_2-B_2X_3C_2)D_2(A_3-B_3X_4C_3),$$

where X_1, X_2, X_3 and X_4 are variable matrices. Discuss the relations between some matrix sets generated by the linear matrix function

$$\Phi_5=\{D_3-A_1X_1-X_2B_1\}$$

and the nonlinear matrix function

$$\Phi_6=\{D_1+(A_2-B_2X_3C_2)D_2(A_3-B_3X_4C_3)\}.$$

The contained and interacting relations between the above two sets are characterized. Then we obtain some applications:

To begin with, we characterize the relations of

$$\Phi_1=\{Y=A_r^-|AYA=A, YAY=Y\},$$

$$\Phi_2=\{X|BXB=B\}.$$

$$X=X_0-L_BV_1-V_2R_B,$$
$$Y=(A^\dagger-L_AV_3)A(A^\dagger-V_4R_A).$$

In addition, if we put

$$M_{n-1,n}=\begin{bmatrix} B_1 & 0 & \cdots & 0 & B_n\\ 0 & B_2 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & B_{n-1} & B_n \end{bmatrix},$$

$$N_{n-1,n}=\begin{bmatrix} B_1 & 0 & \cdots & 0\\ 0 & B_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & B_{n-1}\\ B_n & 0 & \cdots & B_n \end{bmatrix},$$

$$P_{n+1,n}=\left[\begin{array}{c} diag(B_1,B_2,\cdots,B_n)\\ \hline \begin{bmatrix} A & 0 & \cdots & 0 & A\end{bmatrix}\end{array}\right],$$

$$Q_{n,n+1} = \left[\begin{array}{c|c} diag(B_1, B_2, \cdots, B_n) & \begin{matrix} A \\ 0 \\ \vdots \\ A \end{matrix} \end{array} \right],$$

Define

$$\Phi_n = \left\{ \sum_{i=1}^{n} B_i^- \middle| B_i B_i^- B_i \right\},$$

$$\Phi_{10} = \{Y = A_r^- | AYA = A, YAY = Y\}.$$

Then (a) $\Phi_n \bigcap \Phi_{10} \neq 0$ if and only if

$$r \begin{bmatrix} 0 & Q_{n,n+1} \\ P_{n+1,n} & diag(B_1, \cdots, B_n, A) \end{bmatrix} - r(Q_{n,n+1}) - r(P_{n+1,n}) \leq 0,$$

(b) $\Phi_{10} \subseteq \Phi_n$ if and only if

$$r \begin{bmatrix} 0 & Q_{n,n+1} \\ P_{n+1,n} & diag(B_1, \cdots, B_n, A) \end{bmatrix} - 2r(A) - r(M_{n-1,n}) - r(N_{n,n-1}) \leq 0.$$

Invited talks

BIPARTITE GRAPHS IN THE PROBLEM OF NFA MINIMIZATION AND ALGORITHMS FOR CALCULATING THEIR NUMBER

Abramyan M.E.

Southern Federal University, Rostov-on-Don, Russia

1. The problem statement. The main stage of the state minimization algorithm for a nondeterministic finite automaton [1, 2, 3] is an analysis of the special relation # connecting the subsets of the state sets X and Y of two canonical automata constructed on the basis of the given nondeterministic finite automaton K [2, Section 3.3].

The relation # can be described as a bipartite undirected graph G whose edges connect the elements of the sets X and Y and, due to the properties of the relation #, none of the vertices of the graph G is isolated, and the sets of adjacent vertices are different for any two different vertices of G.

The relation # can also be specified as a matrix A, which is the adjacency matrix of the graph G. We assume that the set X corresponds to the rows of the matrix, and the set Y corresponds to its columns. The elements of the matrix A are logical values 0 (false) and 1 (true); the element A_{xy} is equal to 1 if the vertices x and y of the graph G are are in the relation #, i. e. $x\#y$ holds. Due to the properties of the relation #, the matrix A has the additional property (*): it does not contain 0-valued or identical rows and columns.

On the set of all matrices satisfying the specified properties, we can introduce the equivalence relation as follows. Matrices A and B are assumed to be equivalent if the matrix B can be obtained from the matrix A by swapping some of its rows and / or columns. Equivalent matrices match the same pair of canonical automata, for which the order of states is specified differently.

The problem is to determine the number of different pairwise nonequivalent $M \times N$-matrices defining different relations #.

2. Brute force algorithm. Lets start with describing a simple algorithm, in which a comparison is made for the equivalence of all possible matrices of the given size $M \times N$ (note that the total number of such matrices is 2^{MN}).

The input to the algorithm is the numbers M and N; it is assumed, for definiteness, that $M \leq N$.

First, sets of all possible permutations of order M and N are generated; each permutation of order K if stored as an array of K elements with values from 0 to $K-1$.

Then all matrices of $M \times N$ size are generated; it is enough to go through all the numbers from 0 to $2^{MN}-1$ assuming that the binary representations of these numbers correspond to the values of the matrix elements (0 or 1) in row-major order.

All created matrices, for which the condition (*) is fulfilled, are saved in the `List<BMatrix>` collection, where `BMatrix` is the class of Boolean-valued matrices.

The main method connected with matrix validation for equivalence is the `A.Equiv(B)` method, which checks matrices `A` and `B` of `BMatrix` type. We use the simplest algorithm, which iterates through all combinations of permutations of rows and columns of a matrix B, until a matrix is obtained that matches the matrix A (in this case, the metrod returns `true`), or until all combinations of permutations are enumerated (in this case, `false` is returned).

The final `FindEquivMatr(matrices)` method analyzes the `matrices` collection of `List<BMatrix>` type, selects nonequivalent matrices from it and returns the resulting collection of pairwise nonequivalent matrices. When checking for each matrix `A` from the `matrices` collection, the `A.Equiv(B)` method is called for all matrices `B`, which are already included in the resulting collection of nonequivalent matrices.

The described algorithm was implemented in C# and tested on matrices of small size; the results are shown in Table 1. The Time column shows the working time in seconds, the Matr column contains the number of all matrices (of the given size) that satisfy the condition (*), the Result column contains the number of pairwise nonequivalent matrices.

Table 1. The results of the brute force algorithm.

	Time	Matr	Result
3×3	0.042	174	8
3×4	0.070	840	10
4×4	48.144	24 360	66
4×5	8415.5	335 160	168

Although the number of pairwise nonequivalent matrices of 4×5 order is only three times as large as the corresponding number for matrices of 4×4 order, the number of matrices to be compared among themselves increased about 13 times, and the calculation time increased 175 times. It is clear that in this way it will not be possible to obtain results for larger matrices in a reasonable time. Therefore, we should consider various heuristics (see, for example, [4]) that accelerate the basic algorithm.

3. Using matrix properties. The main operation time of the basic algorithm is spent on pairwise equivalence checking of matrices from the source collection. In the case of nonequivalent matrices, the `A.Equiv(B)` method will iterate over all possible row and column permutations of matrix `B`, and only after that returns `false`. However, in some situations, the checking process can be accelerated by analyzing such matrix properties as the total number of `true`-valued elements in each row and each column. Two matrices may be equivalent only if they have the same set of such properties for the rows and for the columns (since these sets do not change with row and column permutations of the matrix).

To speed up the comparison, it is convenient to order a set of required properties for the rows and for the columns of a given matrix and then combine these two ordered sets into an `Info` array of integers (of size $M+N$). At the beginning of the `A.Equiv(B)` method, the `Info` arrays of the matrix `A` and `B` are compared, and if these arrays are different then the method immediately returns `false`.

Table 2 shows the working time of the modified algorithm for matrices of various size.

Table 2. The working time of the modified algorithm.

	Time
3×4	0.012
4×4	1.867
4×5	147.6

Thus, the processing time of 4×4 and 4×5 matrices accelerated by 26 times and by 57 times respectively. However, such acceleration is not enough for matrices of higher dimensions, due to the rapid growth of the number of matrices satisfying the condition (*). For example, there are 3 553 200 such matrices of 4×6 size and 15 198 120 matrices of 5×5 size .

4. Consideration of ordered matrices. Consider another heuristic that allows us to speed up the algorithm by reducing the number of matrices being analyzed. Since the belonging of a matrix to the same set of equivalent matrices does not depend on the order of its rows or columns, we can analyze only those matrices in which the rows and columns are lexicographically sorted in descending order of their elements. Examples of such ordered 3×4-matrices are as follows:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

These matrices satisfy the condition (*) and, moreover, each next row of them is “less” than the previous one, if we assume that the rows are compared element by element and that the value 0 (`false`) is less than the value 1 (`true`). In addition, the same ordering property holds for their columns.

It is clear that the search algorithm for pairwise inequivalent matrices is sufficient to implement only for matrices with the specified ordering property. It can even be assumed that all the various ordered matrices are pairwise nonequivalent and therefore it is sufficient to obtain the set of all ordered matrices to solve the main problem. Unfortunately, this assumption is incorrect. For example, it is easy to verify that two ordered matrices mentioned above are equivalent.

Nevertheless, it is obvious that the number of different ordered matrices with the property (*) significantly less than the total number of all such matrices of the same dimension.

As an algorithm for generating all ordered matrices, we use a variant of the backtracking algorithm with the main recursive function `step(i, j, val)`, where `i` and `j` define the indices of the matrix element being defined and `val` is its value. The first call to this function is of the form `step(0, 0, true)`, since the initial element of an ordered matrix cannot be `false`. Subsequent recursive calls to the step function are organized so that the matrix is filled in columns. Calling `step(i, j, false)` can be done anyway, calling `step(i, j, true)` can be done only if the element to be added maintains the ordering of the already constructed parts of the previous row and column. After filling in the next row or column, the method also checks that this row (column) satisfies additional properties, namely: it does not consist entirely of `false` values and it is not equal the previous row (column).

The “Order” heuristic, whish is based on the use of ordered matrices, can be combined with previously considered “Info” heuristic, which accelerates the matrix equivalence test due to the analysis of their additional `Info` property.

Table 3 shows the results of testing new modifications of the algorithm.

This table contains the work time of the algorithm (Time) and the total number of processed matrices (Matr). The number of found pairwise nonequivalent matrices is indicated in the first column after the matrix size. For comparison, the table also includes the early described results for the previous versions of the algorithm.

Table 3. The results of the various versions of the algorithm.

	Basic algorithm	“Info” heuristic	“Order” heuristic	“Info”+“Order” heuristics
3×4	Time=0.070	Time=0.012	Time=0.001	Time=0.031
10	Matr=840		Matr=19	
4×4	Time=48.144	Time=1.867	Time=0.187	Time=0.016
66	Matr=24 360		Matr=185	
4×5	Time=8415.5	Time=147.56	Time=6.926	Time=0.281
168	Matr=335 160		Matr=706	
4×6	–	–	Time=142.60	Time=5.117
282			Matr=1639	
5×5	–	–	Time=3950.9	Time=52.603
1394			Matr=9109	

5. Concusion. Thus, the use of two implemented heuristics allowed us to process matrices of higher dimension in an acceptable time.

With a further increase in the matrix dimension, new problems arise. For example, for matrices of size 6×6 and 7×7, the number of ordered matrices is equal to 1 271 091 and 505 051 770 respectively. It is clear that in such a situation one should not create the entire set of analyzed matrices, and then proceed to their pairwise comparison: it is necessary to immediately check each obtained matrix and save only the set of pairwise nonequivalent matrices already found. Given the exponentially growing number of different possible permutations, one can also implement a more “intelligent” version of matrix checking, which analyzes only permutations of rows and columns with the same number of `true`-valued elements. Of course, some acceleration can also be achieved by parallelizing the algorithm, in which the set of found matrices is distributed for analysis across several threads connected with different cores of a multi-core processor.

However, even with these modifications, one can expect that the analysis of matrices of order 8 will require some months of calculations.

Bibliography

1. *Melnikov B.* Once more on the edge-minimization of nondeterministic finite automata and the connected problems // Fundamenta Informaticae. 2010. Vol. 104. No. 3. P. 267–283.

2. *Melnikov B.F.* Regular languages and nondeterministic finite automata (in Russian). Moscow, RSSU Edition, 2018. 175 p.

3. *Melnikov B.* The complete finite automaton // International Journal of Open Information Technologies. 2017. Vol. 5. No. 10. P. 9–17.

4. *Hromkovic J.* Algorithmics for Hard Problems. Introduction to Combinatorial Optimization, Randomization, Approximation, and Heuristics. Springer, 2004. 538 p.

ABOUT SOME NEW APPROACHES TO TEXT SENTIMENT ANALYSIS AND TOXICITY ANALYSIS USING DEEP LEARNING METHODS

Abramyan M.E., Litovchenko D.E.

Southern Federal University, Rostov-on-Don, Russia

Natural language processing (NLP) is a branch of computer science, computer linguistic and artificial intelligence that deals with the interaction between computers and humans using any natural language. Along with the development of social networks and other Internet resources, the amount of content generated by users of these online services is increasing very quickly. Every minute, the number of reviews in online stores all around the world, news comments on news web resources, user's photos and videos on Facebook, Twitter, Instagram and other social networks is increasing at an enormous speed. That's why text data in the Internet cannot be processed manually, but it may contain important and useful information for society or commercial and non-profit organizations. Timely processing of information about disasters in a certain place can help save lives. Processed patient complaints allow specialists to find some similar symptoms that the patient didn't mention. Analyzed customer testimonials can help find the advantages and significant drawbacks of products and predict the outflow or influx of customers.

That is why the tasks from the field of natural language processing are extremely relevant and important problems of machine learning and computational linguistics. Representatives of this family of tasks are text sentiment analysis and text toxicity analysis, the results of which contain important information about the author's mood, his motives and further behavior. The paper considers two important natural language processing problems: text sentiment analysis and text toxicity analysis.

1. Sentiment analysis and toxicity analysis. Text sentiment analysis (tonality analysis) is a class of natural language processing tasks in computational linguistics that is necessary for automatic detection of the constructed system of emotions (for example, negative, neutral, positive) in the text and the emotional evaluation of the authors of texts in relation to one way or another it is said (mentioned) in a given text data set.

In general, sentiment (tonality) is a certain attitude of the author, expressed by emotions, to some object, reflected in the text. These objects are people, social (cultural, political) events, film releases, music albums and more. The emotional component, which is expressed at the level of a lexeme or fragment of text, is called lexical tonality. In the simplest case, the tonality of the entire text is defined as a certain function of the input text data and the combination rules of the units of this text.

In practice, the problem of analyzing the sentiment of a text is a non-trivial task, with which a number of difficulties are associated. Basically, these difficulties are justified by the ambiguity of natural languages, which often includes free word order in a sentence, words with the same spelling and different meanings (homonyms), spelling errors in words, mistypes, the use of sarcasm (irony), professional slang, jargon and various abbreviations.

This paper presents several practical methods for the implementation of this task, along with a comparison of the prediction accuracy at the moment for different approaches. It's important to say that the approaches described in this work are directly related to machine learning and deep learning, which involves the use of neural networks with a large number of neurons and internal connections between them. Also, some hypotheses were formulated and tested, allowing to increase the accuracy of the models in certain cases.

It is important to understand that at the moment there are no universal approaches that would allow achieving high accuracy on any text dataset of the chosen language with any stylistics of text, data volume, etc. Most models are aimed at analyzing the tonality of the text within a certain category of texts. A possible goal of implementing the approaches described in this task is, for example, their integration into the service, one of the components of which would be the ability to determine customer loyalty to the chosen company, relying on information from various available sources of textual data.

As for text toxicity analysis, a toxic message is a text snippet, compiled in one of natural languages and being disrespectful/offensive/rude/sarcastic to anyone or anything.

2. Generation of train datasets. With the increasing adoption of artificial intelligence (AI) by companies around the world across all industries, developing a strategy for machine learning is imperative to gain a competitive advantage. A key component of this strategy is the data used to train machine learning-based solutions. Machine learning is a form of AI that uses large datasets to teach computers how to respond to and act like humans — allows businesses to optimize operations, deliver better customer experiences, enhance security and more [1].

It is unbelievable hard to find good enough dataset in Russian containing enough examples to use it for deep learning algorithm. For sentiment analysis task it was decided to use “Twitter mokoron” dataset for binary sentiment classification created by Yulia Rubtsova [2], the largest one among datasets in the Russian language but still it lacks accuracy and contains a lot of outliers, It contains about 300000 marked (as positive/negative) texts from Twitter social network. For toxicity analysis task it was decided to create new dataset in the Russian language. It is impossible to collect and mark properly enough data during this research. Thus, it was decided to use “Kaggle Toxic Comment Classification Challenge Dataset” [3] (dataset in English) stored on open machine learning competition platform “Kaggle”. In the process, this dataset was partly translated into Russian using “Google Translate” service. Then some parts of

Rusentiment dataset [4] were added. Also it was pre-proccesed: clearing unnecessary characters, deleting stop-words, etc. With this approach, a new unique dataset (totally 10000 texts) was formed.

3. Words representation. Any neural network works with a set of vectors of real numbers, and not with text data or images, so the most important task is to convert the input data to a vector representation. The word word representation (word embedding) is a parameterized function that maps words from some natural language w to number vectors of a certain dimension, this function is defined as follows: $W : w \to R^n$. There are a lot of algorithms of creating vectors from words that will reflect the measure of similarity between words: Word2vec, FastText, Glove, etc.

The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention.

Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances. Those guesses can be used to establish a word's association with other words (e.g. "man" is to "boy" whereas "woman" is to "girl"), or cluster documents and classify them by topic. Those clusters can form the basis of search, sentiment analysis and recommendations in such diverse fields as scientific research, legal discovery, e-commerce and customer relationship management.

The output of the Word2vec neural net is a vocabulary in which each item has a vector attached to it, which can be fed into a deep-learning net or simply queried to detect relationships between words.

The vectors we use to represent words are called neural word embeddings, and representations are strange. One thing describes another, even though those two things are radically different.

One choice in the task of representing the words for these tasks is to teach the Word2vec model on textual data from the source dataset. Another option is to teach Word2Vec models on text data from a large text corpus with similar styling and text size. Then it is possible to use pre-trained vectors (in this case, on texts from the social network VK.com). The rationale behind using pre-trained word embeddings in natural language processing is very much the same as for using pre-trained convnets in image classification: we don't have enough data available to learn truly powerful features on our own, but we expect the features that we need to be fairly generic, i.e. common visual features or semantic features. In this case it makes sense to reuse features learned on a different problem.

In this work, all three options were tested and compared to each other.

4. Implementation using convolutional neural networks. A convolutional neural network (CNN) is a deep learning algorithm which can take in an input data, assign importance (learnable weights and biases) to various aspects/objects in the object and be able to differentiate one from the other.

Pre-processing algorithm required in a convolutional neural network is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics. The architecture of a CNN is analogous to that of the connectivity pattern of neurons in the human brain and was inspired by the organization of the visual cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. A collection of such fields overlap to cover the entire visual area.

In 2012, Convolutional Neural Networks made a breakthrough in computer vision (CV). The solution, based on neural networks with this architecture, took first place in the competition ImageNet 2012. Subsequently it turned out that convolutional neural networks have great potential in solving other problems, including the problems of processing natural languages, which are discussed in this work. The peculiarity of the architecture of such networks is that their work is similar to the mechanisms of the visual cortex of the brain, which can highlight the most important details in the input information flow, which is important not only for image processing, but also for text processing.

CNN model has been impelemented, compiled and trained using Keras library (with Tesorflow library in back-end). It was trained on 70% of each dataset (divided into 3 folds) and tested on 30% of each dataset.

5. Implementation using bidirectional long-short term memory neural networks with self-attention. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections that make it a “general purpose computer” (that is, it can compute anything that a Turing machine can). It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

In problems where all timesteps of the input sequence are available, bidirectional LSTM neural networks train two instead of one LSTM neural networks on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

Attention is the idea of freeing the encoder-decoder architecture from the fixed-length internal representation. This is achieved by keeping the intermediate outputs from the encoder LSTM from each step of the input sequence and training the model to learn to pay selective attention to these inputs and relate them to items in the output sequence. Put another way, each item in the output sequence is conditional on selective items in the input sequence.

6. Comparison of models. After training, all CNN and LSTM models (for sentiment and toxicity analysis, with and without pre-trained embeddings from VK social network) were tested on test dataset containing 30% objects of each dataset.

Test results are given in Tables 1 and 2, they contain a comparison of models for accuracy and f1 metrics. Table 1 is devoted to the text toxicity analysis (with and without pre-trained word embeddings), Table 2 is devoted to the text sentiment analysis (with and without pre-trained word embeddings).

Table 1. Toxicity analysis

Metrics	CNN	Bi-LSTM	CNN (PTE)	Bi-LSTM (PTE)
Accuracy (test)	73.07%	75.18%	83.13%	84.40%
F1-metrics (test)	73.44%	74.90%	82.52%	83.97%

Table 2. Sentiment analysis

Metrics	CNN	Bi-LSTM	CNN (PTE)	Bi-LSTM (PTE)
Accuracy (test)	71.43%	72.01%	73.07%	72.90%
F1-metrics (test)	70.07%	70.90%	73.44%	73.01%

7. Conclusion. LSTM neural networks mostly show results that are similar to CNN's results. In some cases, LSTM neural networks gain 1.5–3 percent prediction accuracy. Binary sentiment classification and markup errors affect the neural network's possibilities of generalizing predictions. The used method of automatic translation and marking of data manually allows you to expand the amount of data in Russian, which is so lacking for solving problems of natural languages processing.

Bibliography

1. *Carole L.* An Introduction to Machine Learning Training Data. 2018. URL: https://appen.com/blog/an-introduction-to-machine-learning-training-data/

2. *Rubtsova Y.* Russian tweets sentiment dataset. 2015. URL: https://study.mokoron.com/

3. Kaggle Toxic Comment Classification Challenge Dataset. 2018. URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

4. *Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A.* RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. Proceedings of COLING. 2018. URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

ABOUT SOME FEATURES OF THE HEURISTIC INPLEMENTATION OF BRANCH AND BOUND METHOD FOR THE PROBLEM OF NFA MINIMIZATION

Abramyan M.E.[*], Melnikov B.F.[**]

[*] *Southern Federal University, Rostov-on-Don, Russia*

[**] *Shenzhen MSU – BIT University, Shenzhen, China*

The paper considers the problem of state minimization of nondeterministic finite automata (NFA). This problem has been formulated in the 1960s (see, for instance, [1]). In 1993, it was proved to be NP-hard [2], and therefore it is an important task to describe heuristic algorithms for this problem, i. e. algorithms that yield an acceptable near-optimal (quasi-optimal) solution at a reasonable time. Among such algorithms, the so-called anytime algorithms [3] occupy a prominent place; they are usually based on iterative techniques and work in real time.

1. Problem statement. The NFA state minimization algorithm considered in this paper is based on the results of [4, 5]. Its main stage is the analysis of the special relation # that connects the subsets of the sets X and Y, corresponding to the states of two canonical automata constructed on the basis of the given nondeterministic finite automaton K [5, Section 3.3]. We consider the matrix A defining the relation # in assumption that the set X corresponds to the rows of the matrix, and the set Y corresponds to its columns. The elements of the matrix are logical values 0 (false) and 1 (true); the element $A[x, y]$ is equal to 1 if the relation $x \# y$ is satisfied). By virtue of the properties of the relationship #, the matrix A does not contain 0-valued or identical rows and columns.

A set of rows X_0 and columns Y_0 of the matrix A is called a *grid* if intersection of these rows and columns contains only 1-valued elements. The grid may include non-adjacent rows and/or columns of the given matrix. A grid is called a *complete grid* if it cannot be expanded by adding a new row or column.

If the set of complete grids includes all the 1-valued elements of the given matrix, then we will call this set a *cover of the matrix*, and the number of complete grids will be called a *size of the cover*. Our main problem is to find the matrix cover of a minimum size. As shown in [5, Chapter 6], it is possible (with certain additional conditions being satisfied) to construct a minimum finite automaton equivalent to the given automaton K using the minimal cover of the matrix A.

2. Description of the basic algorithm. The algorithm for solving the main problem formulated in Section 1 is based on the branch and bound method (BBM, see, for example, [3]). The implementation of the algorithm is made in C# 6.0 for the .NET Framework.

A detailed description of the basic algorithm with an overview of the corresponding C# classes was given in [6]. In this paper, the focus will be on additional heuristics (see Sections 3, 4), as well as the results of computational experiments (see Sections 5, 6). However, we first briefly describe the basic algorithm itself.

To generate complete grids, the `MakeGridRnd` method is used, in which some row and column are selected and then a loop starts, and attempts are made to expand the grid by adding a new row or column. The search for a suitable row begins with a randomly selected row, after which all rows are visited cyclically until a suitable one is found or all rows are analyzed. The matrix columns are processed in the same manner.

The main part of the basic algorithm is the creation and processing of a sequence of *subtasks*, each of which is one of the sets of the complete grids for the given matrix (this set of grids, perhaps, does not form the matrix cover yet). Along with the creation of a subtask collection named `subtasks`, a collection of complete grids named `grids` is being created, whose elements are used to generate new subtasks based on existing ones.

Two sets of complete grids are associated with each subtask: the first set (named `Yes`) contains grids that are already included in this subtask, the second set (named `No`) includes those that cannot be selected for this subtask and for all descendants of it. The `No` set allows to avoid duplication of subtasks when they are generated from existing ones.

New subtasks are created by splitting one of the existing subtasks into two new ones. This action is the *main step* of the basic algorithm and it is implemented as the `MainStep` method. Splitting is performed on the basis of the new complete grid, which, therefore, plays the role of *separating element* of the BBM algorithm. The grid is selected from the `grids` collection in such a way as to include the maximum number of new 1-valued elements of the given matrix.

For the first of two new subtasks, this grid is included in its `Yes` set, and for the second one, it is included in its `No` set. If the `grids` collection does not include grids that can be used as a separating element, then a new grid is added to this collection by means of the `MakeGridRnd` method described above.

For processing by the `MainStep` method, a subtask with a minimum *weight* is always selected. The weight of subtask characterizes the chances of building the next quasi-optimal solution on the basis of this subtask. During the implementation of the algorithm, an additional study was carried out to find the optimal coefficients of these three characteristics (see Section 5).

The second subtask of the new ones obtained in the `MainStep` method is always returned to the `subtasks` collection, and the first one is checked as follows. If this subtask is not yet a solution to the main problem, then it also returns to the `subtasks` collection, if this subtask is a solution to the main problem, then it is checked for optimality, and in the case when the subtask is optimal from the previously analyzed subtasks, it is considered to be the next quasi-optimal solution.

To launch an algorithm of quasi-optimal solution search, two methods were used: `StepRun(steps)` and `TimeRun(seconds)`. The `StepRun` method performs the number of calls of the `MainStep` method specified in the `steps` parameter. The duration of the TimeRun method (and thus the number of calls of the MainStep method) is determined by the seconds parameter as follows: the TimeRun method terminates if no new quasi-optimal solution is found within the specified number of seconds.

3. Subtask cutting off. A situation often arises when during building a subtask the number of grids added to it (that is, the size of the Yes set) becomes equal or larger than the size of the quasi-optimal solution already found. It is obvious that in this situation it is not worth analyzing this subtask (and, in particular, to generate new subtasks on its basis). Moreover, one can immediately terminate the processing of a subtask extracted from the subtasks collection if this subtask contains OptSize – 1 grids, where OptSize is the size of the quasi-optimal solution already found. The subtask should also be cut off if the size of the Yes set becomes OptSize – 1 after adding the next grid to it, and the subtask is not yet a solution to the main problem.

It should be emphasized that the cut-off algorithm is better in terms of memory usage; this is due to the fact that early termination of subtasks prevents their subsequent splitting into new subtasks and, ultimately, reduces the size of the collection of subtasks that are awaiting processing. In addition, due to the early cutting off, it is possible to process a larger number of subtasks with the same number of the MainStep method calls.

4. Additional randomization. The new subtasks are created by means of a set of complete grids generated by the MakeGridRnd method using a random number generator. If the other seed values of the random number generator will be used then the another initial set of grids can be created, on the basis of which a smaller quasi-optimal solution can be obtained.

The idea related to the additional randomization of the algorithm was used for its following modification: the BigStep method was added to the algorithm that implements a new "big step" level, on which the main steps are performed (by the MainStep method). The main feature of the big step is that a complete clearing of both the subtasks collection and the grids collection of existing complete grids is performed at the beginning of each big step.

The BigStep method consists of three stages. On the first of them, initialization actions are performed: the grids and subtasks collections are cleared, a random number generator is created with a new seed value (which is greater by 1 than the previous one), and an initial empty subtask is created.

On the second stage, the first quasi-optimal solution is built on the basis of a new set of complete grids. At this stage, no more than OptSize + 1 calls are made to the MainStep method, where OptSize is the size of the best quasi-optimal solution obtained in the previous big steps. This is due to the previously noted fact that the first constructed pseudo-optimal solution can be further improved only in small limits. If the first solution is not received for the specified number of the MainStep method calls, then its construction is interrupted and a new big step begins.

If some solution is built at the second stage, then one can try to improve it by executing an *additional number* of the `MainStep` method calls. This is the third stage of a big step.

The `StepRun(steps)` method for this "BigStep" modification still terminates as soon as the number of the `MainStep` method calls reaches the value of the steps parameter (regardless of the big step stage, during which this happened). The check of the completion of the `TimeRun` method is performed after the completion of each big step.

5. Subtask weight calculation. As noted in Section 2, the choice of the next subtask from the collection of available subtasks is based on its weight, which characterizes the chances of building the next quasi-optimal solution on the basis of this subtask (the lower the weight, the greater the chances). Therefore, the optimal determination of the subtask weight is one of the conditions for the efficiency of the algorithm.

In the implemented algorithm, the weight of a subtask includes three characteristics: (1) the number of true-valued elements of the given matrix, which are not included in the grids of the subtask, (2) the size of the `Yes` set of the subtask, (3) the size of the `No` set of the subtasks. Indeed, a small value of the characteristic 1 indicates that the subtask can be quickly completed, a small value of the characteristic 2 means that the size of the obtained solution would be small, and a small value of the characteristic 3 means that there are more available grids to complete the solution.

Numerical analysis shows that when calculating the weight, it is necessary to take into account all three characteristics, and the best results are achieved in the case when the characteristics 1 and 2 are taken with a multiplier equal to 1, and the characteristic 3 is taken with a multiplier greater than 1. The results of computational experiments given in this paper (see Section 6) were obtained for coefficients 1, 1, 10.

6. Results of computational experiments. Computational experiments related to the basic algorithm and its modifications were performed for a set of 100 matrices of size 30 by 40, which were randomly generated by adding 35 initial grids. Each of these matrices was processed using the following methods (see Section 2): `StepRun(500)`, `StepRun(5000)`, and `TimeRun(10)`. Table 1 shows the average value of size of the best quasi-optimal solution for 100 processed matrices and (in brackets) the minimum and maximum values. The last three rows of the table show the results of three versions of the modification of the cut-off algorithm with additional "BigStep" randomization (see Section 4), in which N denotes the number of the MainStep method calls performed at the third stage of the big step.

The calculations were carried out on a computer with an AMD A10-6700 processor (3.70 GHz). The average time of the modification BigStep, $N = 100$, is 0.08 seconds for `StepRun(500)`, 0.90 seconds for `StepRun(5000)`, and 15.02 seconds for `TimeRun(10)`.

The results of computational experiments indicate that the implementation of the algorithm described in Section 2 gives good near-optimal solutions if it is

Table 1. The size of the best quasi-optimal solution: *average* (*min–max*).

	StepRun(500)	StepRun(5000)	TimeRun(10)
The basic algorithm	54.63 (44–69)	51.49 (41–66)	49.09 (40–63)
The cut-off algorithm	55.61 (44–71)	52.12 (43–66)	49.02 (39–60)
BigStep, $N = 25$	53.03 (43–67)	49.14 (40–60)	45.90 (38–56)
BigStep, $N = 100$	53.00 (43–67)	48.27 (39–59)	45.06 (37–54)
BigStep, $N = 500$	–	47.69 (39–59)	44.57 (36–54)
BigStep, $N = 1000$	–	47.31 (39–60)	44.23 (36–54)

provided with additional randomization (due to the “big step” level) and cutting off subtasks at the early stages of their generation.

Bibliography

1. *Aho A.V., Ullman J.D.* The theory of parsing, translation, and compiling. Vol. 1, Parsing. Prentice Hall, 1972. 613 p.

2. *Jiang T., Ravikumar B.* Minimal NFA problems are hard // SIAM J. Comput. 1993. Vol. 22. No. 6. P. 1117–1141.

3. *Melnikov B.F.* Multiheuristic approach to discrete optimization problems // Cybernetics and Systems Analysis. 2006. Vol. 42. No. 3. P. 335–341.

4. *Melnikov B.* Once more on the edge-minimization of nondeterministic finite automata and the connected problems // Fundamenta Informaticae. 2010. Vol. 104. No. 3. P. 267–283.

5. *Melnikov B.F.* Regular languages and nondeterministic finite automata (in Russian). Moscow, RSSU Edition, 2018. 175 p.

6. *Abramyan M.E.* On one approach to implementing the branch and bound method for optimization problems (in Russian) // Information Technologies in Modeling and Management: Approaches, Methods, Solutions. Proceedings of II Scientific Conference (April 22–24, 2019). Part 1. Togliatti, 2019. P. 56–64.

FOURIER ANALYSIS OF MULTIGRID METHOD WITH HSS METHODS AS SMOOTHERS[1]

Andreeva E.M., Muratova G.V.

Southern Federal University, Rostov-on-Don, Russia

The local Fourier analysis of multigrid method with a HSS smoothers is considered.The skew-Hermitian splitting iteration methods is effective to solve non-Hermitian positive definite linear systems. HSS methods have been used as the smoothers of the multigrid method for the solution of linear algebraic equation systems with a strongly nonsymmetric matrix obtained after difference approximation of the convection-diffusion equation with dominant convection.

I Introduction

Multigrid methods are proving themselves as very successful tools for the solution of the algebraic equation systems associated with discretization of boundary-value problems. MGM is not a fixed multigrid algorithm. There is rather a multigrid technique fixing only the framework of the algorithm. The efficiency of the multigrid algorithm depends on the adjustment of its components to the problem in question [1, 2]. Important components of multigrid method are a smoothing procedure or basic iterative method and the coarse-grid correction.

For nonelliptic and nonsymmetrical problems (or there combinations) a strict mathematical theory, generally speaking, is not available, and the correct choice of multigrid components is far from standard approach. In such cases, Fourier analysis is the main tool for quantitative estimates of MGM convergence. According to the basic idea of Fourier analysis the error (or residual) can be decomposed into a sum of some periodic functions called Fourier components.

In this paper we consider a local Fourier analysis (LFA). In LFA the basic discrete operators with constant coefficients are considered to be formally extended to infinite grid. Consequently, the boundary conditions are neglected. So, according to general assumptions, any discrete operator with, nonlinear, nonregular coefficients can be locally linearized and locally replaced by (freezing coefficients) by the operator with constant coefficients. This approach demonstrates the wide range of LFA applicability and its local nature.

The simplest version of the local Fourier analysis is smoothing analysis. In this case we research the procedure of smoothing. The effect of the coarse-grid correction is neglected or in other words we consider the "ideal" operator of coarse-grid correction.

For detail understanding the concept and structure of the multigrid method we should research two-grid LFA. This analysis gives more information than the

[1] This research was supported by the Russian Foundation for Basic Research under Grants N 19-51-53013 GFEN-a

analysis of smoothing. In this paper a Fourier analysis of multigrid method with a skew-Hermitian splitting iteration methods as smoothers is considered. We used the technique of Fourier analysis presented in [3].

The one-grid local Fourier analysis (or smoothing analysis) and the two-grid Fourier analysis (LFA) were used for the suggested MGM modification.

II Model Problem

We consider the model problem of the steady-state convection-diffusion process with dominant convection in domain $\Omega = [0,1] \times [0,1]$:

$$\begin{array}{ll} \frac{1}{2}\left(\sum\limits_{k=1}^{2} v_k(\mathbf{x})\frac{\partial u(\mathbf{x})}{\partial x_k} + \frac{\partial (v_k(\mathbf{x})u(\mathbf{x}))}{\partial x_k}\right) - \frac{1}{Pe}\sum\limits_{k=1}^{2}\frac{\partial}{\partial x_k}\left(\frac{\partial u(\mathbf{x})}{\partial x_k}\right) = f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega, \end{array} \tag{1}$$

where $\mathbf{x} = (x_1, x_2)$. Equation (1) has a small parameter at the highest derivative. We consider incompressible environments so $div\vec{V} = \sum\limits_{k=1}^{2} \frac{\partial v_k}{\partial x_k} = 0$.

FDM with central differences was used for discretization of (1). We obtain a non-Hermitian and positive definite matrix, the HSS iteration method was proposed to effectively solve this class of linear systems [4].

$$L_h u_h = f_h, \tag{2}$$

considered on a grid

$$G_h = \{(x_1, x_2) : x_k = ih, h = 1/n, n \in \mathbb{N}, k = 1, 2, i \in \mathbb{Z}\},$$

u_h and f_h are grid functions on G_h, and L_h - linear operator

$$L_h : E(G_h) \to E(G_h),$$

where $E(G_h)$ - linear space of grid functions defined on G_h.

The basic strategies in the construction of the HSS iteration method is matrix splitting and stationary iteration. Any non-Hermitian matrix L can be decomposed into its Hermitian and skew-Hermitian parts as

$$L = L_0 + L_1,$$

where $L_0 = \dfrac{1}{2}(L + L^*) = L_0^* > 0$ and $L_1 = \dfrac{1}{2}(L - L^*) = -L_1^*$.

Present the matrix L_1 as:

$$L_1 = K_l + K_u \quad \text{and} \quad K_u = -K_l^*,$$

where K_l and K_u – respectively, lower and upper triangular parts of skew-symmetric matrix L_1.

To solve the system (2) we suggest using MGM, where the skew-Hermitian splitting iteration methods is used as the smoothers of MGM [5].

Consider the structure of HSS methods. Any iterative method can be written in canonical form:

$$B\frac{u_{n+1}-u_n}{\tau}+Lu_n=f, \quad n=0,1,2,... \tag{3}$$

The choice of operator B defines the class of skew-Hermitian splitting iteration methods. For a standard HSS methods operator B is constructed as follows:

$$B=E+2\tau K_l \quad \text{or} \quad B=E+2\tau K_u, \tag{4}$$

for TIM1:

$$B=\alpha E+2K_l \quad \text{or} \quad B=\alpha E+2K_u, \tag{5}$$

for TIM2:

$$B=B_c+2K_l \quad \text{or} \quad B=B_c+2K_u, \tag{6}$$

where τ – scalar iteration parameter, $\alpha=\|M\|$, elements of $B_c=\{b_{c_{ii}}\}_0^n=\sum\limits_{j=0}^{n}|m_{ij}|, \quad i=0,...,n,$ where $M=\{m_{ij}\}_0^n$ – symmetric matrix, which is constructed by $M=L_0+K_u-K_l$, n – the dimension of the matrix L.

Any method from this class behaviors in the same way as Gauss-Seidel one: it quickly reduces the high-, but not low- frequency components of error frequencies. This is the necessary property of the smoother of MGM, that's why we have used these methods as the smoothers [6].

The convergence of proposed MGM modifications with skew-Hermitian splitting methods as smoothers is researched [7].

III The basic principles of Fourier analysis

The basic idea of the Fourier analysis is to present the solution error or residual as the sum of some periodic functions named Fourier components or harmonics. Fourier analysis gives us a possibility to estimate influence of MGM components on the every Fourier-decomposition component.

In this paper the one-grid local Fourier analysis or the smoothing analysis and the two-grid Fourier analysis (LFA) were used for the research of MGM modification with HSS-smoothers [3]. The one-grid analysis allows to estimate the efficiency of the smoothing procedure. The asymptotic convergence factors are considered for the convergence estimation of the multigrid method.

LFA goal is to determine the smoothing coefficients and two-grid convergence coefficients. Denote these coefficients as $\mu_{loc}(S_h)$, $\rho_{loc}(M_h^{2h})$, where $M_h^{2h}=S_h^{\nu_2}K_h^{2h}S_h^{\nu_1}$ – transition operator of two-grid method, $K_h^{2h}=I_h-P_{2h}^{h}L_{2h}^{-1}R_h^{2h}L_h$

– operator of the coarse-grid correction, S_h – transition operator of smoothing method, P^h_{2h} – prolongation operator, $R^{2h}_h L_h$ – restriction operator.

We consider fine grid G_h and coarse grid G_{2h}, obtained from G_h standard coarsening (ie, increasing the mesh size by half for all directions) $G_{2h} = \{\mathbf{x} = (x_1, x_2), x_k = i2h, k = 1, 2, i \in \mathbb{Z}\}$. We also assume that the operators L_h, R^{2h}_h, L_{2h} and P^{2h}_h are represented by patterns for G_h and G_{2h}.

The smoothing factors have been received for an estimation of smoothing property of the MGM -modification with triangular skew-symmetric smoothers. The asymptotic convergence factors are numerically received for the convergence estimation of the two-grid method with HSS smoothers. Fourier analysis results allow to compare MGM-modifications with different smoothers such as Gauss-Seidel method, Jacobi method and suggested HSS-smoothers [7].

IV Fourier smoothing analysis

The results of Fourier smoothing analysis, for MGM with different smoothers: Gauss-Seidel method, Jacobi, HSS iteration methods with large Peclet numbers and the coefficients of the convective term $v_1 = 1$, $v_2 = 1$ are presented in table 1.

Table 1. The smoothing factors μ_{loc}.

Pe	$TIM1-2$	TIM	$\omega - Jac$	$GS - LEX$
1000	0.8875	0.8762	0.9983	0.9999
10000	0.9876	0.9873	0.9999	0.9999
100000	0.9987	0.9987	0.9999	0.9999

V Two-grid analysis

We perform Fourier two-grid analysis, so that the effect of the coarse grid correction and the transfer operators is taken into account.

We calculate $\rho_{loc}(M^{2h}_h)$, using the technique from [3].

The results of Fourier two-grid analysis with various smoothers for large Peclet numbers and the coefficients of the convective terms of the $v_1 = 1$, $v_2 = 1$ are presented in Table 2. The table 2 contains of the rate of asymptotic convergence two-grid method, ρ_{loc} for various numbers of smoothing iterations of Ns. Symbol $-$ means that the value of the coefficient $\rho_{loc} \gg 1$.

VI Conclusions

The numerical results show that the smoothing factor μ for Gauss-Seidel and Jacobi methods is close to unity. For skew-Hermitian splitting methods

Table 2. The asymptotic convergence factors ρ_{loc}.

Pe	Ns	$TIM1-2$	TIM	$\omega-Jac$	$GS-LEX$
1 000	5	0.6782	0.5692	2.6591	–
	10	0.3673	0.3240	2.6241	–
	15	0.2035	0.1844	2.5895	–
10 000	5	0.9787	0.9434	3.7215	–
	10	0.9152	0.8901	3.7211	–
	15	0.8582	0.8397	3.7207	–
100 000	5	>1	0.9713	3.762821	3.7383
	10	>1	0.9435	3.762817	3.7139
	15	0.9970	0.9164	3.762813	3.6897

smoothing coefficient factor μ is less than unity for all these cases. This fact demonstrates the efficiency of HSS methods as MGM smoothers for convection-diffusion equation with dominant convection.

Two-grid Fourier analysis allow us to define the optimal number of smoothing iterations for efficient solving problems with large Peclet numbers.

According to the Fourier analysis results we can conclude that the suggested skew symmetric methods are effective MGM smoothers for solving the convection-diffusion problem with dominant convection.

Bibliography

1. W. Hackbusch. *Multigrid method and application.* – Springer - Verlag, Berlin, 1985.

2. P. Wesseling. *An introduction in multigrid methods.* – Wiley, Chichester, 1992.

3. U. Trottenberg, C.W. Oosterlee, A. Schuller *Multigrid.* – Academic Press, New York, 2001.

4. Z.-Z. Bai, G.H. Golub, M.K. Ng *Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems.* SIAM J. Matrix Anal. Appl. 24 (2003) pp.603-626.

5. Krukier L.A., Chikina L.G., Belokon T.V. *Triangular skew-symmetric iterative solvers for strongly nonsymmetric positive real linear system of equations.* Appl. Num. Math., 41 (2002). pp. 89-105.

6. Galina V. Muratova, Evgeniya M. Andreeva. *Multigrid method for solving convection-diffusion problems with dominant convection.* Journal of Computational and Applied Mathematics 226. - 2009. - pp.77-83.

7. Krukier L.F., Muratova G.V., Andreeva E. M. *Multigrid method of the solution of problems of a convection-diffusion with a dominant convection.* Publishing house of Southern federal university, Rostov-on-Don, 2011, 195p.

THE SECOND ORDER DIFFERENCE SCHEME OF THE CONVECTION-DIFFUSION PROBLEM WITH THE ROBIN'S BOUNDARY CONDITIONS[1]

Chikina L.G.*, Shabas I.N.*, Chikin A.L.**

** Southern Federal University, Rostov-on-Don, Russia,*

*** Southern Scientific Centre of The Russian Academy of Sciences, Rostov-on-Don, Russia*

Convection-diffusion equations are widely used in modeling a diverse range of problems. These mathematical models consist in a partial differential equation or system with initial and boundary conditions, which depends on the phenomena being studied. In the modeling, boundary conditions may be neglected and unnecessarily simplified, or even is misunderstood, causing a model does not reflect the reality adequately, making qualitative and/or quantitative analysis more difficult.

Numerical simulations, by means of the Finite Difference Method, are used in order to exemplify the boundary conditions' impact.

Considering an initial-boundary value problem in $u = u(x,t)$, $x = x_1, ..., x_n) \in R_n$, we have three different boundary condition types: Dirichlet, Neumann and Robin, each of these being separated in homogeneous, if it does not involve values beyond u, or non homogeneous, if it does.

Let u is the solution of the boundary value problem, f, g, and h are arbitrary functions and a and b arbitrary parameters, which may or may not depend on (x,t). A Dirichlet, or first kind, condition specifies the value of u on the boundary, either being zero or any function f that may depend or not on other variables. A Neumann, or second kind, condition, on the other hand, specifies the derivative of the solution u along the boundary, more precisely the directional derivatives in the direction of the external unitary normal vector n. A Robin, or third kind, condition involves both the value of u and its derivative, specifying an equation that must be valid along the boundary.

We consider only one spatial variable (x). All analyzes are similar for two, three or even n-dimensional tasks.

We introduce the steady-state convection-diffusion equation

$$-\frac{d}{dx}(\mu\frac{d}{dx}) + v\frac{d}{dx} + \beta s = f, \quad for \quad x \in (0,l), \tag{1}$$

[1] Supported by Ministry of Science and Higher Education of the Russian Federation (basic part, project N1.5169.2017/8.9) and by the Russian Foundation for Basic Research: grant 18-05-80010 "Dangerous phenomena"

with Robin's boundary conditions

$$\xi(0)\frac{ds}{dx} + \chi(0)s = r(0), \tag{2}$$

$$\xi(l)\frac{ds}{dx} + \chi(l)s = r(l), \tag{3}$$

where $s(x)$ is the unknown function, $s(x), \xi(x), \chi(x)$ are given sufficiently smooth functions. The term $u(x)$ can represent vorticity, temperature, mass concentration, or other physical quantities that are transferred inside the physical system by convection and diffusion. In the following, $v = v(x)$ stands for the convection velocity while $\mu = \mu(x) > 0$ represents the diffusion coefficient.

The interior spatial points. To develop the finite difference scheme, the interested domain is discretized into grid points for numerical calculation. These grid points are labeled sequentially in space as $x_0, x_1, ..., x_N$. For a uniform grid with space step size h, the discrete grid points within a given spatial domain [0, l] are calculated as:

$$x_i = ih, i = 0, 1, ..., N, \quad \text{where} \quad h = l/N.$$

The quantity $s(x_i)$ represents the exact solution at (x_i). The term s_i is used to represents the numerical solution at (x_i). The discretization in space is required to obtain a system of equations for the nodal values of the approximate solution. Of course, the number of equations should be the same as the number of unknowns.

For example, the discretized CDR equation (1) for the nodal value s_i can be written as

$$L_{Dh}s_h + L_{Ch}s_h + \beta s_h = f_h, \tag{4}$$

where $s_h = \{s_i\}$ denotes the mass matrix, $L_{Ch} = \{L_{Cij}\}$ is the discrete transport operator, $L_{Dh} = \{L_{Dij}\}$ is the discrete diffusion operator, and $f_h = \{f_i\}$ is the vector of discretized source or sink terms.

As a rule, the matrices L_C, and L_D are sparse. That is, most of their entries are equal to zero and do not need to be stored. The sparsity pattern of the discrete operators depends on the type of the underlying mesh (structured or unstructured) and on the numbering of nodes. Ideally, the discrete diffusion operator L_D should also be symmetric, as required by the properties of its continuous counterpart . The discrete convection operator L_C is nonsymmetric since the flow direction must be taken into account. For example, this matrix can be skew-symmetric $(C = -C^T)$ or upper/lower triangular.

The finite difference method (FDM) is the oldest among the discretization techniques for partial differential equations. Many modern numerical schemes for transport phenomena trace their origins to finite difference approximations developed in the late 1950s through early 1980s.

Taylor series expansions or polynomial fitting techniques are used to approximate all space derivatives in terms of s_i and/or solution values at a number of neighboring nodes. Higher order approximations to the first derivative can be obtained by using more Taylor series, more terms in the Taylor series, and appropriately weighting the various expansions in a sum.

$$f(x+h) = f(x) + hf^{'}(x) + h^2\frac{f^{''}(x)}{2!} + h^3\frac{f^{'''}(x)}{3!} + h^4\frac{f^{(4)}(x)}{4!} + h^5\frac{f^{(5)}(x)}{5!} + ... \qquad (5)$$

$$f(x-h) = f(x) - hf^{'}(x) + h^2\frac{f^{''}(x)}{2!} - h^3\frac{f^{'''}(x)}{3!} + h^4\frac{f^{(4)}(x)}{4!} - h^5\frac{f^{(5)}(x)}{5!} + ... \qquad (6)$$

$$f(x+2h) = f(x) + 2hf^{'}(x) + 4h^2\frac{f^{''}(x)}{2!} + 8h^3\frac{f^{'''}(x)}{3!} + 16h^4\frac{f^{(4)}(x)}{4!} + 32h^5\frac{f^{(5)}(x)}{5!} + ... \qquad (7)$$

$$f(x-2h) = f(x) - 2hf^{'}(x) + 4h^2\frac{f^{''}(x)}{2!} - 8h^3\frac{f^{'''}(x)}{3!} + 16h^4\frac{f^{(4)}(x)}{4!} - 32h^5\frac{f^{(5)}(x)}{5!} + ... \qquad (8)$$

Forward difference approximation with second order error $(7) - 4 * (5)$

$$f^{'}(x) = \frac{-f(x+2h) + 4f(x+h) - 3f(x)}{2h}.$$

Backward difference approximation with second order error $(8) - 4 * (6)$

$$f^{'}(x) = \frac{f(x-2h) - 4f(x-h) + 3f(x)}{2h}$$

and centered difference approximations are (5)+ (6)+ (7)+ (8)

$$f^{''}(x) = \frac{f(x-2h) + f(x-h) - 4f(x) + f(x+h) + f(x+2h)}{5h^2}.$$

The discretized CDR equation (1) for the nodal value $s_i(h = h_x)$ can be written as

$$-\frac{a_{i-2}s_{i-2} + a_{i-1}s_{i-1} - 4a_is_i + a_{i+1}s_{i+1} + a_{i+2}s_{i+2}}{5h_x^2} +$$
$$+v^+\frac{3s_i - 4s_{i-1} + s_{i-2}}{2h_x} + v^-\frac{-3s_i + 4s_{i+1} - s_{i+2}}{2h_x} + \beta s_i = f_i, \qquad (9)$$

where $v^+ = \frac{1}{2}(u_i + |u_i|), \quad v^- = \frac{1}{2}(v_i - |v_i|)$. Robin's boundary conditions (2)-(3):

$$\frac{-3\xi_0 + 2\chi_0 h_x}{2h_x}s_0 + \frac{2\xi_0}{h_x}s_1 - \frac{\xi_0}{2h_x}s_2 = r_0, \qquad (10)$$

$$\frac{3\xi_N + 2\chi_N h_x}{2h_x}s_N + \frac{2\xi_N}{h_x}s_{N-1} + \frac{\xi_N}{2h_x}s_{N-2} = r_N. \qquad (11)$$

Five-point scheme (9)

$$B_iS_{i-2} + BC_iS_{i-1} + BD_iS_i + E_iS_{i+1} + FS_{i+2} = f_i. \qquad (12)$$

In scheme (12) with $i = 1$ there is a node -1 and with $i = N - 1$ there is a node $N + 1$. To exclude fictitious nodes, we ground the first derivative

S_{i-2}	$\frac{v^+}{2h_x} - \frac{a_{i-2}}{5h_x^2}$	B_i	$\leq 0(h_x \leq \frac{2a_{i-2}}{5v^+})$
S_{i-1}	$\frac{-4v^+}{2h_x} - \frac{a_{i-1}}{5h_x^2}$	C_i	≤ 0
S_i	$\frac{3v^+}{2h_x} - \frac{3v^-}{2h_x} + \frac{4a_{i-2}}{5h_x^2}$	D_i	≥ 0
S_{i+1}	$\frac{4v^-}{2h_x} - \frac{a_{i+1}}{5h_x^2}$	E_i	≤ 0
S_{i+2}	$\frac{v^+}{2h_x} - \frac{a_{i-2}}{5h_x^2}$	B_i	$\leq 0(h_x \leq \frac{2a_{i+2}}{-5v^-})$

in the boundary conditions by central differences. The compact notation of the differential analogue of the third kind of boundary condition (2)-(3) will look like:

$$g_\Theta \check{s} + p_\Theta \hat{s} + q_\Theta \tilde{s} = r_\Theta, \tag{13}$$

where $\check{s}, \hat{s}, \tilde{s}$ are the concentrations of a substance, respectively, in some boundary and two corresponding border nodes, the symbol Θ denotes a five-point template node. Express in (13) the boundary nodes through the border and substitute them in (12).

Contributions of the boundary conditions entered the coefficients of the difference scheme pattern, violating the positivity of the elements of the main diagonal and the negativeness of the side diagonals, and hence the diagonal dominance of the SLAE matrix (12). This effect is not observed in the case of boundary conditions of the first kind.

Bibliography

1. *L. G. Chikina, I. N. Shabas and T. S. Martynova Simulation of the transfer process of multiphase substances in water // "JP Journal of Heat and Mass Transfer", Vol. 16 No. 1. 2019. P. 69–94*

POSITIVE DEFINITENESS AND M-MATRIX CONDITIONS OF A CONVECTION-DIFFUSION DIFFERENCE OPERATOR WITH BOUNDARY CONDITIONS OF THE THIRD KIND[1]

Chikina L.G., Shabas I.N., Martynova T.S.

Southern Federal University, Rostov-on-Don, Russia

I Introduction

A 3-dimensional stationary convection-diffusion equation is considered. Finite differences are used to approximate the first derivatives.

The properties of the difference convection-diffusion operator depend on the fact, what differences are – central or upwind – the convective terms are approximated. Thus, in the case of boundary conditions of the I kind, the central-difference approximation of the convective terms gives a positive definite difference operator if the symmetric form for the convective terms is chosen [1, 2], and the upwind approximation of the convective terms provides the M-matrix property, if the non-divergent form is chosen [3, 4]. But the presence of the III kind boundary conditions can violate these properties.

We investigate the influence of the boundary conditions of the III kind on the properties of the convection-diffusion difference operator. We need sufficient conditions for the positive definiteness and M-matrix property, since the presence of these properties in matrices arising after approximation of the problem significantly affects the convergence of most iterative methods.

II Formulation of the problem

A 3-dimensional stationary convection-diffusion equation in the domain $\bar{\Omega}$, $\bar{\Omega} = \Omega \cup \Gamma$ that describes the substance transfer process in an incompressible medium is considered

$$-\sum_{i=1}^{3} \frac{\partial}{\partial x_i}\left(\nu_i \frac{\partial S}{\partial x_i}\right) + \gamma \sum_{i=1}^{3} \frac{\partial}{\partial x_i}(v_i S) + (1-\gamma)\sum_{i=1}^{3} v_i \frac{\partial S}{\partial x_i} + \beta S = f, \qquad (1)$$

$$div\bar{\mathbf{v}} = 0, \qquad (2)$$

[1] Supported by Ministry of Science and Higher Education of the Russian Federation (basic part, project N1.5169.2017/8.9) and RFBR, grant N19-51-53013, GFENa

where $S = S(\mathbf{x})$ is the substance concentration; $\{\nu_i\}$ are coefficients of turbulent diffusion; βS is the source term of the equation; $\beta = \beta(\mathbf{x}) \geq 0$; $\bar{\mathbf{v}} = \{v_1, v_2, v_3\}$ is the velocity vector; γ is the parameter.

The convective terms in (1) can be written in the non-divergent form ($\gamma = 0$) and in the divergent form ($\gamma = 1$). The condition (2) allows to write them in the equivalent symmetric form ($\gamma = 1/2$) [1, 2].

The system (1)-(2) is complemented by boundary conditions

$$\mu(\mathbf{x})\frac{\partial S(\mathbf{x})}{\partial \bar{n}} + \chi(\mathbf{x})S(\mathbf{x}) = r(\mathbf{x}), \quad \mathbf{x} \in \Gamma, \tag{3}$$

where $\mu(\mathbf{x}), \chi(\mathbf{x}), r(\mathbf{x})$ are piecewise smooth functions. Depending on the equality of the zero functions $\mu(\mathbf{x}), \chi(\mathbf{x}), r(\mathbf{x})$ it is possible setting conditions of I, II and III kind.

III Finite difference approximation

The uniform grid $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$ with vector parameter $h = (h_x, h_y, h_z)$ has been introduced in the domain $\bar{\Omega}$. Here Ω_h is the set of internal grid nodes, Γ_h is the set of the boundary nodes.

When the system is approximated, it is necessary to preserve the properties of the original differential operators. Therefore, when approximating the equations (1), the convective terms of which are written in the symmetric form, a central-difference scheme is chosen, and for non-divergent form of the convective terms, the upwind scheme is chosen.

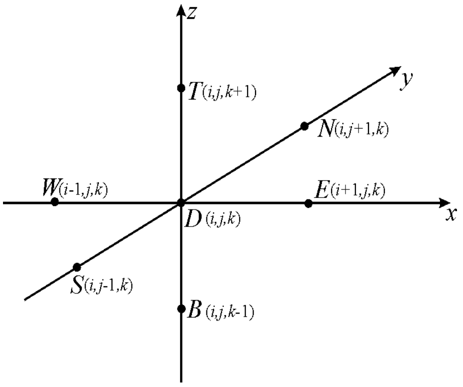


Figure 1. 7-point stencil

Consider, for example, the difference analogue on a 7-point stencil on the axis WE (Fig.1) boundary conditions of the III kind. It looks like this:

$$-\mu_W \frac{s_{1jk} - s_{0jk}}{h_x} + \chi_W s_{0jk} = r_W, \quad \mu_E \frac{s_{Njk} - s_{N-1jk}}{h_x} + \chi_E s_{Njk} = r_E,$$

or after the transformation

$$\frac{(\mu_{_W}+\chi_W h_x)}{h_x}s_{0jk}-\frac{\mu_{_W}}{h_x}s_{1jk}=r_{_W},\;\frac{(\mu_{_E}+\chi_{_E}h_x)}{h_x}s_{Njk}-\frac{\mu_{_E}}{h_x}s_{N-1jk}=r_{_E}.$$

Thus, in the general, the approximation of the III kind boundary condition will have the form:

$$g_{_\Theta}\check{s}+p_{_\Theta}\hat{s}=r_{_\Theta},$$

where $\check{s}$, $\hat{s}$ are the concentration values of the substance, respectively, at some boundary and near-boundary node. For the coefficients $g_{_\Theta}, p_{_\Theta}, r_{_\Theta}$ is true

$$g_{_\Theta}=\frac{(\mu_{_\Theta}+\chi_{_\Theta}h_\alpha)}{h_\alpha},\quad p_{_\Theta}=-\frac{\mu_{_\Theta}}{h_\alpha},\quad r_{_\Theta}=r_{_\Theta},$$

where Θ is replaced by symbols W, S, B, T, N, E, if the boundary of the region falls on the node $(i-1),(j-1),(k-1),(k+1),(j+1),(i+1)$ of the 7-points stencil, respectively, $\alpha=x,y,z$.

Thus, the difference analogue is put in correspondence to the problem (1)-(3):

$$L_h s_h = f_h(x),\quad x\in\Omega_h, \tag{4}$$

$$g_{_\Theta}\check{s}+p_{_\Theta}\hat{s}=r_{_\Theta},\quad x\in\Gamma_h, \tag{5}$$

where

$$L_h = L_{Dh}+L_{Ch}+L_{\beta h},$$

L_{Dh} is the difference diffusion transfer operator, L_{Ch} is the difference convective transfer operator, $L_{\beta h}$ is the difference analog of the interaction function of substances.

Eliminating the solution at the boundary points of the domain $\bar{\Omega}_h$ we obtain

$$\bar{L}_h s_h = \bar{f}_h. \tag{6}$$

Here $\bar{L}_h$ is the operator L_h, taking into account the boundary conditions.

IV Positive definiteness of the difference operator

For the difference operators $\bar{L}_h^c$ obtained as a result of approximation of the system (1)-(2) with the central-difference approximation of the convective terms and $\gamma=1/2$, under the presence of boundary conditions of the III kind, we prove a sufficient condition for the positive definiteness of the difference operator under consideration.

The following difference scheme written on the 7-point stencil corresponds to the operator $\bar{L}_h^c$ in the grid domain:

$$\begin{aligned} W_{ijk}^c s_{i-1jk} + S_{ijk}^c s_{ij-1k} + B_{ijk}^c s_{ijk-1} + D_{ijk}^c s_{ijk} + \\ + T_{ijk}^c s_{ijk+1} + N_{ijk}^c s_{ij+1k} + E_{ijk}^c s_{i+1jk} = f_{ijk}. \end{aligned} \qquad (7)$$

In the above notation we have

$$\Theta_{ijk}^c = \Theta_{Dijk}^c + \Theta_{Cijk}^c,$$

where the diffusion component Θ_{Dijk}^c corresponds to the operator $\bar{L}_{Dh}^c$, and convective component Θ_{Cijk}^c corresponds to the operator L_{Ch}^c. Note that, the contributions of boundary conditions of the III kind enter the diagonal of the operator $\bar{L}_{Dh}^c$ and the operator L_{Ch}^c save the skew-symmetry.

Theorem IV.1 *Let in the convection-diffusion equation* (1)-(2)*, written in the symmetric form (*$\gamma = 1/2$*), with the boundary conditions of the III kind and* $\beta = \beta(x, y, z, t) \geq 0$*, convective terms are approximated by the central differences. In order to the operator* $\bar{L}_h^c$ *be a difference analogue stationary problem of convection-diffusion was a positive definite matrix, it suffices to satisfy inequalities*

$$\beta_{ijk} - \sum_{\Theta=W,S,B,T,N,E} \delta_\Theta \left(\left(1 + \frac{p_\Theta}{g_\Theta}\right) \Theta_{Dijk}^c + \frac{p_\Theta}{g_\Theta} \Theta_{Cijk}^c \right) \geq 0, \qquad (8)$$
$$i = 1, ..., N_x - 1, \quad j = 1, ..., N_y - 1, \quad k = 1, ..., N_z - 1,$$

where at least one of the inequalities (8) *is strict.*

Here Θ_{Dijk}^c, Θ_{Cijk}^c *are coefficients of the difference scheme* (7)*;* δ_Θ *is the Kronecker symbol for the corresponding boundary;* g_Θ, p_Θ *are coefficients from* (5)*,* $g_\Theta \neq 0$.

V M-matrix property of the difference operator

For the difference operators $\bar{L}_h^p$ obtained as a result of the approximation of the equations (1)-(2) with upwind approximation of the convective terms in convection-diffusion equation for $\gamma = 0$ in the presence of boundary conditions of the III kind, a sufficient condition for the M-matrix of the difference operator is proved.

Theorem V.1 *Let in the convection-diffusion equation* (1)-(2)*, written in a non-divergent form (*$\gamma = 0$*), with boundary conditions of the III-kind* $\beta = \beta(x, y, z, t) \geq 0$ *convective terms are approximated by upwind differences. Then* *in order to the operator $\bar{L}_h^p$ (a difference analogue of the stationary convection-diffusion problem) is a nonsingular M-matrix, it suffices to satisfy inequalities*

$$\beta_{ijk} - \sum_{\Theta=W,S,B,T,N,E} \delta_\Theta \left(1 + \frac{p_\Theta}{g_\Theta}\right) \Theta_{ijk}^p \geq 0, \qquad (9)$$

$$i = 1, ..., N_x - 1, \quad j = 1, ..., N_y - 1, \quad k = 1, ..., N_z - 1,$$

where at least one of the inequalities (9) is strong.

Here Θ_{ijk}^p are elements of the matrix A^p of the operator $\bar{L}_h^p$ in the near-boundary node for the corresponding boundary; δ_Θ is the Kronecker symbol for the corresponding boundary; g_Θ, p_Θ are coefficients from the (5), $g_\Theta \neq 0$.

VI Conclusion

The presence of the properties of positive definiteness and M-matrixity in the matrix is very important when solving systems of linear algebraic equations by iterative methods. The results presented above are necessary when choosing an iterative method for solving a system of linear equations arising from the approximation of convection-diffusion problems.

Bibliography

1. *Samarskii A.A., Vabishchevich P.N.* Numerical methods for the solution of Convection-Diffusion Problems, Moscow. URSS. 1998. 272 P. (In Russian).

2. *Krukier L.A.* Implicit Difference Schemes and an Iterative Method for Their Solution for One Class of Quasilinear Systems of Equations // Izvestija Vuzov, Mathematics. 1979. N 7, P. 41–52. (In Russian).

3. *Richtmyer R.D., Morton K.W.* Difference Methods for Initial-Value Problems. 1994. Krieger Publishing Company.

4. *Roache P.J.* Computational Fluid Dynamics. M. Mir, 1980. 616 P. (In Russian).

A REGILARIZATION METHOD ON THE BASE AUGMENTED SYSTEM FOR IDENTIFICATION PROBLEM FOR INTENSITY OF ATMOSPHERIC POLLUTION SOURCE

Chubatov A.A.[*], **Karmazin V.N.**[**]

[*] *Armavir State Pedagogical University, Armavir, Russia*

[**] *Kuban State University, Krasnodar, Russia*

In present study a special case of the identification problem for intensity of the source is studied in application to the modelling of the transport of air pollution [1]. The considered approach uses as input parameters the set of known sensitivity coefficients and corresponding pollution measured in given locations $\bar{M}_j = (x_j, y_j, z_j)$: $c_{ji} = q(\bar{M}_j, t_i)$, $j = 1, \ldots, J$, $i = 1, \ldots, N$, where J is the number of sensors, N is the number of time steps. Then, the source identification problem is represented by the following approximate matrix equation

$$A_h \cdot g = f_\delta, \tag{1}$$

where $A_h \in \mathbb{R}^{(N \cdot J) \times N}$, $g \in \mathbb{R}^N$, $f_\delta \in \mathbb{R}^{N \cdot J}$, g is unknown intensity of the source, A_h, A are the approximate and the exact matrices of sensitivity coefficients [1] and f_δ, f are the approximate and the exact measured data written in terms of sensitivity, h and δ are the maximal allowable errors in approximations of A and f. Approximate matrix and approximate measured data satisfy following inequalities $||A_h - A|| \leq h$, $||f_\delta - f|| \leq \delta$, where h and δ are unknown.

In the following the optimization problem for the residual of (1) $\mu = \inf \|f - A \cdot g\| = \|f - A \cdot g^*\|$ is considered. However, the presence of errors h and δ in (1) overdetermines ($\mu > 0$) the matrix equation of this ill-posed problem. It is more convenient to solve (1) with respect to its pseudo-solution determined as

$$g := g^* = \arg\min \|f_\delta - A_h \cdot g\| = A_h^+ \cdot f_\delta,$$

where $A^+ = (A^T \cdot A)^{-1} \cdot A^T$.

The equation (1) is transformed to the form

$$A_h \cdot g + r = f_\delta,$$

where $r = f_\delta - A_h \cdot g$ is the residual and both values g and r are unknown. Applying the least squares and considering that

$$A^T \cdot r = 0$$

the following augmented system is obtained

$$R_\omega \cdot x = d,$$

$$R_\omega = \begin{pmatrix} \omega \cdot E & A_h \\ A_h^T & O \end{pmatrix}, \quad x = \begin{pmatrix} \omega^{-1} \cdot \hat{r} \\ \hat{g} \end{pmatrix}, \quad d = \begin{pmatrix} f_\delta \\ 0 \end{pmatrix}, \tag{2}$$

where ω is the scaling factor, $\hat{r}$, $\hat{g}$ are the approximations of residual and intensity, $E \in \mathbb{R}^{(N\cdot J)\times(N\cdot J)}$, $O \in \mathbb{R}^{N\times N}$ are the identity and the zero matrices, respectively.

The matrix equation (2) is solved by applying the standard Tikhonov regularization [2] and by applying the method of an imaginary shift of the spectrum (using $R_\omega^T = R_\omega$) proposed by Faddeeva [3]. For the choice of the regularization parameter α two approaches are used here: a priori and a posteriori.

For numerical computation of the solution of ill-posed problem (2) the singular value decomposition (SVD) [4] is applied

$$A = U \cdot S \cdot V^T,$$

where $A, S \in \mathbb{R}^{(N\cdot J)\times N}$, $U \in \mathbb{R}^{(N\cdot J)\times(N\cdot J)}$, $V \in \mathbb{R}^{N\times N}$, U,V are the unitary matrices, $S = \mathrm{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_N)$ is the diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$ on its diagonal.

Then the augmented system (2) is written as follows

$$R_\omega \cdot x = d,$$
$$R_\omega = \begin{pmatrix} \omega \cdot E & S_h \\ S_h^T & O \end{pmatrix}, \quad x = \begin{pmatrix} \omega^{-1} \cdot \rho \\ w \end{pmatrix}, \quad d = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{3}$$

where $w = V^T \cdot \hat{g}$, $\rho = U^T \cdot \hat{r}$, $b = U^T \cdot f_\delta$. This system of equations can be solved efficiently since the matrix R_ω is sparse (tridiagonal matrix).

We use the fact proved by Morozov V.A. and Gilyazov S.F. [2] that $\|g(\alpha) - g^*\| = O(h + \delta)$ for $\mu = 0$ and $\alpha = h$. The a priori choice of α for which $\alpha_{apriory} = h$ guarantees the asymptotic convergence to the exact solution $g(\alpha) \underset{h,\delta\to 0}{\to} g^* = \bar{g}$, where $\bar{g}$ is the exact solution.

In the case of large fixed values of h and δ, parameter α is determined a posteriori using the generalized discrepancy principle as the root of equation corresponding to (3)

$$\varphi_\omega(\alpha) - \psi_\omega(\alpha) = 0,$$

where $\varphi_\omega(\alpha) = \sqrt{\|b - S_h \cdot w(\alpha) - \rho\|^2 + \|\omega^{-1} \cdot S_h^T \cdot \rho\|^2}$ and $\psi_\omega(\alpha) = \delta + \sqrt{2} \cdot h \cdot \sqrt{\|w(\alpha)\|^2 + \|\omega^{-1} \cdot \rho\|^2}$.

The Newton method provides the fast convergence solving this root-search problem with the following initial approximation $\alpha_0 = 10 \cdot h$.

The quality of the choice of the parameter α is controlled using the value

$$\eta_{eff}(\alpha) = \|g(\alpha) - \bar{g}\| \,/\, \|g(\alpha_{best}) - \bar{g}\|,$$

where α_{best} is chosen as $\|g(\alpha_{best}) - \bar{g}\| = \min\limits_{\alpha} \|g(\alpha) - \bar{g}\|$.

The time-efficient algorithm developed and used in this work provides a stable numerical solution of the considered source identification problem. The solution obtained by applying the regularization approach with singular decomposition (SVD) is numerically approved by authors considering numerical experiments. Two considered approaches for the choice of regularization parameter α are found to have following efficiency control parameters: $\eta_{eff}(\alpha_{apriory}) < 2$ for the a priori choice of α, $\eta_{eff}(\alpha_{apost}) < 1.3$ for the a posteriori choice of α.

Bibliography

1. *Chubatov A.A., Karmazin V.N.* The operative control of the atmospheric pollution source on the base of function specification method //Vestn. Sam. state. tech. univ-ty. 2008. V. 2. P. 210–214.

2. *Morozov V.A.* Algorithmic bases of methods of the solution of ill-posed problems //Vychisl. methods and programming. 2003. V. 45. P. 130–141.

3. *Faddeeva V.N.* Shift for systems with badly posed matrices //Zh. Vychisl. Mat. Mat. Fiz. 1965. V. 5(5). P. 907–911.

4. *Golub G.H., Van Loan C.F.* Matrix Computations. JHU Press, 1996. 728 p.

PHYSICAL AND MATHEMATCAL MULTIFUNCTIONAL MODELS, SPECIALIZED NUMERICAL ALGORITHMS AND SOFTWARE FOR MODELING OF PROPAGATION OF CONTAMINATION, TAKING INTO ACCOUNT THREE-DIMENSIONAL URBAN OR INDUSTRIAL BUILDING

Dzama D.V., Sorokovikova O.S., Asfandiyarov D.G.

Nuclear Safety Institute, Russian Academy of Science, Moscow, Russia

Industrial objects with using nuclear energy are potentially dangerous due to probability of arising emergency situ, accompanying atmospheric release and further propagation of radionuclides on the territory of industrial object and further. So, problem of safety justification arises. At safety justification needed forecast of atmospheric propagation of radionuclides. Today several mathematical models of atmospheric dispersion of contamination intended for different scales of distances exist: simples models, such as Gaussian and Lagrangian, narrowly specialized robust models, based on numerical solving of Reynolds averaged Navier-Stokes equations with some simplifications and special boundary conditions near solid surfaces of ground and buildings (so-called misroscale metheorological models, MMM), and engineer models of general use, which require compression of computational grid near solid surfaces. In simplest models wind field is uniform in space or vary in vertical direction. However, buildings with significant sizes influence on three-dimensional wind field formation. Scale of plume of contamination in the vicinity of the source comparable with sizes of buildings. Different aerodynamical effects may significantly influence on the propagation of contamination – local magnitude and direction of the wind may significantly differ from macroscale averaged values. Thus, application of simplest models in the local scale of distances (in scale of industrial or city buildings) may be unacceptable. In turn, application of engineer model is complicated by need of condensation of computational grid near solid surfaces, in case of complex building. So, in NSI RAS has been developing narrowly specialized multi-functional model (further RANS model) and robust computer code, which using Cartesian rectangular computational grid. Distinctive feature of the model – is using uniform grid, condensation of space integration step do not applied. In this paper presented some of main verification results of three-dimensional RANS computer code, developing in NSI RAS. Description of the model and some of the verification results is given in papers [1, 2, 3, 4].

Test calculations with hypothetical building show, that concentration field are sensitive to such parameters as geometry of the building, source location, and wind direction. Comparison of RANS with Gaussian (Smith-Hosker parametrization) model has been made. Depending on source location and wind direction, values of concentration may differ on three orders of magnitude in the vicinity of the source. But Gaussian model do not applicable in such distances (<100 m). Regarding calculation results, concentration trend, calculated using RANS may significantly differs from trend, calculated using Gaussian model. Along the line, where application of Gaussian model gives max values of near-ground concentration, application of RANS model gives small values (on two orders of magnitude lower). Vice versa, along the line, where application of RANS model gives max values, application of Gaussian model gives values on several orders of magnitude lower. The largest concentrations, obtained by RANS model several times exceed max values, obtained using Gaussian model. Max values line, obtained with RANS model very differs from max values line, obtained with Gaussian model.

Validation of hydrodynamic and advection-diffusion modules has been carried out using the set of full-scale and laboratory experiments. Verification of advection-diffusion module has been carried out using analytical solution of advection-diffusion equation. Verification of deposition module has been carried out using quasi-analytical solution (allowing to get solution with arbitrary accuracy). For verification of advection-diffusion and deposition modules, task with homogeneous wind and diffusion fields and boundary condition of full reflection/absorption on the ground has been used. Verification of doses calculation module has been carried out using quasi-analytical solutions too.

Developed three-dimensional computer code, to which is dedicated this paper, is multi-functional software. It able to calculate: three-dimensional wind and turbulent viscosity (diffusion) coefficient fields, three-dimensional field of volume concentration and deposition of contamination on the ground and building, doses of external exposure from plume (three-dimensional cloud), contaminated surface, and inhalation.

In current paper some of main results of verification and validation of developed RANS three-dimensional model and computer code are presented.

For verification of the module of concentration propagation calculation, based on CABARET scheme, has been considered tasks with homogeneous fields of wind and turbulent diffusion coefficients, which allow analytical solution. Convergence of numerical to exact analytical solution has been shown while decreasing space/time integration steps. First task is propagation with no advection and non-zero turbulent diffusion, and with full reflection of contamination from ground $(\partial C/\partial z|_{z=0} = 0)$, where C – is volume concentration of contamination. Second task – is propagation of initial narrow (4 cells) Gaussian profile with flow without turbulent diffusion. Third task – is task with non-zero advection and turbulent diffusion. The task was solved as with boundary condition of full reflection, as boundary condition of full absorption $(C|_{z=0} = 0)$.

For validation of RANS model international open database CEDVAL [5], especially developed and recommended by European scientific community for validation MMM, has been used. Experiments have been carried out in aerodynamic tube with reduced copies of real objects. Inflow wind was created in accordance with atmospheric boundary layer, but in smaller scale of heights. Two sets of experiments and experiment with large amount of obstacles (COST-MUST), are accessible in CEDVAL database. First set of experiments called A1, second – B1. A1 set consists from experiments with one obstacle. B1 set consists from experiments with several obstacles. Within CEDVAL database validation of wind velocity components calculation in the RANS model has been carried out using 8 experiments (A1-1, A1-2, A1-3, A1-4, A1-6, A1-7, B1-1, COST-MUST). Validation of concentration has been carried out using experiment A1-5 and COST-MUST. Special statistical quantitative parameters of the quality of simulation have been used. For calculation of these parameters an array of pairs, where first value – is the value, observed in experiment, and second – is value, obtained in numerical simulation, is needed. First parameter – is so-called *FA*-2, which shows the proportion of the total number of measurement points for which the condition below is met: $(C_{obs} \leq W)$ or $(1/2 \leq C_{calc}/C_{obs} \leq 2)$. Second parameter – is so-called *Hit rate*, which shows the proportion of the total number of measurement points for which other condition is met: $(|C_{obs} - C_{calc}| \leq W)$ or $(|C_{calc} - C_{obs}| / C_{obs} \leq D)$. C_{obs} – is observed value in current measurement point, C_{calc} – is obtained by simulation value in the cell, closest to the measurement point. For wind velocity components recommended values of W and D are 0,34 and 0,25 respectively. In each experiment amount of measurement points was sufficient for calculation of these parameters for wind velocity components: total amount of such points was $\sim 14,000$. Acceptance criterion of the quality of simulation is: *FA*-2 $\geq 66\%$, *Hit rate* $\geq 55\%$. General values of *FA*-2 and *Hit rate*, calculated using data of all listed experiments are presented in Table 1.

Table 1. *FA*-2 and *Hit rate* for longitudinal (u), transverse (v), and vertical (w) velocity components in CEDVAL experiments.

A1 and B1 set	u	v	w	COST-MUST	u	w
FA-2, %	87	96	93	*FA*-2, %	89	29
Hit rate, %	76	82	75	*Hit rate*, %	73	20

For validation of RANS model (hydrodynamic and advection-diffusion modules) used tunnel experiment COST-MUST – laboratory analog of MUST field experiment. Experiment COST-MUST has been carried out using reduced in 75 times copies of original buildings. As an imitation of buildings quasi-ordered array of 119 identical containers was used. Containers was arranged in scheme of 12×10 exclude 1. Each container has sizes (full scale) $12,2 \times 2,42 \times 2,54$ m^3. Mean distance between containers was 12,9 in X and 7,9 m in Y direction. Horizontal components of wind velocity and concentration of the passive tracer have been measured. In Figure 1 presented containers and wind field around one container, simulated by RANS. *FA*-2 and *Hit rate*, obtained by RANS model

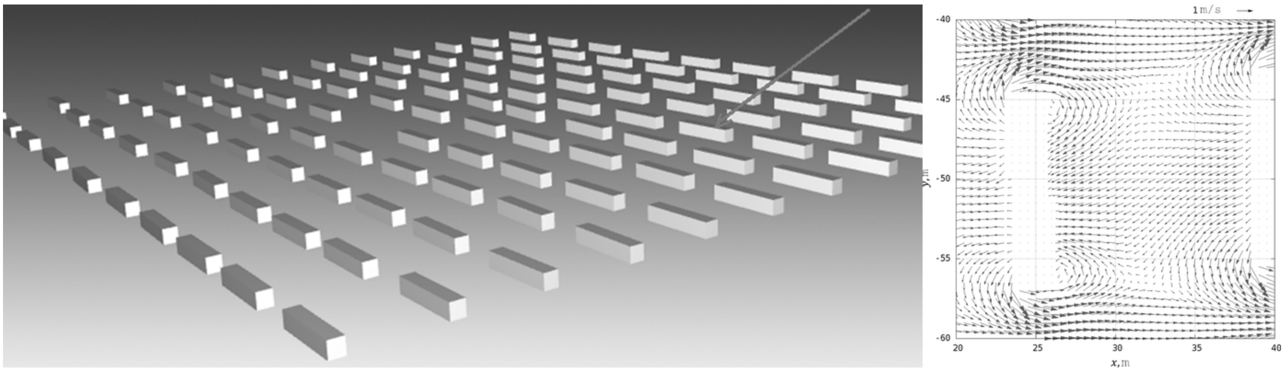


Figure 1. Containers and wind field on the height of 25 cm above the ground, obtained by RANS model, in numerical COST-MUST experiment

(see Table 1), are at the level of values, obtained by commercial codes *MISCAM* (*MMM*) and *ANSYS* (engineer code of general use).

Consider results of validation of concentration calculation. Simulated values have been compared with data of laboratory experiments. In the experiment A1-5 four vent flows from lee side of the obstacle near the ground were created constant flow of passive tracer. Hydrodynamic part of staging of the experiment was the same as in A1-5. As an obstacle considered cuboid with sizes $20 \times 30 \times 25$ m^3. Tunnel experiment has been carried out with reduced in 200 times copy of the obstacle. Measurements of tracer concentration carried out in 6 planes: $z = 2$, $z = 7$, $y = 0$, $y = -12$, $y = -15{,}2$, $x = 10{,}2$ m. Total number of measurement points was 1195. With probability of 95% the ratio of simulated and measured concentration lies in the range from 1/9,7 to 9,7. For calculation of this probability, all numerical and measured values were limited to the bottom by 1% from max measured concentration. Max values on specific distance have better agreement with experiment: with probability of 95% the ratio of max simulated and measured concentration on specific distance lies in the range from 1/4,5 to 4,5.

In the experiment COST-MUST was 244 points, where were made measurements of concentration of the tracer. All points were located on the same height – 1,28 m above the ground. With probability of 95% the ratio of simulated and measured concentration lies in the range from 1/13 to 13. As in the experiment A1-5, all numerical and measured values were limited to the bottom by 1% of max measured concentration.

Verification of deposition calculation module has been carried out using two tasks with different boundary condition on the ground. Deposition on the ground has been calculated. First type of boundary condition – is the condition full absorption. Wind and diffusion fields were homogeneous in space. Flow of the contamination to the ground in this case defined by diffusion flow: $J = D\,\partial C/\partial z|$, where D is diffusion coefficient. Zero concentration was maintained near the ground. In case of point and instantaneous source of contamination this task allows solution in form of integral from analytical expression over the time. This integral was computed numerically with high accuracy. This solution was compared with solution, obtained by the RANS model. Second boundary condition is more realistic. Zero diffusion flow is assumed, deposition occurring from viscous sub-layer: $J = Cu_d$ $(\mathrm{kg/m^2/s})$, where u_d – is the velocity of dry deposition. If $C = \mathrm{const}$, deposition increases linear with the time, what has been confirmed by simulation results, obtained by RANS model.

In the case of radioactive contamination developed computer code calculates external and internal radiation doses. Radiation doses calculate taking into account shielding effect of the γ-radiation by buildings. Module of calculation of doses from plume has been verified on the task with spherically-symmetrical distribution of contamination concentration with and without spherical obstacle. Without obstacle doses from plume were obtained by numerical integration of analytical expression. Obtained values were compared with values, obtained by RANS model. With spherical obstacle shielded region were calculated analytically. In the RANS model shielded region calculates using universal effective special method, which suit for obstacles of arbitrary shape. Task with spherical obstacle allowed to verify this method. Module of calculation of doses from contaminated buildings was verified on the task with conic surface. This task allows solution in form of integral from analytical expression over the conic surface. This integral was calculated with high accuracy, and obtained values were compared with values, obtained by RANS model.

So, all sub-modules (hydrodynamic, advection-diffusion, deposition, radiation doses calculation) of the computer code has been either verified or verified and validated.

Bibliography

1. *Blagodatskykh D.V., Dzama D.V., Sorokovikova O.S.* Simulations of radioactive contamination within an industrial site, Proceedings of fourth China-Russia conference Numerical algebra with applications. 2015. P 84–87.

2. *Sorokovikova O.S., Dzama D.V., Asfandiyarov D.G.* Specialized robust CFD RANS microscale meteorological model for modelling atmospheric processes and contamination transport in urban and industrial areas. Trudy ISP RAN/Proc. ISP RAS, V. 30. Issue 5. 2018. P. 213–234. DOI: 10.15514/ISPRAS-2018-30(5)-13

3. *O. Sorokovikova, D. Dzama, D Asfandiyarov, D. Blagodatskykh* Mathematical apparatus for designing specialized robust CFD RANS microscale meteorological models. Communication on Applied Mathematics and Computation, V. 32, №4, P. 925-940. DOI 10.3969/j.issn.1006-6330.2018.04.019

4. *Dzama, D.V., Sorokovikova, O.S.*Verification of the mathematical model for calculating the irradiation dose from a radioactive cloud of arbitrary shape taking account of screening by buildings, 2016, Atomic Energy 120(6). P. 432-436

5. CEDVAL at Hamburg University Compilation of Experimental Data for Validation of Microscale Dispersion Models. Web reference: https://mi-pub.cen.uni-hamburg.de/index.php?id=433

INVESTIGATION OF THE STABILITY OF TWO-DIMENSIONAL FLOWS CLOSE TO THE SHEAR

Kirichenko O.V., Revina S.V.
Southern Federal University, Rostov-on-Don, Russia

We consider the two-dimensional ($\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$) viscous incompressible flow driven by an external forces field $\boldsymbol{F}(\boldsymbol{x}, t)$ that is periodic in x_1 and x_2 with periods ℓ_1 and ℓ_2, respectively. The flow is described by the Navier-Stokes equations

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}, \nabla)\boldsymbol{v} - \nu \Delta \boldsymbol{v} = -\nabla p + \boldsymbol{F}(\boldsymbol{x}, t), \quad \operatorname{div} \boldsymbol{v} = 0,$$

where $\nu = 1/Re$ is the kinematic viscosity and Re is the Reynolds number. The period $\ell_1 = 2\pi$, and the ratio of the periods is characterized by the wave number α: $\ell_2 = 2\pi/\alpha$, $\alpha \to 0$. Let $\langle f \rangle$ denote the average with respect to x_1, while $\langle\!\langle f \rangle\!\rangle$ denote the average over the period rectangle $\Omega = [0, \ell_1] \times [0, \ell_2]$:

$$\langle f \rangle = \frac{1}{\ell_1} \int\limits_0^{\ell_1} f(\boldsymbol{x}, t)\, dx_1, \quad \langle\!\langle f \rangle\!\rangle(t) = \frac{1}{|\Omega|} \int\limits_\Omega f(\boldsymbol{x}, t)\, dx_1\, dx_2.$$

The spatial average velocity is assumed to be given: $\langle\langle \boldsymbol{v} \rangle\rangle = \boldsymbol{q}$. The velocity field is assumed to be periodic in x_1, x_2 with the same periods ℓ_1, ℓ_2 as the field of external forces.

A longwave asymptotics ($\alpha \to 0$) is constructed for the stability problem of the steady flow close to the shear, which will be called the basic flow:

$$\boldsymbol{V} = (\alpha V_1(x_2), V_2(x_1)), \quad \langle V_2 \rangle \neq 0. \tag{1}$$

The class of flows under consideration generalizes the Kolmogorov flow with a sinusoidal velocity profile

$$\boldsymbol{V} = (0, \gamma \sin(x_1)). \tag{2}$$

In [1], the review of some published works on the Kolmogorov flow is given. The problem of investigating the stability of a two-dimensional flow under the action of spatially periodic force was proposed by A.N. Kolmogorov in his seminar. In [2], nonstationary time-periodic structures are obtained using long-wave perturbations of the Kolmogorov flow. The long-wave asymptotic behavior of the stability problem for two-dimensional parallel flows of general form

$$\boldsymbol{V} = (0, V_2(x_1)), \quad \langle V_2 \rangle \neq 0$$

was considered in [3]. Research [4] is devoted to the study of self-oscillations arising in the loss of stability of parallel flows of a viscous fluid affected by long wavelength perturbations. In [5], the main terms of the asymptotics of the secondary self-oscillatory regimes in the case of the basic flow close to parallel were found, but general rules in coefficient expressions were not obtained.

In [6], recurrence formulas for finding the kth term of the long wavelength asymptotic for the stability of steady shear flows were derived in the case of nonzero average. The coefficients of the expansions are explicitly expressed in terms of some Wronskians, as well as integral operators of Volterra type. In the particular case, when the deviation of the velocity from its mean value $V_2(x) - \langle V_2 \rangle$ is an odd function of x, the coefficients of expansion of the eigenvalues in series in powers of α, starting from the third, are zero and the eigenvalues can be found exactly: $\sigma_{1,2} = \pm im\langle V\rangle\alpha, m \neq 0$. In [7], recurrence formulas for finding the kth term of the long-wave asymptotics for the stability of two-dimensional basic shear flows of a viscous incompressible fluid with zero average are derived.

The linear inviscid damping phenomenon for the linearized Euler equations around the Kolmogorov flow is proved in [8]. Kolmogorov flow for 2D Navier-Stokes equation on a torus is considered in [9].

The aim of this research is to generalize the results [6] related to shear flows in the case of basic flows close to shear.

Looking for a solution $(\widetilde{\mathbf{v}}, \widetilde{p})$ linearized on the basic flow (1) perturbation equation in the form of normal oscillations, we obtain the linear eigenvalue problem (here and below, $x = x_1$, $z = \alpha x_2$):

$$\sigma\varphi_1 + \alpha^2\varphi_2\frac{dV_1}{dz} + \alpha V_1(z)\frac{\partial\varphi_1}{\partial x} + \alpha V_2(x)\frac{\partial\varphi_1}{\partial z} - \nu\left(\frac{\partial^2\varphi_1}{\partial x^2} + \alpha^2\frac{\partial^2\varphi_1}{\partial z^2}\right) = $$

$$= -\frac{\partial P}{\partial x}, \tag{3}$$

$$\sigma\varphi_2 + \varphi_1\frac{dV_2}{dx} + \alpha V_1(z)\frac{\partial\varphi_2}{\partial x} + \alpha V_2(x)\frac{\partial\varphi_2}{\partial z} - \nu\left(\frac{\partial^2\varphi_2}{\partial x^2} + \alpha^2\frac{\partial^2\varphi_2}{\partial z^2}\right) = $$

$$= -\alpha\frac{\partial P}{\partial z}, \tag{4}$$

$$\frac{\partial\varphi_1}{\partial x} + \alpha\frac{\partial\varphi_2}{\partial z} = 0, \quad \int\limits_0^{2\pi}\varphi_1(x,z)dz = 0, \quad \langle\varphi_2\rangle = 0. \tag{5}$$

The value of the parameter ν at which one or several eigenvalues σ lie on the imaginary axis is called critical. The unknown perturbations of velocity $\boldsymbol{\varphi}(x,z)$, the function $P(x,z)$, the eigenvalues σ and the critical viscosity ν are sought in the form of series in powers of α:

$$\sigma(\alpha) = \sum_{k=0}^{\infty}\sigma_k\alpha^k, \quad \nu = \nu_* + \sum_{k=1}^{\infty}\nu_k\alpha^k, \tag{6}$$

$$\boldsymbol{\varphi} = \sum_{k=0}^{\infty}\boldsymbol{\varphi}^k\alpha^k, \quad P = \sum_{k=0}^{\infty}P^k\alpha^k. \tag{7}$$

Up to α^k, $k \geqslant 1$, from (3) – (5) we derive the following system of equations:

$$\nu_* \frac{\partial^2 \varphi_1^k}{\partial x^2} = \frac{\partial P^k}{\partial x} + \sum_{j=1}^{k} \sigma_j \varphi_1^{k-j} - \sum_{j=1}^{k-1} \nu_j \frac{\partial^2 \varphi_1^{k-j}}{\partial x^2} - \sum_{j=0}^{k-2} \nu_j \frac{\partial^2 \varphi_1^{k-2-j}}{\partial z^2} +$$

$$+ V_2(x) \frac{\partial \varphi_1^{k-1}}{\partial z} + \frac{dV_1}{dz} \varphi_2^{k-2} + V_1(z) \frac{\partial \varphi_1^{k-1}}{\partial x}, \tag{8}$$

$$\nu_* \frac{\partial^2 \varphi_2^k}{\partial x^2} = \sum_{j=1}^{k} \sigma_j \varphi_2^{k-j} - \sum_{j=1}^{k} \nu_j \frac{\partial^2 \varphi_2^{k-j}}{\partial x^2} - \sum_{j=0}^{k-2} \nu_j \frac{\partial^2 \varphi_2^{k-2-j}}{\partial z^2} +$$

$$+ \{W(\varphi_1^k, \theta'')\} + \langle V_2 \rangle \frac{\partial \varphi_2^{k-1}}{\partial z} + V_1(z) \frac{\partial \varphi_2^{k-1}}{\partial x} + \frac{\partial \{P^{k-1}\}}{\partial z}, \tag{9}$$

$$\frac{\partial \varphi_1^k}{\partial x} + \frac{\partial \varphi_2^{k-1}}{\partial z} = 0, \quad \int\limits_0^{2\pi} \varphi_1^k dz = 0, \quad \langle \varphi_2^k \rangle = 0. \tag{10}$$

For a non-parallel basic flow (1), the coefficients of expansion of the critical viscosity ν and the eigenvalues σ have the following structure:

$$\nu_k = [\nu_k] + \widetilde{\nu_k}, \quad \sigma_{k+2} = [\sigma_{k+2}] + \widetilde{\sigma_{k+2}}, \tag{11}$$

where the square brackets are used to denote the coefficients of viscosity and eigenvalues in the case of basic shear flow, and the wave is used to denote additional term. If $V_1(z) = 0$ then that additional term is equal to zero. It will be shown later that components of eigenfunctions and pressure φ_1^k, P^k at $k = 1, 2, 3$ and φ_2^k at $k = 1, 2$ have the same structure:

$$\varphi_1^k = [\varphi_1^k] + \widetilde{\varphi_1^k} + \langle \varphi_1^k \rangle, \quad \varphi_2^k = [\varphi_2^k] + \widetilde{\varphi_2^k}, \quad P^k = [P^k] + \widetilde{P^k} + \langle P^k \rangle.$$

while for large values of k, additional terms appear in the expressions of the eigenfunctions $\boldsymbol{\varphi}$ and the pressure P. These terms depend on the mean values of previous coefficients.

The coefficients of the expansion in a series of eigenfunctions have the following structure:

$$\varphi_1^k = [\varphi_1^k] + \widetilde{\varphi_1^k} + \varphi_1^1(\langle \varphi_1^{k-1} \rangle) + \langle \varphi_1^k \rangle,$$
$$\varphi_2^k = [\varphi_2^k] + \widetilde{\varphi_2^k} + \varphi_2^1(\langle \varphi_1^{k-1} \rangle) + \varphi_2^0(\langle \varphi_1^k \rangle),$$

and the coefficients of decomposition in a series of pressure have the following structure

$$P^k = [P^k] + \widetilde{P^k} + P^1(\langle \varphi_1^{k-1} \rangle) + \langle P^k \rangle.$$

Here $k = 1, 2, 3, 4$. The expressions in square brackets $[\varphi_i^k]$, $[P^k]$ are the solutions of the same equations, as in the case of shear basic flow and coincide with corresponding coefficients if $\widetilde{\nu_j} = 0$, $\widetilde{\sigma_j} = 0$.

The first terms of the asymptotics, found analytically, are used to calculate on Maple the eigenfunctions and critical viscosity for two types of flows. These are cases of the main shear flow and the flow close to the shear flow. As an example of a shear flow, we consider a flow that differs from the Kolmogorov flow (2) by the presence of a non-zero mean $\boldsymbol{V} = (0, \sin x + 1)$. As an example of a flow close to a shear one, we consider a flow close to the Kolmogorov flow $\boldsymbol{V} = (\alpha, \sin x + 1)$. For the indicated flows, as well as for other similar flows, the graphs of neutral curves ($Re = \frac{1}{\nu(\alpha)}$) have the following form (Fig. 1, Fig. 2):

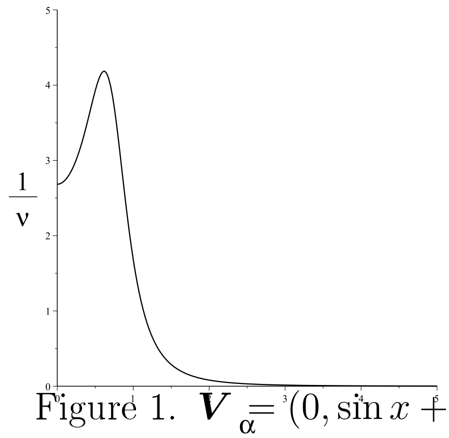


Figure 1. $\boldsymbol{V} = (0, \sin x + 1)$

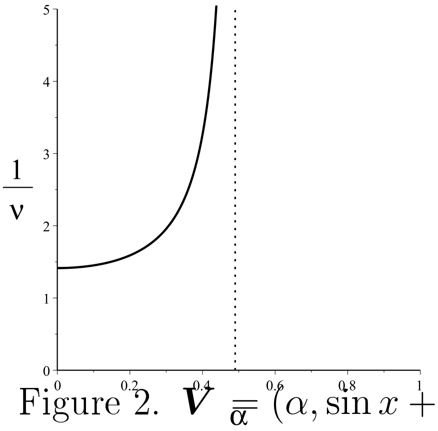


Figure 2. $\boldsymbol{V} = (\alpha, \sin x + 1)$

The asymptotics found allows to investigate the trajectories of the motion of passive impurity particles in the secondary self-oscillatory flow [5]. The trajectories of particles in the linear approximation satisfy the equation:

$$\dot{\mathbf{x}} = \mathbf{V}(\mathbf{x}) + \mathbf{u}(\mathbf{x}, t), \quad \mathbf{u}(\mathbf{x}, t) = \varphi e^{i\omega t} + \varphi^* e^{-i\omega t}.$$

For $\mathbf{V} = (-\alpha \sin z, -\sin x - 2\cos x - 4\cos 2x + 1), m = 1$, phase portrait in coordinates $\mathbf{x} = (x_1, y), y = x_2 - \langle V_2 \rangle$, has the form as shown in Fig. 3. For general flows, as well as for special case considered above, a similar qualitative behavior is detected.

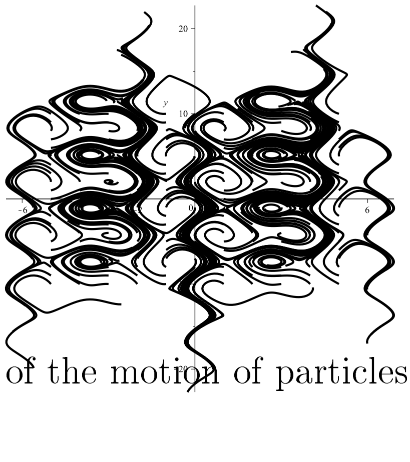


Figure 3. Trajectories of the motion of particles of a passive admixture

Bibliography

1. *Fylladitakis E.D.* Kolmogorov Flow: Seven Decades of History // Journal of Applied Mathematics and Physics, 2018, No 6, pp. 2227—2263. DOI: https://doi.org/10.4236/jamp.2018.611187.

2. *Maxim Kalashnik, Michael Kurgansky* Nonlinear dynamics of long-wave perturbations of the Kolmogorov flow for large Reynolds numbers // Ocean Dynamics, 2018, vol. 68, pp. 1001—1012.

3. *Yudovich V. I.* Instability of viscous incompressible parallel flows with respect to spatially periodic perturbations in Numerical Methods for Problems in Mathematical Physics // Numerical Methods for Problems in Mathematical Physics, 1966, Moscow, pp. 242—249. (in Russian)

4. *Yudovich V. I.* Natural oscillations arising from loss of stability in parallel flows of a viscous liquid under longwavelength periodic disturbances // Fluid Dynamics, 1973, vol. 8, pp. 26—29.

5. *Melekhov A. P., Revina S. V.* Onset of self-oscillations upon the loss of stability of spatially periodic two dimensional viscous fluid flows relative to long-wave perturbations // Fluid Dynamics, 2008, vol. 43, No 2, pp. 203—216.

6. *Revina S. V.* Recurrence formulas for long wavelength asymptotics in the problem of shear flow stability // Computational Mathematics and Mathematical Physics, 2013, vol. 53, No 8, pp. 1207—1220. DOI: 10.1134/S096554251306016X.

7. *Revina S. V.* Stability of the Kolmogorov flow and its modifications // Computational Mathematics and Mathematical Physics, 2017, vol. 57, No 6, pp. 995—1012. DOI: 10.1134/S0965542517020130.

8. *Dongyi Wei, Zhifei Zhang, Weiren Zhao.* Linear inviscid damping and enhanced dissipation for the Kolmogorov flow, 2017, arXiv:1711.01822.

9. *Zhiwu Lin, Ming Xu.* Metastability of Kolmogorov flows and inviscid damping of shear flows // Archive for Rational Mechanics and Analysis, 2019, vol. 231, No 3, pp. 1811—1852. arXiv:1707.00278. DOI: https://doi.org/10.1007/s00205-018-1311-8.

SOME COMPUTATIONAL ALGEBRAIC PROBLEMS, REDUCIBLE TO THE KNAPSACK PROBLEM, AND THEIR SOLUTION USING LUNAR ARITHMETIC

Melnikov B.F.*, Korabelshchikova S.Yu.**

* *Shenzhen MSU–BIT University, Shenzhen, China*

** *Northern (Arctic) Federal University named after M.V. Lomonosov, Arkhangelsk, Russia*

1. Introduction.

A large number of various theoretical and applied problems are reduced to the knapsack problem. We present two problems that can be reduced to this problem: counting the number of cyclic codes and counting the number of roots from special languages.

Consider the *first* problem. Finite field of q will be denoted by F_q, q is a degree of prime. A cyclic (n,k)-code over a finite field F_q (where n is the length and k is the number of information symbols) is uniquely determined by the generating normed polynomial $g(x)$ over F_q, satisfying two the following conditions:

- degree of polynomial $g(x)$ equals $n-k$;
- the polynomial x^n-1 is divided by $g(x)$ in a ring of polynomials $F_q[x]$.

Let the decomposition of polynomial x^n-1 on irreducible normalized factors over a field F_q be known; note that it is sufficient that the degrees of these polynomials are known. Also note that if n and q are co-prime integers, then all factors in the decomposition are different, otherwise the multiplicity of each factor equals v which is the greatest common divisor of n and q. We count the ways of representing the number $n-k$ as the sum of the degrees of polynomials from the expansion, taken no more than v times.

This task is equivalent to the special case of the knapsack problem. The result will be equal to the number of different cyclic codes with fixed parameters n, k and q. Note that to find the degrees of the polynomials in the decomposition of the polynomial x^n-1 into irreducible normalized factors over the field F_q, there exists an efficient algorithm. A detailed description of the solution to this problem can be found in [1].

Let us formulate the *second* problem. For the given language A over alphabet Σ and given $n \in \mathbb{N}$, we have to construct all the languages B such that $A=B^n$. In this case, we call language B by the root of nth degree of language A.

Let M be a finite subset of $\mathbb{N}$; for this problem, we shall consider languages $\Sigma(M)$ only, which contain words over Σ of the length $i \in M$.

Let us consider the problem of finding all roots of the n-th degree from the language $\Sigma(t_1,t_2)$ containing all sorts of words over the alphabet Σ of length from t_1 to t_12 (where $t_1 \leqslant t_2$). As we noted in [2, 3], the operation of extracting the root is an ambiguous function: for example, not only the "obvious" square root $\Sigma(1,5)$ is extracted from the language $\Sigma(2,10)$, but also $\Sigma(M)$, where $M=\{1,2,4,5\}$. This follows from the fact that any natural number from 2 to 10 can be represented as the sum of two (not necessarily different) terms from the set $\{1,2,4,5\}$.

We note also an obvious fact: in order for the n-th root of the $\Sigma(t_1,t_2)$ language to be extracted, it is necessary and sufficient that t_1 and t_2 be divided by n. Besides, let $t_1=n\cdot n_1$ and $t_2=n\cdot n_2$. The language $\Sigma(M)$ is a root of n-th degree from the language $\Sigma(t_1,t_2)$ if and only if M is a subset of the set $\{n_1,n_1+1,,n_2\}$ satisfying the following condition: each number from t_1 to t_2 is the sum of n terms from the set M (not necessarily different).

Thus, the task of checking this condition is equivalent to a special case of the problem of an unbounded backpack, for which a classical solution is known that finds the exact answer in polynomial time.

2. Preliminaries: versions of setting the knapsack problem.

Problem 1. Given N items, W is the capacity of the knapsack. Let w_1, w_2, $\ldots$, w_N be the positive weights of the respective items, and c_1, c_2, $\ldots$, c_N be the costs of the corresponding items. We need to find a set of binary values b_1, b_2, $\ldots$, b_N, where $b_i=1$, if item i is included in the set, and $b_i=0$, if item i is not included in the set, such that:

1. $b_1w_1+b_2w_2+\cdots+b_Nw_N\leqslant W$;
2. $b_1c_1+b_2c_2+\cdots+b_Nc_N\to\max$.

Such problem is the formulation of one of several types of knapsack tasks. This type of task is sometimes called "Backpack 0-1". There are also other kinds of knapsack problems.

We consider a generalization of the previous problem, when any object can be taken any number of times. A special case of this problem is the equivalence of the values of all items. Therefore, in the future, the cost of items can be ignored. Also we sometimes add a condition that we must take exactly M items. Let us formulate the last problem.

Problem 2. The problem of the unlimited knapsack. Given N items, W is the capacity of the knapsack, M is the number of items we have to take. Let w_1, w_2, $\ldots$, w_N be the positive weights of the respective items. It is necessary to find a set of non-negative values x_1, x_2, ..., x_N, where x_i is the number of items of this type taken into the set, such that:

1. $x_1+x_2+\cdots+x_N=M$;
2. $x_1w_1+x_2w_2+\cdots+x_Nw_N\leqslant W$;
3. $x_1w_1+x_2w_2+\cdots+x_Nw_N\to\max$.

3. Preliminaries: solving the problem using the dynamic programming.

The classical solution to Problem 2 is a solution using dynamic programming. We introduce a binary value $D_{i,j}$, where $D_{i,j} = 1$ if we can obtain the weight i using j any items. If it is impossible, then $D_{i,j} = 0$.

Suppose that now we are in the admissible state $D_{i,j} = 1$. Then we shall try to add the item k to the knapsack from the current state $D_{i,j}$. Let $q = w_k$ be the weight of the k-th object. Since the current state $D_{i,j} = 1$, then, when we add an item with the number k, we shall move to the state $D_{i+q,j+1}$ and it will also be equal to 1.

The initial state of the binary value D will be $D_{0,0} = 0$, since it is always possible to gain zero weight without using a single object (i.e., we can always take nothing). The answer to the problem is

$$R = \max\{i \,|\, i \leqslant W \,\&\, D_{i,M} = 1\};$$

this value shows the maximum weight we can gain using exactly M items.

We can show, that the asymptotic complexity of corresponding algorithm is $O(M \cdot N \cdot W)$.

4. Algorithm for solving the problem using the lunar arithmetic.

In lunar arithmetic, the operations of addition and multiplication over the numbers are replaced with the operations of finding the maximum and minimum, respectively. More information about the features of lunar arithmetic can be found in [4]. In the binary case, the addition corresponds to the disjunction, and the multiplication coincides with the usual multiplication of binary numbers.

Let us consider the weights of objects in the form of a binary vector u. Denote by $\tilde{w}$ the set of weights of all objects. Let u_i be the binary value in the binary vector u located at position i. Then in the vector u, for all u_i the following condition holds:

$$u_i = \begin{cases} 1, & i \in \tilde{w} \\ 0, & i \notin \tilde{w}. \end{cases}$$

Binary vectors in the usual way define polynomials in lunar arithmetic. Let us move from the binary vector u to the polynomial y and write it in the form

$$y = \{x^{w_i} \,|\, 1 \leqslant i \leqslant N\} = x^{w_1} + \cdots + x^{w_N}. \tag{1}$$

We assign a new polynomial z which is equal to the polynomial y raised to the M degree, i.e. z $= y^M$. We also denote by z_k the coefficient for the degree x^k in the polynomial z.

We can obtain the following theorem.

Theorem 1. In the polynomial z, all nonzero z_k denote that there is some set of exactly M objects, not necessarily different, for which the sum their weights is equal to k. □

By combining the theory proposed above, as well as a simple algorithm for rapid exponentiation in lunar arithmetic, we can obtain the following algorithm for solving the backpack problem.

Algorithm 1.

1. According to the set of weights of all objects, we determine the polynomial y n the form (1).

2. We raise y to M degree using binary exponentiation in binary lunar arithmetic. This can be done, since in the binary lunar arithmetic, the multiplication of polynomials is an associative operation.

3. We find the answer of the problem as $R = \max\{i \,|\, y_i^M = 1\}$. □

We can show, that the asymptotic complexity of this algorithm matches to the algorithm considered before.

5. Conclusion.

Thus, as we said before, a large number of various theoretical and applied problems are reduced to the knapsack problem. It is applied in various fields of knowledge: in mathematics, computer science, cryptography, economics, etc. Some applications of the backpack problem for calculating the number of error-correcting cyclic codes were considered in the paper cited before. According to the authors, the material presented in the paper combines two different approaches to the description and solution of discrete optimization problems; it can be used in other variants and in other tasks.

Bibliography

1. *Korabelshchikova S.Yu., Chesnokov A.I.* Sequence time coding for data compression // Vektor Nauki of Togliatti State University. 2013. No. 4(26). P. 25–26. (In Russian.)

2. *Zyablitseva L.V., Korabelshchikova S.Yu., Chesnokov A.I.* Linear error correction codes and their calculation algorithms // Heuristic algorithms and distributed calculations. 2014. Vol. 1. No. 3. P. 47–59. (In Russian.)

3. *Melnikov B.F., Korabelshchikova S.Yu., Dolgov V.N.* On the task of extracting the root from the language // International Journal of Open Information Technologies. 2019. Vol. 7. No. 3. P. 1–6.

4. *Applegate D., LeBrun M., Sloane N.* Dismal Arithmetic // http://arxiv.org/pdf/1107.1130v2.pdf. 2011. 33 p.

NUMERICAL STUDY OF THE SKEW-SYMMETRIC PRECONDITIONERS[1]

Pichugina O.A.

Regional Mathematical Center, Southern Federal University, Rostov-on-Don

The described below method for solving strongly non-symmetrical system of linear equations (SLE) was proposed by prof. Krukier in 1979 [2]. The main feature of the method is it takes triangular part of skew-symmetric initial system as inverse matrices. Such approach leaded to a class of iterative methods that have simple enough structure and at the same time suitable exactly for efficient solving of strongly non-symmetrical systems. Construction of the skew-symmetric part of initial matrix does not require large amount of operations and preliminary analytical actions. In major cases it appears explicitly at a stage of constructing of discrete model. Its use in the inverse operator of the method allows to take into account the structure of changes exactly this part of matrix. It is especially important when skew-symmetric part of matrix dominates. The only restriction for these methods is requirement of initial matrix to be positive real.

In this paper the skew-symmetric methods are used as preconditioner to improve the Krylov subspace methods for solving strongly non-symmetrical SLE.

We use the convection-diffusion problem for numerical study of skew-symmetric preconditioners because its a meaningful test for established or new computational methods. Consider two-dimensional steady convection-diffusion equation in Ω with homogeneous Dirichlet boundary conditions. The convective terms are described by «symmetric form» [3].

$$-Pe^{-1}L_2u + L_1(u) + L_0u = f, \quad u(x,y)|_{\partial\Omega} = u_0, \tag{1}$$

$$L_2u = \sum_{\alpha,\beta=1}^{2} \frac{\partial}{\partial x_\alpha}(K_{\alpha\beta}\frac{\partial u}{\partial x_\beta}), \tag{2}$$

$$L_1u = \frac{1}{2}(K_1\frac{\partial u}{\partial x_1} + \frac{\partial K_1 u}{\partial x_1} + K_2\frac{\partial u}{\partial x_2} + \frac{\partial K_2 u}{\partial x_2}), \tag{3}$$

$$L_0u = K_0u, \tag{4}$$

Pe is Peclet number (diffusion coefficient), $K_{\alpha\beta} = K_{\alpha\beta}(x_1,x_2)$, $\alpha,\beta = 1,2$, $K_\gamma = K_\gamma(x_1,x_2)$, $\gamma = 0,1,2$, $K_0(x_1,x_2) \geq 0$, $K_{12} = K_{21}$, $f = f(x_1,x_2)$, $(x_1,x_2) \in \Omega$.

[1] Supported by Ministry of Science and Higher Education of the Russian Federation (project N 1.13558.2019/13.1)

Suppose, the ellipticity condition is satisfied

$$c_1 \sum_{\alpha=1}^{2} \xi_\alpha^2 \le \sum_{\alpha,\beta=1}^{2} K_{\alpha\beta}(x)\xi_\alpha\xi_\beta \le c_2 \sum_{\alpha=1}^{2} \xi_\alpha^2.$$

where $c_1 > 0$, $c_2 > 0$ are constants, $\xi = (\xi_1, \xi_2)$ is arbitrary vector . At first assume $\xi_1 = 1$, $\xi_2 = 0$, then $\xi_1 = 0$, $\xi_2 = 1$ find out $0 < c_1 \le K_{\alpha\alpha} \le c_2$, $\alpha = 1, 2$.

Also assume $DivK = 0$, $K = \{K_1, K_2\}$.

Define a uniform grid in the Ω

$$\omega_h = \{x_{ij} = (ih_1, jh_2), \quad 0 \le i \le N_1, \quad 0 \le j \le N_2, \quad h_\alpha N_\alpha = l_\alpha, \quad \alpha = 1, 2\}.$$

Approximate the problem (1)-(4) on the grid ω_h.

Approximate operator $L_{\alpha\beta}u$ according to [4] by difference operator

$$\Lambda_{\alpha\beta}u = \frac{1}{2}[(K_{\alpha\beta}u_{\overline{x}\beta})_{x\alpha} + (K_{\alpha\beta}u_{x\beta})_{\overline{x}\alpha}],$$

determined by $\alpha \ne \beta$ on 7-point scheme:

$$(x_1, x_2), (x_1 \pm h_1, x_2), (x_1, x_2 \pm h_2), (x_1 - h_1, x_2 + h_2), (x_1 + h_1, x_2 - h_2).$$

Where $u_{\overline{x}_i} = (u(x_i) - u(x_i - h_i))/h_i$ is a left side derivative at x_i, $u_{x_i} = (u(x_i + h_i) - u(x_i))/h_i$ is a right side derivative at x_i.

We use the central difference scheme

$$u_{x_i^0} = (u(x_i + h_i) - u(x_i - h_i))/2h_i.$$

In [2] was proved that skew-symmetric operator appears. The applying of the central difference scheme for approximation convection-diffusion problem eases the found of the solution by iterative methods [3].

Set constants $K_{11} = K_{22} = 1$, $K_0 = 0$ for system (1)-(4). The condition $K_{12} = K_{21}$ is necessary for the symmetry of the diffusion operator, set $K_{12} = K_{21} = 1$. Thus, using finite difference approximation and special arrangement of grid nodes we get a seven-diagonal ribbon matrix

$$Ax = b. \tag{5}$$

The resulting system was solved by BiCG and GMRES(m) methods with preconditioning using MatLAB.

We use triangular skew-symmetric iterative methods (TS) [2]

$$B(\omega) = B_C + 2\omega K_L \tag{6}$$

and product triangular skew-symmetric iterative methods(PTS) [1]:

$$B(\omega) = (B_C + \omega K_L)\, B_C^{-1}\, (B_C + \omega K_U)\,. \tag{7}$$

In equalities (6) and (7) matrix B_C is symmetric, ω is real numerical parameter.

Diagonal elements B_C for TS- and PTS-preconditioners were determined as

$$b_{Cii} = \frac{1}{2}\{\max_i \sum_{j=1}^{n} (|a_{0ij}| + |a_{1ij}|)\}, \quad i = 1, ..., n, \tag{8}$$

where a_{0ij} are elements of matrix A_0, a_{1ij} are elements of matrix $(K_U - K_L)$.

We consider preconditioners without parameters ($\omega = 2$) because of clear practical advantage.

Computational experiments were carried out for the problems with different velocity coefficients (see Table 1).

Table 1. Velocity coefficients

N Problem	K_1	K_2
1	1	-1
2	$1 - 2x_1$	$2x_2 - 1$
3	$x_1 + x_2$	$x_1 - x_2$
4	$\sin(2\pi x_1)$	$-2\pi x_2 \cos(2\pi x_1)$

Table 2. GMRES(m) with preconditioners

Pe	GMRES	GMRES+TS(2)	GMRES+PTS(2)
Problem 1, $K_1 = 1$, $K_2 = -1$			
10	31 (4)	31 (6)	28 (9)
10^3	22 (2)	13 (2)	8 (10)
10^5	953 (2)	525 (10)	172(1)
Problem 2, $K_1 = 1 - 2x_1$, $K_2 = 2x_2 - 1$			
10	38 (8)	37 (5)	37 (10)
10^3	33 (3)	22 (9)	11 (4)
10^5	1170 (9)	579 (5)	307 (1)
Problem 3, $K_1 = x_1 + x_2$, $K_2 = x_1 - x_2$			
10	27 (7)	27 (3)	27 (4)
10^3	37 (10)	20 (9)	8 (10)
10^5	1095 (4)	454 (8)	188 (2)
Problem 4, $K_1 = \sin(2\pi x_1)$, $K_2 = -2\pi x_2 \cos(2\pi x_1)$			
10	24 (9)	30 (3)	27 (7)
10^3	71 (8)	35 (1)	18 (3)
10^5	3346 (4)	1519 (3)	710 (2)

The test function K to cover various functions. The first problem: the module and direction of the velocity vector is independent of a point on the plane, we have constants for convective terms. The second problem contains separable coefficients and the third task contains linear coefficients for convective terms.

Table 3. BiCG with preconditioners

Pe	BiCG	BiCG+TS(2)	BiCG+PTS(2)
Problem 1, $K_1 = 1$, $K_2 = -1$			
10	181	173	172
10^3	191	127	35
10^5	1247	1539	645
Problem 2, $K_1 = 1 - 2x_1$, $K_2 = 2x_2 - 1$			
10	184	182	182
10^3	212	153	93
10^5	881	917	638
Problem 3, $K_1 = x_1 + x_2$, $K_2 = x_1 - x_2$			
10	203	197	199
10^3	364	189	91
10^5	2450	2538	1173
Problem 4, $K_1 = \sin(2\pi x_1)$, $K_2 = -2\pi x_2 \cos(2\pi x_1)$			
10	208	200	189
10^3	628	381	185
10^5	1760	1768	1055

The fourth problem: the velocity field models the vortex motion. The last two problems are the most difficult for numerical solution.

The right hand side is chosen so that $u(x_1, x_2) = e^{x_1 x_2} \sin(\pi x_1) \sin(\pi x_2)$ is a analytical solution of (1)–(4). We use 32×32 grid for the finite difference approximation. The Peclet number varies from 10 to 10^5. This leads to the system of linear algebraic equations with strongly non-symmetric matrix (the skew-symmetric component of the matrix is much larger than its symmetric component in a certain norm [2]) for $Pe > 10^3$.

Numerical results are presented in Tables 2 and 3.

Numerical experiments for the GMRES(m) and BiCG methods demonstrate efficiency for solving the system (5). They converge even when the system matrix loses diagonal dominance and becomes strongly non-symmetric. However, the product triangular skew-symmetric preconditioners improve the properties of the resulting system, they reduce the matrix condition number by 2 times (we are talking about the ratio of the maximum eigenvalue to the minimum, it easily computed in MatLAB), hence convergence properties of the methods improve. Product triangular preconditioners speed up method BiCG by 2-3 times and even more for method GMRES(m).

Unfortunately, there is no good enough results for triangular skew-symmetric preconditioners. It was found that preconditioning does not improve the matrix condition number in this instance.

General conclusion: we recommend the product triangular skew-symmetric preconditioners for accelerating Krylov subspace methods solving the convection-diffusion problems.

Bibliography

1. *Bochev M.A., Krukier L.A.* The solution of strictly non-symmetric systems of linear algebraic equations by iterative methods// Vychislitelnaya Matematika i Matematicheskaya Fizika. - 1997. V. 37. N 11. P. 1283-1293. (in Russian)

2. *Krukier L.A.* Implicit difference schemes and an iterative method for solving them for one class of systems of quasilinear equations // Izvestiya VU-Zov. Matematika. 1979. N 7. P. 41-52.

3. *Krukier L.A., Martynova T.S.* On the influence of the form of the convection-diffusion equation on the convergence of the upper relaxation method // Vychislitelnaya Matematika i Matematicheskaya Fizika. 1999. V. 39. N 11. P. 1821-1827.

4. *Samarskiy A.A.* Theory of difference schemes. Moscow: Nauka, 1989. 616 p.

NUMERICAL ALGEBRA WITH APPLICATIONS

Proceedings of Eighth China-Russia Conference.

Executive editors: Zhong-Zhi Bai, Galina V. Muratova

Technical editors: Irina N. Shabas, Olga A. Pichugina

Signed for printing 18.06.2019.
Offset paper. Offset printing.
Cond. sheet 8,37. Publ. sig. 7,5.
100 copies. Order № 7095.

Southern Federal University Publishing
Stachki Ave., 200/1, 344090 Rostov-on-Don, Russia