Ammosov North-Eastern Federal University

YAKUTIAN MATHEMATICAL JOURNAL

PUBLISHED SINCE 2014 SCIENTIFIC JOURNAL 4 ISSUES PER YEAR

Vol. 22, No. 2 (86) **April—June, 2015**

CONTENTS

Mathematics

Mathematical Modeling

Address for correspondence:
Ammosov North-Eastern Federal University, Belinskii St., 58, Yakutsk, 677000
Phone: 8(4112)32-14-99, Fax: 8(4112)36-43-47;
http://s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/niim/mzsvfu/
e-mail: prokopevav85@gmail.com; madu@ysu.ru: ivanegorov51@mail.ru

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 517.958:539.3(6):517.968.72:517.956.3

LINEAR INSTABILITY OF SOLUTIONS TO A MATHEMATICAL MODEL THAT DESCRIBES THE FLOWS OF POLYMERS IN AN INFINITE CHANNEL

A. M. Blokhin, D. L. Tkachev, and A. V. Yegitov

Abstract. We study the new rheological model that describes the flow of an incompressible viscoelastic polymer fluid. We establish the linear Lyapunov instability of an analog of the Poiseuille flow for the Navier–Stocks system in an infinite flat channel.

Keywords: incompressible viscoelastic polymer fluid, rheological relation, Brownian particle, dumbbell, Poiseuille-type solutions, well-posedness of the mixed problem, linear instability.

1. Introduction

In the article we study the new rheological model accounting for nonlinear effects in a moving polymer medium being a suspension of noninteracting elastic dumbbells [1]. Each dumbbell is formed by two Brownian particles connected by an elastic force and moving in an anisotropic fluid formed by a solvent and other dumbbells.

This model based on a new rheological relation establishing the connection between the kinematic characteristics of a flow and interior thermodynamics parameters is a modification of the celebrated Pokrovskiĭ–Vinogradov model [2, 3]. In the author's opinion of the models, the model demonstrates its high effectiveness under the numerical study of polymer flows in domains with complex geometry [4, 5].

In the article we examine the question of linear stability of an experimentally observable analog of the Poiseuille flow for the Navier–Stokes system.

2. Statement of the Problem, Auxiliary Facts, and Statement of the Main Results

In [1] there is given the new mathematical model that describes flows of an incompressible viscoelastic polymer fluid. In the plane case the nonstationary flows of polymer media are described with the help of the following rheological model (in dimensionless form):

$$u_x + v_y = 0, \tag{2.1}$$

$$\frac{du}{dt} + p_x = \frac{1}{Re}\{(a_{11})_x + (a_{12})_y\}, \tag{2.2}$$

$$\frac{dv}{dt} + p_y = \frac{1}{Re}\{(a_{12})_x + (a_{22})_y\}, \tag{2.3}$$

$$\frac{da_{11}}{dt} - 2A_1u_x - 2a_{12}u_y + K_Ia_{11} = -\beta\big(a_{11}^2 + a_{22}^2\big), \tag{2.4}$$

$$\frac{da_{12}}{dt} - A_1 v_x - A_2 u_y + \widetilde{K}_I a_{12} = 0, \tag{2.5}$$

$$\frac{da_{22}}{dt} - 2A_2 v_y - 2a_{12} v_x + K_I a_{22} = -\beta\big(a_{12}^2 + a_{22}^2\big). \tag{2.6}$$

Here t is time, u and v are the components of the velocity in a Cartesian coordinate system (x, y), while p is the hydrostatic pressure, a_{ij} is the symmetric anisotropy tensor of the second rank, and $\frac{d}{dt} = \frac{\partial}{\partial t} + (u, \nabla)$ is the substantial derivative.

The remaining quantities are defined as follows: $I = a_{11} + a_{22}$ is the first invariant of the anisotropy tensor, $\bar{k} = k - \beta$, k, β are the scalar phenomenological parameters of the rheological model ($0 < \beta < 1$), η_0 and τ_0 are the initial values of the shear viscosity and the relaxation time,

$$A_1 = a_{11} + \frac{1}{W}, \quad A_2 = a_{22} + \frac{1}{W},$$

$$K_I = \frac{1}{W} + \frac{\bar{k}}{3} I, \quad \widetilde{K}_I = \frac{1}{W} + \frac{\hat{k}}{3} I = K_I + \beta I,$$

$$\hat{k} = k + 2\beta = \bar{k} + 3\beta,$$

$$Re = \frac{\rho u_H l}{\eta_0} \text{ is the Reynolds number,}$$

ρ ($=$ const) is the density of a medium, u_H is the characteristic velocity, l is the characteristic length, and $W = \frac{\tau_0 u_H}{l}$ is the Weissenberg number (see [5]).

Remark 1. The Reynolds and Weissenberg numbers occur in the rheological model (2.1)–(2.6) as well as the phenomenological parameters k and β defining the process of a physical experiment. It follows from [6] that the most adequate relation in experiments with polymer fluids is the equality $k = 1.2\beta$.

The linear system of equations was obtained in [7] arising as linearization of the system (2.1)–(2.6) with respect to a chosen stationary solution (in what follows its components are furnished with $\hat{}$) in the case of a fluid in an infinite flat channel.

In vector form it is written as follows: In the domain

$$G = \{(t, x, y) \mid t > 0,\ (x, y) \in \Pi = \{(x, y) \mid |x| < \infty,\ 0 < y < 1\}\},$$

the problem is to find a solution to the system of equations

$$U_t + \widehat{B} U_x + \widehat{C} U_y + \widehat{R} U + F = 0, \tag{2.7}$$

$$\Delta\Omega = \frac{1}{Re}\{\sigma_{xx} + 2(a_{12})_{xy}\} - 2\widehat{\omega} v_x. \tag{2.8}$$

Here $U = \begin{pmatrix} u \\ v \\ a_{11} \\ a_{12} \\ a_{22} \end{pmatrix}$ is an unknown vector-function, $\sigma = a_{11} - a_{22}$, $\Omega = p - \frac{1}{Re} a_{22}$, the matrices $\widehat{B} = B(\widehat{U})$, $\widehat{C} = C(\widehat{U})$, $\widehat{R} = R(\widehat{U})$ are written out with the use of the components of the stationary solution $\widehat{U}(y)$ as follows:

$$\widehat{U}(y)=\begin{pmatrix}\hat{u}(y)\\0\\\hat{a}_{11}(y)\\\hat{a}_{12}(y)\\\hat{a}_{22}(y)\end{pmatrix},\quad \widehat{B}=\begin{pmatrix}\hat{u}&0&-\frac{1}{Re}&0&0\\0&\hat{u}&0&-\frac{1}{Re}&0\\-2\widehat{A}_1&0&\hat{u}&0&0\\0&-\widehat{A}_1&0&\hat{u}&0\\0&-2\hat{a}_{12}&0&0&\hat{u}\end{pmatrix},$$

$$\widehat{C}=\begin{pmatrix}0&0&0&-\frac{1}{Re}&0\\0&0&0&0&-\frac{1}{Re}\\-2\hat{a}_{12}&0&0&0&0\\-\widehat{A}_2&0&0&0&0\\0&-2\widehat{A}_2&0&0&0\end{pmatrix},\quad \widehat{R}=\begin{pmatrix}0&\widehat{\omega}&0&0&0\\0&0&0&0&0\\0&\hat{a}'_{11}&R_{33}&R_{34}&R_{35}\\0&\hat{a}'_{12}&R_{43}&R_{44}&R_{45}\\0&\hat{a}'_{22}&R_{53}&R_{54}&R_{55}\end{pmatrix}, \tag{2.9}$$

where

$$\widehat{A}_1=\hat{a}_{11}+\frac{1}{W},\quad \widehat{A}_2=\hat{a}_{22}+\frac{1}{W},$$

$$R_{33}=\frac{1}{W}+\frac{\bar{k}}{3}\hat{I}+\frac{k+5\beta}{3}\hat{a}_{11},\quad R_{34}=-2(\widehat{\omega}-\beta\hat{a}_{12}),\quad \widehat{\omega}=\hat{u}_y,\quad R_{35}=\frac{\bar{k}}{3}\hat{a}_{11},$$

$$R_{43}=\frac{\hat{k}}{3}\hat{a}_{12},\quad R_{44}=\frac{1}{W}+\frac{\hat{k}}{3}\hat{I},\quad R_{45}=-\widehat{\omega}+\frac{\hat{k}}{3}\hat{a}_{12},$$

$$R_{53}=\frac{\bar{k}}{3}\hat{a}_{22},\quad R_{54}=2\beta\hat{a}_{12},\quad R_{55}=\frac{1}{W}+\frac{\bar{k}}{3}\hat{I}+\frac{k+5\beta}{3}\hat{a}_{22},$$

$F=\begin{pmatrix}p_x\\p_y\\0\\0\\0\end{pmatrix}$, and Δ designates the Laplace operator.

We assume the fulfillment of the boundary conditions

$$u|_{y=0}=v|_{y=0}=u|_{y=1}=v|_{y=1}=0; \tag{2.10}$$

$$\Omega_y=\frac{1}{Re}(a_{12})_x \quad \text{for } y=0,1, \tag{2.11}$$

$$\|U(t,x,y)\|=(U,U)^{\frac{1}{2}}\to 0,\quad p(t,x,y)\to 0,\quad p_x(t,x,y)\to 0\quad \text{as } |x|\to\infty \tag{2.12}$$

on the boundary of G and the initial conditions

$$U|_{t=0}=U_0(x,y),\quad p|_{t=0}=p_0(x,y), \tag{2.13}$$

with the initial data satisfying (2.8) and (2.12).

Remark 2. As the basic solution, we can take, for example, a solution that similar to the Poiseuille solution for the Navier–Stokes system (see [4, 8, 9]), which is symmetric with respect to the axis $y=\frac{1}{2}$ of the channel (in this case $\hat{p}(x,y)=\frac{1}{Re}\hat{a}_{22}(y)+\hat{p}_0=\widehat{A}x$, $\hat{p}_0$ is the value of the pressure on the axis and $\widehat{A}$ is a parameter connected with the dimensionless change of the pressure on the segment h).

Remark 3. It is proven in [7] that the system (2.7) for a given pressure $p(t,x,y)$ is t-hyperbolic [10] whenever $\widehat{A}_1>0$, $\widehat{A}_2>0$ and $\widehat{A}_1\widehat{A}_2-\hat{a}_{12}^2>0$ (see the representation (2.9) of the matrices $\widehat{B}$ and $\widehat{C}$). These inequalities are valid, in particular, when the «Poiseuille solution» is taken as the basic solution (for $k=\beta$, this fact

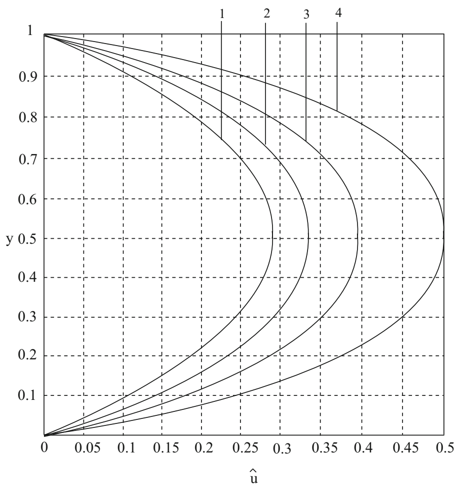

Fig. 1

is verified directly and, for $k \neq \beta$, numerically). The information about the roots of the characteristic equation plays an essential role in posing mixed problems for t-hyperbolic systems.

In view of the geometry of Π, the system of equations (2.7) and the Poisson equation (2.8) admit the Fourier transform in the variable x. Therefore, we consider the problem (2.7), (2.8), (2.10)–(2.13) assuming that $u, v \in D'_{+,a}\big(P'_x(R), C^1_y[0,1]\big)$, the pressure p and the components of the anisotropy tensor a_{11}, a_{12}, a_{22} belong to the class $D_{+,a}\big(P'_x(R), C^2_y[0,1]\big)$, where $D'_{+,a}\big(P'_x(R), C^1_y[0,1]\big)$, and $D_{+,a}\big(P'_x(R), C^2_y[0,1]\big)$ are the spaces of distributions $u(t,x,y)$ vanishing for $t<0$ and such that $u(t,x,y)e^{-\sigma t} \in P_{+,t}$ for all $\sigma > a$, $P'_+ = D'_+ \cap P$, D'_+ is the collection of distributions from $D'(R)$ vanishing for $t<0$, P is the space of tempered distributions [11, 12] in the variables x belonging to the spaces $C^1_y[0,1]$ and $C^2_y[0,1]$, respectively, in the variable y. The index in the notation of the space, for example in $P_x(R)$ denotes the active variable.

Thus, the mixed problem (2.7), (2.8), (2.10)–(2.13) is understood to be the boundary value problem for generalized functions in the variables t, y, and x, and the initial data (2.13) are fulfilled in the sense of passing to the limit as $t \to +0$ [11, 12].

The following are valid:

Theorem 1. *The mixed problem* (3.4)–(3.6) *has the unique solution in* $D'_{+,a}(C_y[0,1])$ *for every real parameter* ξ *(*ξ *is the dual variable to* x*).*

Theorem 2. *A solution to the mixed problem* (3.4)–(3.6) *as* $|\xi| \to \infty$ *does not belong to the space* $D'_{+,a}(C_y[0,1])$ *for every positive* a. *Thus, the problem is not well-posed in* $D_{+,a}$.

3. Statement of the One-Dimensional Problem with a Parameter. Proof of Theorem 1

Consider (2.8) together with the boundary conditions (2.11), (2.12) and apply the Fourier transform in x to this problem. We obtain the boundary value problem

$$\widetilde{\Omega}_{yy} - \xi^2\widetilde{\Omega} = -\xi^2\frac{1}{Re}(\tilde{a}_{11} - \tilde{a}_{22}) - \frac{2i\xi}{Re}(\tilde{a}_{12})_y + 2i\xi\widehat{\omega}\tilde{v}, \quad 0 < y < 1, \tag{3.1}$$

$$\widetilde{\Omega}_y = -\frac{i\xi}{Re}\tilde{a}_{12} \quad \text{for } y = 0, 1 \tag{3.2}$$

(it is assumed that the basic stationary solution depends only on y, in what follows the symbol ˜, used to denote the Fourier images of functions, is omitted).

The Green's function of the boundary value problem (3.1), (3.2) (ξ is a real parameter, $\xi \neq 0$) is of the form

$$G(y,\eta) = \begin{cases} -\frac{1}{2\xi(e^{2\xi}-1)}(e^{\xi\eta} + e^{-\xi\eta}e^{2\xi})(e^{\xi y} + e^{-\xi y}), & 0 \le y \le \eta, \\ -\frac{1}{2\xi(e^{2\xi}-1)}(e^{\xi\eta} + e^{-\xi\eta})(e^{\xi y} + e^{-\xi y}e^{2\xi}), & \eta < y \le 1. \end{cases} \tag{3.3}$$

Applying the Fourier transform, we can find a solution to (3.1), insert it in the right-hand side F, and derive the system

$$U_t + \widetilde{C}U_y + (-i\xi\widetilde{B} + \widehat{R})U + F = 0, \quad 0 < y < 1, \tag{3.4}$$

where

$$\widetilde{C} = \begin{pmatrix} 0 & 0 & 0 & -\frac{1}{Re} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -2\hat{a}_{12} & 0 & 0 & 0 & 0 \\ -\widehat{A}_2 & 0 & 0 & 0 & 0 \\ 0 & -2\widehat{A}_2 & 0 & 0 & 0 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} \hat{u} & 0 & -\frac{1}{Re} & 0 & \frac{1}{Re}a_{22} \\ 0 & \hat{u} & 0 & -\frac{1}{Re} & 0 \\ -2\widehat{A}_1 & 0 & \hat{u} & 0 & 0 \\ 0 & -\widehat{A}_1 & 0 & \hat{u} & 0 \\ 0 & -2\hat{a}_{12} & 0 & 0 & \hat{u} \end{pmatrix},$$

and the components $-i\xi p$ and p_y of $F(t,\xi,y) = \begin{pmatrix} -i\xi p \\ p_y \\ 0 \\ 0 \\ 0 \end{pmatrix}$ are determined with the use of the Green's function (3.3).

Moreover, the components u and v of the velocity satisfy the boundary conditions

$$u|_{y=0} = v|_{y=0} = u|_{y=1} = v|_{y=1} = 0, \tag{3.5}$$

and the unknown vector-function $U(t,x,y)$ the initial condition

$$U|_{t=0} = U_0(\xi, y). \tag{3.6}$$

Simplify (3.4) reducing the matrix $\widetilde{C}$ to upper Jordan form [13]. Note that the eigenvalues of $\widetilde{C}$ are such that

$$\lambda_{1,2,3} = 0, \quad \lambda_{4,5} = \pm\sqrt{\frac{\widehat{A}_2}{Re}} \tag{3.7}$$

(it is assumed that the condition of t-hyperbolicity of (2.7) is fulfilled and thereby $\widehat{A}_2 > 0$ on $[0,1]$ as it noted in Remark 3).

Direct calculations demonstrate that the Jordan form of $\widetilde{C}$ is of the form

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\frac{\widehat{A}_2}{Re}} & 0 \\ 0 & 0 & 0 & 0 & -\sqrt{\frac{\widehat{A}_2}{Re}} \end{pmatrix}. \tag{3.8}$$

After the change

$$U = TZ \tag{3.9}$$

of the unknown vector-function, (3.4) is transformed equivalently to the system with a block-diagonal matrix K (see (3.8)) in front of the derivative Z_y as

$$Z_t + KZ_y - i\xi LZ + [T^{-1}\widetilde{C}T_y + M]Z + G = 0, \quad t > 0,\ 0 < y < 1, \tag{3.10}$$

where the matrices L and M depend only on a stationary solution while the vector $G = \begin{pmatrix} 0 \\ 0 \\ g_3 \\ g_4 \\ g_5 \end{pmatrix}$ has, for example, the component

$$g_3 = -2\widehat{A}_2\Bigg\{\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\int\limits_0^y \cosh(\xi\eta)\bigg(\frac{\xi}{Re}\bigg(Z_1 - Z_2 + 2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_4 + 2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_5\bigg) + \frac{i\widehat{\omega}}{\widehat{A}_2}Z_3\bigg)\,d\eta$$
$$+\frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1 \cosh(\xi(\eta-1))\bigg(\frac{\xi}{Re}\bigg(Z_1 - Z_2 + 2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_4 + 2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_5\bigg) + \frac{i\widehat{\omega}}{\widehat{A}_2}Z_3\bigg)\,d\eta\Bigg\}$$
$$+\frac{2\widehat{A}_2}{Re}\frac{i\xi}{\sinh\xi}[(Z_4(t,\xi,0) + Z_5(t,\xi,0))\sinh(\xi(y-1)) - (Z_4(t,\xi,1) + Z_5(t,\xi,1))\sinh(\xi y)]. \tag{3.11}$$

The boundary conditions (3.5) are reduced to

$$\begin{cases} Z_3(t,\xi,0) = Z_3(t,\xi,1) = 0, \\ Z_4(t,\xi,0) = Z_5(t,\xi,0), \\ Z_4(t,\xi,1) = Z_5(t,\xi,1), \end{cases} \tag{3.12}$$

respectively, the initial condition (3.6) for $U(t,\xi,y)$ to

$$Z|_{t=0} = T^{-1}U_0 = Z_0(\xi,y). \tag{3.13}$$

Next, we study the mixed problem (3.10), (3.12), (3.13). Applying the Laplace transform technique and employing the form of K, we can express the components Z_1 and Z_2 through Z_3, Z_4, and Z_5.

In view of (3.10) the function $Z_3(t,\xi,y)$ meets the integral equation

$$Z_3(t,\xi,y) = e^{i\xi\hat{u}(t)}Z_{30}(\xi,y) + \int\limits_0^t e^{i\xi\hat{u}(t-\tau)}\bigg[\frac{2\widehat{A}_2}{Re}i\xi Z_4 + \frac{2\widehat{A}_2}{Re}i\xi Z_5 - g_3\bigg]\,d\tau, \tag{3.14}$$

and (3.11) implies that $g_3(t,\xi,y)$ depends on $Z_3(t,\xi,y)$, $Z_4(t,\xi,y)$, and $Z_5(t,\xi,y)$ (we can take into account the possibility of integrating by parts in the integral of $\frac{\partial Z_3}{\partial y}$ with respect to η and the first two boundary relations in (3.12)).

Applying the method of successive approximations for the available boundary values $Z_4(t,\xi,0)$ and $Z_5(t,\xi,1)$, we can uniquely determine $Z_3(t,\xi,y)$ from (3.15) expressing this function through the components $Z_4(t,\xi,y)$ and $Z_5(t,\xi,y)$.

Thus, to solve the problem (3.10), (3.12), (3.13), we need to define the two components $Z_4(t,\xi,y)$ and $Z_5(t,\xi,y)$ with given initial data $Z_{40}(t,\xi,y)$ and $Z_{50}(t,\xi,y)$ and the unknown boundary values $Z_4(t,\xi,0)$ and $Z_5(t,\xi,1)$.

Assume that the boundary values are available as well. In this case, accounting for the structure of K (see (3.8)), integrating two last equations of the system (3.10) along the characteristics, and applying the method of successive approximations again, we can uniquely determine the unknowns $Z_4(t,\xi,y)$ and $Z_5(t,\xi,y)$ moving on "layers" in the half-strip $t>0$, $0\le y\le 1$ [10].

Hence, to find the functions $Z_4(t,\xi,y)$ and $Z_5(t,\xi,y)$, it suffices to know the boundary values $Z_4(t,\xi,0)$ and $Z_5(t,\xi,1)$. Assume some analog of the consistency conditions to be fulfilled, i.e.,

$$Z_{30}(\xi,0)=Z_{30}(\xi,1)=0 \tag{3.15}$$

(see the first two relations in (3.12)).

Put $y=0$ and $y=1$ in (3.14). In view of (3.15), we obtain the two relations

$$\begin{aligned}&\int\limits_0^t e^{i\xi\hat{u}(0)(t-\tau)}\frac{8\widehat{A}_2}{Re}i\xi Z_4(\tau,\xi,0)\,d\tau=0,\\ &\int\limits_0^t e^{i\xi\hat{u}(1)(t-\tau)}\frac{8\widehat{A}_2}{Re}i\xi Z_4(\tau,\xi,1)\,d\tau=0\end{aligned} \tag{3.16}$$

which imply that

$$Z_4(t,\xi,0)=Z_5(t,\xi,0)=0,\quad Z_4(t,\xi,1)=Z_5(t,\xi,1)=0.$$

Thus, all components of the unknown vector-function $U(t,\xi,y)$ are determined for all real parameters ξ. Theorem 1 i proven.

REMARK 4. The fulfillment of (3.15) is not a necessary condition of unique solvability of (3.4)–(3.6). They are adopted for simplicity of the exposition. In the general case the transfer to the boundary conditions in (3.14) leads to a system of Volterra equations of the first kind which is uniquely solvable [9].

4. Proof of Theorem 2

Represent Z_3 as

$$\begin{aligned}&Z_{3t}-i\xi\hat{u}Z_3-i\xi\frac{2\widehat{A}_2}{Re}Z_4-i\xi\frac{2\widehat{A}_2}{Re}Z_5-2\widehat{A}_2\biggl(\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\int\limits_0^y\biggl\{\cosh(\xi\eta)\frac{\xi}{Re}\\ &\times\biggl[-(P_1+L_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}+(P_2+L_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\biggr]-\frac{\xi}{Re}\biggl(\cosh(\xi\eta)\\ &\times\biggl\{\biggl[\biggl(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\biggr)\frac{1}{\sqrt{D}}+\frac{K_1}{\sqrt{D}}\biggr]\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-\biggl[\biggl(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\biggr)\frac{1}{\sqrt{D}}+\frac{K_2}{\sqrt{D}}\biggr]\\ &\times\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\biggr\}\biggr)_\eta\biggr\}\biggr)Z_3(\tau,\xi,\eta)\,d\tau d\eta-\frac{\xi^2}{Re}\biggl\{\biggl[\biggl(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\biggr)\frac{1}{\sqrt{D}}+\frac{K_1}{\sqrt{D}}\biggr]\\ &\times\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-\biggl[\biggl(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\biggr)\frac{1}{\sqrt{D}}+\frac{K_2}{\sqrt{D}}\biggr]\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\biggr\}Z_3(\tau,\xi,y)\,d\tau\end{aligned}$$

$$
\begin{aligned}
&+\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\int\limits_0^y \cosh(\xi\eta)\frac{i\omega}{\widehat{A}_2}Z_3\,d\eta+\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\Bigg[\int\limits_0^y\Bigg\{\cosh(\xi\eta)\frac{\xi}{Re}\\
&\times\Bigg[(M_1+G_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-(M_2+G_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg]\Bigg\}Z_4(\tau,\xi,\eta)\,d\tau d\eta\\
&+\int\limits_0^y\Bigg\{\cosh(\xi\eta)\frac{\xi}{Re}\Bigg[(K_1+E_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-(K_2+E_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg]\Bigg\}\\
&\times Z_5(\tau,\xi,\eta)\,d\tau d\eta\Bigg]-\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\Bigg\{\int\limits_0^y\Bigg\{\cosh(\xi\eta)\frac{\xi}{Re}\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_1}{\sqrt{D}}\Bigg]\\
&\times e^{(i\xi\hat{u}+\lambda_1)t}-\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_2}{\sqrt{D}}\Bigg]e^{(i\xi\hat{u}+\lambda_2)t}\Bigg\}Z_{20}(\xi,\eta)\,d\eta\Bigg\}+\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\\
&\times\int\limits_0^y\Bigg\{\cosh(\xi\eta)\frac{\xi}{Re}\Bigg\{\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)R_{53}\frac{1}{K_1\sqrt{D}}+\frac{1}{\sqrt{D}}\Bigg]e^{(i\xi\hat{u}+\lambda_1)t}\\
&-\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)R_{53}\frac{1}{K_2\sqrt{D}}+\frac{1}{\sqrt{D}}\Bigg]e^{(i\xi\hat{u}+\lambda_2)t}\Bigg\}Z_{10}(\xi,\eta)\Bigg\}\,d\eta+\frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\\
&\times\int\limits_0^y\cosh(\xi\eta)\bigg(2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_4+2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_5\bigg)\,d\eta+\frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1\Bigg\{\cosh(\xi(\eta-1))\frac{\xi}{Re}\Bigg[-(P_1+L_1)\\
&\times\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}+(P_2+L_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg]-\frac{\xi}{Re}\bigg(\cosh(\xi(\eta-1))\\
&\times\Bigg\{\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_1}{\sqrt{D}}\Bigg]\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_2}{\sqrt{D}}\Bigg]\\
&\times\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg\}\bigg)_\eta\Bigg\}Z_3(\tau,\xi,\eta)\,d\tau d\eta+\frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1\cosh(\xi(\eta-1))\frac{i\widehat{\omega}}{\widehat{A}_2}Z_3(t,\xi,\eta)\,d\eta\\
&+\frac{\xi\sinh(\xi y)}{\sinh\xi}\Bigg[\int\limits_0^y\Bigg\{\cosh(\xi(\eta-1))\frac{\xi}{Re}\Bigg[(M_1+G_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}-(M_2+G_2)\\
&\times\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg]\Bigg\}Z_4(\tau,\xi,\eta)\,d\tau d\eta+\int\limits_y^1\Bigg\{\cosh(\xi(\eta-1))\frac{\xi}{Re}\Bigg[(K_1+E_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}\\
&-(K_2+E_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}\Bigg]\Bigg\}Z_5(\tau,\xi,\eta)\,d\tau d\eta\Bigg]-\frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1\cosh(\xi(\eta-1))\\
&\times\Bigg\{\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_1}{\sqrt{D}}\Bigg]e^{(i\xi\hat{u}+\lambda_1)t}-\Bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)\frac{1}{\sqrt{D}}+\frac{K_2}{\sqrt{D}}\Bigg]
\end{aligned}
$$

$$\times e^{(i\xi\hat{u}+\lambda_2)t}\Big\}Z_{20}(\xi,\eta)\,d\eta + \frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1 \cosh(\xi(\eta-1))\frac{\xi}{Re}\bigg\{\bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)$$

$$\times R_{53}\frac{1}{K_1\sqrt{D}}+\frac{1}{\sqrt{D}}\bigg]e^{(i\xi\hat{u}+\lambda_1)t}-\bigg[\bigg(R_{35}-\frac{2\hat{a}_{12}}{\widehat{A}_2}R_{45}\bigg)R_{53}\frac{1}{K_2\sqrt{D}}+\frac{1}{\sqrt{D}}\bigg]e^{(i\xi\hat{u}+\lambda_2)t}\bigg\}$$

$$\times Z_{10}(\xi,\eta)\,d\eta + \frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_0^y \cosh(\xi(\eta-1))\bigg(2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_4+2\frac{\hat{a}_{12}}{\widehat{A}_2}Z_5\bigg)\,d\eta. \tag{4.1}$$

Here P_1, P_2, M_1, M_2, G_1, G_2, E_1, E_2, K_1, K_2, L_1, and L_2 are coefficients depending only on a stationary solution. Consider the parts of the integrals connected with the component $Z_3(t,\xi,y)$. Namely, we have

$$I_1 = \frac{\xi\sinh(\xi(y-1))}{\sinh\xi}\int\limits_0^y\bigg\{\cosh(\xi\eta)\frac{\xi}{Re}\bigg[-(P_1+L_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}$$

$$+(P_2+L_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}\bigg]\bigg\}Z_3(\tau,\xi,\eta)\,d\tau d\eta + \frac{\xi\sinh(\xi y)}{\sinh\xi}\int\limits_y^1\bigg\{\cosh(\xi(\eta-1))\frac{\xi}{Re}$$

$$\times\bigg[-(P_1+L_1)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}+(P_2+L_2)\int\limits_0^t e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}\bigg]\bigg\}Z_3(\tau,\xi,\eta)\,d\tau d\eta. \tag{4.2}$$

We can use the Taylor formula expanding the integrands at y. We infer

$$I_1 = \int\limits_0^t\bigg\{\Gamma'(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}$$

$$+\Gamma'''(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}\frac{1}{\xi^2}$$

$$+\Gamma^V(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}\frac{1}{\xi^4}+\cdots\bigg\}Z_3(\tau,\xi,y)\,d\tau$$

$$+\int\limits_0^t\bigg\{\Gamma(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}$$

$$+\Gamma''(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}\frac{1}{\xi^2}$$

$$+\Gamma^{IV}(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}\frac{1}{\xi^4}+\cdots\bigg\}Z_{3y}(\tau,\xi,y)\,d\tau$$

$$+\int\limits_0^t\bigg\{\Gamma'(y,t-\tau)\frac{\sinh\xi-\sinh(\xi y)-\sinh(\xi(1-y))}{\sinh\xi}\frac{1}{\xi^2}+\cdots\bigg\}Z_{3yy}(\tau,\xi,y)\,d\tau+\cdots, \tag{4.3}$$

where

$$\Gamma(y,t-\tau)=\frac{1}{Re}[-(P_1+L_1)e^{(i\xi\hat{u}+\lambda_1)(t-\tau)}+(P_2+L_2)e^{(i\xi\hat{u}+\lambda_2)(t-\tau)}], \tag{4.4}$$

and the derivatives of this function are taken with respect to y.

Note that the summands

$$\frac{\sinh(\xi(y-1))}{\sinh\xi}y + \frac{\sinh(\xi y)}{\sinh\xi}(1-y), \quad 0<y<1,$$

in (4.4), decaying exponentially as $|\xi|\to\infty$, are omitted.

Arguing similarly in the case of the remaining integrals leads to the spectral equation:

$$\begin{aligned}
&s - i\xi\hat{u} - 2\widehat{A}_2\Bigg\{\Bigg\{\Bigg[-(P_1+L_1)'\frac{1}{s-i\xi\hat{u}(y)-\lambda_1} + (P_2+L_2)'\frac{1}{s-i\xi\hat{u}(y)-\lambda_2}\\
&-(P_1+L_1)\frac{i\xi\hat{u}'(y)+\lambda_1'}{(s-i\xi\hat{u}(y)-\lambda_1)^2} + (P_2+L_2)\frac{i\xi\hat{u}'(y)+\lambda_2'}{(s-i\xi\hat{u}(y)-\lambda_2)^2}\Bigg] + \frac{1}{\xi^2}\Bigg[-\frac{(P_1+L_1)'''}{s-i\xi\hat{u}(y)-\lambda_1}\\
&+\frac{(P_2+L_2)'''}{s-i\xi\hat{u}(y)-\lambda_2} - 3(P_1+L_1)''\frac{i\xi\hat{u}'(y)+\lambda_1'}{(s-i\xi\hat{u}(y)-\lambda_1)^2} + 3(P_2+L_2)''\frac{i\xi\hat{u}'(y)+\lambda_2'}{(s-i\xi\hat{u}(y)-\lambda_2)^2}\\
&-3(P_1+L_1)'\frac{i\xi\hat{u}''(y)+\lambda_1''}{(s-i\xi\hat{u}(y)-\lambda_1)^2} + 3(P_2+L_2)'\frac{i\xi\hat{u}''(y)+\lambda_2''}{(s-i\xi\hat{u}(y)-\lambda_2)^2}\\
&-\frac{(P_1+L_1)(i\xi\hat{u}'''+\lambda_1''')}{(s-i\xi\hat{u}(y)-\lambda_1)^2} + \frac{(P_2+L_2)(i\xi\hat{u}'''+\lambda_2''')}{(s-i\xi\hat{u}(y)-\lambda_2)^2} - 6\frac{(P_1+L_1)(i\xi\hat{u}'+\lambda_1')(i\xi\hat{u}''+\lambda_1'')}{(s-i\xi\hat{u}(y)-\lambda_1)^3}\\
&+6\frac{(P_2+L_2)(i\xi\hat{u}'+\lambda_2')(i\xi\hat{u}''+\lambda_2'')}{(s-i\xi\hat{u}(y)-\lambda_2)^3} - 6\frac{(P_1+L_1)'(i\xi\hat{u}'+\lambda_1')^2}{(s-i\xi\hat{u}(y)-\lambda_1)^3}\\
&+6\frac{(P_2+L_2)'(i\xi\hat{u}'+\lambda_2')^2}{(s-i\xi\hat{u}(y)-\lambda_2)^3} - 6\frac{(P_1+L_1)(i\xi\hat{u}'+\lambda_1')^3}{(s-i\xi\hat{u}(y)-\lambda_1)^4} + 6\frac{(P_2+L_2)(i\xi\hat{u}'+\lambda_2')^3}{(s-i\xi\hat{u}(y)-\lambda_2)^4}\Bigg]\\
&+\frac{1}{\xi^4}[\widehat{\Gamma}^V(y,s)+\cdots] + i\Bigg[\Bigg(\frac{\widehat{\omega}}{\widehat{A}_2}\Bigg)'\frac{1}{\xi} + \Bigg(\frac{\widehat{\omega}}{\widehat{A}_2}\Bigg)'''\frac{1}{\xi^3}+\cdots\Bigg]\Bigg\}\Bigg\} = 0. \qquad (4.5)
\end{aligned}$$

Expanding the fractions $1/(s-i\xi\hat{u}-\lambda_1)^k$, $1/(s-i\xi\hat{u}-\lambda_2)^k$, $k=1,2,\dots$, as $|\xi|\to\infty$ in the powers of $1/(s-i\xi\hat{u})$ and equating the coefficients of the same powers we can obtain a formal asymptotic expansion for the roots of (4.5).

The method of indefinite coefficients allows us to justify the following decomposition of the roots of the equation (3.14) in the powers of $\xi^{\frac{1}{3}}$ as $|\xi|\to\infty$:

$$s = i\xi\hat{u} + \sqrt[3]{Q(y)}\xi^{\frac{2}{3}} + R(y)\xi^{\frac{1}{3}} + \cdots. \qquad (4.6)$$

REMARK 5. The main point connected with the decomposition (4.6) is as follows: at least one of the roots satisfies the property Re $s\to+\infty$ as $|\xi|\to\infty$.

Differentiating the expressions for Z_3 with respect to y, we obtain an integral equation for $Z_{3y}(t,\xi,y)$ through the higher order derivatives $Z_{3yy}, Z_{3yyy},\dots$, the components $Z_4(t,\xi,y)$, $Z_5(t,\xi,y)$ and their first derivatives, and the initial data $Z_{10}(\xi,y)$, $Z_{20}(\xi,y)$, and $Z_{30}(\xi,y)$. Arguing by analogy, we can state that Z_{3y} also satisfies some spectral equation. The formal asymptotic expansions of the roots of this equation are found with the use of the Newton diagram [15–17].

Thus, the derivative $Z_{3y}(t,\xi,y)$ is determined through the higher order derivatives $Z_{3yy}(t,\xi,y), Z_{3yyy}(t,\xi,y),\dots$, and the above-mentioned data.

Arguing by induction and inserting the values of the derivatives of the component $Z_3(t,\xi,y)$ into the right-hand side of (4.1), we arrive at a solution to the Cauchy problem for the integro-differential equation (4.2) in the form of a formal asymptotic series as $|\xi|\to\infty$.

Theorem 2 is proven.

REFERENCES

1. *Altukhov Yu. A., Golovicheva I. E., and Pyshnograi G. V.* Molecular approach in linear polymer dynamics: Theory and numerical experiment // Fluid Dynamics. 2000. V. 35, N 1. P. 1–9.

2. *Pyshnograi G. V., Pokrovskii V. N., Yanovskii Yu. G., Obraztsov I. F., and Karnet Yu. N.* Constitutive equation of nonlinear viscoelastic (polymeric) media in zero approximation with respect to molecular-theory parameters and the consequences of shear and tension // Doklady Physics. 1994. V. 39. P. 889–892.

3. *Volkov V. S. and Vinogradov G. V.* Molecular theories of nonlinear viscoelasticity of polymers // Rheol. Acta. 1984. V. 23, N 3. P. 231–237.

4. *Altukhov Yu. A., Gusev A. S., Makarova M. A., and Pyshnograi G. V.* Generalization of the Poiseuille law for a plane-parallel flow of viscoelastic media // Mekh. Kompozit. Mater. Konstruktsii. 2007. V. 13, N 4. P. 581–590.

5. *Altukhov Yu. A. and Pyshnograi G. V.* Inlet flows of linear polymer fluid flow in a 4 : 1 channel // Mekh. Kompozit. Mater. Konstruktsii. 2001. V. 7, N 1. P. 16–23.

6. *Altukhov Yu. A., Gusev A. S., and Pyshnograi G. V.* Introduction to Mesoscopic Theory of Fluctuating Polymer Systems [in Russian]. Barnaul: AltGPA, 2012.

7. *Blokhin A. M. and Bambaeva N. V.* Finding the solutions of Poiseuille and Kette type for equations of an incompressible viscoelastic polymer fluid // Vestn. Novosib. Gos. Univ., Ser. Mat. Mekh. Inform. 2011. V. 11, N 2. P. 3–14.

8. *Wassner E., Schmidt M., and Münstedt H.* Entry flow of a low-density-polyethylene melt into a slit die: An experimental study by Laser Doppler Velocimetry // J. Rheol. 1999. V. 43, N 6. P. 1339–1353.

9. *Loĭtsyanskiĭ L. G.* Fluid Mechanics [in Russian]. Moscow: Nauka, 1978.

10. *Godunov S. K.* Equations of Mathematical Physics. Moscow: Nauka, 1979.

11. *Vladimirov V. S.* Equations of Mathematical Physics. New York: Marcel Dekker, Ins., 1971.

12. *Vladimirov V. S.* Generalized Functions in Mathematical Physics [in Russian]. Moscow: Nauka, 1979.

13. *Gantmakher F. R.* The Theory of Matrices. AMS Chelsea Publishing: Amer. Math. Soc., 2000.

14. *Fedoryuk M. V.* Asymptotics: Integrals and Series [in Russian]. Moscow: Nauka, 1987.

15. *Newton I.* Mathematical Works [Russian translation]. Moscow; Leningrad: Gos. Izdat. Tekh.-Teor. Lit., 1937.

16. *Puiseux V. J.* Recherches sur les fonctions algébriques // J. Math. Pures Appl. 1850. V. 15. P. 365–480.

17. *Vainberg M. M. and Trenogin V. A.* Theory of Branching of Solutions of Non-Linear Equations [in Russian]. Moscow: Nauka, 1969.

August 30, 2015

A. M. Blokhin; D. L. Tkachev
Sobolev Institute of Mathematics, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
`blokhin@math.nsc.ru; tkachev@math.nsc.ru`

A. V. Yegitov
Sobolev Institute of Mathematics, Novosibirsk, Russia
`eav15@mail.ru`

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 517.518.87

OPTIMAL DISTRIBUTION OF NODES OF A QUADRATURE FORMULA WITH WEIGHT

E. N. Bulgatova and E. B. Pavlova

Abstract. The authors consider the quadrature formulas with weight. Some method is given for finding the asymptotically optimal distribution of nodes of these formulas.

Keywords: weighted quadrature formula, optimal distribution of nodes

We consider a distribution of nodes of a weighted quadrature formula in dependence on properties of the weight and the behavior of the integrand from a certain function space.

Assume that $g(x) \in L_{p'}$ is a weight, $1 < p \leq \infty$, $f \in W_p^m$, and we need to calculate the integral

$$If = \int_0^1 g(x)f(x)\,dx.$$

Divide the integration interval $[0,1]$ into parts $[x_{\beta-1}, x_\beta]$, $\beta = 1,2,\dots,N$, $x_0 = 0$, $x_N = 1$, and consider on each of the parts the Lagrange interpolation formula

$$L_m(x - x_{\beta-1}) = \sum_{\gamma=0}^{m} \frac{\omega_m(x - x_{\beta-1})}{\omega'_m(x_{\beta-1+\gamma})(x - x_{\beta-1+\gamma})} f(x_{\beta-1+\gamma})$$

with an arbitrary distribution of the nodes $x_{\beta-1} < x_\beta < x_{\beta+1} < \dots < x_{\beta-1+m}$, $\beta = 1,2,\dots,N$, and $\omega_m(x-x_{\beta-1}) = (x-x_{\beta-1})(x-x_\beta)(x-x_{\beta+1})\dots(x-x_{\beta-1+m})$.

The integral If is representable as the sum

$$If = \sum_{\beta=1}^{N} \int_{x_{\beta-1}}^{x_\beta} g(x)f(x)dx.$$

The integral over each of the parts is calculated by the formula

$$\int_{x_{\beta-1}}^{x_\beta} g(x)f(x)\,dx \approx \int_{x_{\beta-1}}^{x_\beta} g(x)L_m(x-x_{\beta-1})\,dx = h^* \sum_{\gamma=0}^{m} C_\gamma(\beta)f(x_{\beta-1+\gamma}),$$

where the coefficients are determined as follows:

$$C_\gamma(\beta) = \int_{x_{\beta-1}}^{x_\beta} \frac{\omega_m(x-x_{\beta-1})g(x)}{\omega'_m(x_{\beta-1+\gamma})(x-x_{\beta-1+\gamma})}\,dx.$$

Using the error of the Lagrange interpolation formula, we infer

$$f(x) - L_m(x - x_{\beta-1}) = \frac{(x - x_{\beta-1})(x - x_\beta)(x - x_{\beta+1})\dots(x - x_{\beta-1+m})}{(m+1)!} f^{m+1}(\xi),$$

where $\xi \in (x_{\beta-1}, x_{\beta-1+m})$.

Note that $g(x + x_\beta) = g(x_\beta) + o(1)$, $\max_{x\in[x_{\beta-1+\gamma}, x_{\beta+\gamma}]} |x - x_{\beta+\gamma}| = |x_{\beta+\gamma} - x_{\beta-1+\gamma}| + o(1)$, $\gamma = 0, 1, 2, \dots, m-1$ as $N \to \infty$, and $x \in [x_{\beta-1}, x_\beta]$. We can estimate the error

$$\int\limits_{x_{\beta-1}}^{x_\beta} g(x)f(x)\,dx - h^* \sum_{\gamma=0}^{m} C_\gamma(\beta) f(x_{\beta-1+\gamma})$$

$$\leq g(x_\beta)\frac{|x_{\beta+\gamma} - x_{\beta-1+\gamma}|^{m+1}}{(m+1)!} \max_{x\in[x_{\beta-1+\gamma}, x_{\beta+\gamma}]} |f^{(m+1)}(x)|(1 + o(1)).$$

Assume that $f(x) \in W^m_\infty$ and $\max\limits_{x\in[0,1]} |f^{(m+1)}| \leq M$. In this case the total error is equal to

$$R = \sum_{\beta=1}^{N} g(x_\beta)|x_\beta - x_{\beta-1}|^{m+1} \frac{M}{(m+1)!}(1 + o(1)). \tag{1}$$

Consider two increasing number sequences

$$x_0 = 0 < x_1 < x_2 < \dots < x_\gamma < \dots < x_N = 1,$$

$$0 < h < 2h < \dots < \gamma h < \dots < Nh = 1.$$

Involving these sequences, we can construct a differentiable function $x = \varphi(t)$ with values $x_\gamma = \varphi(h\gamma)$, $\gamma = 1, 2, \dots, N$, such that $x(0) = 0$ and $x(1) = 1$.

Theorem. *Assume that* $f \in W^m_\infty$, $\max_{x\in[0,1]} |f^{(m+1)}(x)| \leq M$, $g(x) \in L_1(0, 1)$, *and*

$$\int\limits_0^1 g(x)f(x)\,dx \approx \sum_{\beta=1}^{N}\sum_{\gamma=0}^{m} C_\gamma(\beta) f(x_{\beta-1+\gamma})$$

is a weighted formula. Then the optimal distribution of the nodes x_β, $\beta = 1, 2, \dots, N$, *is defined by the function* $x = \varphi(t)$ *satisfying the differential equation*

$$\frac{d}{dt}(g(\varphi(t))(\varphi'(t))^{m+1} = 0,$$

the initial conditions $\varphi(0) = 0$ *and* $\varphi(1) = 1$ *and given (in implicit form) by the integral*

$$\int\limits_0^x (g(x))^{\frac{1}{m+1}}\,dx = \int\limits_0^1 (g(x))^{\frac{1}{m+1}}\,dx \cdot t$$

.

Proof. Take the values $\varphi(\frac{\beta}{N})$ of the twice differentiable function $\varphi(t)$ such that $\varphi(0) = 0$ and $\varphi(1) = 1$ at x_β.

By continuity of $\varphi(t)$, we have

$$x_\beta - x_{\beta-1} = \varphi\Big(\frac{\beta}{N}\Big) - \varphi\Big(\frac{\beta-1}{N}\Big) = \varphi'\Big(\frac{\beta}{N}\Big)\frac{1}{N} + o\Big(\frac{1}{N^{m+1}}\Big).$$

The formula (1) takes the form

$$R = \frac{1}{N^{m+1}} \sum_{\beta=0}^{N-1} \left[\left(\varphi' \left(\frac{\beta}{N} \right) \right)^{m+1} g(x_\beta) \frac{M}{(m+1)!} \right] + o\left(\frac{1}{N^{m+1}} \right).$$

The expression in brackets is a Riemann quadrature sum for the integral

$$\int_0^1 (\varphi'(t))^{m+1} g(\varphi(t)) \frac{M}{(m+1)!} \, dt. \tag{2}$$

In view of (2), we infer

$$R = \frac{1}{N^m} \int_0^1 (\varphi'(t))^{m+1} g(\varphi(t)) \frac{M}{(m+1)!} \, dt + o\left(\frac{1}{N^{m+1}} \right). \tag{3}$$

To determine the optimal distribution of nodes, we minimize the main term

$$A = \int_0^1 (\varphi'(t))^{m+1} g(\varphi(t)) \, dt$$

in (3). Take the function φ as a new independent variable in the integral A. Then

$$A = \int_0^1 (t'(\varphi))^{-m} g(\varphi) \, d\varphi.$$

Write out the Lagrange function

$$F(t(\varphi) + \lambda \tau(\varphi)) = \int_0^1 (t'(\varphi) + \lambda \tau'(\varphi))^{-m} g(\varphi) \, d\varphi,$$

where $\tau(0) = 0$, $\tau(1) = 0$. Calculate the derivative at $\lambda = 0$ and put it equal to zero, i.e.,

$$F'(t(\varphi)) = \int_0^1 \left(\frac{-m}{(t'(\varphi))^{m+1}} g(\varphi) \tau'(\varphi) + (t'(\varphi))^{-m} \frac{dg(\varphi)}{d\lambda} \right) d\varphi = 0. \tag{4}$$

Note that $g(\varphi)$ is independent of λ and so $\frac{dg(\varphi)}{d\lambda} = 0$.

Integrating by parts in (4) yields

$$\frac{d}{d\varphi} \left[\frac{1}{(t'(\varphi))^{m+1}} g(\varphi) \right] = 0 \tag{5}$$

or $g(\varphi)(\varphi'(t))^{m+1} = C_0$.

The initial conditions implies that

$$\int_0^x (g(x))^{\frac{1}{m+1}} \, dx = \int_0^1 (g(x))^{\frac{1}{m+1}} \, dx \cdot t$$

The theorem is proven.

Next, we examine the simplest weight $g(x) = |x|^s$, $-1 < s < 1$. In this case (5) takes the form

$$\frac{d}{d\varphi}\left[\frac{1}{(t'(\varphi))^{m+1}}\varphi^s\right] = 0.$$

Integrating the equation and using the initial conditions we conclude that $x = t^{\frac{m+1}{m+s+1}}$. In this case an optimal node distribution corresponds to the points

$$x_\beta = \left(\frac{\beta}{N}\right)^{\frac{m+1}{m+s+1}}, \quad \beta = 0, 1, \ldots, N-1.$$

Assume that we need to calculate $\int_0^1 x^s\varphi(x)\,dx$, $-1 < s < 1$, with $\varphi(0) \neq 0$. The integration interval is divided into the parts $[x_{\beta-1}, x_\beta]$, $\beta = 1, 2, \ldots, N$, $x_0 = 0$, $x_N = 1$. In this case we have

$$\int\limits_0^1 x^s\varphi(x)\,dx = \sum_{\beta=1}^N \int\limits_{x_{\beta-1}}^{x_\beta} x^s\varphi(x)\,dx.$$

Each of the integrals is calculated by the formula

$$\int\limits_{x_{\beta-1}}^{x_\beta} x^s\varphi(x)\,dx \approx h^*\sum_{\gamma=0}^m C_\gamma(\beta)\varphi(x_{\beta-1+\gamma}),$$

where $h^* = x_\beta - x_{\beta-1}$ and the coefficients are determined from the system

$$\int\limits_0^1 (x_{\beta-1} + h^*x)^s x^\alpha\,dx = \sum_{\gamma=0}^m C_\gamma(\beta)(x_{\beta-1+\gamma})^\alpha, \quad \alpha = 0, 1, 2, \ldots, m,\ \beta = 1, 2, \ldots, N.$$

The integral over $[0, 1]$ is equal to

$$\int\limits_0^1 x^s\varphi(x)\,dx \approx \sum_{\beta=1}^N (x_\beta - x_{\beta-1})\sum_{\gamma=0}^m C_\gamma(\beta)\varphi(x_{\beta-1+\gamma}).$$

Arrange the weighted formula for $s = -\frac{1}{2}$ and $m = 2$. The coefficients $C_\gamma(\beta)$ are determined from the system

$$\int\limits_0^1 (x_{\beta-1} + h^*x)^{-\frac{1}{2}} x^\alpha\,dx = \sum_{\gamma=0}^2 C_\gamma(\beta)(x_{\beta-1+\gamma})^\alpha, \quad \alpha = 0, 1, 2,\ \beta = 1, 2, \ldots, N.$$

Assume that $\varphi(t)$ is a continuously differentiable function on $[0, 1]$, $\varphi(0) = 0$ and $\varphi(1) = 1$, $x_\beta = \left(\frac{\beta}{N}\right)^{\frac{6}{5}} = (h\beta)^{\frac{6}{5}}$, $\max_{x\in[0,1]}|\varphi^m(x)| \leq M$. The system for the coefficients is representable as

$$\int\limits_0^1 (((\beta-1)h)^{\frac{6}{5}} + ((\beta h)^{\frac{6}{5}} - ((\beta-1)h)^{\frac{6}{5}})x)^{-\frac{1}{2}} x^\alpha\,dx = \sum_{\gamma=0}^2 C_\gamma(\beta)((\beta-1+\gamma)h)^{\frac{6\alpha}{5}},$$

$\alpha = 0, 1, 2$, $\beta = 1, 2, \ldots, N$, $h = \frac{1}{N}$.

The integral is approximately equal to

$$\int\limits_0^1 x^{-\frac{1}{2}}\varphi(x)\,dx \approx \sum_{\beta=1}^N ((\beta h)^{\frac{6}{5}} - ((\beta-1)h)^{\frac{6}{5}})\sum_{\gamma=0}^2 C_\gamma(\beta)\varphi\left(\left(\frac{\beta-1+\gamma}{N}\right)^{\frac{6}{5}}\right).$$

REFERENCES

1. *Bakhvalov S. N.* Numerical Methods: Analysis, Algebra, Ordinary Differential Equations. Moscow: Mir, 1977.

2. *Shoinzhurov Ts. B.* Estimation of the norm of error functional of cubature formulas in various function spaces. Ulan-Ude: Izdat. BNTs SO RAN, 2005.

August 14, 2015

E. N. Bulgatova; E. B. Pavlova
North-Eastern State University of Technology and Control, Ulan-Ude, Russia
`belena77@mail.ru`; `pavlova2607@mail.ru`

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 517.518.87

AN UPPER ESTIMATE FOR THE ERROR FUNCTIONAL OF QUADRATURE FORMULAS WITH A SYMMETRIC BOUNDARY LAYER

E. G. Vasil′eva and N. B. Tsyrenzhapov

Abstract. We obtain an upper estimate for the error functional of the quadrature formulas with a symmetric boundary layer. We singled out the constant in this estimate in explicit form.

Keywords: estimate, error functional, quadrature formula

Cubature formulas with a regular boundary layer for a domain Ω and the corresponding error functionals are defined in [1].

To begin with, we choose an error functional for a quadrature formula on $(0, 1)$ in the set of error functionals with a regular boundary layer, and estimate its norm from above in the $L_p^m(E_1)$ space.

Put

$$l(x) = \varepsilon_{(0,1)}(x) - \sum_{\gamma=0}^{m} C_\gamma \delta(x-\gamma), \quad \langle l, x^\alpha \rangle = 0,\ \alpha = 0, 1, \dots, m,$$

$$\|l\|_{C^*} = 1 + \sum_{\gamma=0}^{m} |C_\gamma| < \infty,$$

$$l_1(x) = \varepsilon_{(0,m)}(x) - \sum_{\gamma=0}^{m} F_\gamma \delta(x-\gamma), \quad \langle l_1, x^\alpha \rangle = 0,\ \alpha = 0, 1, \dots, m,$$

$$\|l_1\|_{C^*} = m + \sum_{\gamma=0}^{m} |F_\gamma| < \infty, \quad (a, b) = [0, 1), \quad \frac{1}{N} = h.$$

Summing the elementary functionals $l\left(\frac{x}{h} - \beta\right)$, $\beta = 0, 1, \dots, N-m-1$, and $l_1\left(\frac{x}{h} - N + m\right)$, we can construct the error functional of a quadrature formula with a regular boundary layer for the half-interval $[0, 1)$ as follows:

$$l^h_{(0,1)}(x) = \sum_{\beta=0}^{N-m-1} l\left(\frac{x}{h} - \beta\right) + l_1\left(\frac{x}{h} - N + m\right).$$

By construction, $l^h_{(0,1)}(x) \in L_p^{m*}$.

Theorem. *Assume that $l^h_{(0,1)}(x)$ is an error functional of the quadrature formula with a regular boundary layer for the half-interval $(0,1)$, $\operatorname{supp} l^h_{(0,1)}(x) \subseteq [0,1]$ and $l^h_{(0,1)}(x) \in L_p^{m*}$. Then the norm of $l^h_{(0,1)}(x)$ as $h \to 0$ satisfies the estimate*

$$\left\| l^h_{(0,1)}(x) \right\|_{L_p^{m*}} \le h^m \left[\int_0^1 \left| \sum_{\beta \ne 0} \frac{e^{2\pi i \beta x}}{(2\pi i \beta)^m} \right|^{p'} dx \right]^{\frac{1}{p'}}$$

$$+ h^m \left[\frac{(m+2) + \sum\limits_{\gamma=0}^{m} (|F_\gamma| + 2|C_\gamma|)}{2(m-1)!} \right] m^{m+1-\frac{1}{p}} h^{1-\frac{1}{p}}.$$

PROOF. Transform the periodic error functional $\tilde{l}_0(\frac{x}{h})$ as follows:

$$\tilde{l}_0\left(\frac{x}{h}\right) = \sum_{\beta=-\infty}^{\infty} l\left(\frac{x}{h} - \beta\right)$$

$$= \sum_{\beta=-\infty}^{-1} l\left(\frac{x}{h} - \beta\right) + \sum_{\beta=0}^{N-m-1} l\left(\frac{x}{h} - \beta\right) + \sum_{\phi=N-m}^{\infty} l\left(\frac{x}{h} - \beta\right)$$

$$= \sum_{\beta=0}^{N-m-1} l\left(\frac{x}{h} - \beta\right) + l^h_{(0,1)*}(x), \tag{1}$$

where

$$l^h_{(0,1)^*}(x) = \sum_{\beta=-\infty}^{-1} l\left(\frac{x}{h} - \beta\right) + \sum_{\beta=N-m}^{\infty} l\left(\frac{x}{h} - \beta\right).$$

Equality (1) implies that the above error functional with a regular boundary layer is representable as

$$l^h_{(0,1)}(x) = \tilde{l}_0\left(\frac{x}{h}\right) + l_1\left(\frac{x}{h} - N + m\right) - l^h_{(0,1)^*}(x). \tag{2}$$

By construction, the support of $l^h_{(0,1)}(x)$ coincides with $\operatorname{supp} l^h_{(0,1)}(x) = [0,1]$. In this case the norm of the error functional is written out explicitly [2] and

$$\left\| l^h_{(0,1)}(x) \right\|_{L_p^{m*}} = \left[\int\limits_{-\infty}^{\infty} \left| \varepsilon_{2m}^{(m)}(x) * l^h_{(0,1)}(x) \right|^{p'} dx \right]^{\frac{1}{p'}},$$

If $x \in (-\infty, 0) \cup (1, \infty)$ and $y \in \operatorname{supp} l^h_{(0,1)}(x)$ then the expression $(x-y)$ has constant sign. Hence, $\varepsilon_{2m}^{(m)}(x) * l\left(\frac{x}{h} - \beta\right) = 0$ for all $h, \beta \in [0,1)$. Therefore, we can state that

$$\left\| l^h_{(0,1)}(x) \right\|_{L_p^{m*}} = \left[\int\limits_0^1 \left| \varepsilon_{2m}^{(m)}(x) * l^h_{(0,1)}(x) \right|^{p'} dx \right]^{\frac{1}{p'}}.$$

The representation (2) of $l^h_{(0,1)}(x)$, the relation $\operatorname{supp} l_1\left(\frac{x}{h} - N + m\right) \subset [hN - hm, hN]$, and the Minkowski inequality yield

$$\big\|l^h_{(0,1)}(x)\big\|_{L_p^{m*}}$$

$$\leq \left[\int_0^1 \left|\varepsilon_{2m}^{(m)}(x) * \tilde{l}_0\left(\frac{x}{h}\right)\right|^{p'} dx\right]^{\frac{1}{p'}} + \left[\int_{hN-hm}^{hN} \left|\varepsilon_{2m}^{(m)}(x) * l_1\left(\frac{x}{h} - N + m\right)\right|^{p'} dx\right]^{\frac{1}{p'}}$$

$$+\left[\int_0^1 \left|\sum_{\beta=-\infty}^{-1} \varepsilon_{2m}^{(m)}(x) * l\left(\frac{x}{h} - \beta\right)\right|^{p'} dx\right]^{\frac{1}{p'}} + \left[\int_0^1 \left|\sum_{\beta=N-m}^{\infty} \varepsilon_{2m}^{(m)}(x) * l\left(\frac{x}{h} - \beta\right)\right|^{p'} dx\right]^{\frac{1}{p'}}$$

$$= I_1 + I_2 + I_3 + I_4. \tag{3}$$

Estimate I_1 as follows:

$$I_1 = \left[\sum_{h\gamma\in[0,1)} \int_{\Delta_{h\gamma}} \left|\sum_{\beta\neq 0} \frac{e^{2\pi i h^{-1}\beta}x}{(2\pi i h^{-1}\beta)^m}\right|^{p'} dx\right]^{\frac{1}{p'}} = \left\langle \begin{array}{c} x \to h\gamma + x \\ x \to hx \\ dx \to hdx \end{array} \right\rangle$$

$$= h^m \left[\int_0^1 \left|\sum_{\beta\neq 0} \frac{e^{2\pi i\beta x}}{(2\pi i\beta)^m}\right|^{p'} dx\right]^{\frac{1}{p'}} = h^m J_m, \tag{4}$$

where $\Delta_{h\gamma} = \{x \in E_1,\ h\gamma \leq x < h\gamma + h\}$.

We can transform the convolution

$$\varepsilon_{2m}^{(m)}(x) * l_1\left(\frac{x}{h} - N + m\right) = \left\langle \frac{y}{h} \to y \right\rangle$$

$$= \int_{-\infty}^{\infty} \frac{(x-hy)^{m-1}\operatorname{sgn}(x-hy)}{2(m-1)!} l_1(y - N + m)h\, dy = \langle y - N + m \to y\rangle$$

$$= \int_0^m \frac{(x - hy + hN - hm)^{m-1}}{2(m-1)!} \operatorname{sgn}(x - hy + hN - hm) h l_1(y)\, dy. \tag{5}$$

In view of (5), I_2 in (3) is estimated as

$$I_2 = \left[\int_{hN-hm}^{hN} \left|\int_0^m \frac{(x - hy + hN - hm)^{m-1}}{2(m-1)!} \operatorname{sgn}(x - hy + hN - hm) l_1(y) h\, dy\right|^{p'} dx\right]^{\frac{1}{p'}}$$

$$= \langle x \to hx + hN - hm\rangle = h^{m+\frac{1}{p'}} \left[\int_0^m \left|\int_0^m \frac{(x-y)^{m-1}\operatorname{sgn}(x-y)}{2(m-1)!} l_1(y)\, dy\right|^{p'} dx\right]^{\frac{1}{p'}}$$

$$\leq h^{m+1-\frac{1}{p}} \left[\int_0^m \left|\max_{x,y\in[0,m]} |x-y|^{m-1} \|l_1\|_{C^*} \frac{1}{2(m-1)!}\right|^{p'} dx\right]^{\frac{1}{p'}}$$

$$\leq h^{m+1-\frac{1}{p}} m^{m+1-\frac{1}{p}} \frac{m + \sum\limits_{\gamma=0}^{m} |F_\gamma|}{2(m-1)!}. \tag{6}$$

Taking the equality

$$\sum_{\beta=-\infty}^{-m} \varepsilon_{2m}^{(m)} * l\left(\frac{x}{h} + \beta\right) = 0 \text{ for } x \in [0, h(m-1))$$

into account, we can transform the convolution

$$\sum_{\beta=-m+1}^{-1} \int \frac{(x-y)^{m-1}\operatorname{sgn}(x-y)}{2(m-1)!} l\Big(\frac{x}{h}+\beta\Big)\,dy \quad = \Big\langle \frac{y}{h} \to y,\ y-\beta \to y \Big\rangle$$

$$= \sum_{\beta=-m+1}^{-1} \int \frac{(x-hy+h\beta)^{m-1}\operatorname{sgn}(x-hy+h\beta)}{2(m-1)!} l(y)h\,dy. \tag{7}$$

To estimate I_3 on the base of (7), we derive that

$$I_3 = \Bigg[\int\limits_0^{h(m-1)} \Bigg| \sum_{\beta=-m+1}^{-1} \int\limits_{h\beta}^{h\beta+hm} \frac{(x-hy+h\beta)^{m-1}}{2(m-1)!} \operatorname{sgn}(x-hy+h\beta) l(y)h\,dy \Bigg|^{p'} dx\Bigg]^{\frac{1}{p'}}$$

$$= \langle x \to hx,\ x-\beta \to x\rangle$$

$$= h^{m+\frac{1}{p'}} \Bigg[\int\limits_0^{h(m-1)} \Bigg| \sum_{\beta=-m+1}^{-1} \int\limits_0^{m} \frac{(x-y)^{m-1}\operatorname{sgn}(x-y)}{2(m-1)!} l(y)\,dy \Bigg|^{p'} dx\Bigg]^{\frac{1}{p'}}$$

$$\le h^{m+1-\frac{1}{p}} \Bigg[\int\limits_0^{m-1} \Bigg| \sum_{\beta=-m+1}^{-1} \max_{x,y\in[0,m]} |x-y| \frac{\|l\|_{C^*}}{2(m-1)!} \Bigg|^{p'} dx\Bigg]^{\frac{1}{p'}}$$

$$\le h^{m+1-\frac{1}{p}} m^{m+1-\frac{1}{p}} \frac{1+\sum\limits_{\gamma=0}^{m} |C_\gamma|}{2(m-1)!}. \tag{8}$$

Since $x \in [0, 1+h(m-1)]$ and $\varepsilon_{2m}^{(m)} * l\big(\frac{x}{h}-\beta\big) = 0$ for all $\beta > N+m-1$, similar arguments validate the inequality

$$I_4 \le h^{m+1-\frac{1}{p}} m^{m+1-\frac{1}{p}} \frac{1+\sum\limits_{\gamma=0}^{m} |C_\gamma|}{2(m-1)!}. \tag{9}$$

Collecting (3), (4), (6), (8), and (9), we infer that

$$\big\| l_{(0,1)}^h \big\|_{L_p^{m*}} \le h^m J_m + \frac{(m+2)+\sum\limits_{\gamma=0}^{m}(|F_\gamma|+2|C_\gamma|)}{2(m-1)!} m^{m+1-\frac{1}{p}} h^m h^{1-\frac{1}{p}}.$$

The theorem is proven.

REFERENCES

1. *Sobolev S. L.* Cubature Formulas and Modern Analysis: An Introduction. Montreux: Gordon and Breach Science Publishers, 1974.
2. *Shoinzhurov Ts. B.* Estimation of the norm of an error functional of cubature formulas in various function spaces. Ulan-Ude: Izdat. BNTs SO RAN, 2005.

August 25, 2015

E. G. Vasil′eva
North-Eastern State University of Technology and Control, Ulan-Ude, Russia
vasil_eg@mail.ru

Tsyrenzhapov N. B.
North-Eastern Institute of Culture, Ulan-Ude, Russia
nimac@mail.ru

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 517.55

THE $\bar\partial\partial$–EQUATION ON A POSITIVE CURRENT

T. N. Nikitina

Abstract. We study the induced $\bar\partial\partial$-equation on a positive current on a complex manifold. We show that L^2-estimates hold for the $\bar\partial\partial$-equation on a positive closed $(1,1)$-current in a pseudoconvex domain in $\mathbb{C}^n$. We also discuss currents of higher bidegree.

Keywords: $\bar\partial\partial$-equation, positive current, differential form, complex manifold, primitive form, definite quadratic form, differential operator on a current, existence theorem for $\bar\partial\partial$ on a closed current, current of higher bidegree

1. Introduction

Let M be a complex manifold and let T be a positive current on M. If u and f are smooth differential forms on M then we say that

$$\bar\partial\partial u = f \text{ on } T \text{ if } \bar\partial\partial u \wedge T = f \wedge T.$$

Initially, the $\bar\partial\partial$-operator is thus defined only on smooth forms but it can be extended (in various ways) to the forms defined only on T. The present article deals with the following question: Can the $\bar\partial\partial$-equation be solved on T and, if so, what kind of estimates can be found for its solution?

The solvability of $\bar\partial\partial$-equations is classical (see [1–8]). We can also similarly consider smooth $(1,1)$-currents that are strictly positive in a subdomain D in M and vanish outside D, which means that we study our equation in D.

Let V be a complex vector space of dimension n. A (q,q)-form v is strictly positive if it belongs to the cone generated by the forms $i\alpha_1 \wedge \bar\alpha_1 \wedge \dots \wedge i\alpha_q \wedge \bar\alpha_q$, where $\alpha_j \in \Lambda^{1,0}(V^*)$.

A form $u \in \Lambda^{p,p}(V^*)$ is positive if and only if $u \wedge v$ is positive for every strictly positive (q,q)-form v with $q+p=n$. On a complex manifold M, a differential form $u \in C^\infty_{p,p}(M)$ is strictly positive (respectively, positive) if so is $u(z)$ positive for every $z \in M$ as an element of $\Lambda^{p,p}(T^*M)$.

The space $\mathscr{D}'_{(r,s)}(M)$ of (r,s)-currents on M is by definition equal to the space $\mathscr{D}_{(r,s)}(M)$ of test (r,s)-forms on M with respect to the usual inductive limit topology on the space of test forms.

A (p,p)-current T is positive (strictly positive) if $\langle T, u\rangle \geq 0$ for all test forms $u \in \mathscr{D}_{(p,p)}(M)$ that are strictly positive (positive).

Put $c_q = (-1)^{q(q+1)/2} i^q = (-i)^{q^2}$.

The operators ∂ and $\bar\partial$ act on currents. A current T is closed if $dT = 0$.

2. Linear Algebra and the L^2-Spaces on $(1,1)$-Currents

For the $\bar\partial\partial$-problem on currents of higher bidimension, we will first discuss in more detail the linear algebra of forms on a current. This is necessary for developing a version of Kähler identities on a current, which we will later use in proving an a priori Kodaira–Nakano–Hörmander estimate.

Let us begin with discussing forms and currents at a fixed point; i.e., we will consider T—a nonnegative element in $\Lambda^{1,1}(\mathbb{C}^{n+1})$, and forms $f \in \Lambda^{*,*}(\mathbb{C}^{n+1})$. The space of (p,q)-forms on T, denoted by $\Lambda_T^{p,q}$, is defined as the space of all $f \in \Lambda^{p,q}(\mathbb{C}^{n+1})$ modulo the subspace of forms such that $f \wedge T = 0$. To avoid burdensome notation, we use the same symbol for an element of $\Lambda^{p,q}(\mathbb{C}^{n+1})$ and the corresponding element in $\Lambda_T^{p,q}$.

On a manifold, the space of $(0,q)$-forms is the exterior algebra of the space of $(0,1)$-forms but it is important to clearly understand that this is not so in our case [9].

In any case, for defining norms on Λ_T, we also need an auxiliary $(1,1)$-form $\omega > 0$ which will define a metric on T. Let $\omega_k = \omega^k/k!$.

Let σ_T be the trace of T with respect to ω regarded as a form of maximal degree; i.e., $\sigma_T = T \wedge \omega_n$. This means that $\sigma_T = \mathrm{tr}(T)\omega_{n+1}$, where $\mathrm{tr}(T)$ is the trace consider as a number.

It can be proved (see, for example, [10, p. 170]) that a k-form f on an n-manifold is primitive if and only if $k \le n$ and $f \wedge \omega_{n-k+1} = 0$. This condition makes sense on Λ_T.

Definition 1 [9]. A k-form f is *primitive on* T if $k \le n$ and $f \wedge \omega_{n-k+1} \wedge T = 0$.

The following proposition is a key ingredient of the proof of the a priori inequality for the $\bar\partial\partial$-operator.

Let $e_1, \dots, e_{n+1}$ be a basis for the space of $(1,0)$-forms on $\mathbb{C}^{n+1}$. Write $\gamma = \sum \gamma_{JK} e_J \wedge \bar e_K$ and partition γ into the sum $\tau + \sigma$ depending on whether J belongs to K (the τ-part) or not:

$$
\begin{aligned}
\gamma = &\sum_{j_1 \in K} e_{j_1} \dots \sum_{j_p \in K} \gamma_{JK} e_{j_p} \wedge \bar e_K \\
&+ \Bigg(\sum_{r=1}^{p-1} \sum_{|M|=r} \sum_{j_1 \notin K} e_{j_1} \dots \sum_{j_{m_1-1} \notin K} e_{j_{m_1-1}} \sum_{j_{m_1} \in K} e_{j_{m_1}} \sum_{j_{m_1+1} \notin K} e_{j_{m_1+1}} \\
&\dots \sum_{j_{m_r-1} \notin K} e_{j_{m_r-1}} \sum_{j_{m_r} \in K} e_{j_{m_r}} \sum_{j_{m_r+1} \notin K} e_{j_{m_r+1}} \dots \sum_{j_p \notin K} \gamma_{JK} e_{j_p} \wedge \bar e_K \\
&+ \sum_{j_1 \notin K} e_{j_1} \dots \sum_{j_p \notin K} \gamma_{JK} e_{j_p} \wedge \bar e_K \Bigg) = \tau + \left(\sum_{r=1}^{p-1} \sigma_r + \sigma_0 \right) = \tau + \sigma.
\end{aligned}
$$

Proposition 1. *The quadratic form*

$$[\gamma, \gamma]\sigma_T = c_{q+p} \gamma \wedge \bar\gamma \wedge \omega_{n-q-p} \wedge T, \tag{1}$$

defined on the space of primitive forms in $\Lambda_T^{p,q}$ *splits into positive definite forms* $[\sigma_r, \sigma_r]\sigma_T$ *if* $(-1)^{p+r} = -1$ *and into negative definite forms* $[\tau, \tau]\sigma_T$, $[\sigma_r, \sigma_r]\sigma_T$, $1 \le r \le p-1$, *if* $(-1)^{p+r} = 1$ *(If* $p = 0$ *then the form* $[\tau, \tau]\sigma_T$ *is positive definite; for* $p = 2k+1$, *the form* $[\sigma_0, \sigma_0]\sigma_T$ *is negative definite, and for* $p = 2k$, *it is positive definite.)*

Proof. Let us first choose a basis $e_1, \dots, e_{n+1}$ for the space of $(1,0)$-forms in $\mathbb{C}^{n+1}$ that diagonalizes both ω and T. Put $dV_j = i e_j \wedge \bar e_j$ and $dV_J = \bigwedge_J dV_j$. Then $\omega = \sum dV_j$, $T = \sum \lambda_j dV_j$, and

$$T \wedge \omega_{n-q-p+1} = \sum_{|K|=n-q-p+2} \lambda_K dV_K$$

if we put $\lambda_J = \sum_J \lambda_j$.

It is easy to check that

$$[\sigma,\sigma] = \sum_{r=0}^{p-1}\sum_{t=0}^{p-1}[\sigma_r,\sigma_t] = \sum_{r=0}^{p-1}[\sigma_r,\sigma_r]$$

since $[\sigma_r,\sigma_t]\sigma_T = 0$ for $r \neq t$. Consider

$$[\sigma_r,\sigma_r]\sigma_T = c_{q+p}\sum_{|K|=q-r}\sum_{|M|=r}\sum_{j_1\notin K} e_{j_1}\dots\sum_{j_p\notin K}\sigma^r_{JK}e_{j_p}[M]\wedge dV_{J_M}\wedge\bar e_K$$
$$\wedge\sum_{|L|=q-r}\sum_{|P|=r}\sum_{s_1\notin L}\bar e_{s_1}\dots\sum_{s_p\notin L}\overline{\sigma^r_{SL}}\bar e_{s_p}[P]\wedge dV_{S_P}\wedge e_L\wedge\omega_{n-q-p}\wedge T$$
$$= (-1)^{p+r}\sum_{|K|=q-r}\sum_{|M|=r}\sum_{|P|=r}\sum_{J_M\cap S_P=\varnothing}\sigma^r_{JK}$$
$$\times\overline{\sigma^r_{(j_1,\dots,j_{m_1-1},j_{m_1+1},\dots,s_{p_1},\dots,s_{p_r},\dots,j_{m_r-1},j_{m_r+1},\dots,j_p)K}}\sum_{|L|=n-q-p+1}\lambda_L dV_{L\cup J\cup S_P\cup K},$$
$$1\le r\le p-1,$$

$$[\sigma_0,\sigma_0]\sigma_T = (-1)^p\sum|\gamma_{JK}|^2\sum_{|L|=n-q-p+1}\lambda_L dV_{L\cup J\cup K}.$$

Here the notation $(j_1,\dots,j_{m_1-1},j_{m_1+1},\dots,s_{p_1},\dots,s_{p_r},\dots,j_{m_r-1},j_{m_r+1},\dots,j_p)$ means that, in the index S, the expression $S[P] = (s_1,\dots,s_{p_1-1},s_{p_1+1},\dots,s_{p_r-1},s_{p_r+1},\dots,s_p)$ is replaced by $J[M]$.

The condition that σ_r, $r\ge 1$, is primitive (σ_0 is always primitive since

$$\sum_{j_1\notin K}e_{j_1}\dots\sum_{j_p\notin K}\gamma_{JK}e_{j_p}\wedge\bar e_K\wedge\omega_{n-q-p+1}\wedge T$$
$$= \sum_{j_1\notin K}e_{j_1}\dots\sum_{j_p\notin K}\gamma_{JK}e_{j_p}\wedge\bar e_K\wedge\sum_{|L|=n-q-p+2}\lambda_L dV_L = 0)$$

means that

$$\sigma_r\wedge\omega_{n-q-p+1}\wedge T = \sum_{|K|=q-r}\sum_{|M|=r}\sum_{|L|=n-q-p+2}\sum_{j_1\notin K}e_{j_1}\dots\sum_{j_p\notin K}\sigma^r_{JK}$$
$$\times\lambda_L e_{j_p}[M]\wedge dV_{J_M\cup L}\wedge\bar e_K = 0.$$

We have

$$[\sigma_r,\sigma_r] = (-1)^{p+r}\sum_{|K|=q-r}\sum_{|M|=r}\sum_{|P|=r}\sum_{J_M\cap L_P=\varnothing}$$
$$\times\sigma^r_{JK}\bar\sigma^r_{(j_1,\dots,j_{m_1-1},j_{m_1+1},\dots,l_{p_1},\dots,l_{p_r},\dots,j_{m_r-1},j_{m_r+1},\dots,j_p)K}\lambda_{(J\cup L_P\cup K)^c},\quad 1\le r\le p-1,$$
$$[\sigma_0,\sigma_0] = (-1)^p\sum|\gamma_{JK}|^2\lambda_{(J\cup K)^c}.$$

Fix K, $j_1,\dots,j_{m_1-1},j_{m_1+1},\dots,j_{m_r-1},j_{m_r+1},\dots,j_p$ and rename the remaining indices as $1,\dots,N$. Put

$$\hat\lambda_{J_M} = \sum_1^N\lambda_i - (\lambda_{j_{m_1}}+\cdots+\lambda_{j_{m_r}}) = \sum_1^N\lambda_i - \lambda_{J_M}$$

and

$$\hat{\lambda}_{J_M L_P} = \sum_1^N \lambda_i - \lambda_{J_M} - \lambda_{L_P}.$$

We can prove that

$$\text{if } \sum \sigma^r_{J_M} \hat{\lambda}_{J_M} = 0 \text{ then } \sum_{J_M \cap L_P = \varnothing} \sigma^r_{J_M} \sigma^{\bar{r}}_{L_P} \hat{\lambda}_{J_M L_P} \leq 0.$$

We have

$$[\tau, \sigma] = \sum_{r=0}^{p-1} [\tau, \sigma_r] = 0$$

because

$$[\tau, \sigma_r]\sigma_T = c_{q+p} \sum_{|K|=q-p} \tau_{JK} dV_J \wedge \bar{e}_K$$
$$\wedge \sum_{|P|=r} \sum_{s_1 \notin L} \bar{e}_{s_1} \dots \sum_{s_{p_1-1} \notin L} \bar{e}_{s_{p_1-1}} \sum_{s_{p_1} \in L} \bar{e}_{s_{p_1}} \sum_{s_{p_1+1} \notin L} \bar{e}_{s_{p_1+1}}$$
$$\dots \sum_{s_{p_r-1} \notin L} \bar{e}_{s_{p_r-1}} \sum_{s_{p_r} \in L} \bar{e}_{s_{p_r}} \sum_{s_{p_r+1} \notin L} \bar{e}_{s_{p_r+1}} \dots \sum_{s_p \notin L} \bar{\gamma}_{SL} \bar{e}_{s_p} \wedge e_L \wedge \omega_{n-q-p} \wedge T = 0.$$

The primitivity of τ means that

$$\tau \wedge \omega_{n-q-p+1} \wedge T = \sum_{|K|=q-p} \sum_{|L|=n-q-p+2} \tau_{JK} \lambda_L dV_{J \cup L} \wedge \bar{e}_K = 0,$$

and

$$[\tau, \tau]\sigma_T = \sum_{|K|=q-p} \sum_{J \cap S = \varnothing} \tau_{JK} \bar{\tau}_{SK} \sum_{|L|=n-q-p+1} \lambda_L dV_{L \cup J \cup S \cup K}.$$

Hence,

$$[\tau, \tau] = \sum_{|K|=q-p} \sum_{J \cap L = \varnothing} \tau_{JK} \bar{\tau}_{LK} \lambda_{(J \cup L \cup K)^c}.$$

Fix K and rename the remaining indices as $1, \dots, N$. Put

$$\hat{\lambda}_J = \sum_1^N \lambda_i - (\lambda_{j_1} + \dots + \lambda_{j_p}) = \sum_1^N \lambda_i - \lambda_J$$

and

$$\hat{\lambda}_{JL} = \sum_1^N \lambda_i - \lambda_J - \lambda_L.$$

We can prove that if $\sum \tau_J \hat{\lambda}_J = 0$ then $\sum_{J \cap L = \varnothing} \tau_J \bar{\tau}_L \hat{\lambda}_{JL} \leq 0$. $\square$

Proposition 5.7 in [9] is a particular case of Proposition 1 for $p = 1$.

DEFINITION 2. Let $f \in \Lambda^{p,q}(\mathbb{C}^{n+1})$. The *norm* of f on T is defined as

$$|f|^2_{\omega,T} \sigma_T = c_{q+p} f \wedge \overline{\widehat{f}} \wedge \omega_{n-q-p} \wedge T, \tag{2}$$

where

$$\widehat{f} = f_0 \wedge \omega^p + \sum_{k=1}^p \widehat{f}_k \wedge \omega^{p-k}$$

and $f_k \in \Lambda_T^{k,q-p+k}$ are primitive forms,

$$\widehat{f_k} = -\tau^k - \sum_{r=1}^{k-1} (-1)^{k+r} \sigma_r^k + (-1)^k \sigma_0^k.$$

Recall that the norm of f in $\mathbb{C}^{n+1}$, measured in the ω-metric, is defined as

$$|f|_\omega^2 \omega_{n+1} = c_{q+p} f \wedge \overline{\widehat{f}} \wedge \omega_{n-q-p+1}.$$

Therefore, $(n+1)|f|_{\omega,\omega}^2 = (n-q-p+1)|f|_\omega^2$ if $T = \omega$. In the general case, since $T \le \operatorname{tr}(T)\omega$, we obtain $|f|_{\omega,T}^2 \le (n-q-p+1)|f|_\omega^2$. Finally, polarizing we get the inner product such that

$$(f,f)_{\omega,T} = |f|_{\omega,T}^2,$$

and, in what follows, we will omit the dependence on ω and T.

The norms and inner products on $\Lambda_T^{q,p}$ are of course defined similarly, and so $(f,g) = \overline{(\bar f, \bar g)}$. In particular, if $f, g \in \Lambda_T^{q,p}$ then

$$(f,g)\sigma_T = \bar c_{q+p} f \wedge \widehat{\bar g} \wedge \omega_{n-q-p} \wedge T.$$

Let us define norms on $\Lambda_T^{n-p,q}$. To this end, observe that every $f \in \Lambda_T^{n-p,q}$ defines the linear form $L_f(g)\sigma_T = g \wedge f \wedge T$ on $\Lambda_T^{p,n-q}$.

Definition 3. If $f \in \Lambda_T^{n-p,q}$ then

$$|f|_{\omega,T} = \|L_f\| = \sup_{|g|_{\omega,T} \le 1} |L_f(g)|.$$

Equivalently, L_f can be represented as the inner product with an element $f' \in \Lambda_T^{p,n-q}$, and so

$$g \wedge f \wedge T = L_f(g)\sigma_T = (g,f')\sigma_T = c_{n-q+p} g \wedge \overline{\widehat{f'}} \wedge \omega_{q-p} \wedge T. \tag{3}$$

Then $|f|_{\omega,T} = |f'|_{\omega,T}$.

Recall that the Hodge $*$-operator on a Kähler (or Riemann) manifold is defined as follows: $h \wedge \overline{*g} = (h,g)dV$ if h and g are k-forms and dV is the volume element. Similarly, define the $*$-operator $* : \Lambda_T^{n-q,p} \to \Lambda_T^{n-p,q}$ by setting

$$h \wedge \overline{*g} \wedge T = (h,g)\sigma_T. \tag{4}$$

Since

$$(h,g)\sigma_T = \overline{c_{n-q+p}} h \wedge \widehat{\bar g} \wedge \omega_{q-p} \wedge T,$$

this means that $*g = c_{n-q+p} \overline{\widehat{\bar g}} \wedge \omega_{q-p}$ on $\Lambda_T^{n-q,p}$.

In the same manner, (4) defines $* : \Lambda_T^{n-p,q} \to \Lambda_T^{n-q,p}$. Since then the inner product is defined as $(h,g) = (\tilde h, \tilde g)$, we find

$$h \wedge \overline{*g} \wedge T = \overline{c_{n-q+p}} \tilde h \wedge \widehat{\bar{\tilde g}} \wedge \omega_{q-p} \wedge T = \overline{c_{n-q+p}} h \wedge \widehat{\bar{\tilde g}} \wedge T;$$

therefore, $*g = c_{n-q+p} \overline{\widehat{\bar{\tilde g}}}$ on $\Lambda_T^{n-p,q}$.

The following proposition is connected with the Lefschetz isomorphism in $\mathbb{C}^{n+1}$ and will play a key role when we approximate general currents by smooth forms in the sequel.

Proposition 2. *Let $T \in \Lambda^{1,1}(\mathbb{C}^{n+1})$ be strictly positive. Assume further that $F \in \Lambda^{n-p+1,q+1}(\mathbb{C}^{n+1})$, $0 \le p \le q \le n$. Then there exists a unique form $\widetilde{F} \in \Lambda^{n-q,p}(\mathbb{C}^{n+1})$ such that*

$$F = \widetilde{F} \wedge \omega_{q-p} \wedge T.$$

In particular, F can be represented as $F = f \wedge T$ with $f \in \Lambda^{n-p,q}(\mathbb{C}^{n+1})$.

Proposition 5.4 in [9] is the particular case of Proposition 2 for $p = 0$.

Proposition 3. *Let $f \in \Lambda_T^{q,p}$. Then there are uniquely defined primitive forms $f_0 \in \Lambda_T^{q-p,0}$, $f_1 \in \Lambda_T^{q-p+1,1}, \dots, f_p \in \Lambda_T^{q,p}$ such that*

$$f = \sum_{k=0}^{p} f_k \wedge \omega^{p-k}. \tag{5}$$

Proof. Induct on p. $\square$

Proposition 5.6 in [9] the particular case of Proposition 3 for $p = 1$.

Similarly, we of course have a primitive decomposition of (p,q)-forms. The following proposition says that we have in fact obtained a norm for forms on T.

Proposition 4. *Suppose that $\gamma \in \Lambda^{p,q}(\mathbb{C}^{n+1})$ and $|\gamma|^2_{\omega,T} = 0$. Then $\gamma \wedge T = 0$.*

Proposition 5.2 in [9] is the particular case of Proposition 4 for $p = 0$.

Now, let $T \ge 0$ be a $(1,1)$-current in $\mathbb{C}^{n+1}$. Such a current can be written as $T = i \sum T_{j\bar{k}} dz_j \wedge d\bar{z}_k$, where the coefficients are absolutely continuous measures with respect to the trace measure $\sigma_T = T \wedge \omega_n$. Let $\mathrm{tr}(T)$ be the $(0,0)$-current defined as $\mathrm{tr}(T)\omega_{n+1} = \sigma_T$. Then T can be written as $T = \widetilde{T}\,\mathrm{tr}(T)$, where $\widetilde{T}$ is a form with coefficients defined almost everywhere with respect to σ_T. Since the coefficients of $\widetilde{T}$ constitute a semidefinite matrix with unit trace, Cauchy's inequality implies that

$$T = i \sum \widetilde{T}_{j\bar{k}} dz_j \wedge d\bar{z}_k \,\mathrm{tr}(T),$$

where $|\widetilde{T}_{j\bar{k}}| \le 1$.

If f is a smooth or just continuous (p,q)-form in $\mathbb{C}^{n+1}$ then define the L^2-norm of f on T:

$$\|f\|^2_{\omega,T} = \int |f|^2_{\omega,\widetilde{T}} \sigma_T. \tag{6}$$

Equality (6) means that

$$\|f\|^2_{\omega,T} = c_{p+q} \int f \wedge \bar{\bar{f}} \wedge \omega_{n-p-q} \wedge T$$

because

$$c_{p+q} f \wedge \bar{\bar{f}} \wedge \omega_{n-p-q} \wedge T = c_{p+q} f \wedge \bar{\bar{f}} \wedge \omega_{n-p-q} \wedge \widetilde{T}\,\mathrm{tr}(T) = |f|^2_{\omega,\widetilde{T}} \sigma_T,$$

and $\mathrm{tr}(\widetilde{T}) = 1$.

Define the L^2-spaces of (p,q)-forms on T, denoted by $L^2_{p,q}(T)$, as the completion of smooth (p,q)-forms with respect to L^2-norms. Thus, the smooth forms are dense in the L^2-spaces by definition.

If, finally, φ is a Borel weight function then define $L^2_{p,q}(T, e^{-\varphi})$ as the space of those $f \in L^2_{p,q,\mathrm{loc}}$ that satisfy

$$\|f\|^2_{\omega,T,\varphi} = \int |f|^2_{\omega,\widetilde{T}} e^{-\varphi} \sigma_T < \infty.$$

3. Differential Operators on T

Suppose that T is closed.

DEFINITION 4. Given $u \in L^2_{p,q,\mathrm{loc}}(T)$, we say that $\bar\partial\partial_w u = f$ on T if $f \in L^2_{p+1,q+1,\mathrm{loc}}(T)$ and $\bar\partial\partial(u \wedge T) = f \wedge T$ in the sense of currents.

The strong extension of $\bar\partial\partial$ is defined as follows:

DEFINITION 5. If $u \in L^2_{p,q,\mathrm{loc}}(T)$ and $f \in L^2_{p+1,q+1,\mathrm{loc}}(T)$ then $\bar\partial\partial_s u = f$ if there exists a sequence of (C^2-)smooth (p,q)-forms u_n such that $u_n \to u$ in $L^2_{\mathrm{loc}}(T)$ and $\bar\partial\partial u_n \to f$ in $L^2_{\mathrm{loc}}(T)$.

Now, let φ be a Borel measurable weight function. Then we obtain closed densely defined operators $\bar\partial\partial_w$ and $\bar\partial\partial_s$ on $L^2_{p,q}(T, e^{-\varphi})$ with the domains consisting of all u such that $\|\bar\partial\partial u\|_{T,\varphi} < \infty$ with $\bar\partial\partial = \bar\partial\partial_w$ or $\bar\partial\partial_s$.

Let us now define the formal adjoints. If f is a (p,q)-form such that $*f$ is smooth then we put $\vartheta f = \varepsilon_{p,q} * \partial * f$, where $\varepsilon_{p,q}$ is chosen so that

$$(g, \vartheta f)_{\omega,T} = (\bar\partial_w g, f)_{\omega,T} \tag{7}$$

if f has compact support. If φ is a weight function then we put $\vartheta_\varphi = e^{\varphi}\vartheta e^{-\varphi}$, and so $(g, \vartheta_\varphi f)_{\omega,T,\varphi} = (\bar\partial_w g, f)_{\omega,T,\varphi}$.

4. A Priori Estimates for $\bar\partial\partial$

The main technical result of this section is the following generalization of the Kodaira–Nakano–Hörmander identity. In the statement of the result, we use the notation $\partial_{-\varphi} = e^{-\varphi}\partial e^{\varphi}$ for the twisted $\bar\partial$-operator.

Theorem 1. *Let $T \geq 0$ be a $(1,1)$-current in a domain D in $\mathbb{C}^{n+1}$ such that $i\partial\bar\partial T$ has measurable coefficients. Let ω be a Kähler form in D. Let, finally, g be a test (p,q)-form with support in D and suppose that $\varphi \in C^2(D)$. Then*

$$\begin{aligned}
\int c_{p+q+1}\partial g \wedge \overline{\widehat{\partial g}} \wedge i\partial\bar\partial\varphi \wedge \omega_{n-p-q-2} \wedge Te^{\varphi} - \int c_{p+q+1}\partial g \wedge \overline{\widehat{\partial g}} \wedge \omega_{n-p-q-2} \wedge i\partial\bar\partial Te^{\varphi}& \\
+c_{p+q+2}\int \bar\partial\partial g \wedge \overline{\widehat{\bar\partial\partial g}} \wedge \omega_{n-p-q-2} \wedge Te^{\varphi}& \\
-c_{p+q+2}\int (\partial_{-\varphi}\partial g)_{p+2} \wedge \overline{(\partial_{-\varphi}\widehat{\partial g})}_{p+2} \wedge \omega_{n-p-q-2} \wedge Te^{\varphi}& \\
-c_{p+q}\int \widehat{\vartheta_{-\varphi}\widehat{\partial g}} \wedge \overline{\widehat{\vartheta_{-\varphi}\partial g}} \wedge \omega_{n-p-q} \wedge Te^{\varphi}& \\
= (\vartheta_{-\varphi}\widehat{\bar\partial\partial g}, \widehat{\partial g})_{\omega,T,-\varphi} + \overline{(\vartheta_{-\varphi}\widehat{\widehat{\bar\partial\partial g}}, \partial g)}_{\omega,T,-\varphi}&.
\end{aligned} \tag{8}$$

In particular, if $i\partial\bar\partial T$ is strictly positive and $i\partial\bar\partial\varphi \geq \omega$ then

$$\begin{aligned}
(n-p-q-1)\|\partial g\|^2 + c_{p+q+2}\int \bar\partial\partial g \wedge \overline{\widehat{\bar\partial\partial g}} \wedge \omega_{n-p-q-2} \wedge Te^{\varphi}& \\
-c_{p+q+2}\int (\partial_{-\varphi}\partial g)_{p+2} \wedge \overline{(\partial_{-\varphi}\widehat{\partial g})}_{p+2} \wedge \omega_{n-p-q-2} \wedge Te^{\varphi}& \\
-c_{p+q}\int \widehat{\vartheta_{-\varphi}\widehat{\partial g}} \wedge \overline{\widehat{\vartheta_{-\varphi}\partial g}} \wedge \omega_{n-p-q} \wedge Te^{\varphi} \leq (\vartheta_{-\varphi}\widehat{\bar\partial\partial g}, \widehat{\partial g}) + \overline{(\vartheta_{-\varphi}\widehat{\widehat{\bar\partial\partial g}}, \partial g)}&.
\end{aligned} \tag{9}$$

If, moreover, $dT = 0$, *then*

$$(n-p-q-1)\|\partial g\|^2 - c_{p+q+2}\int (\partial_{-\varphi}\partial g)_{p+2}\wedge \overline{(\partial_{-\varphi}\widehat{\partial g})}_{p+2}\wedge \omega_{n-p-q-2}\wedge Te^{\varphi}$$
$$-c_{p+q}\int \widehat{\vartheta_{-\varphi}\widehat{\partial g}}\wedge \overline{\widehat{\vartheta_{-\varphi}\partial g}}\wedge \omega_{n-p-q}\wedge Te^{\varphi} \le (\widehat{\bar\partial\partial g}, \bar\partial\widehat{\partial g}). \tag{10}$$

Proof. Clearly, (9) and (10) follow from (8) since

$$(\bar\partial\partial g, \widehat{\bar\partial\widehat{\partial g}}) = \overline{(\widehat{\bar\partial\widehat{\partial g}}, \bar\partial\partial g)}.$$

For proving (8), we follow the Bochner–Kodaira method (see [11]). □

Theorem 1 has a duplicate for $(n-p,q)$-forms.

Theorem 2. *Under the notation and assumptions of Theorem* 1, *let* f *be a test* $(n-p,q)$*-form with support in* D *such that* $*f$ *is* (C^2-)*smooth. If* $i\partial\bar\partial T \le 0$, $i\partial\bar\partial\varphi \ge \omega$, *and* $dT = 0$, *then*

$$(q-p-1)\|\partial_{\varphi}\overline{*f}\|^2 - c_{n-q+p+2}\int (\partial\partial_{\varphi}\overline{*f})_{p+2}\wedge \overline{(\widehat{\partial\partial_{\varphi}\overline{*f}})_{p+2}}\wedge \omega_{q-p-2}\wedge Te^{-\varphi}$$
$$-c_{n-q+p}\int \widehat{\vartheta\widehat{\partial_{\varphi}\overline{*f}}}\wedge \overline{\widehat{\vartheta\partial_{\varphi}\overline{*f}}}\wedge \omega_{q-p}\wedge Te^{-\varphi} \le (\widehat{\bar\partial_{\varphi}\partial_{\varphi}\overline{*f}}, \bar\partial_{\varphi}\widehat{\partial_{\varphi}\overline{*f}}).$$

Proof. Apply Theorem 1 to $g = \overline{*f}e^{-\varphi}$. □

5. Existence Theorems for $\bar\partial\partial$ on Closed $(1,1)$-Currents

Theorem 3. *Let* $T \ge 0$ *be a closed* $(1,1)$*-current in* $\mathbb{C}^{n+1}$ *and let* $\omega = i\partial\bar\partial|z|^2$ *be a Kähler form in the Euclidean metric in* $\mathbb{C}^{n+1}$. *Let* φ *be a plurisubharmonic function in* $\mathbb{C}^{n+1}$ *satisfying* $i\partial\bar\partial\varphi \ge \omega$. *Then, for every* $\bar\partial_w$*-closed* $(n-p,q)$*-form* f *on* T *with* $q-p-1 \ge 1$, *there exists a* $(n-p-1,q-1)$*-form* u *on* T *such that* $\bar\partial\partial_w u = f$ *on* T *and*

$$\int |\partial u|^2_{\omega,T}\sigma_T e^{-\varphi} \le \frac{1}{q-p-1}\int |f|^2_{\omega,T}\sigma_T e^{-\varphi}.$$

Let us first prove the theorem on assuming that T is smooth and strictly positive and then obtain the general theorem from the approximation of T by such currents $T_{(\varepsilon)}$. After that we must approximate the form f defined only on T by global forms closed on $T_{(\varepsilon)}$. This turns out surprisingly easy: Instead of regularizing T and f, we separately regularize $f\wedge T$ and then use Proposition 2 to write $(f\wedge T)_{\varepsilon} = f_{(\varepsilon)}\wedge T_{(\varepsilon)}$.

To this end, choose a nonnegative test function χ supported by the unit ball such that $\int \chi = 1$, and let $\chi_{\varepsilon}(z) = \varepsilon^{-2n}\chi(z/\varepsilon)$. For any form or a current α, denote the convolution $\alpha * \chi_{\varepsilon}$ by α_{ε}.

Proof of Theorem 3. Suppose first that T is strictly positive, while T and φ are smooth. For proving the theorem, we must show that

$$|(f,\bar\vartheta_{\varphi}\alpha)|^2 \le \frac{1}{q-p-1}\|\partial_{\varphi}\bar\partial_{\varphi} * \alpha\|^2 = \frac{1}{q-p-1}\|\vartheta_{\varphi}\overline{\vartheta}_{\varphi}\alpha\|^2 \tag{11}$$

if α is a test $(n-p+1,q)$-form and normalize it so that $\|f\|^2 = 1$. (If (11) is fulfilled then the Riesz representation theorem implies that we can find a form u on T such that

$$(f,\bar\vartheta_{\varphi}\alpha) = (\partial_w u, \vartheta_{\varphi}\overline{\vartheta}_{\varphi}\alpha) \text{ and } \|\partial_w u\| \le \frac{1}{q-p-1}.$$

Then $\bar\partial\partial_w u = f$, and we are done.)

The proof of (11) is carried out in a standard manner.

We also have versions of Theorem 3 for a pseudoconvex domain in $\mathbb{C}^{n+1}$ and general Kähler metrics. We further have versions of these theorems for some compact Kähler manifolds.

6. Currents of Higher Bidegree

The key ingredient of the proof was Proposition 1 by which the quadratic form

$$[\gamma,\hat\gamma]\sigma_T = c_{q+p}\gamma\wedge\overline{\hat\gamma}\wedge T\wedge\omega_{n-q-p},$$

is definite on the space of (q,p)-forms on T satisfying $\gamma\wedge\omega_{n-p-q+1}\wedge T = 0$ (i.e., for "primitive" forms). This fails for $(2,2)$-currents even if they are strictly positive [9].

Let T be the (s,s)-form

$$T = \sum_{l=0}^{1} dV_{sl+1,sl+2,\ldots,s(l+1)}$$

in $\mathbb{C}^{2s}$ (where $dV_{jk} = dV_j\wedge dV_k$, $dV_j = i dz_j\wedge d\bar z_j$).

It is easy to observe that there is no local solvability for $\bar\partial\partial u$ on $(s,s-1)$-forms for such choice of T. Take

$$f = \sum_j \sum_{k_1<k_2<\cdots<k_{s-1}} f_j^{k_1k_2\ldots k_{s-1}} dz_j\wedge dV_{k_1k_2\ldots k_{s-1}},$$

and so

$$f\wedge T = \sum_j \sum_{k_1<k_2<\cdots<k_{s-1}} \sum_{l=0}^{1} f_j^{k_1k_2\ldots k_{s-1}} dz_j\wedge dV_{k_1k_2\ldots k_{s-1},sl+1,sl+2,\ldots,s(l+1)}.$$

Then $\bar\partial f\wedge T = 0$ means that

$$\sum \frac{\partial f_j}{\partial \bar z_j} = 0, \tag{12}$$

where $f_j = \sum_{k_1<k_2<\cdots<k_{s-1}} f_j^{k_1k_2\ldots k_{s-1}}$.

If $f\wedge T = \bar\partial\partial u\wedge T$ then, for a $(s-1,s-2)$-form u, we may write

$$u = \sum_{j=1}^{2s} \sum_{k_1<\cdots<k_{s-2}} u_j^{k_1\ldots k_{s-2}} dz_j\wedge dV_{k_1\ldots k_{s-2}}.$$

Now

$$u\wedge T = \sum_{j=1}^{2s} \sum_{k_1<\cdots<k_{s-2}} \sum_{l=0}^{1} u_j^{k_1\ldots k_{s-2}} dz_j\wedge dV_{k_1\ldots k_{s-2},sl+1,sl+2,\ldots,s(l+1)},$$

and we can show that

$$\bar\partial\partial u\wedge T = \sum_{j=1}^{2s} \sum_{k_1<\cdots<k_{s-2}} \sum_{k_{s-1}} \sum_{l=0}^{1} i\left(\frac{\partial^2 u_j^{k_1\ldots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_{k_{s-1}}} - \frac{\partial^2 u_{k_{s-1}}^{k_1\ldots k_{s-2}}}{\partial \bar z_{k_{s-1}}\partial z_j}\right)$$
$$\times dz_j\wedge dV_{k_1\ldots k_{s-1},sl+1,\ldots,s(l+1)}.$$

Therefore, the equation $\bar\partial\partial u\wedge T = f\wedge T$ splits into

$$\sum_{1\le k_1<\dots<k_{s-1}\le s}\sum_{k_{s-1}=1}^{s} i\Big(\frac{\partial^2 u_j^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_{k_{s-1}}}-\frac{\partial^2 u_{k_{s-1}}^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_j}\Big)=f_j,\quad 1\le j\le s,$$

and a similar for $u_{s+1}^{k_1\dots k_{s-2}}$, $u_{s+2}^{k_1\dots k_{s-2}},\dots u_{2s}^{k_1\dots,k_{s-2}}$, $s+1\le k_1<\dots<k_{s-2}\le 2s$, $s+1\le k_{s-1}\le 2s$. It is solvable only if

$$\frac{\partial f_1}{\partial\bar z_1}+\frac{\partial f_2}{\partial\bar z_2}+\dots+\frac{\partial f_s}{\partial\bar z_s}=0, \tag{13}$$

which is not assumed by (12).

Let T be the (s,s)-form

$$T=\sum_{k_s<k_{s+1}<\dots<k_{2s-1}} dV_{k_sk_{s+1}\dots k_{2s-1}}$$

in $\mathbb{C}^{2s}$. It is easy to observe that the local solvability for $\bar\partial\partial u$ on $(s,s-1)$-forms holds for this choice of T. We have

$$f\wedge T=\sum f_j dz_j\wedge d\widehat{V}_j,$$

where $f_j=\sum_{k_1<\dots<k_{s-1}} f_j^{k_1\dots k_{s-1}}$. Then $\bar\partial f\wedge T=0$ means that

$$\sum\frac{\partial f_j}{\partial\bar z_j}=0. \tag{14}$$

If $f\wedge T=\bar\partial\partial u\wedge T$ then

$$u\wedge T=\sum_{j=1}^{2s}\sum_{k_1<\dots<k_{s-2}}\sum_{k_s<\dots<k_{2s-1}} u_j^{k_1\dots k_{s-2}}dz_j\wedge dV_{k_1\dots k_{s-2}k_s\dots k_{2s-1}},$$

and we can show that

$$\bar\partial\partial u\wedge T=\sum_{j=1}^{2s}\sum_{k_1<\dots<k_{s-2}}\sum_{k_{s-1}}\sum_{k_s<\dots<k_{2s-1}} i\Big(\frac{\partial^2 u_j^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_{k_{s-1}}} -\frac{\partial^2 u_{k_{s-1}}^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_j}\Big)dz_j\wedge dV_{k_1\dots k_{2s-1}}.$$

Therefore, the equation $\bar\partial\partial u\wedge T=f\wedge T$ splits into the system

$$\sum_{k_1<\dots<k_{s-2}}\sum_{k_{s-1}}\sum_{k_s<\dots<k_{2s-1}} i\Big(\frac{\partial^2 u_j^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_{k_{s-1}}}-\frac{\partial^2 u_{k_{s-1}}^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_j}\Big)=f_j,\quad 1\le j\le 2s.$$

It is solvable only if

$$\sum\frac{\partial f_j}{\partial\bar z_j}=\sum_j\sum_{k_1<\dots<k_{s-2}}\sum_{k_{s-1}}\sum_{k_s<\dots<k_{2s-1}} i\Big(\frac{\partial^3 u_j^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_{k_{s-1}}\partial\bar z_j}-\frac{\partial^3 u_{k_{s-1}}^{k_1\dots k_{s-2}}}{\partial\bar z_{k_{s-1}}\partial z_j\partial\bar z_j}\Big)=0, \tag{15}$$

which is assumed by (14).

Let T be the (s,s)-form $\sum_{j=0}^{1} dV_{sj+1,sj+2,\dots,s(j+1)}$ in $\mathbb{C}^{2s}$. Since T has bidimension (s,s), a primitive 2-form must satisfy

$$\gamma \wedge \omega^{s-1} \wedge T = 0. \tag{16}$$

In particular, take $\gamma = \sum\limits_{j=1}^{2s} \gamma_j dV_j$. Then

$$\gamma \wedge \omega = \sum_{j<k} \gamma_{jk} dV_{jk}, \text{ where } \gamma_{jk} = \gamma_j + \gamma_k,$$

$$\vdots$$

$$\gamma \wedge \omega^{s-1} = (s-1)! \sum_{j<k_1<\dots<k_{s-1}} \gamma_{jk_1\dots k_{s-1}} dV_{jk_1\dots k_{s-1}},$$

where $\gamma_{jk_1\dots k_{s-1}} = \gamma_j + \gamma_{k_1} + \dots + \gamma_{k_{s-1}}$, (this is proved by induction); therefore, equality (16) exactly means that $(s-1)! \sum_1^{2s} \gamma_j = 0$.

On the other hand,

$$\gamma \wedge \bar\gamma = 2\,\mathrm{Re} \sum_{j<k} \gamma_j \bar\gamma_k dV_{jk};$$

therefore,

$$\gamma \wedge \bar\gamma \wedge \omega = 2\,\mathrm{Re} \sum_{j<k_1<k_2} (\gamma_j(\bar\gamma_{k_1} + \bar\gamma_{k_2}) + \gamma_{k_1}\bar\gamma_{k_2}) dV_{jk_1k_2},$$

$$\vdots$$

$$\gamma \wedge \bar\gamma \wedge \omega^{s-2} = (s-2)! 2\,\mathrm{Re} \sum_{j<k_1<k_2<\dots<k_{s-1}} (\gamma_j(\bar\gamma_{k_1} + \dots + \bar\gamma_{k_{s-1}})$$
$$+\gamma_{k_1}(\bar\gamma_{k_2} + \dots + \bar\gamma_{k_{s-1}}) + \dots + \gamma_{k_{s-3}}(\bar\gamma_{k_{s-2}} + \bar\gamma_{k_{s-1}}) + \gamma_{k_{s-2}}\bar\gamma_{k_{s-1}}) dV_{jk_1k_2\dots k_{s-1}}$$

(this is proved by induction),

$$\gamma \wedge \bar\gamma \wedge \omega^{s-2} \wedge T = (s-2)! 2\,\mathrm{Re}[\gamma_1(\bar\gamma_2 + \dots + \bar\gamma_s) + \gamma_2(\bar\gamma_3 + \dots + \bar\gamma_s) +$$
$$\dots + \gamma_{s-1}\bar\gamma_s + \gamma_{s+1}(\bar\gamma_{s+2} + \dots + \bar\gamma_{2s}) + \dots + \gamma_{2s-2}(\bar\gamma_{2s-1} + \bar\gamma_{2s}) + \gamma_{2s-1}\bar\gamma_{2s}] dV_{1\dots 2s}.$$

This form is obviously indefinite since we obtain different signs for $\gamma = (1,\dots,1,-1,\dots,-1)$ and $\gamma = (1,-1,\dots,1,-1)$.

Let T be the (s,s)-form $\sum_{j_1<\dots<j_s} dV_{j_1,\dots,j_s}$ in $\mathbb{C}^{2s}$. Then (16) means that

$$(s-1)! \binom{2s-1}{s-1} \sum_1^{2s} \gamma_j = 0.$$

On the other hand,

$$\gamma \wedge \bar\gamma \wedge \omega^{s-2} \wedge T = (s-2)! \binom{2s-2}{s-2} 2\,\mathrm{Re}\left[\sum_{j=1}^{2s-1} \gamma_j(\bar\gamma_{j+1} + \dots + \bar\gamma_{2s})\right] dV_{1\dots 2s}$$
$$= (s-2)! \binom{2s-2}{s-2} 2\,\mathrm{Re}\left[\sum_{j=1}^{2s-1} -|\gamma_j|^2 - \gamma_j(\bar\gamma_1 + \dots \bar\gamma_{j-1})\right] dV_{1\dots 2s} \le 0.$$

The proof is carried out by induction.

REFERENCES

1. *Beloshapka V. K.* Functions pluriharmonic on a manifold // Math. USSR-Izv. 1978. V. 12, N 3. P. 439–447.
2. *Chirka E. M.,* Flows and some of their applications, in: Holomorphic Chains and Their Boundaries [Russian translation], Harvey R. Moscow: Mir, 1979. P. 122–158.
3. *Nikitina T. N.* Removable Singularities in the Boundary and $\bar\partial$-Closed Extension of CR-Forms with Singularities on the Generic Manifold [in Russian]. Novosibirsk: Nauka, 2008.
4. *Nikitina T. N.* The $\bar\partial\partial$-equation on a positive current // The 14th General Meeting of European Women in Mathematics. Book of Abstracts Part II. University of Novi Sad (Serbia), 2009. P. 15–16.
5. *Nikitina T. N.* The $\bar\partial$ and $\bar\partial\partial$-equation on a positive current // Abstracts: International Conference "Contemporary Problems of Analysis and Geometry." Novosibirsk: Sobolev Institute of Mathematics, 2009. P. 81.
6. *Nikitina T. N.,* The Amper–Monge equation on a positive current // Abstracts: International Conference "Differential Equations. Function Spaces. Approximation Theory." Novosibirsk: Sobolev Institute of Mathematics, 2013. P. 402.
7. *Nikitina T. N.* The $\bar\partial$- and $\bar\partial\partial$-Closed Form Continuation. Saarbrucken, Germany: LAP, 2014.
8. *Nikitina T. N.* The Amper–Monge equation on a positive current // Abstracts: International Conference "Differential Equations and Mathematical Modeling." Ulan-Ude; Novosibirsk: Sobolev Institute of Mathematics, 2015. P. 209–211.
9. *Berndtsson B. and Sibony N.* The $\bar\partial$-equation on a positive current // Invent. Math. 2002. V. 147. P. 371–428.
10. *Wells R. O.,* Differential Analysis on Complex Manifolds. Englewood Cliffs, N.J.: Prentice Hall, 1967.
11. *Siu Y.-T.* Complex-analiticity of harmonic maps, vanishing and Lefschetz theorems // J. Diff. Geom. 1982. V. 17. P. 55–138.

August 10, 2015

T. N. Nikitina
Siberian Federal University
Institute of Mathematics and Fundamental Informatics
Krasnoyarsk, Russia
AANick@yandex.ru

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 519.632.4:550.832.7

PARALLEL ALGORITHMS FOR DIRECT ELECTRICAL LOGGING PROBLEMS

I. V. Surodina

Abstract. We consider direct electrical logging problems and describe fully parallel algorithms for GPU architecture.

Keywords: Poisson equation, logging, simulation, Krylov subspace methods, preconditioner

Introduction

The fast solution of direct problems can serve as a foundation for inverting certain problems in geophysics. One of the prospective current directions for speeding up solutions to these problems is the use of parallel calculations on graphical processors (GPU). To maximize gain while using GPU, it is necessary to implement fully parallel calculation taking advantage of the GPU architecture. Solving direct electrical logging problems by the finite-difference method or finite-element method leads to the large sparse linear systems that are rather often solved using conjugate direction methods. Without suitable preconditioners, solutions to these systems converge slowly. The efficiency of implementation depends mainly on the degree of parallelization of the preconditioner. This article realizes the algorithm we proposed in [1] to construct a parallel preconditioner that approximates the inverse matrix. This algorithm requires small expenses, or none at all, on the construction of the preconditioning matrix and is fully parallel. The implementation uses the linear algebra function library CUBLAS NVIDIA. Depending on the mesh dimension and geophysical properties of real models, we improve the calculation time by the factor of 10 to 50 in comparison with sequential software versions.

1. The 2-Dimensional Direct Electrical Logging Problem

Consider the electrical logging problem on the example of lateral electrical logging probing problem (LELP). Consider an axially symmetric distribution $\sigma = \sigma(r,z)$ of the specific electric conductivity in cylindrical coordinates. The problem of modeling the sonde readings of the LELP problem reduces to Poisson's equation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(\sigma r\frac{\partial U^a}{\partial r}\right)+\frac{\partial}{\partial z}\left(\sigma\frac{\partial U^a}{\partial z}\right)=\frac{1}{r}\frac{\partial}{\partial r}\left((\sigma_0-\sigma)r\frac{\partial U^0}{\partial r}\right)+\frac{\partial}{\partial z}\left((\sigma_0-\sigma)\frac{\partial U^0}{\partial z}\right) \quad (1)$$

for the anomalous electric potential $U^a = U - U^0$, where U is the total required electric potential, U^0 is the electric potential of the pointlike source at the origin in a homogeneous medium with specific electric conductivity, $U^0 = \frac{I}{4\pi\sigma_0 L}$, while I is the current and $L = \sqrt{r^2+z^2}$. The potential decays as $1/L$ away from the source.

Thus, we may impose the zero boundary conditions $U^a|_{r=R} = 0$ and $U^a|_{z=\pm Z} = 0$ on the function U^a away from the source ($r = R$, $z = \pm Z$). Conditions on the well axis are determined from the axial symmetry of the source and medium: $\frac{\partial U}{\partial r} = 0$.

By axial symmetry, consider the half-plane $[0, R] \times [-Z, Z]$ and introduce the rectangular nonuniform coordinate mesh [2]

$$\underset{h}{\overset{\wedge}{\omega}} = \{(r_i, z_j),\ i = 0, \dots, N_r,\ j = -N_z, \dots, N_z\}. \tag{2}$$

On (2) consider the finite-dimensional linear space H^0 of mesh functions vanishing on the boundary equipped with the inner product

$$(u, v) = \sum_{i=0}^{N_r} \sum_{j=-N_z}^{N_z} u_{ij} v_{ij} \hbar_i^{(r)} \hbar_j^{(z)} r_i, \tag{3}$$

where

$$\hbar_i^{(r)} = \big(h_i^{(r)} + h_{i+1}^{(r)}\big)/2, \quad h_i^{(r)} = r_i - r_{i-1},\ i = 1, \dots, N_r,$$

$$\hbar_j^{(z)} = \big(h_j^{(z)} + h_{j+1}^{(z)}\big)/2, \quad h_j^{(z)} = z_j - z_{j-1},\ j = -N_z + 1, \dots, N_z.$$

Define the difference operator A on H^0 as

$$AV = -\frac{1}{r}(\bar{r} a V_{\bar{r}})_{\hat{r}} - (b V_{\bar{z}})_{\hat{z}}, \tag{4}$$

where $V, a, b \in H^0$,

$$a(i, j) = \sigma\big(r_i - h_i^{(r)}/2, z_j + h_j^{(z)}/2\big), \quad b(i, j) = \sigma\big(r_i + h_i^{(r)}/2, z_j - h_j^{(z)}/2\big),$$

$$(V)_{\bar{r}}(i, j) = (V_{i,j} - V_{i-1,j})/h_i^{(r)}, \quad (V)_{\hat{r}}(i, j) = (V_{i+1,j} - V_{i,j})/\hbar_i^{(r)},$$

$$(V)_{\bar{z}}(i, j) = (V_{i,j} - V_{i,j-1})/h_i^{(z)}, \quad (V)_{\hat{z}}(i, j) = (V_{i,j+1} - V_{i,j})/\hbar_i^{(z)}.$$

Using this discretization, replace (1) with the difference equation

$$AV = F, \tag{5}$$

where

$$F = \frac{1}{r}\big(\bar{r}(a - \sigma_0) U^0_{\bar{r}}\big)_{\hat{r}} + \big((b - \sigma_0) U^0_{\bar{z}}\big)_{\hat{z}}.$$

2. Rearrangement of the Linear System Convenient for Applying the Conjugate Gradient Method

When we write the two-dimensional vectors V and F, for instance, in columns, as one-dimensional arrays, we express (5) as a system of linear algebraic equations with five-diagonal matrix A and vectors V and F of size n. In H^0 the operator A is selfadjoint, but it is wasteful to solve the system of equations in this space due to the inner product (4). Pass to the space $\mathbb{R}^n$ with the inner product $(u, v) = \sum_{i=1}^{n} u_i v_i$. In $\mathbb{R}^{n\times n}$ the matrix A is not symmetric.

The Krylov subspace methods are often used to solve systems with sparse matrices, for instance, the stabilization method Bicg-Stab of biconjugate gradients [3, 4]. In this case it is possible to symmetrize the matrix by a diagonal transformation and use more efficient algorithms. To symmetrize the matrix, apply the algorithm of [5].

Put $l = 2N_z - 1$ and $m = N_r - 1$. A necessary and sufficient condition for symmetrizability is the cyclicity of the matrix entries of A,

$$\begin{aligned} & a_{(j+1)m+i+1,jm+i+1} a_{jm+i+1,jm+i} a_{jm+i,(j+1)m+i} a_{(j+1)m+i,(j+1)m+i+1} \\ & = a_{jm+i+1,(j+1)m+i+1} a_{(j+1)m+i+1,(j+1)m+i} a_{(j+1)m+i,jm+i} a_{jm+i,jm+i+1}, \end{aligned} \tag{6}$$

which approximation (4) meets. The transformation $B^{-1/2}AB^{1/2}$ with $B = \operatorname{diag}(b_1, \dots, b_n)$ leads to a symmetric matrix $\overline{A}$ with the entries $\overline{a_{ij}}$. The entries of B satisfy the recurrence

$$\begin{aligned} & b_0 = 1, \\ & b_{jm+i+1} = b_{jm+i} \frac{a_{jm+i+1,jm+i}}{a_{jm+i,(j+1)m+i}}, \quad i = 1, \dots, m-1,\ j = 1, \dots, l-1, \\ & b_{(j+1)m+i} = b_{jm+1} \frac{a_{(j+1)m+1,jm+1}}{a_{jm+i,(j+1)m+1}}, \quad i = m,\ j = 1, \dots, l-1. \end{aligned} \tag{7}$$

This yields the algebraic system

$$\overline{A}X = \overline{F}, \tag{8}$$

where $X = B^{-1/2}V$ and $\overline{F} = B^{-1/2}F$.

3. The 3-Dimensional Direct Electrical Logging Problem

In cylindrical coordinates consider an arbitrary distribution of specific electric conductivity $\sigma = \sigma(r, \varphi, z)$. The problem of modeling the LELP sonde readings reduces to the Dirichlet problem for Poisson's equation

$$\begin{aligned} & \frac{1}{r}\frac{\partial}{\partial r}\left(\sigma r \frac{\partial U^a}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left(\sigma \frac{\partial U^a}{\partial \varphi}\right) + \frac{\partial}{\partial z}\left(\sigma \frac{\partial U^a}{\partial z}\right) \\ & = \frac{1}{r}\frac{\partial}{\partial r}\left((\sigma_0 - \sigma) r \frac{\partial U^0}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left((\sigma_0 - \sigma)\frac{\partial U^0}{\partial \varphi}\right) + \frac{\partial}{\partial z}\left((\sigma_0 - \sigma)\frac{\partial U^0}{\partial z}\right) \end{aligned} \tag{9}$$

for the anomalous potential U^a with boundary conditions $U^a|_{r=R} = 0$ and $U^a|_{z=\pm Z} = 0$ and periodicity condition $U^a|_{\varphi=0} = U^a|_{\varphi=2\pi}$. To avoid the singularity arising as $r \to 0$, we use the mesh that is shifted along r away from $r = 0$, as suggested in [6]. In the cylinder

$$G = \{0 \le r \le R,\ 0 \le \varphi \le 2\pi,\ -Z \le z \le Z\}$$

introduce an arbitrary mesh [6] which is nonuniformly rectangular with respect to r and z and uniform with respect to φ:

$$\mathop{\overset{\wedge}{\omega}}_{h} = \{(r_i, \phi_k, z_j),\ i = 0, \dots, N_r,\ k = 0, \dots, N_k,\ j = -N_z, \dots, N_z\}. \tag{10}$$

Considering (10), take the linear finite-dimensional space H^0 of mesh functions equipped with the inner product

$$\begin{gathered} (u, v) = \sum_{i=0}^{N_r} \sum_{k=0}^{N_k} \sum_{j=-N_z}^{N_z} u_{ikj} v_{ikj} \hbar_i^{(r)} \hbar_k^{(\varphi)} \hbar_j^{(z)} r_i, \\ \hbar_i^{(r)} = \big(h_i^{(r)} + h_{i+1}^{(r)}\big)/2, \quad h_i^{(r)} = r_i - r_{i-1},\ i = 1, \dots, N_r, \\ \hbar_k^{(\varphi)} = \big(h_k^{(\varphi)} + h_{k+1}^{(\varphi)}\big)/2, \quad h_k^{(\varphi)} = \varphi_k - \varphi_{k-1},\ k = 1, \dots, N_\varphi, \\ \hbar_j^{(z)} = \big(h_j^{(z)} + h_{j+1}^{(z)}\big)/2, \quad h_j^{(z)} = z_j - z_{j-1},\ j = -N_z + 1, \dots, N_z. \end{gathered}$$

Define the difference operator A on H^0 as

$$AV = -\frac{1}{r}(\bar{r}aV_{\bar{r}})_{\hat{r}} - \frac{1}{r^2}(cV_{\overline{\varphi}})_{\widehat{\varphi}} - (bV_{\bar{z}})_{\hat{z}}, \tag{11}$$

where $V, a, b, c \in H^0$,

$$a(i,k,j) = \sigma\big(r_i - h_i^{(r)}/2, \varphi_k + h_k^{(\varphi)}/2, z_j + h_j^{(z)}/2\big),$$

$$b(i,j,k) = \sigma\big(r_i + h_i^{(r)}/2, \varphi_k + h_k^{(\varphi)}/2, z_j - h_j^{(z)}/2\big),$$

$$c(i,j,k) = \sigma\big(r_i + h_i^{(r)}/2, \varphi_k - h_k^{(\varphi)}/2, z_j + h_j^{(z)}/2\big).$$

The operators $(V)_{\bar{r}}$, $(V)_{\hat{r}}$, $(V)_{\bar{z}}$, $(V)_{\hat{z}}$, $(V)_{\overline{\varphi}}$, and $(V)_{\widehat{\varphi}}$ are defined by analogy with the two-dimensional case. Finally, we obtain the equation

$$AV = F, \tag{12}$$

where

$$F = \frac{1}{r}\big(\bar{r}(a-\sigma_0)U^0_{\bar{r}}\big)_{\hat{r}} + \frac{1}{r^2}\big((c-\sigma_0)U^0_{\overline{\varphi}}\big)_{\widehat{\varphi}} + \big((b-\sigma_0)U^0_{\bar{z}}\big)_{\hat{z}}.$$

The algorithm of symmetrization generalizes naturally to the three-dimensional case, so that we can also symmetrize (11) using the transformation $\overline{A} = B^{-1/2}AB^{1/2}$. Finally, by analogy with (8), we obtain

$$\overline{A}X = \overline{F}, \tag{13}$$

where $X = B^{1/2}V$ and $\overline{F} = B^{1/2}F$.

4. A Solution Method

In order to solve the linear systems (8) and (13), choose the conjugate gradient method because the matrices of these systems are symmetric and positive definite. Denote by x_n the approximate solution to the system $Ax = b$ at step n. Calculate the corresponding residual $r_n = b - Ax_n$ and an auxiliary vector p_n as

$$r_0 = b - Ax_0, \quad p_0 = r_0, \tag{14}$$

$$r_n = r_{n-1} - \alpha_{n-1}AP_{n-1}, \tag{15}$$

$$p_n = r_n + \beta_{n-1}AP_{n-1}, \quad n = 1, 2, \dots, \tag{16}$$

$$\alpha_n = \frac{r_n^T r_n}{p_n^T A p_n}, \quad \beta_n = \frac{r_{n+1}^T r_{n+1}}{r_n^T r_n}. \tag{17}$$

All operations in these formulas (14)–(17) are matrix-by-vector and parallelize well on GPU. We can do vector additions, multiplications by constants, and inner multiplications of vectors using the standard function library CUBLAS NVIDIA. To efficiently multiply matrices by vectors, we wrote a special procedure. We could, of course, apply a similar procedure of the CUSPARSE NVIDIA library, but in our case storing the matrix in the CSR format is inefficient. It is convenient to store only the values of matrix entries in separate arrays because we know the matrix structure and can use it to multiply the matrix by vectors. However, the rate of convergence in the conjugate gradient method is low. The maximal speedup attainable on GPU is only five-to-sixfold. Thus, we should use this method with a preconditioner.

The idea of preconditioning is to replace the system $Ax = b$ by the system $M^{-1}Ax = M^{-1}b$ or $AM^{-1}y = b$, with $x = M^{-1}y$, where either $M^{-1}A$ or AM^{-1} has a significantly smaller condition number than A itself, and the system

$$Mz = r \tag{18}$$

for an auxiliary vector z must be easily solvable.

Algorithm of the Preconditioned Conjugate Gradient Method

$$\begin{aligned}
&\text{Initialization:}\quad x_0,\quad r_0 = b - Ax_0,\quad Mz_0 = r_0,\quad p_0 = z_0;\\
&(1)\quad q_i = Ap_i,\quad \alpha_i = \frac{z_i^T r_i}{p_i^T q_i};\\
&(2)\quad x_{i+1} = x_i + \alpha_i p_i,\quad r_{i+1} = r_i - \alpha_i q_i;\\
&(3)\quad Mz_{i+1} = r_{i+1};\\
&(4)\quad \beta_i = \frac{z_{i+1}^T r_{i+1}}{z_i^T r_i},\quad p_{i+1} = r_{i+1} + \beta_i p_i.
\end{aligned} \tag{19}$$

In the case of GPU realization we impose on the matrix M the requirement of high parallelization not only of the solution of (18), but also of the construction of M itself. In this article we use the original approach [1] to constructing the preconditioning matrix relying on an approximation of the inverse matrix. Based on the Hotelling–Schulz algorithm [7, 8], this approach is fully parallel. Let us sketch it. Take an initial approximation D_0 to the inverse matrix. If

$$\|R_0\| \le k < 1,\quad R_0 = E - AD_0; \tag{20}$$

then we can construct an iteration approximating the inverse matrix as

$$D_1 = D_0 + D_0(E - AD_0), \tag{21}$$

$$D_2 = D_1 + D_1(E - AD_1) = 2D_1 - D_1AD_1D_{m+1} = D_m + D_m(E - AD_m). \tag{22}$$

This process converges provided that (20) holds, and the rate of convergence is described in [9] as

$$\|D_n - A^{-1}\| \le \|D_0\| \frac{k^{2^n}}{1-k}. \tag{23}$$

An important property in [9] of this process is that it preserves the symmetry of all matrices D_m: if $A = A^T$ and $D_0 = D_0^T$, then $D_m = D_m^T$.

As the initial approximation to the inverse matrix we take the Jacobi preconditioner $D_0 = \operatorname{diag}\left(a_{11}^{-1}, a_{22}^{-1}, \dots, a_{nn}^{-1}\right)$. In our case this is possible since the approximation to (1) and (9) yields matrices with weak diagonal domination. The matrix D_1 is easy to calculate:

$$d_{ii} = \frac{1}{a_{ii}},\quad d_{i,i+1} = \frac{a_{i,i+1}}{a_{i+1,i+1}a_{ii}},\quad d_{i,i+m} = \frac{a_{i,i+m}}{a_{i+m,i+m}a_{ii}}. \tag{24}$$

The structure of D_1 is the same as in the original matrix. This is rather useful since we can apply the already available procedure for multiplying a matrix by a vector. To decrease the number of arithmetic operations in the PCG algorithm, we can scale (8) and (13) beforehand to make the diagonal entries of A equal to 1:

$$a_{ij} = \frac{1}{a_{ii}} \cdot a_{ij} \cdot \frac{1}{a_{jj}},\quad i, j, = 1, \dots, n.$$

This procedure is preferable to the use of the Jacobi preconditioner in the conjugate gradient method because for the same number of iterations it needs fewer arithmetic operations to achieve prescribed accuracy.

The symmetry of the matrix is preserved. The formulas for calculating the preconditioner D_1 simplify. It is clear from (8) that the diagonal entries of D_1 become equal to 1, while the off-diagonal entries become opposite to the corresponding entries of the scaled matrix A. To decrease the number of iterations in the PCG method, we can also apply better preconditioners D_2 and D_3 (for better approximations to the inverse matrix). Note that the matrix D_2 in the two-dimensional case has 25 diagonals, while D_3 has 113 diagonals. It is inadvisable to calculate these matrices and, even worse, to store them on GPU. In the conjugate gradient method we are interested not in the preconditioning matrices, but only in the result of multiplication of a matrix by a vector. Therefore, use (22). Then step 3 of the PCG method requires three matrix-by-vector multiplications and one addition of vectors with multiplication by a constant. For D_3 we have

$$\begin{aligned} D_3 = D_2 + D_2(E - AD_2) &= (2D_1 - D_1AD_1)(2E - A(2D_1 - D_1AD_1)) \\ &= 2(2D_1 - D_1AD_1) - (2D_1 - D_1AD_1)A(2D_1 - D_1AD_1). \end{aligned} \tag{25}$$

This implies that step 3 in the PCG algorithm requires seven matrix-by-vector multiplications and scalar operations like vector addition and multiplication by a constant. All operations are fully parallel, so we use the CUBLAS CUDA NVIDIA library and our previously written matrix-by-vector multiplication procedure. But which preconditioner is preferable? To answer this question, we ran simulations. As a criterion for choosing the optimal preconditioner we took the minimal time for solving the problem with the prescribed accuracy.

5. Simulations

For the two-dimensional problem we tested the Jacobi preconditioner (taking a scaled system), D_1, D_2, and D_3 in the conjugate gradient method.

Consider a typical model with axially symmetric distribution of electric conductivity (Fig. 1).

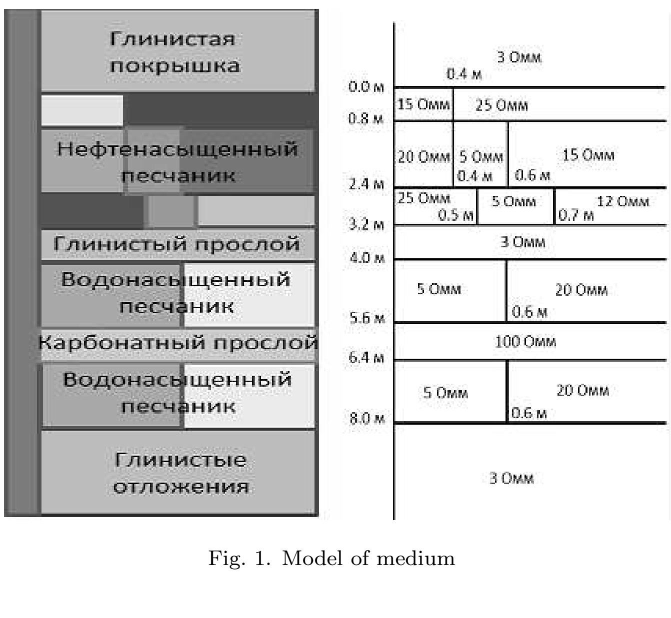

Fig. 1. Model of medium

The medium is divided into laterally inhomogeneous strata by a system of parallel flat boundaries. There is a well of radius 0.108 m with resistance 2 Ohm·m. Some strata could include the drilling fluid zone and the surrounding zone. The resistance of beds varies from 3 to 100 Ohm·m. For one probe position in the well we calculated the condition numbers (Table 1) of the original (symmetrized) matrices A, AD_0, AD_1, AD_2, and AD_3, for $n = 17136$. It is clear from Table 1 that the condition numbers decrease. We should also expect the number of iterations in the PCG method to decreases.

Table 1. The condition number of AM^{-1}

Matrix	Condition number
A	$4.5377 * 10^7$
AD_0	$3.0542 * 10^5$
AD_1	$7.6542 * 10^4$
AD_2	$3.8271 * 10^4$
AD_3	$1.9135 * 10^4$

Table 2. The two-dimensional problem.
Results of calculations by the CG method with various preconditioners

Метод	$n = 17139$		$n = 37846$		$n = 76136$	
	итерации	время, с	итерации	время, с	итерации	время, с
CG (scal)	2063	0.22	3412	0.41	4427	0.65
D_1	1030	0.094	1706	0.20	2274	0.35
D_2 (D_1)	729	0.085	1205	0.18	1577	0.31
D_3 (D_1)	505	0.075	846	0.17	1112	0.33
IC (Cusp.)	96	0.7	160	1.92	212	4.46

We simultaneously calculate LELP sonde readings for several (5 to 7) probes on one mesh. Sometimes not all probes are needed for interpretation; therefore, we use several meshes. The calculations summarized in Table 2 ran for three meshes. Since we have to solve the problem with specified accuracy, the error of at most 3% with respect to exact solutions, available for the radially layered media in a sufficiently wide electric conductivity range, near receiving points we constructed quite dense meshes. The first mesh ($n = 17139$) disregards the shortest (0.3 m) and the longest probes (8 m) of LELP. The second mesh ($n = 37846$) disregards the longest probe, while in the third mesh ($n = 76136$) all probes are described well. All calculations ran on the cluster NKS-30T+GPU with one Xeon X5670 processor (2.93 Ггц) and one NVIDIA Tesla M 2090 videocard on the Fermi architecture (compute capability 2.0).

Consider the results of simulations for one probe position in model 1. Table 2 summarizes the calculations for the conjugate gradient method with various preconditioners. Iterations ran until the relative norm of the residual reaches $1.d - 7$. It is clear from Table 2 that for all N the number of iterations decreases as the quality of preconditioning improves. The time spent on solving the system decreases too, but for $n = 76136$ for the conjugate gradient method with D_3 it increases. The reason is that on step 3 seven matrix-by-vector multiplications are necessary. These operations become more time-consuming than the use of the preconditioner D_2 giving more iterations but requiring three matrix-by-vector multiplications on this step. The last row of the table reports calculations with the IC preconditioner (incomplete Cholesky factorization) from the NVIDIA CUSPARSE library. In this variant we used the CSR format for storing the matrix and the matrix-by-vector multiplication procedure from the CUSPARSE library. It is clear from the table that the high-quality IC preconditioner beats all previous ones: the number of iterations is 5 times less than with D_3. However, the calculation time is greater by an order of magnitude than with D_3.

Of practical interest is the calculation at many points of the probe position in the well relative to the model (the calculation mesh is translated together with the probe). During the calculation, at each subsequent point with respect to depth it is reasonable to use as the initial solution the solution already found, which enables us to decrease the number of iterations on the next step. As simulations show, in the majority of cases this is efficient, especially so when the model includes sufficiently heavy strata. Table 3 summarizes the comparative calculations of 155 profile points for model 1 by the conjugate gradient method (with the preconditioner D_3) implemented on GPU and by a direct solution method, by the PARDISO program from the Intel MKL library. Presently PARDISO is one of the best programs as regards speed and accuracy of solution, but for relatively low-dimensional problems.

Table 3. The two-dimensional problem. Calculation for 155 profile points. Comparison with PARDISO

Method	$n = 17139$	$n = 37846$	$n = 76136$
CG (D_3)	13	22	39
Pardiso	13	29	66

Table 3 shows the total solution time for the entire problem (in seconds). Here it is necessary to consider that videocard initialization on the cluster NKS-30T takes 7–8 seconds. Thus, the actual time spent on the solution is less than the tabulated time. If we do not account for the time of videocard initialization then the GPU version for all meshes is twice as fast as the PARDISO program on CPU. Note that calculations ran on GPU with simple precision, but on CPU with double precision.

Table 4 summarizes the results of simulations for the system of equations for the three-dimensional problem with one probe position for model 1, but with inclined well. The calculations ran for two meshes. It is clear from Table 4 that as the preconditioner quality improves, the number of iterations decreases, but already for the preconditioner D_2 the solution time begins to grow.

Table 4. Three-dimensional problem. The results of calculations by the CG method with various preconditioners

Method	$n = 857480$		$n = 1458600$	
	iterations	time, s	iterations	time, s
CG (scal)	1816	1.45	2463	3.09
D_1	877	0.96	1203	2.09
D_2 (D_1)	606	1.07	846	2.39

Using the results of simulations, for two-dimensional problems we choose the preconditioner D_3, and for three-dimensional problems D_1. When the system of linear equations is prescaled, the construction of D_1 requires neither memory nor time resources. As a result of full parallelization, even for sufficiently large number of iterations we managed to obtain efficient algorithms. Note also their simplicity and reliability.

Conclusion

Algorithms and programs for fast GPU calculation of the lateral electrical logging sonde readings are created. This yields a speedup of 10-to-50 times in comparison with the running time of sequential software versions (for CPU) [10, 11] in dependence on mesh dimensions and geophysical properties of real models. Basing on the two-dimensional program, we created an inversion program [12].

REFERENCES

1. *Labutin I. B. and Surodina I. V.* Algorithm for sparse approximate inverse preconditioners in conjugate gradient method // Reliable Computing. 2013. V. 19, N 1. P. 120–126.

2. *Samarskiĭ A. A.* The Theory of Difference Schemes [in Russian]. Moscow: Nauka, 1979.

3. *Barret R., Berry M., Chan T. F., et al.* Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. Philadelphia, PA: SIAM, 1994.

4. *Van der Vorst H. A.* Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems // SIAM J. Sci. Stat. Comput. 1992. V. 13. P. 631–644.

5. *Kuznetsov Yu. I. and Agapitova N. S.* Computer Modeling: Mathematical Background [in Russian]. Yuzhno-Sakhalinsk: Izdat. YuSIEPI, 2003.

6. *Samarskiĭ A. A. and Nikolaev E. S.* Methods for Solving Grid Equations [in Russian]. Moscow: Nauka, 1978.

7. *Hotelling H.* Analysis of a complex of statistical variables into principal components // J. Educ. Psych. 1933. P. 417–441.

8. *Schulz G.* Iterative Berechnung der reziproken Matrix // Z. Angew. Math. Mech. 1933. Bd 13. S. 57–59.

9. *Faddeev D. K. and Faddeeva V. N.* Computational Methods of Linear Algebra [in Russian]. Moscow; Leningrad: Fizmatgiz, 1963.

10. *Dashevskii Yu. A., Surodina I. V., and Èpov M. I.* Three-dimensional mathematical modeling of the monitoring system of electric properties of shink fluid // International Conference "Mathematical Methods for Geophysics". Novosibirsk, 2003. Part 1. P. 268–272.

11. *Dashevskii Yu. A., Surodina I. V., and Èpov M. I.* Quasi-three-dimensional mathematical simulation of diagrams of non-axisymmetric direct current sounds in anisotropic cuts // Sib. Zh. Ind. Mat. 2002. V. 5, N 3. P. 76–91.

12. *Nikitenko M. N., Surodina I. V., Mikhaylov I. V., Glinskikh V. N., and Suhorukova C. V.* Formation evaluation via 2D processing of induction and galvanic logging data using high-performance computing // Abstr. 77th EAGE Conf. Exhibition 2015 (Madrid, Spain, June 1–4, 2015). Madrid, 2015. P. 1–5.

September 4, 2015

I. V. Surodina
Institute of Computational Mathematics and Mathematical Geophysics
Novosibirsk, Russia
sur@ommfao1.sscc.ru

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 519.6

ANALYSIS AND NUMERICAL SOLUTION OF AN INVERSE PROBLEM OF MODELING CIRCULATION IN AQUATORIA WITH LIQUID BOUNDARY

V. I. Agoshkov, D. S. Grebennikov, and T. O. Sheloput

Abstract. In geophysical hydrodynamics the problem exists of modeling physical processes in water areas with the so-called liquid boundaries. One of the approaches to solving the problem is to apply the optimal control theory and data assimilation methods. In this paper under study the problem of finding the unknown function in the boundary condition of the system of shallow-water equations. We propose an iteration algorithm based on the theory of inverse problems and optimal control theory. We also obtain conditions for the unique and dense solvability of the problem and some conditions for the convergence iteration algorithm as well. We present the results of simulations of the Baltic Sea with this algorithm.

Keywords: inverse problem, liquid (open) boundary, ill-posed problem, iteration, shallow water equations

1. Introduction

Among geophysical hydrodynamics often addresses the problem of modeling the processes in water areas (seas, oceans, rivers, and so on) with liquid boundary. For instance, the southern boundaries of the Indian ocean, the northern boundaries of the Barents and Kara Seas, the boundaries going along straits, river mouths, and so on. This article deals with the problem of finding the boundary functions on liquid boundaries more precisely.

We can apply various existing approximations to specify boundary conditions on a liquid boundary. The material boundary approximation is sometimes used: the liquid boundary is regarded as dynamic, with the nonpermeability condition imposed on it [1, pp. 82–141]. The approximation is convenient when the deformation of the model region is not too large. But in this case the boundary is an additional unknown of the problem [2], which complicates the use of many modern numerical methods, algorithms, and tools, as well as theoretical studies. Another common approach is to use the averaged data on the flow through the open boundary [3]. Sometimes it is possible to make a preliminary calculation over the World Ocean on a coarse mesh and use the resulting data as a boundary condition on the liquid boundary. Probably, it is most promising to combine one of these methods with data assimilation methods.

The idea of using optimal control theory and data assimilation methods to solve the liquid boundary problem was studied in [4–6] for instance. In particular, in [5] there is proposed and studied an iterative algorithm for reconstructing from observations the unknown boundary function accounting for the influence of the ocean on the open boundary of the simulated region, where the system of tidal dynamics equations is chosen as the model describing physical processes in the model area. Note that the iterative algorithms of [4, 5] must be implemented at each time.

In this article we study the questions of existence and uniqueness for a solution to the inverse problem of calculating the unknown function in the boundary condition for the equations of shallow water type used to model certain kinds of fluid circulation in basins. The just of our approach to studying these questions is in reducing them to similar questions concerning the boundary function for the wave equation which the original system reduces to under some restrictions. We construct the boundary condition for the wave equation itself basing on the shallow water equations under consideration. In addition, we propose an iterative algorithm and apply it to the Baltic Sea. We make a test in which the liquid boundary passes around the Swedish town of Trelleborg and separates the North sea from the Baltic Sea. By this example the article demonstrates that the proposed algorithm is sufficiently precise.

2. Statement of the Problem

1. Introduce the following notation. In the rectangular system of coordinates (x, y, z), take $(x, y) \in \Omega$, where Ω is a bounded region in $\mathbb{R}^2$. Take the time variable $t \in [0, T]$ with $T < \infty$, and consider the cylinder $Q_T \equiv \Omega \times (0, T)$ over Ω. The boundary $\Gamma \equiv \partial\Omega$ of Ω is piecewise C^2 smooth and satisfies the Lipschitz condition, $\Gamma_T \equiv \Gamma \times (0, T)$ is the lateral surface of Q_T, and $\Gamma_{cT} = \Gamma_c \times (0, T)$, where Γ_c is the liquid boundary. Denote by u and v the components of the fluid velocity along the axes Ox and Oy. Assume that $-\xi(x, y, t) < z < H(x, y)$, where $z = \xi(x, y, t)$ is the equation of the free ocean surface, $z = H(x, y)$ is the floor equation (assume for simplicity that $H(x, y)$ is a smooth function), $g = \text{const}$ is the free fall acceleration, ρ is the density of the fluid, p^a is the atmospheric pressure, τ_1 and τ_2 are the wind friction stresses, and l is the Coriolis parameter.

Consider the system of hydrodynamics equations averaged over depth (the z coordinate) [7, p. 47]:

$$\frac{\partial U}{\partial t} - lV + g\frac{\partial \xi}{\partial x} = -\frac{1}{\rho_0} p_x^a + \frac{1}{H\rho_0}\tau_1 \quad \text{in } Q_T, \tag{1}$$

$$\frac{\partial V}{\partial t} + lU + g\frac{\partial \xi}{\partial y} = -\frac{1}{\rho_0} p_y^a + \frac{1}{H\rho_0}\tau_2 \quad \text{in } Q_T, \tag{2}$$

$$\xi_t + (UH)_x + (VH)_y = 0 \quad \text{in } Q_T, \tag{3}$$

where U and V are the averaged fluid velocity functions along Ox and Oy (henceforth they will be velocities):

$$U = \frac{1}{H}\int_0^H u\,dz, \quad V = \frac{1}{H}\int_0^H v\,dz;$$

$$p_x^a = \frac{\partial p^a}{\partial x}, \quad p_y^a = \frac{\partial p^a}{\partial y}, \quad (UH)_x \equiv \frac{\partial (UH)}{\partial x}, \quad (VH)_y \equiv \frac{\partial (VH)}{\partial y}, \quad \xi_t \equiv \frac{\partial \xi}{\partial t}.$$

Below we neglect the Coriolis force, putting $l \equiv 0$.

Equip the system with the initial and boundary conditions

$$U(x, y, 0) = U_0(x, y), \quad V(x, y, 0) = V_0(x, y), \quad \xi(x, y, 0) = \xi(x, y) \quad \text{in } \Omega, \tag{4}$$

$$(\mathbf{U}, \mathbf{n}) = m_c u_c \quad \text{on } \Gamma_{cT}, \tag{5}$$

where $\mathbf{U} = (U, V)^T$ is the velocity vector, $\mathbf{n}$ is the outer normal, m_c is the characteristic function of Γ_{cT}.

Put

$$f_1 = -\frac{H}{\rho_0}p_x^a + \frac{1}{\rho_0}\tau_1, \quad f_2 = -\frac{H}{\rho_0}p_y^a + \frac{1}{\rho_0}\tau_2, \quad \mathbf{f} = (f_1, f_2)^T.$$

Multiply (1) and (2) by H, differentiate the first equation with respect to x, the second one with respect to y, and (3) with respect to t. Combining then yields

$$-\frac{\partial^2 \xi}{\partial t^2} + \operatorname{div}(gH\nabla\xi) = \operatorname{div}\mathbf{f} \quad \text{in } Q_T.$$

We can put the first two equations in vector form:

$$gH\begin{pmatrix} \partial\xi/\partial x \\ \partial\xi/\partial y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} - \frac{\partial}{\partial t}\begin{pmatrix} UH \\ VH \end{pmatrix} \text{ in } Q_T.$$

Considering these two equations on Γ and taking the inner product with the outer normal $\mathbf{n}$ to Ω, we obtain the boundary condition

$$gH\frac{\partial\xi}{\partial\mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) - \frac{\partial}{\partial t}H(\mathbf{U}\cdot\mathbf{n}) \quad \text{on } \Gamma_T.$$

Thus, we can reformulate (1)–(5) for the shallow water equations as the following problem for the wave equation:

$$\begin{aligned}
\frac{\partial^2\xi}{\partial t^2} - \operatorname{div}(gH\nabla\xi) &= -\operatorname{div}\mathbf{f} \quad \text{in } Q_T,\\
\xi|_{t=0} &= \xi_0 \quad \text{in } \Omega,\\
\left.\frac{\partial\xi}{\partial t}\right|_{t=0} &= -\frac{\partial U_0}{\partial x} - \frac{\partial V_0}{\partial y} \equiv \xi_1 \quad \text{in } \Omega,\\
gH\frac{\partial\xi}{\partial\mathbf{n}} &= (\mathbf{f}\cdot\mathbf{n}) \quad \text{on } (\Gamma\backslash\Gamma_c)\times(0,T),\\
gH\frac{\partial\xi}{\partial\mathbf{n}} &= (\mathbf{f}\cdot\mathbf{n}) - m_cH\frac{\partial u_c}{\partial t} \equiv (\mathbf{f}\cdot\mathbf{n}) + m_cU_c \text{ on } \Gamma_c\times(0,T),
\end{aligned} \tag{6}$$

where $U_c = -m_cH\partial u_c/\partial t$. We impose necessary smoothness and agreement conditions while considering the classical statement of the problem of type (6).

Suppose further that U_c is an additional unknown on $\Gamma_c \times (0,T)$ and introduce the closuring equation

$$m_0\xi = m_0\varphi_{obs} \text{ on } \Gamma_T,$$

where m_0 is the characteristic function of $\Gamma_{oT} \subset \Gamma_T$ with $\Gamma_{oT} \equiv \Gamma_o \times (0,T)$, while φ_{obs} is the observed level of ξ on Γ_{oT}.

2. Consider only real variables, functions, and function spaces. Introduce the Hilbert spaces (see [8] for instance)

$$L_2(Q_T) : (u,v)_{L_2(Q_T)} \equiv (u,v)_{2,Q_T} = \int_0^T\int_\Omega uv\, d\Omega dt,$$

$$W_2^1(Q_T) : (u,v)_{W_2^1(Q_T)} = \int_{Q_T}\left(uv + \sum_{i=1}^2 \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} + \frac{\partial u}{\partial t}\frac{\partial v}{\partial t}\right) d\Omega dt,$$

$$W_{2,T}^1 \equiv \{u :\ u \in W_2^1(Q_T),\ u = 0 \text{ for } t = T\}.$$

Equip each of the spaces with the norm induced by the inner product.

Introduce the space H_o as the subspace of $L_2(\Gamma_T)$ consisting of the elements vanishing on $\Gamma_T \setminus \Gamma_{oT}$. Also introduce the space H_c as the subspace of the space of traces of functions in $W_2^1(Q_T)$ on Γ_T consisting only of the elements vanishing on $\Gamma_T \setminus \Gamma_{cT}$. Taking $\mathbf{f} \in (W_2^1(Q_T))^2$, $\xi_0 \in W_2^1(Q_T)$, $\xi_1 \in L_2(Q_T)$, $0 < \nu \leq gH(x)$, and $\phi_{obs} \in H_o$, consider the **inverse problem**: Find $\xi \in W_2^1(Q_T)$ on Q_T and $U_c \in H_c$ such that

$$\frac{\partial^2 \xi}{\partial t^2} - \operatorname{div}(gH\nabla\xi) = -\operatorname{div}\mathbf{f} \quad \text{a.e. in } Q_T, \tag{7}$$

$$\xi|_{t=0} = \xi_{(0)}, \ \left.\frac{\partial \xi}{\partial t}\right|_{t=0} = \xi_{(1)} \quad \text{a.e. in } \Omega, \tag{8}$$

$$gH\frac{\partial \xi}{\partial \mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) \quad \text{a.e. on } (\Gamma\backslash\Gamma_c)\times(0,T), \tag{9}$$

$$gH\frac{\partial \xi}{\partial \mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) + U_c \quad \text{a.e. on } \Gamma_c\times(0,T), \tag{10}$$

$$m_0\xi = m_0\varphi_{obs} \quad \text{a.e. on } \Gamma_T. \tag{11}$$

To generalize problem (7)–(10), take the inner product of (7) with $\tilde{\xi} \in W_{2,T}^1(Q_T)$ and integrate by parts while accounting for the boundary conditions. This yields

$$a(\xi,\tilde{\xi}) = F(\tilde{\xi}) + b(U_c,\tilde{\xi}) \quad \forall \tilde{\xi} \in W_{2,T}^1(Q_T), \tag{12}$$

where

$$a(\xi,\tilde{\xi}) \equiv \int\limits_{Q_T} \left(-\xi_t\tilde{\xi}_t + gH\nabla\xi\nabla\tilde{\xi}\right) d\Omega dt,$$

$$F(\tilde{\xi}) \equiv \int\limits_{Q_T} \mathbf{f}\cdot\nabla\tilde{\xi}\, d\Omega dt + \int\limits_{\Omega} \xi_1\tilde{\xi}(x,y,0)\, d\Omega, \quad b(U_c,\tilde{\xi}) \equiv \int\limits_{\Gamma_{cT}} U_c\tilde{\xi}\, d\Gamma dt.$$

The generalized statement of (7)–(10) is as follows: Find $\xi \in W_2^1(Q_T)$ satisfying (12) such that $\xi|_{t=0} = \xi_0$ a.e. in Ω. The generalized statement of the inverse problem is in order: Find $\xi \in W_2^1(Q_T)$ and $U_c \in H_c$ satisfying (12) and (11) such that $\xi|_{t=0} = \xi_0$ a.e. in Ω.

Below we understand problems of type (7)–(10) in generalized form, although for clarity we often write down their in classical form (7)–(10).

3. The Optimal Control Problem

We now proceed to the generalized statement of (7)–(11) in which we understand (11) in the sense of least squares: Find $\xi \in W_2^1(Q_T)$ and $U_c \in H_c$ satisfying (7)–(10) and minimizing J_α:

$$\inf_{U_c\in H_c} J_\alpha(U_c,\xi(U_c)),$$

where $\alpha \geq 0$ and

$$J_\alpha(U_c,\xi(U_c)) \equiv \frac{\alpha}{2}\iint\limits_{\Gamma_T} m_c U_c^2\, d\Gamma dt + \frac{1}{2}\iint\limits_{\Gamma_T} m_0(\xi-\varphi_{obs})^2\, d\Gamma dt. \tag{13}$$

It is not difficult to show that for $\alpha > 0$ this functional is strictly convex and the minimization problem has the unique solution. The optimality condition $\delta J_\alpha = 0$ leads to the equation

$$\alpha \iint\limits_{\Gamma_T} m_c U_c \delta U_c \, d\Gamma dt + \iint\limits_{\Gamma_T} m_0(\xi - \varphi_{obs})\delta\xi \, d\Gamma dt = 0, \tag{14}$$

where δU_c and $\delta\xi$ satisfy

$$\begin{gathered}\frac{\partial^2 \delta\xi}{\partial t^2} - \operatorname{div}(gH\nabla\delta\xi) = 0 \quad \text{in } Q_T,\\ \delta\xi|_{t=0} = 0, \quad \left.\frac{\partial\delta\xi}{\partial t}\right|_{t=0} = 0 \quad \text{in } \Omega,\\ gH\frac{\partial\delta\xi}{\partial\mathbf{n}} = m_c\delta U_c \quad \text{on } \Gamma_T.\end{gathered} \tag{15}$$

To rearrange (14), introduce the adjoint problem

$$\begin{gathered}\frac{\partial^2 q}{\partial t^2} - \operatorname{div}(gH\nabla q) = 0 \quad \text{in } Q_T,\\ q|_{t=T} = 0, \quad \left.\frac{\partial q}{\partial t}\right|_{t=T} = 0 \quad \text{in } \Omega,\\ gH\frac{\partial q}{\partial\mathbf{n}} = m_0(\xi - \varphi_{obs}) \quad \text{on } \Gamma_T.\end{gathered} \tag{16}$$

Then

$$0 = \iint\limits_{Q_T} \left(\frac{\partial^2 q}{\partial t^2} - \operatorname{div}(gH\nabla q)\right)\delta\xi \, d\Omega dt$$

$$= \int\limits_{\Omega} \frac{\partial q}{\partial t}\delta\xi\bigg|_0^T d\Omega - \int\limits_{\Omega} q\frac{\partial\delta\xi}{\partial t}\bigg|_0^T d\Omega + \iint\limits_{Q_T} q\underbrace{\left(\frac{\partial^2\delta\xi}{\partial t^2} - \operatorname{div}(qH\nabla\delta\xi)\right)}_{=0} d\Omega dt$$

$$- \iint\limits_{\Gamma_T} \underbrace{\left(gH\frac{\partial q}{\partial\mathbf{n}}\right)}_{=m_0(\xi-\varphi_{obs})} \delta\xi d\Gamma dt + \iint\limits_{\Gamma_T} q\underbrace{\left(gH\frac{\partial\delta\xi}{\partial\mathbf{n}}\right)}_{=m_c\delta U_c} d\Gamma dt.$$

Consequently,

$$\iint\limits_{\Gamma_T} m_0(\xi - \varphi_{obs})\delta\xi \, d\Gamma dt = \iint\limits_{\Gamma_T} q m_c \delta U_c \, d\Gamma dt, \tag{17}$$

and (14) becomes

$$\alpha \iint\limits_{\Gamma_T} m_c U_c \delta U_c \, d\Gamma dt + \iint\limits_{\Gamma_T} q m_c \delta U_c \, d\Gamma dt = 0. \tag{18}$$

Since δU_c is an independent variation, we can express the optimality condition as

$$\alpha m_c U_c + m_c q = 0 \text{ on } \Gamma_T. \tag{19}$$

Now we can write down the complete system of variational equations and state an iterative process of approximate solution of the generalized problem. The system of variational equations is

$$
\begin{aligned}
&\frac{\partial^2 \xi}{\partial t^2} - \operatorname{div}(gH\nabla\xi) = -\operatorname{div}\mathbf{f} \text{ in } Q_T,\\
&\xi|_{t=0} = \xi_0, \quad \left.\frac{\partial \xi}{\partial t}\right|_{t=0} = \xi_1 \text{ in } \Omega,\\
&gH\frac{\partial \xi}{\partial \mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) \text{ on } (\Gamma\backslash\Gamma_c)\times(0,T),\\
&gH\frac{\partial \xi}{\partial \mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) + U_c \text{ on } \Gamma_c\times(0,T),\\
&\frac{\partial^2 q}{\partial t^2} - \operatorname{div}(gH\nabla q) = 0 \text{ in } Q_T,\\
&q|_{t=T} = 0, \quad \left.\frac{\partial q}{\partial t}\right|_{t=T} = 0 \text{ in } \Omega,\\
&gH\frac{\partial q}{\partial \mathbf{n}} = m_0(\xi - \varphi_{obs}) \text{ on } \Gamma_T,\\
&\alpha m_c U_c + m_c q = 0 \text{ on } \Gamma_T.
\end{aligned}
\tag{20}
$$

Before preluding the iterative process, let us study the solvability of the inverse problem.

4. Solvability of the Problem

4.1. Unique solvability. Study the unique solvability of problem (7)–(11).

Suppose that the problem has two solutions $\xi' \neq \xi''$ and $U_c' \neq U_c''$. Then $\xi \equiv \xi' - \xi''$ and $U_c \equiv U_c' - U_c''$ satisfy

$$
\begin{aligned}
&\frac{\partial^2 \xi}{\partial t^2} - \operatorname{div}(gH\nabla\xi) = 0 \text{ in } Q_T,\\
&\xi|_{t=T} = 0, \quad \left.\frac{\partial \xi}{\partial t}\right|_{t=T} = 0 \text{ in } \Omega,\\
&gH\frac{\partial \xi}{\partial \mathbf{n}} = m_c U_c \text{ on } \Gamma\times(0,T),\\
&\xi = 0 \text{ on } \Gamma_o\times(0,T).
\end{aligned}
\tag{21}
$$

In case $\Gamma_c = \Gamma_o$ we can treat (21) as the mixed initial-boundary problem. Theorem (5.1) of [8] implies that (21) has the unique solution $\xi = 0$ in $W_2^1(Q_T)$; consequently, U_c vanishes on Γ_{cT}.

In case $\Gamma_c \neq \Gamma_o$ we come to the problem with homogeneous Cauchy-type boundary conditions with respect to space and time:

$$
\begin{aligned}
&\frac{\partial^2 \xi}{\partial t^2} - \operatorname{div}(gH\nabla\xi) = 0 \text{ in } Q_T,\\
&\xi|_{t=T} = 0, \quad \left.\frac{\partial \xi}{\partial t}\right|_{t=T} = 0 \text{ in } \Omega,\\
&gH\frac{\partial \xi}{\partial \mathbf{n}} = \xi = 0 \text{ on } \Gamma_o\times(0,T).
\end{aligned}
\tag{22}
$$

The unique solvability of this problem is studied in [9] (see Theorem 1.2.1 and Corollary 1.2.5 on pp. 4–10 for instance). Without stating these results here, we observe that the sufficient conditions for the uniqueness of solution to (22) (call them conditions I) include the requirements on the boundary Γ_o that are too strong and often incompatible with practical problems.

4.2. Dense solvability. Proceed to the dense solvability of (7)–(11) (see [10]).

It is clear from (18) that for $\alpha = 0$ the optimality condition is $m_c q = 0$ a.e. on Γ_{cT}, where q is a solution to (16). In case $\Gamma_c = \Gamma_o$ the optimality conditions become

$$\frac{\partial^2 q}{\partial t^2} - \mathrm{div}(gH\nabla q) = 0 \text{ in } Q_T, \tag{23}$$

$$q|_{t=T} = 0, \quad \left.\frac{\partial q}{\partial t}\right|_{t=T} = 0 \text{ in } \Omega, \tag{24}$$

$$gH\frac{\partial q}{\partial \mathbf{n}} = m_o(\xi - \phi_{obs}) \text{ on } \Gamma_o \times (0,T), \tag{25}$$

$$gH\frac{\partial q}{\partial \mathbf{n}} = 0 \text{ on } \Gamma \setminus \Gamma_o \times (0,T), \tag{26}$$

$$q = 0 \text{ on } \Gamma_o \times (0,T). \tag{27}$$

The unique generalized solution to this system vanishes identically; consequently, (25) yields $m_o(\xi - \phi_{obs}) = 0$ and the minimum of J_α for $\alpha = 0$ is also zero, which means the dense solvability of (7)–(11) (see [10]).

In case $\Gamma_c \neq \Gamma_o$ we have a problem with Cauchy-type boundary conditions on Γ_c; therefore, dense solvability requires additional conditions (conditions I, see the previous subsection).

Basing on the above argument, we can state the following:

1. *In case* $\Gamma_c = \Gamma_o$, *problem* (7)–(11) *is uniquely and densely solvable.*
2. *In case* $\Gamma_c \neq \Gamma_o$ *we have unique or dense solvability under conditions* I.

5. Iterative Algorithm

Since the dense solvability of (7)–(11) yields $\inf J_\alpha = J_* \to 0$ as $\alpha \to 0$, for $\alpha > 0$ sufficiently small we can assume that $\xi \cong \xi(\alpha)$ and $U_c \cong U_c(\alpha)$, where $\xi(\alpha)$ and $U_c(\alpha)$ are exact solutions to the minimization problem for J_α; hence, it suffices to construct an approximation to $\xi(\alpha)$ and $U_c(\alpha)$ by a suitable iterative algorithm (see [10]).

Let us state a simple iterative method for the system of variational equations

(20) similar to the gradient descent method for J_α:

$$
\begin{aligned}
&\frac{\partial^2\xi^k}{\partial t^2} - \operatorname{div}(gH\nabla\xi^k) = -\operatorname{div}\mathbf{f} \text{ in } Q_T,\\
&\qquad \xi^k|_{t=0} = \xi_0, \quad \left.\frac{\partial \xi^k}{\partial t}\right|_{t=0} = \xi_1 \text{ in } \Omega,\\
&\qquad gH\frac{\partial\xi^k}{\partial\mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) \text{ on } (\Gamma\backslash\Gamma_c)\times(0,T),\\
&\qquad gH\frac{\partial\xi^k}{\partial\mathbf{n}} = (\mathbf{f}\cdot\mathbf{n}) + U_c^k \text{ on } \Gamma_c\times(0,T),\\
&\frac{\partial^2 q^k}{\partial t^2} - \operatorname{div}(gH\nabla q^k) = 0 \text{ in } Q_T,\\
&q^k|_{t=T} = 0, \quad \left.\frac{\partial q^k}{\partial t}\right|_{t=T} = 0 \text{ in } \Omega,\\
&\qquad gH\frac{\partial q^k}{\partial\mathbf{n}} = m_0(\xi^k - \varphi_{obs}) \text{ on } \Gamma_T,\\
&\qquad U_c^{k+1} = U_c^k - \tau_k(\alpha U_c^k + q^k), \text{ on } \Gamma_c\times(0,T).
\end{aligned}
\tag{28}
$$

Here τ_k is a parameter of the iterative process. The choice of τ_k and the regularization parameter $\alpha \geq 0$ affect the convergence of approximate solutions $\xi^k(\alpha)$ and $U_c^k(\alpha)$ to the solutions ξ and U_c to problem (7)–(11). For instance, [10] implies that *for arbitrary $\alpha > 0$ and sufficiently small $\tau = \tau_k$ the iterative algorithm* (28) *converges.*

Using the theory of extremal problems, we can choose the parameter of the iterative process as [11]

$$\tau_k \simeq \frac{J_\alpha(v^k) - J_*}{\|J'_\alpha(v^k)\|^2},$$

where $\inf J_\alpha = J_*$. The dense solvability implies $J_* \approx 0$, and we can take (see [10])

$$\tau_k \cong \frac{J_\alpha(v^k)}{\|J'_\alpha(v^k)\|^2} = \frac{\|m_o(\xi^k - \phi_{obs})\|^2_{L_2(\Gamma_T)}}{4\|m_c q^k\|^2_{L_2(\Gamma_T)}} \tag{29}$$

as the optimal collection of parameters of the iterative process in this problem.

As we showed above, at each iteration it is necessary to solve the direct and adjoint problem. For a numerical implementation of these problems we can use, for instance, projection-grid methods or finite difference methods.

5. Simulations

Let us present the results of implementation of (28) to the Baltic Sea. For this article, we ran the two series of simulations: the first for test functions (calculation in the real sea area), and the second for certain data on the Baltic Sea which are close to reality. The goal of simulations for test functions was to test and estimate the performance of the developed programs, analyze the convergence of iterations, and estimate the relative error of the solution obtained. Running simulations with data which is close to real made it possible to estimate the performance of the developed method and the feasibility of its practical application. As the liquid boundary in all simulations we chose the boundary near the Swedish town of Trelleborg and separating the North and Baltic Seas. To solve the direct and adjoint problems,

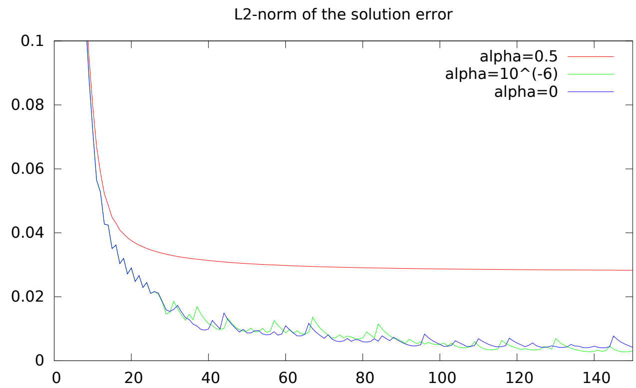

Fig. 1. Relative error of the solution for $t = T$ depending on the number of iterations

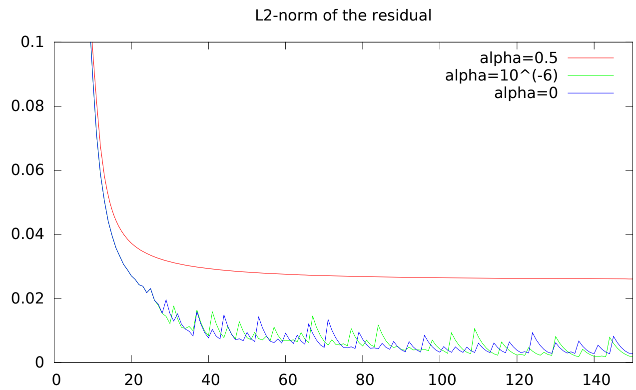

Fig. 2. Relative residue of the solution for $t = T$ depending on the number of iterations

we used the finite difference method (see [8]). The data on the boundary of the Baltic Sea was encoded as a masque of 0's and 1's, while the boundary itself was approximated by segments parallel to the coordinate axes.

To try the programs, we chose the test function $\sin(x/L)\sin(y/L)\sin(t/2T)$, used it to calculate the right-hand side as well as the initial and boundary conditions. Then this function and the boundary function on the liquid boundary were assumed unknown. They were then reconstructed by using the above iterative algorithm. Fig. 1 depicts the dependence of the relative error of the solution on the number of iterations for various values of the regularization parameter α, while Fig. 2 shows the norm of the residual (in other words, the square root of the value of J_α). It is clear from these figures that the algorithm converges sufficiently fast (in 20 iterations) and monotonely for large α; however, the error of solution is then large. For smaller values of the parameter it is expedient either to halt the process after 40–60 iterations, or to increase the accuracy of solution to direct and adjoint problems (in particular, decrease space and time meshsizes). We showed experimentally that the relative error of the resulting solution is uniformly distributed over the whole region; consequently, the resulting solution acceptably reproduces the characteristic test solution. The results also show that for small values of the regularization parameter both the residual and the error of solution can be decreased by a factor of

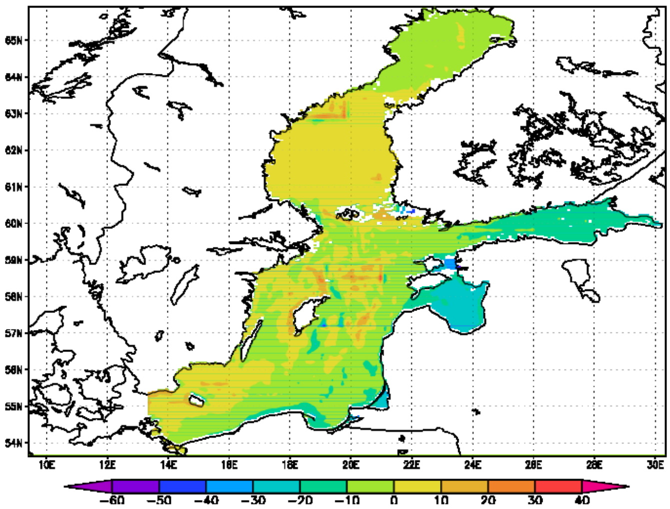

Fig. 3. The level (in cm) for $t = T$ after 3 iterations

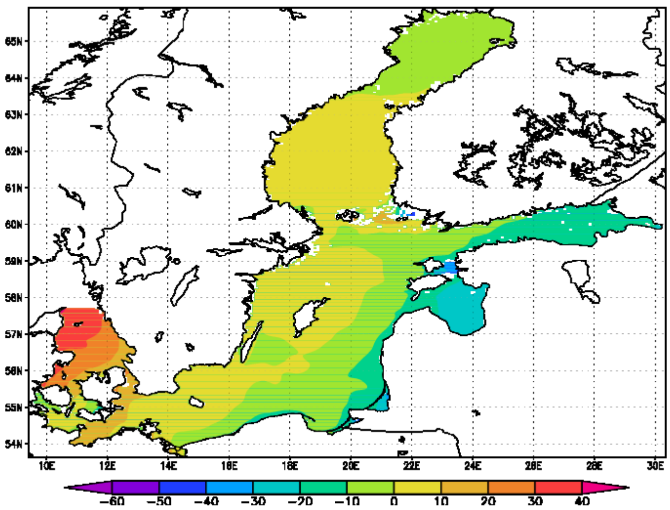

Fig. 4. The level (in cm) obtained as the result of calculating the model [12]

approximately 10^4.

We can draw the following conclusions from the above results. Firstly, the choice of optimal parameter τ_k using (29) guarantees fast convergence of the process (20 iterations). Secondly, the choice of regularization parameter depends on the accuracy of solution to the direct and adjoint problems (in particular, on the time and space meshsize).

Let us present the results of modeling the hydrodynamics of the Baltic Sea taking into account the liquid boundary for the initial data which are close to reality. We took the required data on atmospheric forcing on the ERA-Interim resource, while the initial data and observations were obtained from the results of calculating the three-dimensional model of hydrothermodynamics of the Baltic Sea developed at the Institute of Computational Mathematics [12]. As the initial data we used that for January 1, 2012. We also chose the following parameters of the iterative process: τ_k using (29), and $\alpha = 10^{-3}$. For this choice of parameters the process converges in 3 iterations (i.e., sufficiently fast), which agrees with the theory of [10]. Fig. 3 shows the level (in cm) at the final moment of time obtained on the last iteration. It is clear from the figure that our results acceptably reproduce the data of the model (Fig. 4). The oscillations of the level depicted in Fig. 3 in the central part of the basin are due to large depth variations in this region of the Baltic Sea. The resulting norm of the residual (the difference between the obtained and model level on the liquid boundary) on the last iteration was $6 \cdot 10^{-4}$, which enables us to appreciate the accuracy of the above algorithms for solving the problem of refining the form of boundary conditions on the liquid boundary.

Clearly, the algorithms and approaches of this article can be applied to solve the problem of boundary conditions on the liquid boundary for other water areas.

REFERENCES

1. *Gill A. E.* Atmosphere-Ocean Dynamics. Vol. 1 [Russian translation]. Moscow: Mir, 1986.

2. *Agoshkov V. I.* Investigation of a class of inverse problems on optimal boundaries // Computational Science for the 21st Century (Ed. by M.-O. Bristeau, G. Etgen and others). Chichester; New York; Toronto: John Wiley and Sons, 1997. P. 589–598.

3. *Chernov I. A. and Tolstikov A. V.* Numerical modeling of large-scale dynamics of the White Sea // Tr. Karel. Nauch. Tsentra RAN. 2014. N 4. P. 137–142.

4. *Dementyeva E. V., Karepova E. D., and Shaidurov V. V.* Recovery of a boundary function from observation data for the surface wave propagation problem in an open basin // Sib. Zh. Ind. Mat. 2013. V. 16, N 1. P. 10–20.

5. *Agoshkov V. I.* Inverse problems of the mathematical theory of tides: boundary-function problem // Russ. J. Numer. Anal. Math. Modelling. 2005. V. 20, N 1. P. 1–18.

6. *Agoshkov V. I.,* Application of mathematical methods for solving the problem of liquid boundary conditions in hydrodynamics // Numerical Analysis, Scientific Computing, Computer Science, Special Volume of ZAMM (Proc. ICIAM-95). Berlin, 1996. P. 337–338.

7. *Vol′tsinger N. E. and Pyaskovskii R. V.* Basic Oceanological Problems of the Theory of Shallow Water [in Russian]. Leningrad: Gidrometeoizdat, 1968.

8. *Ladyzhenskaya O. A.* The Boundary Value Problems of Mathematical Physics [in Russian]. New York etc.: Springer-Verlag, 1985.

9. *Isakov V.,* Inverse Source Problems. Providence: Amer. Math. Soc., 1996.

10. *Agoshkov V. I.* Optimal Control Methods and Adjoint Equations in Mathematical Physics Problems [in Russian]. Moscow: IVM RAN, 2003.

11. *Vasil′ev F. P.* Methods for Solving Extremal Problems [in Russian]. Moscow: Nauka, 1981.

12. *Zalesny V. B., Gusev A. V., Chernobay S. Yu., Aps R., Tamsalu R., Kujala P., and Rytkönen J.* The Baltic Sea circulation modelling and assessment of marine pollution // Russ. J. Numer. Anal. Math. Modelling. 2014. V. 29, N 2. P. 129–138.

August 27, 2015

V. I. Agoshkov
Institute of Computational Mathematics, Moscow, Russia
agoshkov@inm.ru

D. S. Grebennikov; T. O. Sheloput
Moscow Physics and Technology Institute, Dolgoprudnyĭ, Russia
dmitry.ew@gmail.com; sheloput@phystech.edu

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 539.3

SIMULATION OF A STRESS–STRAIN STATE IN LAYERED ORTHOTROPIC PLATES

Yu. M. Volchkov and E. N. Poltavskaya

Abstract. Using the modified equations of the elastic layer, we derive some equations of layered orthotropic plates. Numerical simulation is fulfilled of a stress-strain state in single-layer, two-layer, and three-layer plates. Comparison is given of the numerical and analytical solutions.

Keywords: layered orthotropic plates, stress-strain, numerical solution

Introduction. Reducing the three-dimensional elasticity problem to a two-dimensional problem (theory of shells), we either use hypotheses of kinematic and dynamic character [1] or expand solutions to elasticity equations in some complete system of functions [2–6]. The hypotheses of kinematic and dynamic character impose quite strong restrictions on the stress-strain state and thus, as rule, are invoked to construct theory-of-shells equations in the case that stress is given on the front faces of the shell. Solving the contact problems that are based on these equations often leads to nonphysical effects. Applying expansions of solutions to elasticity equations in some system of functions, we can construct the equations of shells in various approximations. Furthermore, one of the main questions is as follows: Which additional assumptions does this or that approximation rely on, namely, how many terms in the expansion should we keep to construct approximation? Since Legendre polynomials constitute a complete system of functions in the $L_2[-1,1]$ space, precisely this system of functions is often used to construct equations of the theory of shells.

Basing on [4–13], we construct the differential equations of layered orthotropic shells.

1. Equations of the planar elasticity problem. Express the equations of the problem in the rectangular Cartesian coordinates x_1, x_2, x_3. Below the indices 1, 2, and 3 correspond to the coordinates x_1, x_2, x_3.

In the planar problem the required functions are as follows: the stress tensor components σ_{11}, σ_{12}, and σ_{22}, the strain tensor components ε_{11}, ε_{12}, and ε_{22}, and the displacement vector components u_1 and u_2. Put the stress tensor components σ_{13} and σ_{23} and the strain tensor components ε_{13} and ε_{23} equal to zero. All required quantities are functions of the independent variables x_1 and x_2. In the problem of planar stress state put the stress tensor component σ_{33} equal to zero, and find the strain tensor component ε_{33} after solving the problem. Write down the equations of the planar elasticity problem.

Express the equilibrium equations for an infinitely small element as

$$\frac{\partial\sigma_{11}(x_1,x_2)}{\partial x_1}+\frac{\partial\sigma_{12}(x_1,x_2)}{\partial x_2}+f_1(x_1,x_2)=0, \tag{1a}$$

$$\frac{\partial\sigma_{21}(x_1,x_2)}{\partial x_1}+\frac{\partial\sigma_{22}(x_1,x_2)}{\partial x_2}+f_2(x_1,x_2)=0. \tag{1b}$$

Express the strain tensor components in terms of the displacement vector components using Cauchy's relation

$$\varepsilon_{11}=\frac{\partial u_1}{\partial x_1},\quad \varepsilon_{12}=\frac{\partial u_1}{\partial x_2}+\frac{\partial u_2}{\partial x_1},\quad \varepsilon_{22}=\frac{\partial u_2}{\partial x_2}. \tag{2}$$

In this article we study the stress-strain state of plates made of an orthotropic material. A material with three mutually orthogonal elastic symmetry planes is called *orthogonally anisotropic* or *orthotropic*. Orthotropic materials are used in industry; for instance, natural wood, rolled plate concrete, metal, and so on [14]. Some types of composite materials are orthotropic. In particular, carbon plastics are orthotropic materials; these are polymer composite materials with carbon fibers lying symmetrically in a polymer matrix, for instance, in epoxy resin. These materials are firm and rigid, although light. They are tougher than steel, but much lighter. Carbon plastics are widely used in industry because of these properties.

In the case of planar stress state we can write Hooke's law for an orthotropic material as [14]:

$$\sigma_{11}(x_1,x_2)-\alpha_1\left(\frac{\partial u_1(x_1,x_2)}{\partial x_1}+\gamma_2\frac{\partial u_2(x_1,x_2)}{\partial x_2}\right)=0, \tag{3a}$$

$$\sigma_{22}(x_1,x_2)-\alpha_2\left(\frac{\partial u_2(x_1,x_2)}{\partial x_2}+\gamma_1\frac{\partial u_1(x_1,x_2)}{\partial x_1}\right)=0, \tag{3b}$$

$$\sigma_{12}(x_1,x_2)-g_{12}\left(\frac{\partial u_1(x_1,x_2)}{\partial x_2}+\frac{\partial u_2(x_1,x_2)}{\partial x_1}\right), \tag{3c}$$

where

$$\alpha_1=\frac{E_1}{1-\nu_{12}\nu_{21}},\quad \alpha_2=\frac{E_2}{1-\nu_{12}\nu_{21}},\quad \gamma_1=\nu_{12},\quad \gamma_2=\nu_{21}.$$

These relations involve the following independent constants of the material: E_1 and E_2 are the elastic moduli in directions 1 and 2, and ν_{12} is the Poisson coefficient characterizing the transverse compression due to expansion in direction 1. Two more constants appear in the expression for the strain ε_{33} in the case we determine it by solving the planar stress state problem. Consider the boundary value problem in the rectangular region (layer) $\Omega:[0\le x_1\le l,-h\le x_2\le +h]$.

On the boundary of the region we impose the following boundary conditions.

On the lateral surface layer:

$$a_1u_1(0,x_2)+b_1\sigma_{11}(0,x_2)=\varphi_0(x_2),\quad a_2u_2(l,x_2)+b_2\sigma_{21}(l,x_2)=\varphi_l(x_2). \tag{4}$$

On the front surface layer:

$$\begin{aligned} c_1u_1(x_1,\pm h)+d_1\sigma_{12}(x_1,\pm h)&=\varphi_{\pm h}(x_1),\\ c_2u_2(x_1,\pm h)+d_2\sigma_{22}(x_1,\pm h)&=\varphi_{\pm h}(x_1)\end{aligned} \tag{5}$$

Therefore, we pose the following problem: *Find the functions* σ_{11}, σ_{12}, σ_{22}, ε_{11}, ε_{12}, ε_{22}, u_1, and u_2 *satisfying equations* (1a), (1b), (2), (3a)–(3c) *and boundary conditions* (4), (5).

2. Passage to dimensionless variables. Introduce the dimensionless variables

$$\begin{aligned}
&\xi = \frac{x_1}{l}, \quad \zeta = \frac{x2}{h}, \quad (\hat{\sigma}_{11}, \hat{\sigma}_{12}, \hat{\sigma}_{22}) = \left(\frac{\sigma_{11}}{\sigma_0}, \frac{\sigma_{12}}{\sigma_0}, \frac{\sigma_{22}}{\sigma_0}\right),\\
&\hat{u}_1 = \frac{u_1}{h}, \quad \hat{u}_2 = \frac{u_2}{h}, \quad \hat{f}_1 = \frac{f_1 h}{\sigma_0}, \quad \hat{f}_2 = \frac{f_2 h}{\sigma_0}, \quad \eta = \frac{h}{l},
\end{aligned} \tag{6}$$

where σ_0 is some characteristic stress.

Write down the equations of the problem in dimensionless variables omitting $\hat{}$ for simplicity.

The equilibrium equations in dimensionless variables are

$$\eta\frac{\partial\sigma_{11}(\xi,\zeta)}{\partial\xi} + \frac{\partial\sigma_{12}(\xi,\zeta)}{\partial\zeta} + f_1(\xi,\zeta) = 0, \tag{7a}$$

$$\eta\frac{\partial\sigma_{21}(\xi,\zeta)}{\partial\xi} + \frac{\partial\sigma_{22}(\xi,\zeta)}{\partial\zeta} + f_2(\xi,\zeta) = 0. \tag{7b}$$

Hooke's law in dimensionless variables is

$$\sigma_{11}(\xi,\zeta) - \alpha_1\left(\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi} + \gamma_2\frac{\partial u_2(\xi,\zeta)}{\partial\zeta}\right) = 0, \tag{8a}$$

$$\sigma_{22}(\xi,\zeta) - \alpha_2\left(\frac{\partial u_2(\xi,\zeta)}{\partial\zeta} + \gamma_1\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi}\right) = 0, \tag{8b}$$

$$\sigma_{12}(\xi,\zeta) - g_{12}\left(\frac{\partial u_1(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2(\xi,\zeta)}{\partial\xi}\right) = 0 \tag{8c}$$

$$\sigma_{21}(\xi,\zeta) - g_{12}\left(\frac{\partial u_1(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2(\xi,\zeta)}{\partial\xi}\right) = 0. \tag{8d}$$

Since the stress tensor is symmetric, (8c) and (8d) amount to the same relation. However, this expression is useful below.

Cauchy's relation is

$$\varepsilon_{11} = \eta\frac{\partial u_1}{\partial\xi}, \quad \varepsilon_{12} = \frac{\partial u_1}{\partial\zeta} + \eta\frac{\partial u_2}{\partial\xi}, \quad \varepsilon_{22} = \frac{\partial u_2}{\partial\varsigma}. \tag{9}$$

3. Approximating stress and displacement by segments of Legendre polynomials. While constructing the planar layer equations, replace the equilibrium equations (7a)–(7b) for infinitely small element in the directions of x_1 and x_2 and unit width in the direction of x_3 with the equilibrium equations for an infinitely small element in the direction of x_1, finite width $2h$ in the direction of x_2 and unit width in the direction of x_3:

$$\int\limits_{-1}^{1}\left(\eta\frac{\partial\sigma_{11}(\xi,\zeta)}{\partial\xi} + \frac{\partial\sigma_{12}(\xi,\zeta)}{\partial\zeta} + f_1(\xi,\zeta)\right)d\zeta = 0, \tag{10}$$

$$\int\limits_{-1}^{1}\left(\eta\frac{\partial\sigma_{11}(\xi,\zeta)}{\partial\xi} + \frac{\partial\sigma_{12}(\xi,\zeta)}{\partial\zeta} + f_1(\xi,\zeta)\right)\zeta\, d\zeta = 0, \tag{11}$$

$$\int\limits_{-1}^{1}(\eta\frac{\partial\sigma_{21}(\xi,\zeta)}{\partial\xi} + \frac{\partial\sigma_{22}(\xi,\zeta)}{\partial\zeta} + f_2(\xi,\zeta))\, d\zeta = 0. \tag{12}$$

Approximate stress and displacement by segments of Legendre polynomials. According to (10)–(12), stress and mass forces are approximated by the following segments of Legendre polynomials:

$$\sigma_{11}(\xi,\zeta) = t_1(\xi) + m_1(\xi)\mathrm{P}_1(\zeta), \tag{13}$$

$$\sigma_{22}(\xi,\zeta) = t_2(\xi) + m_2(\xi)\mathrm{P}_1(\zeta), \tag{14}$$

$$\sigma_{12}(\xi,\zeta) = t_{12}(\xi) + m_{12}(\xi)\mathrm{P}_1(\zeta) + r_{12}(\xi)\mathrm{P}_2(\zeta), \tag{15}$$

$$\sigma_{21}(\xi,\zeta) = t_{12}(\xi), \tag{16}$$

$$f_1(\xi,\zeta) = q_{10}(\xi) + q_{11}(\xi)\mathrm{P}_1, \quad f_2(\xi,\zeta) = q_{20}(\xi). \tag{18}$$

In (13)–(18) $\mathrm{P}_1(\zeta)$ and $\mathrm{P}_2(\zeta)$ are Legendre polynomials comprising an orthonormal system of functions on the closed interval $[-1,1]$.

The stresses σ_{12} and σ_{21} are approximated by different segments of polynomials because the equilibrium equations involve the derivatives of these functions with respect to different coordinates. This approximation accounts for different variability of the stress-strain states with respect to the spatial coordinates in thin-walled constructions.

Choose approximations for displacement so that the expressions in the parentheses in (8a)–(8d) have the same approximation order with respect to ζ as stress. Therefore, approximate (9) as

$$\varepsilon_{11} = \eta\frac{\partial u_1'(\xi,\zeta)}{\partial\xi}, \quad 2\varepsilon_{12} = \frac{\partial u_1''(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2'(\xi,\zeta)}{\partial\xi}, \quad \varepsilon_{22} = \frac{\partial u_2''(\xi,\zeta)}{\partial\zeta}, \tag{19}$$

where

$$u_1'(\xi,\zeta) = u_1^0(\xi) + u_1^1(\xi)\mathrm{P}_1(\zeta), \tag{20}$$

$$u_1''(\xi,\zeta) = u_1^0(\xi) + u_1^1(\xi)\mathrm{P}_1(\zeta) + u_1^2(\xi)\mathrm{P}_2(\zeta) + u_1^3(\xi)\mathrm{P}_3(\zeta), \tag{21}$$

$$u_2'(\xi,\zeta) = u_2^0(\xi), \tag{22}$$

$$u_2''(\xi,\zeta) = u_2^0(\xi) + u_2^1(\xi)\mathrm{P}_1(\zeta) + u_2^2(\xi)\mathrm{P}_2(\zeta). \tag{23}$$

Use two approximations for each of the displacements u_1 and u_2 because Cauchy's relation involves the derivatives of these functions with respect to both ξ and ζ.

Taking the above approximations for stress and displacement into account, replace (8a)–(8d) with

$$\int_{-1}^{1}\left(\sigma_{11}(\xi,\zeta) - \alpha_1\left(\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi} + \gamma_2\frac{\partial u_2(\xi,\zeta)}{\partial\zeta}\right)\right)\mathrm{P}_0(\zeta)\,d\zeta = 0, \tag{24a}$$

$$\int_{-1}^{1}\left(\sigma_{11}(\xi,\zeta) - \alpha_1\left(\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi} + \gamma_2\frac{\partial u_2(\xi,\zeta)}{\partial\zeta}\right)\right)\mathrm{P}_1(\zeta)\,d\zeta = 0, \tag{24b}$$

$$\int_{-1}^{1}\left(\sigma_{22}(\xi,\zeta) - \alpha_2\left(\frac{\partial u_2(\xi,\zeta)}{\partial\zeta} + \gamma_1\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi}\right)\right)\mathrm{P}_0(\zeta)\,d\zeta = 0, \tag{24c}$$

$$\int_{-1}^{1}\left(\sigma_{22}(\xi,\zeta) - \alpha_2\left(\frac{\partial u_2(\xi,\zeta)}{\partial\zeta} + \gamma_1\eta\frac{\partial u_1(\xi,\zeta)}{\partial\xi}\right)\right)\mathrm{P}_1(\zeta)\,d\zeta = 0, \tag{24d}$$

$$\int_{-1}^{1}\left(\sigma_{12}(\xi,\zeta) - g_{12}\left(\frac{\partial u_1(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2(\xi,\zeta)}{\partial\xi}\right)\right)\mathrm{P}_0(\zeta)\,d\zeta = 0, \tag{24e}$$

$$\int_{-1}^{1}\left(\sigma_{12}(\xi,\zeta) - g_{12}\left(\frac{\partial u_1(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2(\xi,\zeta)}{\partial\xi}\right)\right)\mathrm{P}_1(\zeta)\,d\zeta = 0, \tag{24f}$$

$$\int_{-1}^{1}\left(\sigma_{21}(\xi,\zeta) - g_{12}\left(\frac{\partial u_1(\xi,\zeta)}{\partial\zeta} + \eta\frac{\partial u_2(\xi,\zeta)}{\partial\xi}\right)\right)\mathrm{P}_2(\zeta)\,d\zeta = 0. \tag{24g}$$

The boundary conditions on the front faces (5) for the coefficients of Legendre polynomial segments become

$$c_1\big(u_1^0(\xi) \pm u_1^1(\xi) + u_1^2(\xi) \pm u_1^3(\xi)\big) + d_1(t_{12}(\xi) \pm m_{12}(\xi) + r_{12}(\xi)) = \varphi_{\pm h}(\xi), \tag{25}$$

$$c_2\big(u_2^0(\xi) \pm u_2^1(\xi) + u_2^2(\xi)\big) + d_2(t_2(\xi) \pm m_2(\xi)) = \varphi_{\pm h}(\xi). \tag{26}$$

Equations (10)–(12), (24a)–(24g), (25), and (26) amount to a system of differential and algebraic equations on the coefficients of Legendre polynomial segments for stress and displacement:

$$\eta t_1' + m_{12} + q_{10} = 0, \tag{27a}$$

$$\eta m_1' + 3r_{12} + q_{11} = 0, \tag{27b}$$

$$\eta t_{12}' + m_2 + q_{20} = 0, \tag{27c}$$

$$\alpha_1(\gamma_2 v_1 + \eta u_0') - t_1 = 0, \tag{27d}$$

$$\alpha_1(3\gamma_2 v_2 + \eta u_1') - m_1 = 0, \tag{27e}$$

$$\alpha_2(v_1 + \gamma_1\eta u_0') - t_2 = 0, \tag{27f}$$

$$\alpha_2(3v_2 + \gamma_1\eta u_1') - m_2 = 0, \tag{27g}$$

$$g_{12}(u_1 + u_3 + \eta v_0') - t_{12} = 0, \tag{27h}$$

$$m_{12} - 3g_{12}u_2 = 0, \tag{27i}$$

$$r_{12} - 5g_{12}u_3 = 0, \tag{27j}$$

$$\sigma_{22}^{\pm} = t_2 \pm m_2, \quad u_2^{\pm} = u_2^0 \pm u_2^1 + u_2^2, \tag{27k}$$

$$\sigma_{12}^{\pm} = t_{12} \pm m_{12} + r_{12}, \quad u_1^{\pm} = u_1^0 \pm u_1^1 + u_1^2 \pm u_1^3, \tag{27l}$$

where $\sigma_{22}^{\pm}$, $u_2^{\pm}$, $\sigma_{12}^{\pm}$, and $u_1^{\pm}$ are prescribed functions; the prime indicates derivatives with respect to ξ.

Equations (27a)–(27l) reduce to a system of ordinary differential equations for the functions $u_0(\xi)$, $u_1(\xi)$, $v_0(\xi)$, $t_{11}(\xi)$, $m_{11}(\xi)$, and $t_{12}(\xi)$.

Introducing the vector

$$\mathbf{Z} = [u_0,\ u_1,\ v_0,\ t_{11},\ m_{11},\ t_{12}]^{\mathrm{T}},$$

we can express the system of differential equations of the orthotropic layer in matrix form

$$\mathbf{Z}' = \mathbf{HZ} + \mathbf{F}, \tag{28}$$

where $\mathbf{H}$ is a 6×6 matrix and $\mathbf{F}$ is a vector with six components.

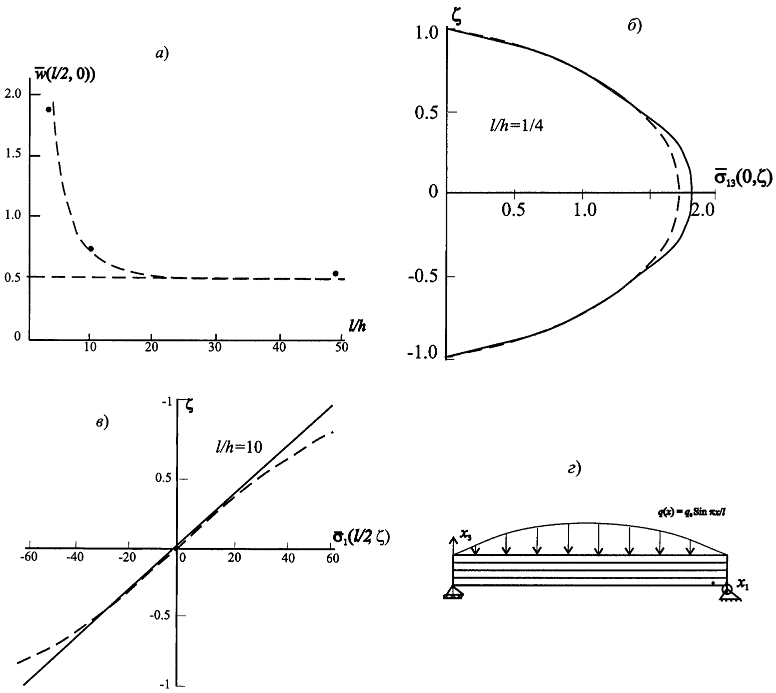

Fig. 1. (a) Dependence of deflection in the middle of the beam on l/h;
(b) distribution of tangent stress at the ends of the cross-section of a beam;
(c) distribution of normal stress in the middle of the cross-section of a beam;
(d) the one-layer beam; solid lines and points show the solution from layer equations,
dashed lines show the analytical solution.

The boundary conditions on the faces of the layer surfaces (4) imply the boundary conditions for $\xi = \xi_0$ and $\xi = \xi_1$ for the (28), which we can express as

$$\mathbf{AX} + \mathbf{BY} = \mathbf{C}, \tag{29}$$

where

$$\mathbf{X} = \begin{Vmatrix} u_0 \\ u_1 \\ v_0 \end{Vmatrix}, \quad \mathbf{Y} = \begin{Vmatrix} t_{11} \\ m_{11} \\ t_{12} \end{Vmatrix},$$

while $\mathbf{A}$ and $\mathbf{B}$ are prescribed 3×3 matrices and $\mathbf{C}$ is a prescribed three-dimensional vector. The matrix $\mathbf{H}$ and the vector $\mathbf{F}$ depend on the form of boundary conditions on the faces of the layer. The components of $\mathbf{Z}$ have the following physical meaning: u_0 is the longitudinal displacement averaged over the thickness of the layer; u_1 is the rotation of the transverse section; v_0 is the transverse displacement averaged over the thickness of the layer; t_1 is the longitudinal force, t_{12} is the shear force, m_1 is the bending moment.

4. Algorithm for calculating the stress-strain states in layered plates. The main advantage of the above equations of an elastic orthotropic layer is that they admit conditions on the faces on both displacement and stress, and the order of the system of differential equations stays the same. This enables us to construct layered

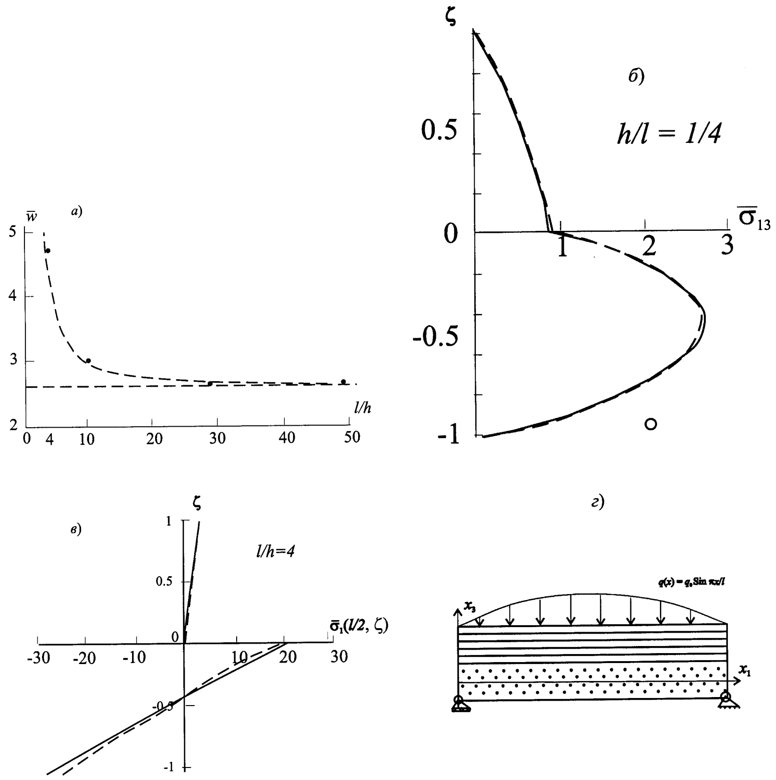

Fig. 2. (a) Dependence of deflection in the middle of the beam on l/h;
(b) distribution of tangent stress at the ends of the cross-section of a beam;
(c) distribution of normal stress in the middle of the cross-section of a beam;
(d) the two-layer beam; solid lines and points show the solution of the layer equations, dashed lines show the analytical solution.

plate equations. In each layer we use (28). We impose the matching conditions on the boundary between layers: the continuity of normal stress and displacement.

For instance, for a 3-layer plate a system of differential equations of order 18 results. The matrices **H** in each layer are different because the conditions on the faces of the layers are of different types. To solve the boundary value problem for the system of differential equations of order 18 we use the orthogonal sweep method.

5. Comparison between numerical and analytical solutions to the problem of stress-strain state of layered orthotropic plate. Analytical solutions to problems of cylindrical deflection of multilayered beams consisting of orthotropic layers are constructed in [15–17]. Consider the problem of cylindrical deflection of a beam with hinged faces under the exterior load $q(x) = q_0 \sin(\pi x/l)$, where q_0 is the load intensity, and L is the length of the beam.

The beam consists of carbon plastic monolayers with the following characteristics (the x-axis coincides with the reinforcement direction): $E_{11} = 1.724 \cdot 10^5$ MPa; $E_{22} = 6895$ MPa; $G_{12} = 3448$ MPa; $G_{23} = 1379$ MPa; $\nu_{12} = 0.25$ MPa.

Fig. 1 depicts the results of calculations for a one-layer beam (the reinforcement direction of the layer coincides with the beam axis x).

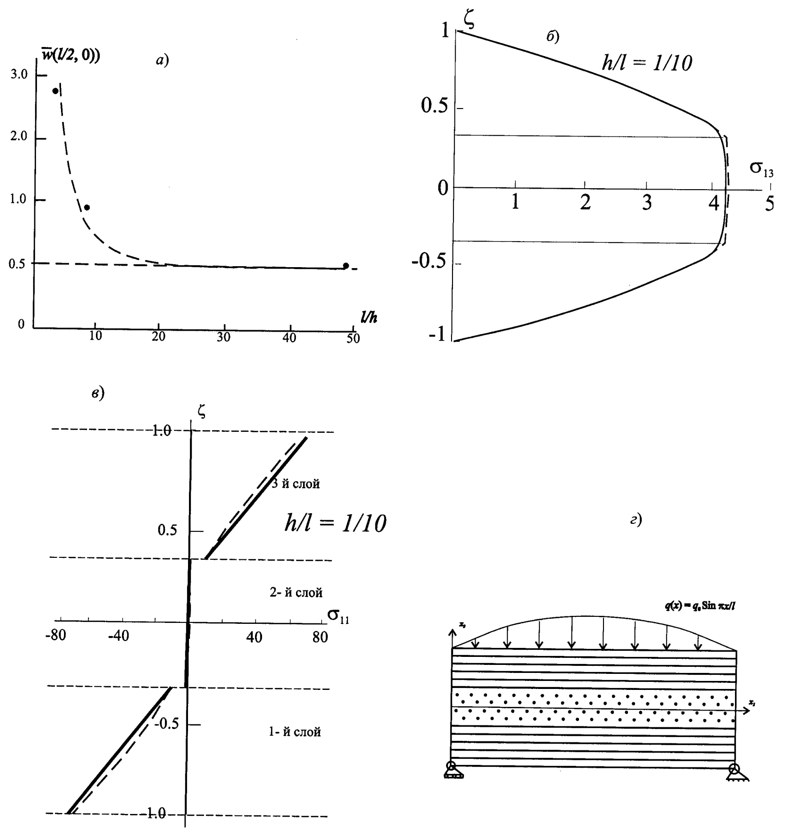

Fig. 3. (a) Dependence of deflection in the middle of the beam on l/h;
(b) distribution of tangent stress at the ends of the cross-section of a beam;
(c) distribution of normal stress in the middle of the cross-section of a beam;
(d) the three-layer beam; solid lines and points show the solution of the layer equations, dashed lines show the analytical solution.

Fig. 2 depicts the results of calculations for a two-layer beam (the reinforcement direction of the first layer coincides with the beam axis x, while the reinforcement direction of the second layer is orthogonal to the beam axis: $E_x^1 = E_{11}$, $G_{xz}^1 = G_{12}$, $E_x^2 = E_{22}$, and $G_{xz}^2 = G_{23}$).

Fig. 3 depicts the results of calculations for a three-layer beam (the reinforcement directions of the first and third layers coincide with the beam axis x, while the reinforcement direction of the second layer is orthogonal to the beam axis: $E_x^k = E_{11}$ and $G_{xz}^k = G_{12}$ for $k = 1, 3$, while $E_x^2 = E_{22}$ and $G_{xz}^2 = G_{23}$).

The load strength is $q_0 = 0,6895\,\mathrm{MPa}$. The figures present the results in the dimensionless variables

$$\widehat{w}(l/2,0) = \frac{100E_{22}h^3\overline{w}(l/2,0)}{q_0 l^4}, \ \hat{u} = \frac{E_{22}\bar{u}(0,z)}{q_0 l^4}, \ \hat{\sigma}_{13} = \frac{\hat{\tau}_{xz}(0,z)}{q_0}, \ \hat{\sigma}_3 = \frac{\hat{\sigma}_z(l/2,z)}{q_0}.$$

For various values of the parameter h/l (where h and l are the thickness and length of the beam) the figures present the distribution of displacements and stress at certain characteristic points and sections of the beam. The maximal error in the calculation of stress on using the above equations is at most 3%.

Conclusion. Using the modified elastic layer equations, we constructed the differential equations of a layered orthotropic plate. We compared the numerical solutions of the stress-strain state of 1-layer, 2-layer, and 3-layer orthotropic beams with the analytical solutions to the corresponding problems. The results of comparison imply that, using the modified elastic layer equations, we can construct the layered plate equations enabling us to determine the stress-strain state in a layered plate with the accuracy sufficient for technical applications.

REFERENCES

1. *Timoshenko S. P. and Woinowsky-Krieger S.* Theory of Plates and Shells. New York: McGraw-Hill, 1959.

2. *Soler A. I.* Higher-order theories for structural analysis using Legendre polynomial expansions // J. Appl. Mech. 1969. V. 36, N 4. P. 757–762.

3. *Ivanov G. V.* Solution of the plane mixed problem of the theory of elasticity in the form of a series in Legendre polynomials // J. Appl. Mech. Tech. Phys. 1976. V. 17, N 6. P. 856–866.

4. *Ivanov G. V.* Theory of Plates and Shells [in Russian]. Novosibirsk: NGU, 1980.

5. *Pelekh B. L. and Laz′ko V. A.* Layered Anisotropic Plates and Shells with Stress Concentrations [in Russian]. Kiev: Naukova Dumka, 1982.

6. *Volchkov Yu. M. and Dergileva L. A.* Solution of problems of an elastic layer on the basis of approximate equations and comparison with the solutions of the elasticity theory // Dynamics of Continuous Media. 1977. N 28. P. 43–54.

7. *Vajeva D. V., Volchkov Yu. M.* The equations for determination of stress-deformed state of multilayered shells // Proc. 9th Russian–Korean Intern. Symp. on Sci. and Technol. Novosibirsk, 26 June–2 July 2005. Novosibirsk: Novosib. State Univ., 2005. P. 547–550.

8. *Volchkov Yu. M.* Finite elements with conjugation conditions on their faces // Dynamics of Continuous Media. 2000. N 116. P. 175–180.

9. *Volchkov Yu. M., Dergileva L. A., and Ivanov G. V.* Numerical simulation of stress states in plane problems of elasticity by the method of layers // J. Appl. Mech. Tech. Phys. 1994. V. 35, N 6. P. 354–359.

10. *Volchkov Yu. M. and Dergileva L. A.* Edge effects in the stress state of a thin elastic interlayer // J. Appl. Mech. Tech. Phys. 1999. V. 40, N 2. P. 354–359.

11. *Alekseev A. E., Alekhin V. V., and Annin B. D.* Plane elastic problem for an inhomogeneous layered body // J. Appl. Mech. Tech. Phys. 2001. V. 42, N 6. P. 1038–1042.

12. *Alekseev A. E. and Annin B. D.* Deformation equations of an inhomogeneous layer elastic body of rotation // Prikl. Mekh. Tekhn. Fiz. 2003. V. 44, N 3. P. 157–163.

13. *Volchkov Yu. M. and Dergileva L. A.* Equations of an elastic anisotropic layer // J. Appl. Mech. Tech. Phys. 2004. V. 45, N 2. P. 301–309.

14. *Ambartsumyan S. A.* Theory of Anisotropic Shells. Washington: NASA, 1964.

15. *Pagano N. J.* Exact solutions for composite laminates in cylindrical bending // J. Composite Materials. 1969. V. 3, N 4. P. 398–409.

16. *Pagano N. J.* Exact solutions for rectangular bidirectional composites and sandwich plates // J. Composite Materials. 1970. V. 4, N 1. P. 20–34.

17. *Pagano N. J.* Elastic behavior of multilayered bidirectional composites // Mechanics of Composite Materials. 1972. V. 10, N 7. P. 931–933.

September 2, 2015

Yu. M. Volchkov; E. N. Poltavskaya
Lavrent′ev Institute of Hydrodynamics, Novosibirsk, Russia
`volk@hydro.nsc.ru`

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 519.622

COMPARISON OF THE GRADIENT AND SIMPLEX METHODS FOR NUMERICAL SOLUTION OF AN INVERSE PROBLEM FOR THE SIMPLEST MODEL OF AN INFECTIOUS DECEASE

S. I. Kabanikhin, O. I. Krivorot′ko, D. V. Ermolenko, and D. A. Voronov

Abstract. The infected human organism releases antibodies that help to cope with deceases. Individual peculiarities of the immunity and the decease which are responsible for the formation of antibodies (for example, viruses or bacteria), resistance of an organism, etc. differ and so does the reaction of each organism with the same decease. Despite this fact, doctors as a rule offer a standard treatment plan which is not always optimal. Hence, it is important to define the individual peculiarities of immunity (the velocity of the immune response or the production of specific antibodies) and those of a decease (the velocity of propagation of viruses and bacteria and so on) for every patient separately by the blood and urina tests, etc.

In the article we study the problem of determining the parameters of an infectious decease in the simplest mathematical model "antigen-antibody" on the measurements of concentrations of antigens and antibodies at fixed times. Some objective functional describing the discrepancy between experimental and model data is examined. We obtain an explicit representation of the gradient of the objective functional with the use a solution to the corresponding adjoint problem. Comparative analysis of a numerical solution to an inverse problem obtained by the gradient method (the Landweber iteration) and the simplex method (the Nelder–Mead method) is exposed. It is demonstrated that the Nelder–Mead method in the model under study defines a collection of local approximate values of the velocities of propagation of the immune response and the production of specific antibodies with a prescribed accuracy. The Landweber iteration calculates the minimizer of the objective functional which is closest to the initial approximation using sufficiently large number of iterations.

Keywords: inverse problem, optimization approach, Landweber iteration, Nelder–Mead method, modeling in immunology

Introduction

The infected human organism releases antibodies that help to cope with deceases. In every particular case the individual peculiarities of the immunity and the decease responsible for the growth of antibodies, resistance of an organism, etc. differ and so does the reaction of every organism to the same decease. Despite this fact, doctors as a rule offer a standard treatment plan which is not always optimal. Hence, it is important to define the individual peculiarities of immunity and those of decease for every patient separately by the blood and urina tests, etc. One of the

The authors were supported by the Russian Foundation for Basic Research (Grant 16–31–00382).

methods for solving this problem is mathematical modeling and numerical solution of an inverse problem.

Mathematical modeling of immunology systems, based on numerical solution of systems of ordinary (generally nonlinear) differential equations, has been actively developed since recently. The immunology models are characterized by their parameters which are coefficients of the differential equations describing the peculiarities of the immunity of a patient, those of a decease, and so on.

Mathematical models of immunology, including numerical solution of direct and inverse problems, were studied by G. I. Marchuk [1], A. A. Romanyukha [2], S. M. Andrew [3], H. W. Engl [4], C. Molina-Paris, G. Lythe [5], and so on. H. T. Banks, S. Hu [6] use direct methods of numerically solving the problem of the least squares with a random distribution of data. G. P. Kuznetsova in [7] employs the method of numerical integration of an inverse problem for the simplest model of the infectious decease which is due to G. I. Marchuk. In [8] the authors exhibit a numerical study of an inverse problem for the simplest mathematical model of iteration of antigens and antibodies by the gradient method. The estimates of convergence of the algorithm are justified and the uniqueness theorem together with local stability are proven. The main aim of this article is to analyze two algorithms of recovering the parameters of the simplest mathematical model which characterizes the character of a decease and the immune response with the use of the blood tests. This problem of recovering parameters is called below *an inverse problem for the simplest immunology model.*

We study a numerical solution to an inverse problem for the simplest model of an infectious decease (the so-called "antigen-antibody" model), consisting of two nonlinear differential equations. This model allows us to describe in details the iteration of antigens and antibodies in an organism. A numerical solution is calculated by the Landweber iteration and the Nelder–Mead method. The articles is organized as follows: In Section 1 we state an inverse problem for the simplest model of an infectious decease. In Sections 2 and 3 the two methods of solving an inverse problem are studied. In particular, in Section 2 the gradient method (the Landweber iteration) is described; and the numerical results, obtained by this method, are exposed. In Section 3 the Nelder–Mead method is studied and the numerical solution of an inverse problem is presented. Section 4 is devoted to a comparative analysis of the Landweber iteration and the Nelder–Mead method.

1. Statement of the Inverse Problem

We study the following Cauchy problem for the simplest model "antigen-antibody" of an infectious decease [8, 9]:

$$\begin{cases} \frac{dN_1(t)}{dt} = N_1(t)(\beta_{11} - \beta_{12}N_2(t)), & t \in (0,T), \\ \frac{dN_2(t)}{dt} = \beta_{21}N_1(t)N_2(t), & t \in (0,T), \\ N_1(0) = N_{10}, \quad N_2(0) = N_{20}, \end{cases} \tag{1}$$

which can be written in vector form

$$\begin{cases} \frac{dN(t)}{dt} = P(N(t),\beta), \quad t \in (0,T), \\ N(0) = N^0. \end{cases} \tag{2}$$

Here $N(t) = (N_1(t), N_2(t))^T$ are the variables of the system (the concentration of antigens and antibodies in an organism), $\beta = (\beta_{11}, \beta_{12}, \beta_{21})^T$ is the vector of parameter characterizing the peculiarities of the immunity, where β_{11} describes the growth of the number of antigens, β_{12} is the velocity of the immune response, β_{21} is the velocity of production of the specific antibodies, and P is a given vector-function.

The problem (2) for given β and N^0 is called the *direct problem*.

Let the concentrations of antigens $N_1(t)$ and antibodies $N_2(t)$ (put $N_i(t) = N_i(t;\beta)$, $i = 1,2$) be measured at fixed times t_k, $k = 1,\dots,K$, i.e.,

$$N_i(t_k;\beta) = \Phi_i(t_k), \quad i = 1,2;\ k = 1,\dots,K. \tag{3}$$

Inverse problem (2), (3) includes the determination of the vector of parameters β with a given function P, the initial data N^0, and the additional information (3). Introduce the operator of the inverse problem (2), (3) as follows: $A : \mathscr{P} \to \mathbb{R}^K$, where $\mathscr{P} := \{\beta \in \mathbb{R}^3 : \beta_{ij} \geq 0,\ i,j = 1,2\}$ is the space of the parameters under consideration.

Rewrite (2), (3) in operator form

$$A(\beta) = \Phi, \quad \Phi = (\Phi_1(t_1),\dots,\Phi_1(t_K),\Phi_2(t_1),\dots,\Phi_2(t_K))^T. \tag{4}$$

The vector Φ is defined, for example, by the blood and the urine tests at t_k, $k = 1,\dots,K$. A solution to (4) is sought by minimizing the objective functional $J(\beta) = \|A(\beta) - \Phi\|^2$ that is defined as

$$J(\beta) = \sum_{k=0}^{K} |N(t_k;\beta) - \Phi(t_k)|^2. \tag{5}$$

This means that a solution to (2) for an optimal β at times t_k, $k = 1,\dots,K$, is closest to the measurements of the states of the system (the concentrations of antigens $N_1(t)$ and antibodies $N_2(t)$) at t_k.

2. Numerical Solution of the Inverse Problem by the Landweber Iteration

We employ the Landweber iteration for solving the problem $\min_{\beta\in\mathscr{P}} J(\beta)$ in which the approximate solution is defined as follows [10, 11]:

$$\beta_{n+1} = \beta_n - \alpha J'(\beta_n), \quad \alpha > 0,\ \beta_0 \in \mathscr{P}, \tag{6}$$

where α is the descent parameter, $J'(\beta) \in \mathbb{R}^3$ is the gradient of the objective functional (5) which is written explicitly [12] as

$$J'(\beta) = -\int_0^T \Psi(t)^T P_\beta(N(t),\beta)\,dt. \tag{7}$$

Here $\Psi(t)$ is a solution to the adjoint problem

$$\begin{cases} \frac{d\Psi(t)}{dt} = -P_N^T(N(t),\beta)\Psi(t), & t \in \bigcup\limits_{k=0}^{K}(t_k,t_{k+1}),\ t_0 = 0,\ t_{K+1} = T, \\ \Psi(T) = 0, \\ [\Psi]_{t=t_k} = 2(N(t_k;\beta) - \Phi(t_k)), & k = 1,\dots,K, \end{cases} \tag{8}$$

where $P_N(N(t),\beta) \in \mathbb{R}^2 \times \mathbb{R}^2$ and $P_\beta(N(t),\beta) \in \mathbb{R}^2 \times \mathbb{R}^3$ are the corresponding Jacobi matrices

$$P_N = \begin{pmatrix} \beta_{11} - \beta_{12}N_2(t) & -\beta_{12}N_1(t) \\ \beta_{21}N_2(t) & \beta_{21}N_1(t) \end{pmatrix},\ P_\beta = \begin{pmatrix} N_1(t) & -N_1(t)N_2(t) & 0 \\ 0 & 0 & N_1(t)N_2(t) \end{pmatrix},$$

$[\Psi]_{t=t_k} := \Psi(t_k + \varepsilon) - \Psi(t_k - \varepsilon)$ is the jump of Ψ at t_k, where $\gamma > 0$ is arbitrarily small.

To solve the direct and adjoint problems numerically, (2) and (8), respectively, we employ the Runge-Kutta method of the forth order of approximation. Construct the uniform grid $\omega := \{t_j = jh_t,\ h_t = T/N_t,\ j \in \overline{0, N_t}\}$. Let the time of modeling T is equal to 4 weeks, $N_t = 100$ is the number of nodes of ω, $\alpha = 0.001$, $\varepsilon_s = 10^{-6}$ is the stopping time parameter for the iteration procedure, $N_0 = (1.8, 1.8)^T$ are the initial data. Choose the vector of parameters $\beta = (0.5, 0.5, 0.6)^T$, describing the immunity of an average man, which is called below an *exact solution* to (2), (3). We use the synthetic data $N_1(t_k, \beta)$ and $N_2(t_k, \beta)$ at times t_k, $k = 1, \ldots, K$ uniformly distributed over the grid ω as the vector of the data Φ.

The algorithm for numerical solution of the inverse problem (2), (3) by the Landweber iteration consists of the following steps:

1. Specify the initial approximation $\beta_0 = (0.1, 0.1, 0.2)^T$ that describes a light form of an infection and solve (2) for a given β_0. Construct the vector $N(t_k; \beta_0)$, $k = 1, \ldots, K$.

2. By induction, show how to compute β_{n+1} on using β_n.

3. Solve (2) for the collection of parameters β_n, i.e., we find $N(t_k; \beta)$, $k = 1, \ldots, K$.

4. If $J(\beta_n) < \varepsilon_s$ then β_n is an approximate solution.

5. If $J(\beta_n) > \varepsilon_s$ then we solve (8) with $\beta = \beta_n$.

6. Determine $J'(\beta_n)$ from (7).

7. Calculate β_{n+1} in accord with (6).

Let us study the discrepancy $|N_i(t; \beta_n) - N_i(t; \beta)|$ of a calculated curve $N_i(t; \beta_n)$ and "experimental" $N_i(t; \beta)$ in dependence on the number of measurements K, $i = 1, 2$. It is displayed in Fig. 1 that the discrepancy of a calculated curve $N_i(t; \beta_n)$ and an experimental $N_i(t; \beta)$ decreases with the growth of the number of measurements. However it is a problem to make 35 measurements for 4 weeks (i.e., to make 35 tests). In accord with [2] we can choose a‘ maximal measure of this discrepancy equal $7 \cdot 10^{-4}$, i.e. $|N_i(t; \beta_n) - N_i(t; \beta)| < 7 \cdot 10^{-4}$. Thus, we make 20 measurements for 4 weeks in calculations below.

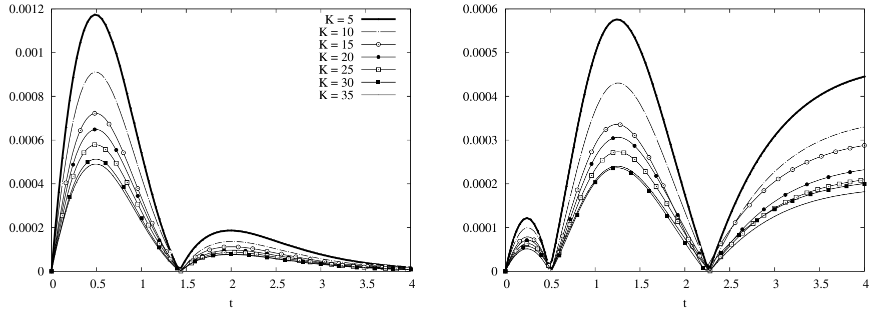

Fig. 1. The graphs $|N_i(t; \beta_n - N_i(t; \beta)|$ in dependence on the number of measurements K: $i = 1$ on the left and $i = 2$ on the right

We now inspect the relative error $|\beta_n - \beta|/|\beta|$ of computations of a solution to the inverse problem which is a dimensionless quantity equal to the ratio of the absolute error and an exact solution to the inverse problem. Fig. 2 shows that the relative error decreases with the growth of the number of measurements, i.e., an approximate solution to the inverse problem approaches an exact solution. Observe the connections between the quality measures of a solution to the inverse problem, namely, the discrepancy between "experimental" and calculated curves $|N_i(t;\beta_n) - N_i(t;\beta)|$ (see Fig. 1), and the relative accuracy $|\beta_n - \beta|/|\beta|$ (see Fig. 2). In what follows, we use the dependence of the relative error on the number of measurements K as the criterion of an optimal number of measurements for a numerical solution of (2), (3).

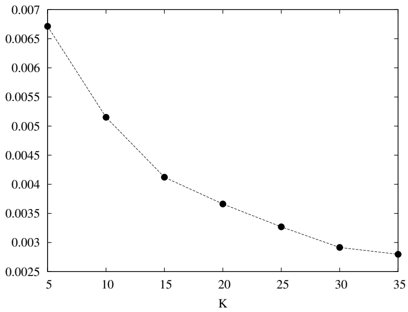

Fig. 2. The graph of the dependence of of the relative error $|\beta_n - \beta|/|\beta|$ on the number of measurements K

The graphs of the dependence of the objective functional $J(\beta_n)$ on the number of iterations n and the absolute error $|\beta_n - \beta|$ are displayed in Fig. 3. We can see in Fig. 3 on the left that for the first two iterations the functional growth rapidly (due to the weak stability of (2), (3)) and beginning with the third iteration decreases monotonically with the velocity $1/n$, the latter shows the convergence of the method. Note that the absolute errors $|\beta_n - \beta|$ decrease monotonically. It is shown in Fig. 3 on the right that the larger absolute error $|\beta_n - \beta|$ (on the first iterations) ensures the larger discrepancy of the model and "experimental" data $J(\beta_n)$.

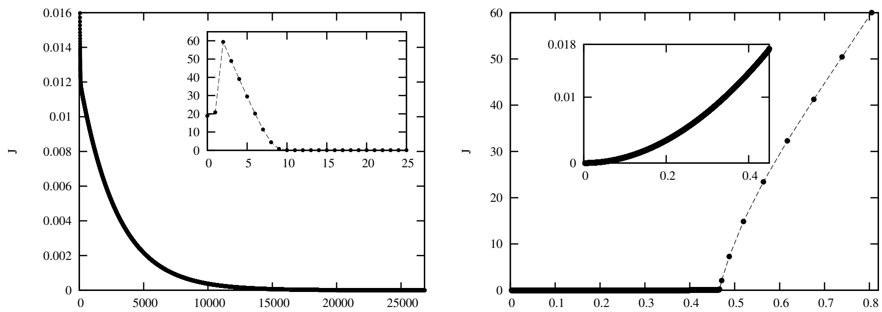

Fig. 3. The graph of $J(\beta_n)$ for the general number of iterations $n = 26836$, $K = 20$ (on the left). The graph of the dependence $J(\beta_n)$ on the absolute error $|\beta_n - \beta|$ for the number of iterations $n = 26836$, $K = 20$ (on the right)

The results of numerical solution of the inverse problem (2), (3) for $K = 20$ are displayed in Fig. 4. Note that we obtain the numerical solution $\beta_{11}^n = 0.49675$, $\beta_{12}^n = 0.49903$, and $\beta_{21}^n = 0.60017$ for 26836 iterations.

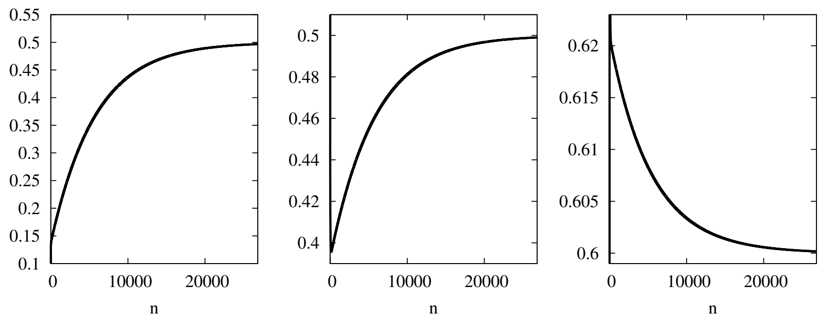

Fig. 4. The graphs of the parameters β_{11}^n (on the left), β_{12}^n (in the center), and β_{21}^n (on the right) in dependence on the number of iterations, $n = 26836$

Note also that for each initial approximation the Landweber iteration method converges to a normal solution to the inverse problem (2), (3) [13]. In the cases $\alpha = 10^{-4}$ and $\alpha = 10^{-5}$ a numerical solution to the inverse problem (2), (3) is very close to the above results and the execution time is essentially larger than in the case of $\alpha = 10^{-3}$.

3. Numerical Solution of the Inverse Problem by the Nelder–Mead Method

The Nelder–Mead method (simplex method) [14] is a method of unconditional optimization of a functional which does not use the gradient. In view of this fact the Nelder–Mead method is easily applicable to noisy and nonsmooth functions. The method consists of a consecutive transmission and deforming an initial approximation (a simplex) around the extremum point. The Nelder–Mead method is widely used for refining parameters in the problems of pharmacokinetics and immunology. The main problem of the method is that it defines a local extremum, while being sensible to the choice of an initial approximation.

A solution to the inverse problem as well as in the case of the gradient method is sought by minimizing the objective functional (5). Thereby, we need to determine an unconditional minimum of the function $J(\beta_{11}, \beta_{12}, \beta_{21})$ of three variables.

In this section we expose the results of numerical calculations obtained by the Nelder–Mead method with the same model parameters as those in the case of the Landweber iteration (the choice of the grid ω, the initial conditions of the direct problem (2), the exact vector of parameters β, and the stopping time ε_s).

For the initial simplex

$$\beta_1 = (0.1, 0.1, 0.2)^T,\ \beta_2 = (0.4, 0.7, 0.8)^T,\ \beta_3 = (0.2, 0.3, 0.4)^T,\ \beta_4 = (0.9, 0.7, 0.1)^T,$$

a numerical solution to (2), (3) by the Nelder–Mead method for 10 measurements (the case of the least relative error) is as follows: β_{11}^n converges to 0.17, β_{12}^n to 0.4, and β_{21}^n to 0.62. We can note that the difference between approximate and exact solutions is sufficiently large. Hence, the Nelder–Mead method for this initial simplex defines a local minimum.

Now we expose a similar calculations for the initial simplex $\beta_1 = (0.15, 0.2, 0.35)^T$, $\beta_2 = (0.05, 0.1, 0.3)^T$, $\beta_3 = (0.05, 0.1, 0.3)^T$, and $\beta_4 = (0.3, 0.3, 0.2)^T$.

The graph of the dependence of the relative error $|\beta_n - \beta|/|\beta|$ on the number of measurements K is displayed in Fig. 5 on the left. We can see that for 25 measurements the relative error is the least but it is a problem to make 25 measurements for 4 weeks. Hence, we take $K = 20$ below, as in Section 2. The Nelder–Mead method converges for a given choice of parameters, i.e., the functional $J(\beta_n)$ vanishes (see Fig. 5 on the right).

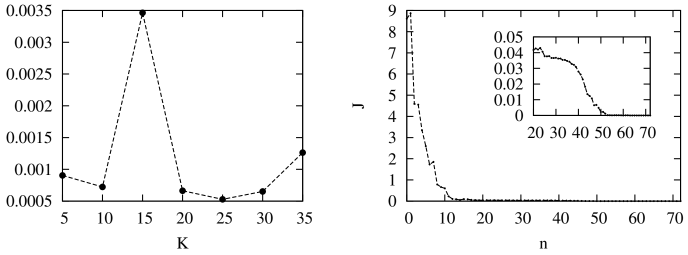

Fig. 5. The graph of the dependence of the relative error $|\beta_n - \beta|/|\beta|$ on the number of measurements K (on the left). The graph of $J(\beta_n)$ with the number of iterations $n = 72$ (on the right)

A solution to the inverse problem in dependence on the number n of iterations is displayed in Fig. 6. Note that the results obtained are close to an exact solution $\beta = (0.5, 0.5, 0.6)^T$. Hence, for a given initial simplex $\beta_1 = (0.15, 0.2, 0.35)^T$, $\beta_2 = (0.05, 0.1, 0.3)^T$, $\beta_3 = (0.05, 0.1, 0.3)^T$, and $\beta_4 = (0.3, 0.3, 0.2)^T$, we find a minimum of the functional J.

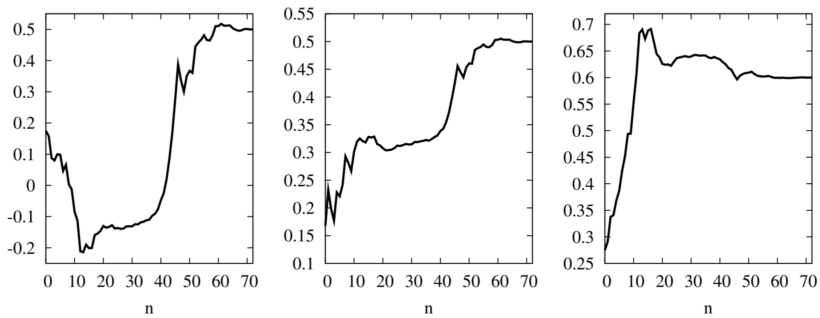

Fig. 6. The graphs of the parameters β_{11}^n (on the left), β_{12}^n (in the center), β_{21}^n (on the right) in dependence the number of iterations $n = 72$

These examples shows that the results, obtained by the Nelder–Mead method, depend on an initial approximation. In dependence of an initial simplex we can obtain a local or global minimum. It is the main problem of this method. The Nelder–Mead method with real data does not ensure determining a global minimum of the objective functional. This problem can be solved by stochastic methods such as the Monte-Carlo method [15], the genetic algorithm [16], etc.

4. Comparative Analysis of the Nelder–Mead Method and the Landweber Interation

Table 1 contains an analysis of a numerical solution to the inverse problem (2), (3) which is obtained by the Nelder–Mead method and the Landweber iteration for 20 measurements. As is easily seen, the Nelder–Mead method allows us to find an answer quicker that the gradient method. The relative error for the Nelder–Mead method is significantly less that in the Landweber iteration. The Landweber iteration converges to a normal solution to (2), (3) with respect to an initial approximation in any case [13]. However, the computation time exceeds that for the Nelder-Mead method by several times in view of the complexity of the algorithm (the computation of the gradient of the objective functional).

Table 1. Analysis of the Nelder–Mead method and the Landweber iteration in solving the inverse problem (2), (3)

	the Nelder–Mead method	The Landweber iteration
K, the number of measurements	20	20
an initial approximation	$\beta^{(1)} = (0.15, 0.2, 0.35)^T$, $\beta^{(2)} = (0.05, 0.1, 0.3)^T$, $\beta^{(3)} = (0.2, 0.07, 0.25)^T$, $\beta^{(4)} = (0.3, 0.3, 0.2)^T$	$\beta_0 = (0.1, 0.1, 0.2)^T$
ε_s, the stopping parameter	10^{-6}	10^{-6}
β^n_{11}	0.50059584	0.49675121
β^n_{12}	0.50013173	0.49902571
β^n_{21}	0.59992131	0.60017153
$\lvert\beta_n - \beta\rvert/\lvert\beta\rvert$, the relative error	0.00066348	0.00366208
n, the number of iterations	72	26836
time of fulfillment of the algorithm (sec.)	0.013	4.882

5. Conclusion

In the article we study a numerical solution to an inverse problem for the simplest mathematical model “antigen-antibody” obtained by the Landweber iteration and the Nelder-Mead method. In the numerical experiments we demonstrate that the Nelder–Mead method determines the set of local velocity approximations of the antigen propagation, the immune response, and the production of specific antibodies with a prescribed accuracy. The Landweber iteration finds the minimizer of the objective functional closest to an initial approximation. Thus, we have constructed a numerical algorithm that allows us to refine parameters of the simplest mathematical model (the velocities of the antigen propagations, the immune response, and the production of the specific antibodies) with 20 measurements of the concentrations of antigens and antibodies for 4 weeks (one for 5 c) on a computer with the processor Intel (R) Core (TM) i3 2.13GHz and RAM 4 gb.

REFERENCES

1. *Marchuk G. I.* Mathematical Models in Immunology [in Russian]. Moscow: Nauka, 1980.
2. *Romanyukha A. A., Rudnev S. G., and Zuev S. M.* Analysis of data and the modeling of infectious deceases // Modern Problems of Numerical Mathematics and Mathematical Modeling. V. 2. Mathematical Modeling (Ed. V. P. Dymnikov) [in Russian]. Moscow: Nauka, 2005. P. 352–404.
3. *Andrew S. M., Baker C. T. H., and Bocharov G. A.* Rival approaches to mathematical modeling in immunology // J. Comput. Appl. Math. 2007. V. 205. P. 669–686.
4. *Engl H. W., Flamm C., Kügler P., Lu J., Müller S., and Schuster P.* Inverse problems in systems biology // Inverse Probl. 2009. V. 25. P. 123014.
5. *Molina-Paris C. and Lythe G.* Mathematical Models and Immune Cell Biology. New York; Dordrecht; Heidelberg; London: Springer-Verlag, 2011.
6. *Banks H. T., Hu Sh., and Clayton Thompson W.* Modeling and Inverse Problems in the Presence of Uncertainty. Boca Raton; London; New York; Washington, DC: CRC Press, 2014 (Monogr. Res. Notes Math.).
7. *Kuznetsova G. P.,* The inverse problem for the Marchuk immunologic "simplest model" // Dal′nevost. Mat. Zh. 2003. V. 4, No. 1. P. 134–140.
8. *Afraites L. and Atlas A.* Parameters identification in the mathematical model of immune competition cells // J. Inverse Ill-posed Probl. 2015. V. 23, N 4. P. 323–337.
9. *Bellouquid A. and Delitala M.* Modeling Complex Biological Systems: A Kinetic Theory Approach. Boston; Basel; Berlin: Birkhäuser, 2006.
10. *Alifanov O. M., Artyukhin E. A., and Rumyantsev S. V.* Optimal Methods for Solving Ill-Posed Problems [in Russian]. Moscow: Nauka, 1988.
11. *Kabanikhin S. I.* Inverse and Ill-Posed Problems. Theory and Applications. Berlin; Boston: Walter de Gruyter GmbH & Co. KG, 2012.
12. *Il′in A. I., Kabanikhin S. I., and Krivorot′ko O. I.* Determining the parameters of models that describe by the systems of nonlinear differential equations // Sib. Elektron. Mat. Izv. 2014. V. 11. P. 1–14.
13. *Vasin V. V.* On convergence of gradient type methods for nonlinear equations // Dokl. Akad. Nauk. 1998. V. 359, N 1. P. 7–9.
14. *Nelder J., Mead R.* A simplex method for function minimization // Comput. J. 1965. V. 7. P. 308–313.
15. *Mikhaĭlov G. A. and Voĭtishek A. V.* Numerical Statistic Modeling. The Monte Carlo Methods [in Russian]. Moscow: Akademiya, 2006.
16. *Gladkov L. A., Kureichik V. V., and Kureichik V. M.* Genetic Algorithms [in Russian]. Moscow: Fizmatlit, 2006.

August 28, 2015

S. I. Kabanikhin; O. I. Krivorot′ko; D. V. Ermolenko; D. A. Voronov
Institute of Computational Mathematics and Mathematical Geophysics
Novosibirsk State University, Novosibirsk, Russia
`ksi52@mail.ru; krivorotko.olya@mail.ru;`
`ermolenko.dasha@mail.ru; dmitriy.voronov@gmail.com`

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 51.7

THE 3D FLOW PROBLEM FOR AN AIRCRAFT MODEL WITH ACTIVE INFLUENCE ON THE FLOW

A. E. Lutskiĭ and Ya. V. Khankhasaeva

Abstract. In the frame of the 3D URANS equations with the Spalart–Allmaras (SA) turbulence model, numerical simulation was conducted of the energy input into the stream in front of an aircraft model with an angle of attack. For the regimes considered it was shown that the energy input before the bow results in a significant reduction of wave resistance and increase in lift. This ensures high efficiency of energy input as a mean of increasing the aerodynamic quality of an aircraft. The effect of the energy input in front of the wings has been studied.

Keywords: computational fluid dynamics, energy input, drag reduction

Introduction

One of the methods for improving the aerodynamic characteristics of prospective aircraft is a controlled action on the oncoming flow. It can be performed in various ways, in particular by using energy input localized in a small closed region. The possibility of remote energy input into a supersonic flow is confirmed in many experiments [1–6]. The high-temperature trace with reduced values of Mach number, total pressure, and impact pressure is formed behind the energy source, which enables us to vary the flow regime. If the energy source and body are of comparable sizes then the flow around the body is quasi-uniform and drag can be reduced by changing straightforwardly the parameters of the oncoming flow. This oncoming flow requires large energy expense and is impractical. However, energy input even in a relatively small space region can lead to realignment of the bow shock-wave structures ahead of the body. The possibility of controlling the airflow around bodies by using a relatively small action on the oncoming flow rests in particular on the well-known nonuniqueness of solution to the problem of flow around a body in classical fluid dynamics [7]. For every blunt body, along with a solution with a detached shock wave, infinitely many solutions are formally possible with a front cone filled with a gas at rest and constant pressure. As a rule, the solution with a detached shock wave is realized in experiments and simulations. However, it is known [8] that the presence of a thin needle protruding ahead of the nose of a blunt body leads to the formation of a cone-shaped region of backward flow. Energy input into the oncoming flow ahead of the nose can create a similar effect.

Much theoretical and experimental work has been done (see [9–13] for instance) to decrease the wave drag of bodies, mostly in the axially symmetric situation. It is shown that energy input into the flow ahead of the nose enables us to decrease drag by a factor of 10 or more due to the formation of a cone-shaped detached region

The authors were supported by the Russian Foundation for Basic Research (Grants 13–01–12043 and 14–08–00624).

ahead of the nose. This method of drag reduction is rather efficient. Power expenses on this energetic influence are substantially smaller than the gain in propulsion power from lower drag. Three-dimensional effects of energy input, in particular the impact of the angle of attack, are studied much less [14, 15].

This article pays most attention to studying the influence of energy input, while solving the problem of flow around a model aircraft in the three-dimensional setting.

Statement of the Problem and Results

We consider the questions of flow realignment as a result of energetic influence on the flow by the example of flow around an ideal model aircraft (Fig. 1).

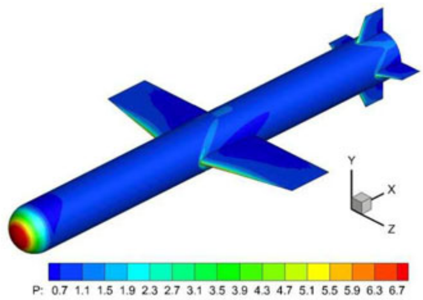

Fig. 1. Model aircraft. Pressure distribution for unperturbed flow

Below we present the results that are obtained in the framework of the mathematical model of the averaged Navier–Stokes equations for a viscous compressible gas with the Spalart–Allmaras turbulence model complemented with a source term in the conservation-of-energy equation:

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z} = S, \quad S = (0,0,0,0,q)^T, \; q = q(x,y,z,t),$$

$$U = (\rho, \rho u, \rho v, \rho w, e)^T, \quad F = F^i + F^v, \quad G = G^i + G^v, \quad H = H^i + H^v,$$

$$F^i = (\rho u, \rho u^2 + p, \rho uv, \rho uw, (e+p)u)^T,$$

$$F^v = (0, -\tau_{xx}, -\tau_{xy}, -\tau_{xz}, -u\tau_{xx} - v\tau_{xy} - w\tau_{xz} - q_x)^T,$$

$$G^i = (\rho v, \rho uv, \rho v^2 + p, \rho vw, (e+p)v)^T,$$

$$G^v = (0, -\tau_{xy}, -\tau_{yy}, -\tau_{zy}, -u\tau_{xy} - v\tau_{yy} - w\tau_{zy} - q_y)^T,$$

$$H^i = (\rho w, \rho uw, \rho vw, \rho w^2 + p, (e+p)w)^T,$$

$$H^v = (0, -\tau_{xz}, -\tau_{yz}, -\tau_{zz}, -u\tau_{xz} - v\tau_{yz} - w\tau_{zz} - q_z)^T,$$

$$e = \rho\varepsilon + \frac{\rho(u^2+v^2+w^2)}{2} = \frac{p}{\gamma-1} + \frac{\rho(u^2+v^2+w^2)}{2}.$$

The components of viscous stress tensor are defined as

$$\tau_{xx} = \frac{2}{3}(\mu+\mu_t)(2u_x - v_y - w_z), \quad \tau_{yy} = \frac{2}{3}(\mu+\mu_t)(2v_y - u_x - w_z),$$

$$\tau_{zz} = \frac{2}{3}(\mu+\mu_t)(2w_z - u_x - v_y), \quad \tau_{xy} = \tau_{yx} = (\mu+\mu_t)(u_y + v_x),$$

$$\tau_{xz} = \tau_{zx} = (\mu+\mu_t)(u_z + w_x), \quad \tau_{yz} = \tau_{zy} = (\mu+\mu_t)(v_z + w_y).$$

Firstly, we consider a version with energy input ahead of the nose of the model. Let us present some results of calculations for the Mach number of the oncoming flow $\mathrm{M} = 2.5$ at the angle $\alpha = 3°$. Pressure and density are relative to these quantities in the oncoming flow, and the diameter of the model is taken as the unit of length. The total power of energy input Q is relative to the power $N = F_x U$ necessary to overcome drag for the unperturbed flow. We assume that energy input is stationary and spatially homogeneous in some region on the symmetry axis of the model.

Table 1. The variants considered

Variants	1	2	3	4	5	6
Q	0	$6.3\%N$	$6.3\%N$	$6.3\%N$	$12.6\%N$	$12.6\%N$
Transverse size		0.2	0.1	0.1	0.2	0.1
Longitudinal size		0.2	0.2	0.2	0.2	0.2
Distance from the model		1.2	1.2	0.5	1.2	0.5

We considered the following variants of size, location of the energy input region, and the quantity of energy put in (Table 1): in variant 1 energy input is absent; in variants 3, 4, and 6 the transverse size is four times smaller than in variants 2 and 5; variant 4 lies two times closer to the body than all others; in variants 5 and 6 energy input is twice as large as in all others.

Consider some simulations results for the Mach number $\mathrm{M} = 2.5$ of the oncoming flow at the angle $\alpha = 3°$. Energy input ahead of the body substantially changes the flow structure. Shock waves issue from the energy input region. The front of the bow shock wave is destroyed by the trace formed behind the energy input region. In the space between the energy input region and the nose there is formed a region with lower pressure in comparison with the oncoming flow. This is illustrated in Figs. 3 and 4.

We observe an interesting effect. The drag in variant 4 is smaller than that in variant 3. They differ only in the distance between the energy input region and the model. The origin of this effect might be that for the closer location of the region the shock waves issuing from the region of energy input interact with the leftover bow shock wave, creating a region where trace concentrates (Fig. 4). The differences also consist in the formation of a backward flow region.

The formation of a backward flow region ahead of the body is an important feature of the flow. We observe this region only for those variants (Fig. 4, var. 4) in which a zone of a positive pressure gradient is formed near the body under the action of thermal trace. In the presence of an angle of attack the thermal trace lies along the velocity vector of the oncoming flow. For certain distance of the energy input region from the model the trace does not enter the deceleration region and pressure decreases monotonely along the current lines issuing from the point of deceleration. Fig. 5 shows that we observe the positive pressure gradient precisely for the variants with a backward flow region.

It is clear also that on the downwind side ($y > 0$) pressure is lower in the case of the energy input than for the unperturbed flow. This fact explains the increase of lift.

Table 2 presents our results on decreasing drag and increasing lift. Here Eff stands for the energy input efficiency: $\mathrm{Eff} = (N(0) - N(Q))/Q$, where $N = F_x U$.

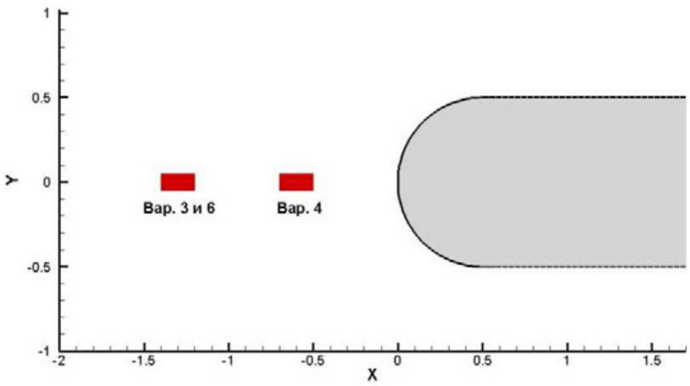

Fig. 2. Location of the energy input region

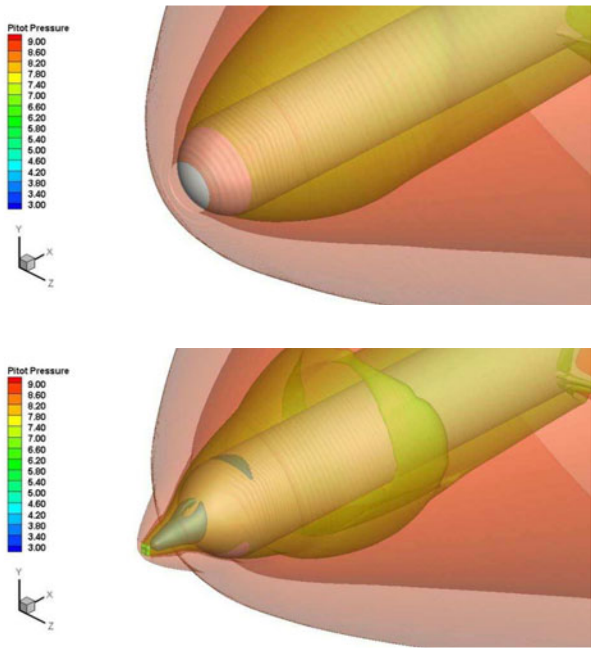

Fig. 3. Total pressure isosurfaces behind the shock wave
without energy input (on the left) and for variant 4 (on the right)

For energy input ahead of the body we observe decrease in drag (row 3 of Table 2) and increase in lift (row 4). Even though drag in variants 5 and 6 is decreased more for the double power of energy input, variant 4 with the energy input region lying closer than in other variants is better from the viewpoints of both energy efficiency and lift.

Table 3 presents the results for various angles of attack. As the angle of attack increases (for fixed Mach number) we observe some decrease in energy input efficiency. In addition, increasing the angle of attack also increases lift. Furthermore,

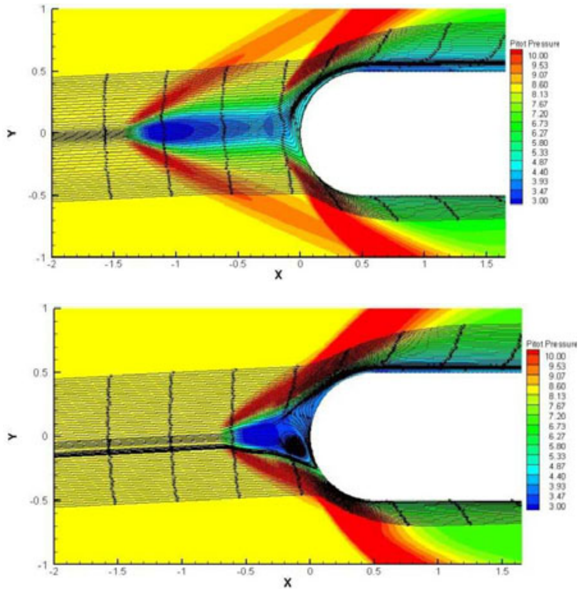

Fig. 4. Total pressure distribution behind the shock wave
for variants 3 (on the left) and 4 (on the right)
for the angle of attack $\alpha = 3^{\circ}$ on the cross-section $z = 0$

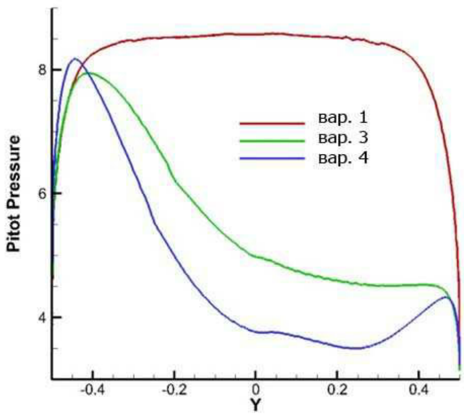

Fig. 5. Total pressure behind the shock wave
for variants 1 (red), 3 (green) and 4 (blue)

Table 2. Drag, lift and energy efficiency
coefficient for different variants

Variants	1	2	3	4	5	6
Eff		1.82592	2.16932	2.75823	1.50756	1.81972
$\Delta C_x/C_{x0}$		−10.70%	−12.72%	−16.17%	−17.67%	−21.33%
$\Delta C_y/C_{y0}$		+1.66%	+2.41%	+3.93%	+2.49%	+3.44%

Table 3. Drag, lift and energy efficiency coefficient for different angles of attack

	Варианты	1	3	4
$\alpha = 1.5°$	C_x	1.750284	−14.23%	−17.81%
	C_y	0.636395	+3.46%	+4.90%
	Eff		2.43158	3.04412
$\alpha = 3°$	C_x	1.749550	−12.72%	−16.17%
	C_y	1.333832	+2.41%	+3.93%
	Eff		2.16932	2.75823
$\alpha = 5°$	C_x	1.747118	−10.86%	−13.59%
	C_y	2.272646	+1.62%	+2.85%
	Eff		1.85207	2.31777

the greater the angle of attack, the smaller the influence of energy input (the decrease in drag and the increase in lift become smaller as the angle of attack decreases).

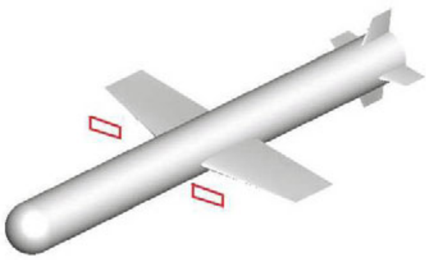

Fig. 6. Location of energy sources ahead of the wings.

We also made calculations with energy input into the flow ahead of the wings, which in our model have blunt front edge. The sizes of the region are $Lx = 0.1$, $Ly = 0.02$, and $Lz = 0.5$; the variants of location are:

(1) distance to the wing 0.28, in the wing plane;

(2) distance to the wing 0.28, by 0.005 above the wing plane;

(3) distance to the wing 0.18, in the wing plane.

We obtain some increase in lift with small decrease in total drag. Variant 3 with the closest location of energy sources to the wings is the most efficient among those considered, which is not surprising because the mechanism of influence is similar to the case of the energy input ahead of the bow, as the wings have blunt front edge. Decrease in drag and increase in lift are not so great here since the wings are small as compared to the fuselage and the hull itself generates lift.

Conclusion

We studied the influence of energy input on the aerodynamic characteristics of a model aircraft for various angles of attack.

1. We showed that energy input ahead of the bow leads to a substantial decrease in drag and increase in lift.

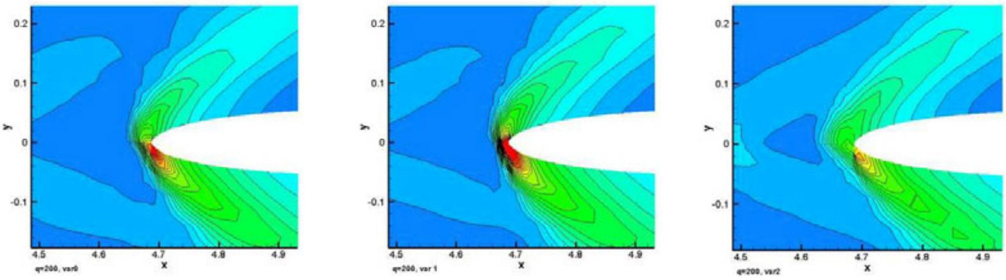

Fig. 7. Pressure ahead of the front edge of the wing. Cross-section at $z = 1.25$. Variants 1–3 from left to right

Table 4. Drag, lift, aerodynamic quality and energy efficiency coefficient for various variants of location

Variant	ΔC_x, %	ΔC_y, %	ΔK, %	Eff
1	-1.81	$+1.22$	$+2.81$	0.38
2	-1.17	$+1.03$	$+2.03$	0.24
3	-2.32	$+0.80$	$+2.89$	0.48

2. Among the variants under consideration we determined the size and location of the region ensuring the most efficient energy input in terms of decreasing drag and increasing lift. We showed in particular that energy input efficiency increases with the region approaches the model and its transverse size.

3. As the angle of attack increases (for a fixed Mach number), we observe some decrease in the efficiency of energy input.

4. Energy input ahead of the wings in the case of the model considered leads to insignificant decrease in drag and increase in lift.

This work shows that the energy input flow has many effects on the spatial flow around an aircraft. Thorough examination of these effects, aiming at further improvements of aerodynamic characteristics of aircraft, will be a subject of further research.

REFERENCES

1. *Mishin G. I., Klimov A. I., and Gridin A. Yu.* Longitudinal electric discharge in supersonic gas flow // Pisma Zh. Tekhn. Fiz. 1992. V. 18, N 15. P. 86–92.
2. *Fomin V. M., Lebedev A. V., and Ivanchenko A. I.* Space-energy characteristics of electric discharge in supersonic gas flow // Doklady Physics. 1998. V. 43, N 7. P. 440–443.
3. *Grachev L. P., Esakov I. I., and Khodataev K. V.* Streamer SVCh discharge in supersonic air flow // Zh. Tekhn. Fiz. 1999. V. 69, N 11. P. 14–18.
4. *Kolesnichenko Yu. F., Brovkin V. G., Azarova O. A., Grudnitsky V. G., Lashkov V. A., and Mashek I. Ch.* MW energy deposition for aerodynamic application // 41st Aerospace Sci. Meeting and Exhibit (Reno, NV, Jan. 6–9, 2003). : AIAA, 2003. P. 1–11.
5. *Tret′yakov P. K., Garanin A. F., Grachev G. N., Krainev V. L., Ponomarenko A. G., Ivanchenko A. I., and Yakovlev V. I.* Control of supersonic flow around bodies by means of high-power recurrent optical background // Doklady Physics. 1996. V. 41, N 3. P. 566–567.
6. *Leonov S. B, Bityurin V. A., Yuriev A., Pirogov S., and Zhukov B.* Problems in energetic method of drag reduction and flow/flight control // 41st Aerospace Sci. Meeting and Exhibit (Reno, NV, Jan. 6–9, 2003). : AIAA, 2003. P. 1–8.

7. *Chernyĭ G. G.* Gas Dynamics [in Russian]. Moscow: Nauka, 1988.
8. *Chang P. K.* Control of Flow Separation: Energy Conservation, Operational Efficiency, and Safety. New York etc.: McGraw-Hill Book Co., 1976.
9. *Georgievskii P. Yu. and Levin V. A.* Supersonic flow past bodies in the presence of external heat release sources // Pisma Zh. Tekhn. Fiz. 1988. V. 14, N 8. P. 684–687.
10. *Georgievskii P. Yu. and Levin V. A.* Control of the flow past bodies using localized energy addition to the supersonic oncoming flow // Fluid Dynamics. 2003. V. 38, N 5. P. 154–167.
11. *Zudov V. N., Tret′yakov P. K., Tupikin A. V., and Yakovlev V. I.* Supersonic flow past a thermal source // Fluid Dynamics. 2003. V. 38, N 5. P. 140–153.
12. *Gordeev V. P., Krasilnikov A. V., Lagutin V. I., and Otmennikov V. N.* Experimental study of the possibility of reducing supersonic drag by employing plasma technology // Fluid Dynamics. 1996. V. 31, N 2. P. 313–317.
13. *Levin V. A., Gromov V. G., and Afonina N. E.* Numerical study of the influence of the local energy supply on aerodynamic drag and heat spherical bluntness in a supersonic air flow // Prikl. Mekh. i Tekhn. Fiz. 2000. V. 41, N 5. P. 171–179.
14. *Korotaeva T. A., Fomin V. M., and Shashkin A. P.* Spatial supersonic flow past a pointed body with energy supply in front of him // Prikl. Mat. i Tekh. Fiz. Novosibirak: Izdat. SO RAN, 1998. P. 116–121.
15. *Fomin V. M., Maslov A. A., Korotaeva T. A., and Shashkin A. P.* Numerical simulation of a supersonic spatial nonuniform flow // Int. J. Comput. Fluid Dyn. Special Issue. 2003. V. 12, N 2. P. 367–382.

August 28, 2015

A. E. Lutskiĭ; Ya. V. Khankhasaeva
Keldysh Institute of Applied Mathematics, Moscow, Russia
allutsky@yandex.ru; hanhyana@mail.ru

Yakutian Mathematical Journal
April—June, 2015. Vol. 22, No. 2

UDC 517.956.3

MATHEMATICAL MODELING OF THE PROPAGATION OF ACOUSTICS–GRAVITY AND SEISMIC WAVES IN A HETEROGENEOUS EARTH–ATMOSPHERE MODEL WITH A WIND–STRATIFIED ATMOSPHERE

A. A. Mikhaĭlov and V. N. Martynov

Abstract. A numerical-analytical algorithm for seismic and acoustic-gravity waves propagation is applied to a heterogeneous Earth–Atmosphere model. Seismic wave propagation in an elastic half-space is described by a system of first-order dynamic equations of elasticity theory. The propagation of acoustic-gravity waves in the atmosphere is described by the linearized Navier–Stokes equations with a wind. The algorithm is based on the integral Laguerre transform with respect to time, the finite integral Fourier transform with respect to a spatial coordinate combined with a finite difference method for the reduced problem. The algorithm is numerically tested for the heterogeneous Earth–Atmosphere model for different source locations.

Keywords: Navier–Stokes equations, finite difference methods, Laguerre transform, acoustic-gravity waves, seismic waves

Introduction

In the mathematical modeling of seismic wave fields in an elastic medium, the surface of the medium is usually assumed to be adjacent to vacuum and boundary conditions on the free surface are prescribed. Therefore, it is assumed that seismic waves are absolutely reflected on the boundary and the generation of acoustic-gravity waves in the atmosphere by the elastic waves and their interaction along the boundary are neglected.

Theoretical and experimental studies of the last decade have showed a high degree of interrelation between waves in the lithosphere and atmosphere. The acoustic-seismic induction effect is described in [1], in which an acoustic wave from a vibrator, owing to refraction in the atmosphere, excites surface seismic waves tens of kilometers away. In turn, the lithosphere seismic waves from earthquakes and explosions generate atmospheric acoustic-gravity waves which are especially strong in the upper layers of atmosphere with small density and the ionosphere. Many articles present theoretical studies of wave processes on the boundary between the elastic half-space and isothermal homogeneous atmosphere; let us mention only the articles [2, 3] that established and studied the properties of Stoneley–Scholte surface waves and modified Lamb waves.

The authors were supported by the Russian Foundation for Basic Research (Grant 14–05–00867).

In this article, using numerical modeling, we continue studying the propagation of seismic and acoustic-gravity waves in the spatially heterogeneous Earth–Atmosphere model basing on the ideas of Mikhaĭlenko, who pioneered and then supported these studies. We consider a numerical algorithm for solving the combined problem of the propagation of acoustic-gravity waves in a wind-stratified atmosphere and seismic waves in a heterogeneous elastic medium in a Cartesian coordinate system. The similar problem for a vertically heterogeneous model in a cylindrical coordinate system was considered in [4] without accounting for the wind. The algorithm for solving the stated problem rests on the Laguerre transform originally proposed in [5]. The propagation of acoustic-gravity waves in isothermal atmosphere is described by the linearized Navier–Stokes system. We assume that the density of the atmosphere and the velocity of the wind depend on height. The propagation of seismic waves in an elastic medium is described by a hyperbolic first-order system in terms of the velocity vector of the displacement and the components of the stress tensor.

The algorithm, presented here, is constructed on using the complexification of integral transforms and the finite difference method. We assume that the parameters of the medium (its density and the speed of longitudinal and transverse waves) depend only on two coordinates, while the medium is homogeneous with respect to the third coordinate. This statement of the problem is known as a 2.5-D problem. We may regard the application of the Laguerre transform with respect to the time coordinate for a numerical solution of the problem as an analog of the well-known spectral method based on the Fourier transform, where instead of the frequency ω we have a parameter p, the degree of the Laguerre polynomial. However, in contrast to the Fourier transform, applying the Laguerre integral transform with respect to time enables us to reduce the original problem to a system of equations in which the separation parameter appears only on the right-hand side and has a recursive nature. This method originated in [5, 6] for solving dynamical problems of elasticity theory and was later developed for viscoelasticity problems [7, 8] and the theory of porous media [9]. These articles show how this method differs from the usual approaches and discuss the advantages of applying the Laguerre integral transform in contrast to the difference method and the Fourier transform with respect to time.

1. Statement of the Problem

The system of equations describing the propagation of acoustic-gravity waves in a heterogeneous not ionized isothermal atmosphere in the Cartesian coordinate system (x, y, z) in the presence of a wind directed along the horizontal x-axis, and vertically stratified along the z-axis, is of the form

$$\frac{\partial u_x}{\partial t} + v_x \frac{\partial u_x}{\partial x} = -\frac{1}{\rho_0}\frac{\partial P}{\partial x} - u_z \frac{\partial v_x}{\partial z}, \tag{1}$$

$$\frac{\partial u_y}{\partial t} + v_x \frac{\partial u_y}{\partial x} = -\frac{1}{\rho_0}\frac{\partial P}{\partial y}, \tag{2}$$

$$\frac{\partial u_z}{\partial t} + v_x \frac{\partial u_z}{\partial x} = -\frac{1}{\rho_0}\frac{\partial P}{\partial z} - \frac{\rho g}{\rho_0}, \tag{3}$$

$$\frac{\partial P}{\partial t} + v_x \frac{\partial P}{\partial x} = c_0^2 \left[\frac{\partial \rho}{\partial t} + v_x \frac{\partial \rho}{\partial x} + u_z \frac{\partial \rho_0}{\partial z}\right] - u_z \frac{\partial P_0}{\partial z} \tag{4}$$

$$\frac{\partial \rho}{\partial t} + v_x \frac{\partial \rho}{\partial x} = -\rho_0 \left[\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z}\right] - u_z \frac{\partial \rho_0}{\partial z} + F(x, y, z, t). \tag{5}$$

Here g is the free fall acceleration, $\rho_0(z)$ is the density of unperturbed atmosphere, $c_0(z)$ is the speed of sound, $v_x(z)$ is the wind speed along the x-axis, $\vec{u} = (u_x, u_y, u_z)$ is the velocity vector of the displacement of air particles, P and ρ are respectively the perturbations of pressure and density under the action of the propagating wave from a source of mass $F(x, y, z, t) = \delta(r - r_0)f(t)$, where $f(t)$ is a specified time signal at the source. Assume that the z-axis is directed upward. The zero subscripts of the physical parameters of the medium indicate that their values are defined for the unperturbed state of the atmosphere. We can determine the dependence of the atmospheric pressure P_0 and the density ρ_0 for the unperturbed state of the atmosphere in the homogeneous gravity field as

$$\frac{\partial P_0}{\partial z} = -\rho_0 g, \quad \rho_0(z) = \rho_1 \exp(-z/H),$$

where H is the height of the homogeneous isothermal atmosphere, while ρ_1 is the density of the atmosphere near the surface of the Earth; i.e., at $z = 0$.

We can express the propagation of seismic waves in an elastic medium as the well-known system of first-order elasticity theory equations via the relation among the components of the velocity vector of displacements and the components of the stress tensor:

$$\frac{\partial u_i}{\partial t} = \frac{1}{\rho_0}\frac{\partial \sigma_{ik}}{\partial x_k} + F_i f(t), \tag{6}$$

$$\frac{\partial \sigma_{ik}}{\partial t} = \mu\left(\frac{\partial u_k}{\partial x_i} + \frac{\partial u_i}{\partial x_k}\right) + \lambda \delta_{ik} \operatorname{div} \vec{u}. \tag{7}$$

Here $\lambda(x_1, x_2, x_3)$ and $\mu(x_1, x_2, x_3)$ are elastic parameters of the medium, $\rho_0(x_1, x_2, x_3)$ is the density of the medium, δ_{ij} is the Kronecker symbol, $\vec{u} = (u_1, u_2, u_3)$ is the velocity vector of displacements, $\sigma_{i\,j}$ are the components of the stress tensor. The function $\vec{F}(x, y, z) = F_1\vec{e}_x + F_2\vec{e}_y + F_3\vec{e}_z$ describes the source distribution localized in space, while $f(t)$ is the prescribed time signal at the source.

Then we can express the combined system of equations for describing the propagation of seismic and acoustic-gravity waves in the Cartesian system of coordinates $(x, y, z) = (x_1, x_2, x_3)$ as

$$\frac{\partial u_i}{\partial t} = \frac{1}{\rho_0}\frac{\partial \sigma_{ik}}{\partial x_k} + F_i f(t) - K_a\left[v_x \frac{\partial u_i}{\partial x_1} + \frac{\rho g}{\rho_0} e_z - u_z \frac{\partial v_x}{\partial x_3} e_x\right], \tag{8}$$

$$\frac{\partial \sigma_{ik}}{\partial t} = \mu\left(\frac{\partial u_k}{\partial x_i} + \frac{\partial u_i}{\partial x_k}\right) + \lambda \delta_{ik} \operatorname{div} \vec{u} - \delta_{ik} K_a\left[v_x \frac{\partial \sigma_{ik}}{\partial x_1} + \rho_0 g u_z\right], \tag{9}$$

$$K_a\left[\frac{\partial \rho}{\partial t} + v_x \frac{\partial \rho}{\partial x} = -\rho_0 \operatorname{div} \vec{u} - u_z \frac{\partial \rho_0}{\partial z}\right]. \tag{10}$$

Here δ_{ij} is the Kronecker symbol, $\rho_0(x, z)$ is the density of the medium, $\lambda(x, z)$ and $\mu(x, z)$ are elastic parameters of the medium, $\vec{u} = (u_1, u_2, u_3)$ is the velocity vector of displacements, σ_{ij} are the components of the stress tensor. The function $\vec{F}(x, y, z) = F_1\vec{e}_x + F_2\vec{e}_y + F_3\vec{e}_z$ describes the source distribution as localized in space, while $f(t)$ is the prescribed time signal at the source. We assume that the medium is homogeneous with respect to the Y-axis.

We obtain system (1)–(5) for the atmosphere from system (8)–(10) if we take $\sigma_{11} = \sigma_{22} = \sigma_{33} = -P$, $\mu = 0$, $\lambda = c_0^2\rho_0$, $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0$, and $K_a = 1$. Putting $K_a = 0$ in (8)–(10), we obtain system (6), (7) for seismic waves propagating in an elastic medium.

In our problem, assume that the interface of the media, the atmosphere and the elastic half-space, lies on the plane $z = x_3 = 0$. In this case we can express the contact condition for the two media at $z = 0$ as

$$u_z|_{z=-0} = u_z|_{z=+0}; \quad \left.\frac{\partial \sigma_{zz}}{\partial t}\right|_{z=-0} = \left.\left(\frac{\partial \sigma_{zz}}{\partial t} + \rho_0 g u_z\right)\right|_{z=+0}; \qquad (11)$$
$$\sigma_{xz}|_{z=-0} = \sigma_{yz}|_{z=-0} = 0.$$

This problem is solved for the zero initial data

$$u_i|_{t=0} = \sigma_{ij}|_{t=0} = P|_{t=0} = \rho|_{t=0} = 0, \quad i = 1,2,3,\ j = 1,2,3. \qquad (12)$$

To apply transforms below, we assume that all functions of the components of the wave field are sufficiently smooth.

2. A Method of Numerical Solution

At the first stage of solution, use the finite cosine-sine Fourier transform with respect to the spatial coordinate y, in the direction of which the medium is regarded as homogeneous. For each component of the system, introduce the corresponding cosine or sine transform

$$\overrightarrow{W}(x,z,n,t) = \int_0^a \overrightarrow{W}(x,y,z,t) \begin{Bmatrix} \cos(k_n y) \\ \sin(k_n y) \end{Bmatrix} d(y), \quad n = 0,1,2,\ldots,N, \qquad (13)$$

with the corresponding inversion formula

$$\overrightarrow{W}(x,y,z,t) = \frac{1}{\pi}\overrightarrow{W}(x,0,z,t) + \frac{2}{\pi}\sum_{n=1}^{N} \overrightarrow{W}(x,n,z,t)\cos(k_n y) \qquad (14)$$

or

$$\overrightarrow{W}(x,y,z,t) = \frac{2}{\pi}\sum_{n=1}^{N} \overrightarrow{W}(x,n,z,t)\sin(k_n y), \qquad (15)$$

where $k_n = \frac{n\pi}{a}$.

Choose a sufficiently large distance a and consider the wave field up to the time $t < T$, where T is the minimal propagation time of the longitudinal wave to the boundary $r = a$. This transformation yields $N + 1$ independent nonstationary problems which are two-dimensional with respect to space.

At the second stage, apply to the resulting $N + 1$ independent problems the Laguerre integral transform with respect to time of the form

$$\overrightarrow{W}_p(x,n,z) = \int_0^\infty \overrightarrow{W}(x,n,z,t)(ht)^{-\frac{\alpha}{2}} l_p^\alpha(ht)\, d(ht), \quad p = 0,1,2,\ldots, \qquad (16)$$

with the inversion formula

$$\overrightarrow{W}(x,n,z,t) = (ht)^{\frac{\alpha}{2}} \sum_{p=0}^{\infty} \frac{p!}{(p+\alpha)!} \overrightarrow{W}_p(x,n,z) l_p^\alpha(ht), \qquad (17)$$

where $l_p^\alpha(ht)$ are the Laguerre orthogonal functions.

The Laguerre functions $l_p^\alpha(ht)$ can be expressed in terms of the classical orthonormal Laguerre polynomials $L_p^\alpha(ht)$ [10]. Here we choose α (the order of Laguerre functions) to be integral and positive. Thus, we have

$$l_p^\alpha(ht) = (ht)^{\frac{\alpha}{2}} e^{-\frac{ht}{2}} L_p^\alpha(ht).$$

To meet the initial condition (12), it is necessary and sufficient to put $\alpha \geq 1$. In addition, we introduce the parameter $h > 0$ of translation, whose meaning is discussed in [6–8] as well as the effectiveness of its applications.

As a result of these transformations, solving the original problem (8)–(12) reduces to solving $N+1$ independent two-dimensional differential problems in the spectral region of the form

$$\frac{h}{2}u_x^p - \frac{1}{\rho_0}\left(\frac{\partial\sigma_{xz}^p}{\partial z} + \frac{\partial\sigma_{xx}^p}{\partial x} + k_n\sigma_{xy}^p\right) + K_a\left[v_x\frac{\partial\, u_x^p}{\partial\, x} - u_z^p\frac{\partial\, v_x}{\partial\, z}\right] = F_x(n)f^p - h\sum_{j=0}^{p-1} u_x^j, \tag{18}$$

$$\frac{h}{2}u_y^p - \frac{1}{\rho_0}\left(\frac{\partial\sigma_{yz}^p}{\partial z} + \frac{\partial\sigma_{xy}^p}{\partial x} - k_n\sigma_{yy}^p\right) + K_a v_x\frac{\partial\, u_y^p}{\partial\, x} = F_y(n)f^p - h\sum_{j=0}^{p-1} u_y^j, \tag{19}$$

$$\frac{h}{2}u_z^p - \frac{1}{\rho_0}\left(\frac{\partial\sigma_{zz}^p}{\partial z} + \frac{\partial\sigma_{xz}^p}{\partial x} + k_n\sigma_{yz}^p\right) + K_a\left[v_x\frac{\partial\, u_z^p}{\partial\, x} + \frac{g}{\rho_0}\bar{\rho}^p\right] = F_z(n)f^p - h\sum_{j=0}^{p-1} u_z^j, \tag{20}$$

$$\frac{h}{2}\sigma_{xx}^p - \lambda\left(\frac{\partial u_z^p}{\partial z} + k_n u_y^p\right) - (\lambda+2\mu)\frac{\partial u_x^p}{\partial x} + K_a\left[v_x\frac{\partial\,\sigma_{xx}^p}{\partial\, x} + \rho_0 g u_z^p\right] = -h\sum_{j=0}^{p-1}\sigma_{xx}^j, \tag{21}$$

$$\frac{h}{2}\sigma_{yy}^p - \lambda\left(\frac{\partial u_z^p}{\partial z} + \frac{\partial u_x^p}{\partial x}\right) - (\lambda+2\mu)k_n u_y^p + K_a\left[v_x\frac{\partial\,\sigma_{yy}^p}{\partial\, x} + \rho_0 g u_z^p\right] = -h\sum_{j=0}^{p-1}\sigma_{yy}^j, \tag{22}$$

$$\frac{h}{2}\sigma_{zz}^p - \lambda\left(\frac{\partial u_x^p}{\partial x} + k_n u_y^p\right) - (\lambda+2\mu)\frac{\partial u_z^p}{\partial z} + K_a\left[v_x\frac{\partial\,\sigma_{zz}^p}{\partial\, x} + \rho_0 g u_z^p\right] = -h\sum_{j=0}^{p-1}\sigma_{zz}^j, \tag{23}$$

$$\frac{h}{2}\sigma_{xy}^p - \mu\left(\frac{\partial u_y^p}{\partial x} + k_n u_x^p\right) = -h\sum_{j=0}^{p-1}\sigma_{xy}^j, \tag{24}$$

$$\frac{h}{2}\sigma_{xz}^p - \mu\left(\frac{\partial u_x^p}{\partial z} - \frac{\partial u_z^p}{\partial x}\right) = -h\sum_{j=0}^{p-1}\sigma_{xz}^j, \tag{25}$$

$$\frac{h}{2}\sigma_{yz}^p - \mu\left(\frac{\partial u_y^p}{\partial z} + k_n u_z^p\right) = -h\sum_{j=0}^{p-1}\sigma_{yz}^j, \tag{26}$$

$$K_a\left[\frac{h}{2}\rho^p + v_x\frac{\partial\rho^p}{\partial\, x} + \rho_0\left(\frac{\partial u_x^p}{\partial x} + k_n u_y^p + \frac{\partial u_z^p}{\partial z}\right) + u_z^p\frac{\partial\rho_0}{\partial\, z} = -h\sum_{j=0}^{p-1}\rho^j\right], \tag{27}$$

where f^p are the Laguerre coefficients of the source function $f(t)$. The coefficients u_x^p, u_y^p, u_z^p, σ_{xx}^p, σ_{yy}^p, σ_{zz}^p, σ_{xy}^p, σ_{xz}^p, σ_{yz}^p, and ρ^p in (18)–(27) are functions of the variables (n, x, z).

It is easy to observe that the parameter p of the Laguerre transform appears only on the right-hand side of the equations and the spectral harmonics for all components of the field are in recursive dependence.

We can express the condition of contact between the two media at $z = 0$ as

$$\begin{gathered}\left.\frac{h}{2}\sigma_{zz}^p + h\sum_{j=0}^{p-1}\sigma_{zz}^j\right|_{z=-0} = \left.\left(\frac{h}{2}\sigma_{zz}^p + h\sum_{j=0}^{p-1}\sigma_{zz}^j + \rho_0 g u_z^p\right)\right|_{z=+0}; \\ u_z^p|_{z=-0} = u_z^p|_{z=+0}; \quad \sigma_{xz}^p|_{z=-0} = \sigma_{yz}^p|_{z=-0} = 0.\end{gathered} \tag{28}$$

To solve (18)–(28), use the finite cosine-sine Fourier transform with respect to the space coordinate x and a finite difference approximation of the second order of accuracy [11] to the derivatives with respect to the z coordinate.

To this end, introduce in the direction of the z-coordinate the region of simulation of the two meshes ωz_i and $\omega z_{i+1/2}$ with meshsize Δz shifted with respect to each other by $\Delta z/2$:

$$\omega z_i = \{z_i = i\Delta z\ ;\ i = 0,\ldots,K\}, \quad \omega z_{i+1/2} = \{z_{i+1/2} = (i+\tfrac{1}{2})\Delta z;\ i = 0,\ldots,K-1\}.$$

On these meshes introduce the operator D_z of differentiation, approximating to the second order of accuracy the derivative $\frac{\partial}{\partial z}$ with respect to the z-coordinate as

$$D_z u(x,z) = \frac{1}{\Delta z}\left[u\left(x, z + \frac{\Delta z}{2}\right) - u\left(x, z - \frac{\Delta z}{2}\right)\right].$$

Define the required components of the solution vector at the following nodes:

$$\rho^p, u_x^p(x,z), u_y^p(x,z), \sigma_{xx}^p(x,z), \sigma_{yy}^p(x,z), \sigma_{zz}^p(x,z), \sigma_{xy}^p(x,z) \in \omega z_i,$$

$$u_z^p(x,z), \sigma_{xz}^p(x,z), \sigma_{yz}^p(x,z) \in \omega z_{i+1/2}.$$

Choose the locations of components at integer and half-integer nodes of the mesh basing on the difference approximation to (18)–(27) and the required boundary condition (28). For the upper and lower boundaries impose boundary conditions of the first and second kind for the corresponding components.

With respect to the x-coordinate, use the finite cosine-sine Fourier transform similar to the previously used transform with respect to the y-coordinate with the corresponding inversion formulas:

$$\overrightarrow{W}_p(x,n,z_i,p) = \frac{1}{\pi}\overrightarrow{W}_0(n,z_i,p) + \frac{2}{\pi}\sum_{m=1}^{M}\overrightarrow{W}(m,n,z_i,p)\cos(k_m x) \tag{29}$$

or

$$\overrightarrow{W}(x,n,z_i,p) = \frac{2}{\pi}\sum_{m=1}^{M}\overrightarrow{W}(m,n,z_i,p)\sin(k_m x), \tag{30}$$

where $k_m = \frac{m\pi}{b}$. We should account for the heterogeneity of the medium in this direction.

This yields a system of linear algebraic equations, expressible for nodes i and $i+\frac{1}{2}$ of the mesh as

$$\frac{h}{2}\bar{u}_x^p - \sum_{s=0}^{M} q_1\left(D_z\bar{\sigma}_{xz}^p - k_s\bar{\sigma}_{xx}^p + k_n\bar{\sigma}_{xy}^p\right) + K_a\sum_{s=0}^{M} r_1\left(v_x k_s \bar{u}_x^p - \bar{u}_z^p D_z v_x\right)$$
$$= F_x f^p - h\sum_{j=0}^{p-1}\bar{u}_x^j, \tag{31}$$

$$\frac{h}{2}\bar{u}_y^p - \sum_{s=0}^{M} q_2\left(D_z\bar{\sigma}_{yz}^p + k_s\bar{\sigma}_{xy}^p - k_n\bar{\sigma}_{yy}^p\right) - K_a\sum_{s=0}^{M} r_2 v_x k_s \bar{u}_y^p = F_y f^p$$
$$-h\sum_{j=0}^{p-1}\bar{u}_y^j, \tag{32}$$

$$\frac{h}{2}\bar{u}_z^p - \sum_{s=0}^{M} q_3\left(D_z\bar{\sigma}_{zz}^p + k_s\bar{\sigma}_{xz}^p + k_n\bar{\sigma}_{yz}^p\right) + K_a\left[\frac{g}{\rho_0}\bar{\rho}^p - \sum_{s=0}^{M} r_2 v_x k_s \bar{u}_z^p\right]$$
$$= F_z f^p - h\sum_{j=0}^{p-1}\bar{u}_z^j, \tag{33}$$

$$\frac{h}{2}\bar{\sigma}_{xx}^p - \sum_{s=0}^{M} q_4\left(D_z\bar{u}_z^p + k_n\bar{u}_y^p\right) - \sum_{s=0}^{M} q_5 k_s \bar{u}_x^p + K_a\left[\rho_0 g\bar{u}_z^p - \sum_{s=0}^{M} r_2 v_x k_s \bar{\sigma}_{xx}^p\right]$$
$$= -h\sum_{j=0}^{p-1}\bar{\sigma}_{xx}^j, \tag{34}$$

$$\frac{h}{2}\bar{\sigma}_{yy}^p - \sum_{s=0}^{M} q_4\left(D_z\bar{u}_z^p + k_s\bar{u}_x^p\right) - \sum_{s=0}^{M} q_5 k_n \bar{u}_y^p + K_a\left[\rho_0 g\bar{u}_z^p - \sum_{s=0}^{M} r_2 v_x k_s \bar{\sigma}_{yy}^p\right]$$
$$= -h\sum_{j=0}^{p-1}\bar{\sigma}_{yy}^j, \tag{35}$$

$$\frac{h}{2}\bar{\sigma}_{zz}^p - \sum_{s=0}^{M} q_4\left(k_s\bar{u}_x^p + k_n\bar{u}_y^p\right) - \sum_{s=0}^{M} q_5 D_z \bar{u}_z^p + K_a\left[\rho_0 g\bar{u}_z^p - \sum_{s=0}^{M} r_2 v_x k_s \bar{\sigma}_{zz}^p\right]$$
$$= -h\sum_{j=0}^{p-1}\bar{\sigma}_{zz}^j, \tag{36}$$

$$\frac{h}{2}\bar{\sigma}_{xy}^p - \sum_{s=0}^{M} q_6\left(k_s\bar{u}_y^p + k_n\bar{u}_x^p\right) = -h\sum_{j=0}^{p-1}\bar{\sigma}_{xy}^j, \tag{37}$$

$$\frac{h}{2}\bar{\sigma}_{xz}^p - \sum_{s=0}^{M} q_7\left(D_z\bar{u}_x^p + k_s\bar{u}_z^p\right) = -h\sum_{j=0}^{p-1}\bar{\sigma}_{xz}^j, \tag{38}$$

$$\frac{h}{2}\bar{\sigma}^p_{yz} - \sum_{s=0}^{M} q_8 \Big(D_z \bar{u}^p_y + k_n \bar{u}^p_z \Big) = -h \sum_{j=0}^{p-1} \bar{\sigma}^j_{yz}, \tag{39}$$

$$K_a \left[\frac{h}{2}\bar{\rho}^p - \sum_{s=0}^{M} r_2 v_x k_s \bar{\rho}^p + \sum_{s=0}^{M} q_9 \left(k_s \bar{u}^p_x + k_n \bar{u}^p_y + D_z \bar{u}^p_z \right) + \bar{u}^p_z D_z \rho_0 = -h \sum_{j=0}^{p-1} \bar{\rho}^j \right], \tag{40}$$

where

$$r_1 = \int_0^b \cos(k_s x)\sin(k_m x)\; dx, \quad r_2 = \int_0^b \sin(k_s x)\cos(k_m x)\; dx,$$

$$q_1 = \int_0^b \frac{1}{\rho_0(x,z_i)} \sin(k_s x)\sin(k_m x)\; dx, \quad q_2 = \int_0^b \frac{1}{\rho_0(x,z_i)} \cos(k_s x)\cos(k_m x)\; dx,$$

$$q_3 = \int_0^b \frac{1}{\rho_0(x,z_{i+1/2})} \cos(k_s x)\cos(k_m x)\; dx,$$

$$q_4 = \int_0^b \lambda(x,z_i)\cos(k_s x)\cos(k_m x)\; dx,$$

$$q_5 = \int_0^b [\lambda(x,z_i) + 2\mu(x,z_i)]\cos(k_s x)\cos(k_m x)\; dx,$$

$$q_6 = \int_0^b \mu(x,z_i)\sin(k_s x)\sin(k_m x)\; dx,$$

$$q_7 = \int_0^b \mu(x,z_{i+1/2})\sin(k_s x)\sin(k_m x)\; dx,$$

$$q_8 = \int_0^b \mu(x,z_{i+1/2})\cos(k_s x)\cos(k_m x)\; dx,$$

$$q_9 = \int_0^b \rho_0(x,z_i)\cos(k_s x)\cos(k_m x)\; dx, \quad k_m = \frac{m\pi}{b}, \quad k_S = \frac{s\pi}{b}.$$

In (31)–(40) we use the notation $\bar{u}^p_x = \bar{u}^p_x(m,n,z_j)$. It works similarly for the other components. The bar over the symbol of a field component means that we consider the coefficients of its Fourier transform with respect to the x-coordinate.

These manipulations lead to $N+1$ systems of linear algebraic equations, where N is the number of harmonics of the Fourier transform with respect to the y-coordinate. Express the required solution vector $\vec{W}$ as

$$\vec{W}(p) = (\vec{V}_0(p), \vec{V}_1(p), \ldots, \vec{V}_K(p))^T,$$

$$\vec{V}_i = (\bar{\rho}^p(m=0,\ldots,M;z_i), \bar{\sigma}^p_{xx}(m=0,\ldots,M;z_i), \bar{u}^p_x(m=0,\ldots,M;z_i), \ldots)^T.$$

Then for each harmonic n, with $n = 0, \dots, N$, we can express the system of linear algebraic equations in vector form as

$$\left(A + \frac{h}{2}E\right)\vec{W}(p) = \vec{F}(p-1). \tag{41}$$

Choose the sequence of components of the wave field in the vector solution $\vec{V}$ taking into account the minimization of the number of diagonals in the matrix A. Furthermore, on the main diagonal of the matrix of the system under solution we intentionally put the components that appear in the equations as the terms with the parameter h as a factor (the Laguerre transform parameter). By the choice of the value of h, it is possible to improve the condition number of the matrix substantially. Solving (41), we can determine the spectral values of all components of the wave field $\vec{W}(m, n, p)$. Then, by the inversion formulas (14), (15), (29), and (30) for the Fourier transform and (17) for the Laguerre transform, we obtain a solution to the original problem (8)–(12).

3. Aspects of Numerical Implementation

In the analytical Fourier and Laguerre transforms, when evaluating functions from their spectrum, we use inversion formulas in the form of infinite series. For a numerical implementation, we should find the required number of terms of the series in order to construct the solution with specified accuracy. Thus, for instance, the number of harmonics in the inversion formulas (14), (15), (29), (30) for the Fourier transform depends on the minimal spatial length of waves in the modeled medium and the size of the simulated region of the reconstructed field, which is given by the finite bounds of the integral transform. In addition, the convergence rate of the series being summed depends on the smoothness of functions of the modeled wave field.

The number of series terms in the expansion into Laguerre functions necessary for determining the field components using (17) depends on the prescribed signal $f(t)$ at the source, the choice of parameter h, and the value of time interval of the modeled wave field. How to find the required number of harmonics and choose the optimal value of h is discussed in detail in [6–8].

Inspection of simulations shows that the main calculation error in the presented algorithm for solving the problem under consideration has to do with numerical approximations of spatial derivatives. Therefore, to approximate the derivatives near the interface of strongly contrasting layers of the medium more precisely, as well as to account better for conditions (11) on the Earth–Atmosphere interface, it is better to use a mesh with variable discretization meshsize. Thus, we can decrease the meshsize to approximate the derivatives in certain parts of the medium, which enables us to obtain a solution with required accuracy for a lower number of nodes of the mesh.

To solve system (41), it turned out most efficient to use the iterative conjugate gradient method [12, 13]. In this case the matrices for systems of large dimension need not be fully stored in memory at once. Another advantage of this method is its fast convergence to the solution provided that the matrix of the system is well-conditioned. Our matrix enjoys this property due to the parameter h. Specifying a suitable value of h, we can substantially speed up the convergence of iterations. The optimal value of h in this case is chosen to minimize the number of Laguerre harmonics in the inversion formula (17) and to decrease the number of iterations required for finding the solution for each harmonic.

The use of the Fourier transform with respect to the space coordinate in the direction of which the medium is regarded as homogeneous enables us to implement efficient parallelization of the solution. In this case each processor solves an independent problem for each Fourier harmonic. In addition, when running calculations on computing clusters with a low amount of memory accessible to one process, to solve large spatial problems (more than 100 wavelengths) we parallelized the solution of the two-dimensional spatial problem. At this stage of calculations we implemented a parallel version of the conjugate gradient method for solving the system of algebraic equations for each Fourier harmonic. At the level of input data, as we prescribe a model of the medium, this is equivalent to decomposing the original region into several subregions of the two-dimensional problem with respect to the z-coordinate. This approach makes it possible to distribute memory during both the prescription of input parameters of the model and the subsequent numerical implementation of the algorithm in the subregions.

4. Numerical Results

In this article we consider the results of simulations for two variants of wave propagation in the Earth–Atmosphere medium in the presence of a wind. In the first variant the velocity of a wind in the atmosphere is constant and independent of height. In the second variant the velocity of a wind in the atmosphere is a function of height. Figs. 1 and 2 show the results of simulating the wave field as snapshots at the fixed time.

Fig. 1 depicts the result of calculating the wave field for the constant velocity of a wind in the atmosphere equal to 50 m/s. We chose this value to obtain the main physical effects of wave propagation without calculations at very large distances. The specified model of a medium consists of a homogeneous elastic layer and an atmospheric layer separated by a flat boundary. The physical characteristics of the layers are as follows:

(1) the atmosphere: the speed of sound is $c_0 = 340\,\mathrm{m/s}$; the density depending on the z-coordinate is calculated by the formula $\rho_0(z) = \rho_1 \exp(-z/H)$, where $\rho_1 = 1.225 * 10^{-3}\,\mathrm{g/cm^3}$ and $H = 6700\,\mathrm{m}$;

(2) the elastic layer: the speed of the longitudinal wave is $c_p = 800\,\mathrm{m/s}$; the speed of the transverse wave is $c_s = 500\,\mathrm{m/s}$; the density is $\rho_0 = 1.5\,\mathrm{g/cm^3}$.

We took a bounded region of a medium of size $(x, y, z) = (80\,km,\ 80\,km,\ 60\,km)$. We modeled the wave field of a point source of pressure center type lying in an elastic medium at depth 1/4 of the longitudinal wave length with coordinates $(x_0, y_0, z_0) = (40\,km,\ 40\,km,\ -0.2\,km)$. The time signal at the source was specified as the Puzyrëv pulse:

$$f(t) = \exp\left(-\frac{(2\pi f_o(t-t_0))^2}{\gamma^2}\right)\sin(2\pi f_0(t-t_0)), \tag{42}$$

where $\gamma = 4$, $f_0 = 1\,\mathrm{Hz}$, and $t_0 = 1.5\,\mathrm{s}$.

Fig. 1 shows the snapshots of the wave field at time $t = 5\,\mathrm{s}$. for the component $u_x(x, y, z)$ in the plane XZ for $y = y_0 = 40\,\mathrm{km}$. The left image is without a wind, the right image is with a wind at 50 m/s. The interface of an elastic medium and atmosphere is shown as the solid line. It is clear from the pictures that in the elastic medium, aside from the spherical longitudinal wave $\boldsymbol{P}$ and conical transverse wave $\boldsymbol{S}$, the "nonray" spherical wave $\boldsymbol{S}^*$ propagates, followed by the Stoneley surface

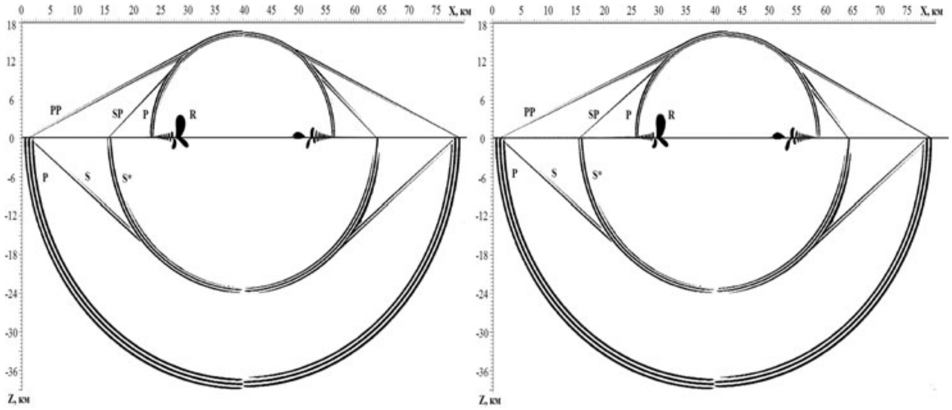

Fig. 1

wave $\boldsymbol{R}$. In the atmosphere, aside from the conical acoustic-gravity waves $\boldsymbol{PP}$ and $\boldsymbol{SP}$ refracted at the boundary, the spherical wave $\boldsymbol{P}$ propagates, followed by the Stoneley surface wave. In the images of the wave field in Fig. 1 we can notice the influence of a wind on the propagation of acoustic-gravity waves in the atmosphere and Stoneley surface waves, as well as on the overall wave portrait. Inspection of the results of simulating the wave field and the influence of a wind on it in the case that the velocity of a wind is constant appeared in [14], which also included a description of the influence of a wind on the propagation of the Stoneley surface wave, an effect discovered as a result of these studies. Previously only the influence of a wind on the propagation of acoustic-gravity waves in the atmosphere was known. It is established that in the presence of a wind, the velocity and amplitude of the spherical wave in the atmosphere and the Stoneley surface wave depend on the direction of propagation of these waves with respect to the velocity vector of the wind.

Fig. 2 shows the result of simulating the wave field for the velocity of a wind depending on height. In this model the physical characteristics of the elastic medium and atmosphere were specified as follows:

(1) the atmosphere: the speed of sound is $c_0 = 340\,\mathrm{m/s}$; the density depending on the z-coordinate is calculated by the formula $\rho_0(z) = \rho_1 \exp(-z/H)$, where $\rho_1 = 1.225 * 10^{-3}\,\mathrm{g/cm^3}$ and $H = 6700\,\mathrm{m}$;

(2) the elastic layer: the speed of the longitudinal wave is $c_p = 450\,\mathrm{m/s}$; the speed of the transverse wave is $c_s = 300\,\mathrm{m/s}$; density is $\rho_0 = 1.5\,\mathrm{g/cm^3}$.

We took a bounded region of medium of size $(x, y, z) = (40\,km,\ 40\,km,\ 33\,km)$. We modeled the wave field of a point source of pressure center type lying in the elastic medium at depth 1/4 of the longitudinal wavelength with coordinates $(x_0, y_0, z_0) = (20\,km,\ 20\,km,\ -0.12\,km)$. The time signal at the source was specified by (42). The velocity of a wind in the atmosphere was specified as the function

$$V(z) = 50 \cdot \exp(-10 \cdot (z - 3800)^2) - 50 \cdot \exp(-10 \cdot (z - 7500)^2)\, m/s.$$

Fig. 2 shows the snapshots of the wave field at time $t = 40\,\mathrm{s}$ for the horizontal component u_x of the velocity of displacements in the plane XZ at $y = y_0 = 20\,\mathrm{km}$. The left image is without a wind, the right image is with a wind. The interface of an elastic medium and atmosphere is shown as the solid line.

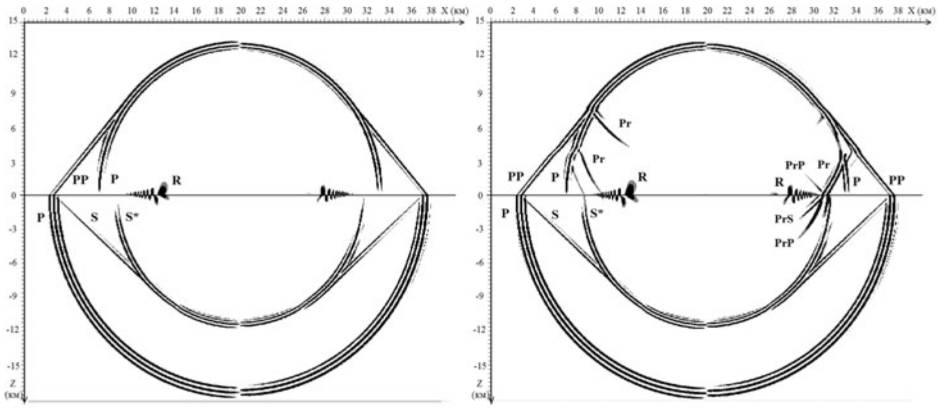

Fig. 2

From the image of a wave field without a wind in Fig. 2 (the left picture) it is clear that the acoustic-gravity conical wave ***PP*** and the spherical wave ***P*** propagate in the atmosphere, followed by the Stoneley surface wave ***R***. In the image of the wave field with a wind (the right picture) it is clear that the refracted acoustic-gravity waves ***Pr*** occur in the atmosphere. Their appearance is explained by the changing velocity of a wind with height. Falling on the atmosphere/lithosphere interface, these waves generate appropriate longitudinal wave ***PrP*** and transverse wave ***PrS*** in the lithosphere and the reflected acoustic-gravity wave in the atmosphere. This phenomenon, known as the acoustic-seismic induction effect, is described in [1] for instance.

Inspection of the results of simulations yields new features of the propagation of acoustic-gravity waves in the presence of a wind in the atmosphere. These studies establish that the distribution of energy in the transmitted and refracted acoustic-gravity waves in the case of wave bending effect depends on the gradient of the velocity of a wind. In the case of a small gradient wave, bending does not occur. The direction of a wind relative to the propagating wave vector also plays a role.

Conclusion

The proposed approach to stating and solving the problem under consideration enables us to model the effects of wave field propagation for a combined Earth–Atmosphere mathematical model and study the processes of appearance of exchange waves on their boundary. Simulating these processes also enables us to study the specific features of the influence of a wind on the propagating acoustic-gravity waves in the atmosphere and Stoneley surface waves. Inspection of the test calculations shows that the algorithm is stable even for the models of media with sharply contrasting interface of the layers or with thin layers of width comparable to the spatial wavelength.

REFERENCES

1. *Alekseev A. S., Glinskii B. M., Dryakhlov S. I. et al.* The effect of acousto-seismic induction in vibroseismic sounding // Dokl. Akad. Nauk. 1996. V. 346. No. 5. C. 664–667.

2. *Gasilova L. A. and Petukhov Yu. V.* On the theory of surface wave propagation along different interfaces in the atmosphere // Izv. RAN. Fizika Atmosfery i Okeana. 1999. V. 35. No. 1. P. 14–23.

3. *Razin A. V.* Propagation of a spherical acoustic delta wavelet along the gas-solid interface // Izv. RAN. Fizika Zemli. 1993. No. 2. P. 73–77.
4. *Mikhailenko B. G. and Reshetova G. V.* Mathematical simulation of propagation of seismic and acoustic-gravity waves for an inhomogeneous earth-atmosphere model // Geologiya i Geofizika. 2006. V. 47. No. 5. P. 547–556.
5. *Mikhailenko B. G.* Spectral Laguerre method for the approximate solution of time dependent problems // Appl. Math. Let. 1999. N 12. P. 105–110.
6. *Konyukh G. V., Mikhailenko B. G., and Mikhailov A. A.* Application of the integral Laguerre transforms for forward seismic modeling // J. Comput. Acoustics. 2001. V. 9, N 4. P. 1523–1541.
7. *Mikhailenko B. G., Mikhailov A. A., and Reshetova G. V.* Numerical modeling of transient seismic fields in viscoelastic media based on the Laguerre spectral method // Pure Appl. Geophys. 2003. N 160. P. 1207–1224.
8. *Mikhailenko B. G., Mikhailov A. A., and Reshetova G. V.* Numerical viscoelastic modeling by the spectral Laguerre method // Geophys. Prospecting. 2003. N 51. P. 37–48.
9. *Imomnazarov Kh. Kh. and Mikhailov A. A.,* Use of the spectral laguerre method to solve a linear 2d dynamic problem for porous media // Sib. Zh. Ind. Mat. 2008. V. 11. No. 2. P. 86–95.
10. *Suetin P. K.* Classical Orthogonal Polynomials [in Russian]. Moscow: Nauka, 1974.
11. *Virieux J.* P-, SV-wave propagation in heterogeneous media: velocity-stress finite-difference method // Geophysics. 1986. N 51. P. 889–901.
12. *Saad Y. and Van der Vorst H. A.* Iterative solution of linear systems in the 20^{th} century // J. Comput. Appl. Math. 2000. N 123. P. 1–33.
13. *Sonneveld P.* CGS, a fast Lanczos-type solver for nonsymmetric linear system // J. Sci. Statist. Comput. 1989. N 10. P. 36–52.
14. *Mikhailenko B. G.and Mikhaĭlov G. A.* Numerical modeling of acoustic-gravity waves propagation in a heterogeneous "Earth–Atmosphere" model with a wind in the atmosphere // Sib. Zh. Vychisl. Mat. 2014. V. 17. No. 2. P. 149–162.

September 10, 2015

A. A. Mikhaĭlov; V. N. Martynov
Institute of Computational Mathematics and Mathematical Geophysics
Novosibirsk, Russia
`alex_mikh@omzg.sscc.ru`; `vnm@nmsf.sscc.ru`