

СОДЕРЖАНИЕ

Том 47, номер 10, 2007 год

Tikhonov regularization for general nonlinear constrained optimization problem <i>Nguyen Buong</i>	1651
Об устойчивости решения задач о внешней и внутренней оценке выпуклого компакта шаром <i>С. И. Дудов, А. С. Дудова</i>	1657
Численный анализ спектра задачи Орра–Зоммерфельда <i>С. Л. Скорыходов</i>	1672
Численное моделирование перехода к хаосу диссипативного осциллятора Дуффинга с двухчастотным возбуждением <i>Т. В. Завражина</i>	1692
Нелинейный источник в диффузионных методах фильтрации изображений <i>Г. В. Борисенко, А. М. Денисов</i>	1701
Необходимые условия для ϵ -равномерной сходимости разностных схем для параболических уравнений с движущимися пограничными слоями <i>Г. И. Шишкин</i>	1706
Spectral discretizations of 3-d elliptic problems and fast domain decomposition methods <i>V. Korneev, A. Rytov</i>	1727
Асимптотика решения временного уравнения Шрёдингера с малой константой Планка <i>А. С. Омуралиев</i>	1746
Консервативная разностная схема для задачи распространения фемтосекундного лазерного импульса с аксиально-симметричным профилем в среде с кубичной нелинейностью <i>А. Г. Волков, В. А. Трофимов</i>	1752
Численный анализ атомной структуры и формы металлических наночастиц <i>А. В. Вахрушев, А. М. Липанов</i>	1774
Задача построения крылового профиля с выдувом реактивной струи навстречу дозвуковому потоку <i>Н. Б. Ильинский, Р. Ф. Марданов</i>	1784
Логические закономерности в задачах распознавания (параметрический подход) <i>В. В. Рязанов</i>	1793

Contents

Vol. 47, No. 10, 2007

Simultaneous English language translation of the journal is available from Pleiades Publishing, Ltd.
Distributed worldwide by Springer. *Computational Mathematics and Mathematical Physics* ISSN 0965-5425.

Tikhonov Regularization for a General Nonlinear Constrained Optimization Problem <i>Nguyen Buong</i>	1651
On the Stability of Inner and Outer Approximations of a Convex Compact Set by a Ball <i>S. I. Dudov and A. S. Dudova</i>	1657
Numerical Analysis of the Spectrum of the Orr–Sommerfeld Problem <i>S. L. Skorokhodov</i>	1672
Numerical Simulation of the Transition to Chaos in a Dissipative Duffing Oscillator with Two-Frequency Excitation <i>T. V. Zavrzhina</i>	1692
Nonlinear Source in Diffusion Filtering Methods for Image Processing <i>G. V. Borisenko and A. M. Denisov</i>	1701
Necessary Conditions for ε -Uniform Convergence of Finite Difference Schemes for Parabolic Equations with Moving Boundary Layers <i>G. I. Shishkin</i>	1706
Spectral Discretizations of 3-d Elliptic Problems and Fast Domain Decomposition Methods <i>V. Korneev and A. Rytov</i>	1727
Asymptotics of Solutions to Time-Dependent Schrödinger Equation with a Small Planck Constant <i>A. S. Omuraliev</i>	1746
Conservative Finite-Difference Scheme for the Problem of a Femtosecond Laser Pulse with an Axially Symmetric Profile Propagating through a Medium with Cubic Nonlinearity <i>A. G. Volkov and V. A. Trofimov</i>	1752
Numerical Analysis of the Atomic Structure and Shape of Metal Nanoparticles <i>A. V. Vakhrouchev and A. M. Lipanov</i>	1774
Design of a Subsonic Airfoil with Upstream Blowing <i>N. B. Il'inskii and R. F. Mardanov</i>	1784
Logical Regularities in Pattern Recognition (Parametric Approach) <i>V. V. Ryazanov</i>	1793

УДК 519.642.8

TIKHONOV REGULARIZATION FOR GENERAL NONLINEAR CONSTRAINED OPTIMIZATION PROBLEM¹⁾

© 2007 г. Nguyen Buong

(Vietnamse Academy of Science and Technology Institute of Information Technology

18, Hoang Quoc Viet, q. Cau Giay, Ha Noi, Vietnam)

e-mail: nbuong@ioit.ac.vn

Received August 07.2006

Регуляризация по Тихонову для общей задачи нелинейного программирования. Исследуется метод регуляризации по Тихонову для решения общей задачи нелинейного программирования, которая предварительно преобразована в задачу безусловной оптимизации. Свойства стабильности рассмотрены, сходимость метода регуляризации доказана. Библ. 21.

Ключевые слова: метод регуляризации по Тихонову, нелинейная оптимизация, непрерывные функции.

1. INTRODUCTION

Let $f_i, i = 1, 2, \dots, m$, $\tilde{\varphi}_j, j = 1, 2, \dots, N - 1$, and φ_N be nonlinear continuous functions defined on the Euclidian space $\mathbb{R}^n$ with the scalar product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ and $\|\cdot\|_{\mathbb{R}^n}$, respectively. Consider the following general nonlinear constrained optimization problem: find an element $\tilde{x}$ in $\mathbb{R}^n$ such that

$$\varphi_N(\tilde{x}) = \min_{x \in S} \varphi_N(x), \quad (1.1)$$

$$S = \{x \in \mathbb{R}^n : f_i(x) = 0, i = 1, 2, \dots, m; \tilde{\varphi}_j(x) \leq 0, j = 1, 2, \dots, N - 1\}.$$

Set

$$S_0 := \{x \in \mathbb{R}^n : f_i(x) = 0, i = 1, 2, \dots, m\},$$

$$S_j := \{x \in \mathbb{R}^n : \tilde{\varphi}_j(x) \leq 0\}, \quad j = 1, 2, \dots, N - 1.$$

Then, $S = \bigcap_{j=0}^{N-1} S_j$. Assume that $S^0 := \{\tilde{x} \in S : \varphi_N(\tilde{x}) = \min_{x \in S} \varphi_N(x)\} \neq \emptyset$.

For the case $S = \mathbb{R}^n$, we have an unconstrained optimization problem, i.e., the problem of finding an element $\tilde{x} \in \mathbb{R}^n$ such that

$$\varphi_N(\tilde{x}) = \min_{x \in \mathbb{R}^n} \varphi_N(x). \quad (1.2)$$

When φ_N is sufficiently smooth with the gradient function $g(x)$, the approximations for $\tilde{x}$ of (1.2) can be obtained by the conjugate gradient method:

$$x^{k+1} = x^k + \alpha_k d^k, \quad k \geq 1, \quad (1.3)$$

where x^k is the k -th iteration, $\alpha_k > 0$ is steplength, and d^k is a search direction. Normally the search direction at the first iteration is the steepest descent direction, namely $d^1 = -g^1, g^k = g(x^k)$. The other search directions can be defined recursively:

$$d^{k+1} = -g^{k+1} + \beta_k d^k, \quad (1.4)$$

¹⁾ Работа выполнена при финансовой поддержке Vietnamese Fundamental Research Program in Natural Science MS 121104.

$\beta_k \in \mathbb{R}^1$ is so chosen that (1.3), (1.4) reduces to the linear conjugate gradient method if φ_N is a strictly convex quadratic function. The differences in choosing the parameter β_k such as

$$\begin{aligned}\beta_k &= \|g^{k+1}\|_{\mathbb{R}^n}^2 / \|g^k\|_{\mathbb{R}^n}^2, \\ \beta_k &= \langle g^{k+1}, y^k \rangle_{\mathbb{R}^n} / \|g^k\|_{\mathbb{R}^n}^2, \\ \beta_k &= \langle g^{k+1}, y^k \rangle_{\mathbb{R}^n} / \langle d^{k+1}, y^k \rangle_{\mathbb{R}^n},\end{aligned}$$

where $y^k = g^{k+1} - g^k$, give the formulaes Fletcher–Reeves, Polak–Ribière–Polyak, and Hestenes–Stiefel (see [1]–[4]). Then, this method is developed in [5]–[7] by Yuan.

For solving (1.2) there exists also the other type of iteration, named the trust region method. The first trust region method is given by Powell [8]. At the k -th iteration, the trial step is computed by

$$\begin{aligned}\min_{d \in \mathbb{R}^n} \langle g^k, d \rangle_{\mathbb{R}^n} + \frac{1}{2} \langle d, B_k d \rangle_{\mathbb{R}^n}, \\ s.t. \|d\|_{\mathbb{R}^n} \leq \Delta_k,\end{aligned}\tag{1.5}$$

where B_k is an $n \times n$ symmetric matrix which approximates the Hessian and $\Delta_k > 0$ is a trust region radius. The convergence of the method are studied in [9], [10].

The two classes of methods (1.4) and (1.5) can be applied to solve equality constrained optimization problem in [11]–[13].

Some other iteration methods are presented in [14, p. 375], [15].

Solving an unconstrained optimization problem is simpler than solving a constrained one, in general. So, the aim of this paper is to transfer the stated general nonlinear optimization problem into unconstrained optimization one depending on parameter, by using Tikhonov regularization method. To do this, set $\varphi_j(x) = \max\{0, \tilde{\varphi}_j(x)\}$, $j = 1, 2, \dots, N-1$. Clearly, φ_j are also continuous,

$$S_j := \left\{ \bar{x} \in \mathbb{R}^n : \varphi_j(\bar{x}) = \min_{x \in \mathbb{R}^n} \varphi_j(x) \right\}, \quad j = 1, 2, \dots, N-1,\tag{1.6}$$

$\varphi_j(x) \geq 0 \ \forall x \in \mathbb{R}^n$ and $\varphi_j(y) = 0 \ \forall y \in S_j$.

When $S_0 = \mathbb{R}^n$, and φ_j , $j = 1, 2, \dots, N$, are convex function, problem (1.1) was studied in [16] on the base of convex analysis by transforming it into an unconstrained vector optimization. The idea in [16] is developed here for the case where φ_j are not necessary to be convex, and maybe S_0 is not the whole space $\mathbb{R}^n$. For this purpose, consider the following vector optimization problem with equality constraints: find an element $\tilde{x} \in \mathbb{R}^n$ such that

$$\begin{aligned}\varphi_j(\tilde{x}) &= \min_{x \in S_0} \varphi_j(x), \quad j = 1, 2, \dots, N, \\ S_0 &= \{x : F(x) = 0\},\end{aligned}\tag{1.7}$$

where $F(x) = (f_1(x), \dots, f_m(x))^T$.

Note that problem (1.7) with $N = 1$ is studied in [17]–[20] for a monotone operator $F : X \rightarrow X^*$, where X denotes a reflexive Banach space and X^* is its dual space.

Further, to solve (1.7), consequently (1.1) (see the proof of theorem 1 in [16]), consider the following unconstrained optimization problem: find an element $x_\alpha \in \mathbb{R}^n$ such that

$$\begin{aligned}\mathcal{F}_\alpha(x_\alpha) &= \min_{x \in \mathbb{R}^n} \mathcal{F}_\alpha(x), \quad \alpha > 0, \\ \mathcal{F}_\alpha(x) &= \|F(x)\|_{\mathbb{R}^n}^2 + \sum_{j=1}^N \alpha^{\mu_j} \varphi_j(x) + \alpha \|x - x^*\|_{\mathbb{R}^n}^2,\end{aligned}\tag{1.8}$$

$$0 \leq \mu_1 < \dots < \mu_N < 1, \quad j = 2, 3, \dots, N-1,$$

where x^* is some element in $\mathbb{R}^n$.

It is well-known [14] that problem (1.8) has a solution x_α for each $\alpha > 0$. Without loss of generality, assume that $\varphi_N(x) \geq 0 \quad \forall x \in \mathbb{R}^n$.

In addition, assume that φ_N has the bounded level sets, i.e., the set $\{x \in \mathbb{R}^n : \varphi_N(x) \leq c\}$ is bounded for every $c > 0$.

2. MAIN RESULTS

Theorem 1. Let $\alpha_k \rightarrow 0$, as $k \rightarrow \infty$. Then every sequence $\{x^k\}$, where $x^k := x_{\alpha_k}$ is a solution of (1.8) with α replaced by α_k , has a convergent subsequence. The limit of every convergent subsequence is a solution of (1.1). If, in addition, the solution $\tilde{x}$ is unique, then

$$\lim_{k \rightarrow \infty} x^k = \tilde{x}.$$

Proof. From (1.8) it follows

$$\begin{aligned} \|F(x^k)\|_{\mathbb{R}^m}^2 + \sum_{j=1}^N \alpha_k^{\mu_j} \varphi_j(x^k) + \alpha_k \|x^k - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \|F(y)\|_{\mathbb{R}^m}^2 + \sum_{j=1}^N \alpha_k^{\mu_j} \varphi_j(y) + \alpha_k \|y - x^*\|_{\mathbb{R}^n}^2, \end{aligned} \quad (2.1)$$

for each fixed element $y \in \mathbb{R}^n$. Taking $y \in S$, we have $F(y) = 0$, and from (1.6) it deduces $\varphi_j(x^k) \geq \varphi_j(y) = 0$, $j = 1, 2, \dots, N-1$. Then, from (2.1) it implies

$$\varphi_N(x^k) \leq \varphi_N(y) + \alpha_k^{1-\mu_N} \|y - x^*\|_{\mathbb{R}^n}^2. \quad (2.2)$$

Since φ_N is level-bounded, then $\{x^k\}$ is bounded. Let $\{x^l\} \subset \{x^k\}$ be such that $x^l \rightarrow \bar{x}$ as $l \rightarrow \infty$. We shall prove that $\bar{x}$ is a solution of (1.1). Indeed, from (2.1) and $\varphi_N(x) \geq 0 \quad \forall x \in \mathbb{R}^n$ we obtain

$$0 \leq \|F(x^l)\|_{\mathbb{R}^m}^2 \leq \alpha_l^{\mu_N} \varphi_N(y) + \alpha_l \|y - x^*\|_{\mathbb{R}^n}^2.$$

Tending $l \rightarrow \infty$ in the last inequality, the continuous property of F gives $F(\bar{x}) = 0$, i.e., $\bar{x} \in S_0$. Now, we prove that $\bar{x} \in S_1$. For any element $y \in S_0$, from (2.1) and $\varphi_j(x) \geq 0 \quad \forall x \in \mathbb{R}^n$, $j = 1, 2, \dots, N-1$, we can write

$$\varphi_1(x^l) + \alpha_l^{1-\mu_1} \|x^l - x^*\|_{\mathbb{R}^n}^2 \leq \varphi_1(y) + \sum_{j=2}^N \alpha_l^{\mu_j - \mu_1} \varphi_j(y) + \alpha_l^{1-\mu_1} \|y - x^*\|_{\mathbb{R}^n}^2.$$

After passing $l \rightarrow \infty$ in the last inequality we obtain $\varphi_1(\bar{x}) \leq \varphi_1(y) \quad \forall y \in S_0$. Note that $\bar{x}$ is a local minimizer of φ_1 on S_0 . But, because of $S_1 \cap S_0 \neq \emptyset$, $\bar{x}$ also in a global minimizer of φ_1 on $\mathbb{R}^n$. It means that $\bar{x} \in S_1$. Further, we prove that $\bar{x} \in S_2$. For any $y \in S_0 \cap S_1$, from (2.1) and $\varphi_j(x) \geq 0 \quad \forall x \in \mathbb{R}^n$, $j = 1, 2, \dots, N$ we have

$$\varphi_2(x^l) + \alpha_l^{1-\mu_2} \|x^l - x^*\|_{\mathbb{R}^n}^2 \leq \varphi_2(y) + \sum_{j=3}^N \alpha_l^{\mu_j - \mu_2} \varphi_j(y) + \alpha_l^{1-\mu_2} \|y - x^*\|_{\mathbb{R}^n}^2.$$

By the similar argument, we obtain $\bar{x} \in S_2$, and $\bar{x} \in S_j$, $j = 3, 4, \dots, N-1$. Consequently, $\bar{x} \in S$. In finall, we have to prove that $\bar{x}$ is a solution of (1.1). As above, for any $y \in S$, from (2.2) it deduces that $\varphi_N(\bar{x}) \leq \varphi_N(y) \quad \forall y \in S$. Hence, $\varphi_N(\bar{x}) = \min_{y \in S} \varphi_N(y)$. Theorem is proved now.

Under the above conditions (1.1) is ill-posed. By this we mean that the set S^0 do not depend continuously on the data $\{f_i, \varphi_j\}$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, N$. Consequently, to obtain approximation solutions we need

to use stable methods. Assume that instead of $\{f_i, \varphi_j\}$ we have their continuous approximations $\{f_i^\delta, \varphi_j^\delta\}$ such that

$$\begin{aligned} |f_i(x) - f_i^\delta(x)| &\leq \delta, \quad i = 1, 2, \dots, m, \\ |\varphi_j(x) - \varphi_j^\delta(x)| &\leq \delta, \quad j = 1, 2, \dots, N, \quad \forall x \in \mathbb{R}^n, \quad \delta \rightarrow 0. \end{aligned} \quad (2.3)$$

In [21] one proposed the regularization method that minimizers the functional

$$\Phi_{\alpha\lambda}(x) = \|F^\delta(x)\|_{\mathbb{R}^m}^2 + \alpha(\varphi_N^2(x) + \lambda\Omega(x)), \quad \alpha, \lambda > 0,$$

on the subset $C := \bigcap_{j=1}^{N-1} S_j$, where $F^\delta(x) = (f_1^\delta(x), \dots, f_m^\delta(x))^T$, α, λ are the small parameters, and $\Omega(x)$ is a stabilizing term. This is a constrained optimization problem, too.

For the case where $\{f_i, \varphi_j\}$ are given approximately by $\{f_i^\delta, \varphi_j^\delta\}$ satisfying (2.3) we consider the following unconstrained optimization problem: find an element $x_\alpha^\delta \in \mathbb{R}^n$ such that

$$\begin{aligned} \mathcal{F}_\alpha^\delta(x_\alpha^\delta) &= \min_{x \in \mathbb{R}^n} \mathcal{F}_\alpha^\delta(x), \quad \alpha > 0, \quad \delta \geq 0, \\ \mathcal{F}_\alpha^\delta(x) &= \|F^\delta(x)\|_{\mathbb{R}^m}^2 + \sum_{j=1}^N \alpha^{\mu_j} \varphi_j^\delta(x) + \alpha \|x - x^*\|_{\mathbb{R}^n}^2. \end{aligned} \quad (2.4)$$

As spoken above, for each $\alpha > 0$ problem (2.4) possesses a solution denoted by x_α^δ .

We have the result.

Theorem 2. Let $\alpha_k, \delta_k \rightarrow 0$ so that $\delta_k/\alpha_k \rightarrow 0$, as $k \rightarrow \infty$. Then every sequence $\{x^k\}$, where $x^k := x_{\alpha_k}^{\delta_k}$ is a solution of (2.4) with α, δ replaced by α_k, δ_k , respectively, has a convergent subsequence. The limit of every convergent subsequence is a solution of (1.1). If, in addition, the solution $\tilde{x}$ is unique, then

$$\lim_{k \rightarrow \infty} x^k = \tilde{x}.$$

Proof. From (2.4) we have

$$\begin{aligned} \|F^\delta(x^k)\|_{\mathbb{R}^m}^2 + \sum_{j=1}^N \alpha_k^{\mu_j} \varphi_j^{\delta_k}(x^k) + \alpha_k \|x^k - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \|F^{\delta_k}(y)\|_{\mathbb{R}^m}^2 + \sum_{j=1}^N \alpha_k^{\mu_j} \varphi_j^{\delta_k}(y) + \alpha_k \|y - x^*\|_{\mathbb{R}^n}^2, \end{aligned} \quad (2.5)$$

for every fixed element $y \in \mathbb{R}^n$. Consequently,

$$\begin{aligned} &\|F^{\delta_k}(x^k)\|_{\mathbb{R}^m}^2 + \alpha_k^{\mu_N} \varphi_N^{\delta_k}(x^k) + \alpha_k \|x^k - x^*\|_{\mathbb{R}^n}^2 \leq \\ &\leq \|F^{\delta_k}(y)\|_{\mathbb{R}^m}^2 + \alpha_k^{\mu_N} \varphi_N^{\delta_k}(y) + \sum_{j=1}^{N-1} \alpha_k^{\mu_j} [\varphi_j^{\delta_k}(y) - \varphi_j^{\delta_k}(x^k) + \varphi_j^{\delta_k}(x^k) - \varphi_j^{\delta_k}(x^n)] + \alpha_k \|y - x^*\|_{\mathbb{R}^n}^2. \end{aligned} \quad (2.6)$$

By taking $y \in S$, we have $F(y) = 0$, and $\varphi_j(x^k) \geq \varphi_j(y) = 0$, $j = 1, 2, \dots, N-1$. Then, from (2.3) and (2.6) in implies

$$\varphi_N^{\delta_k}(x^k) \leq \varphi_N^{\delta_k}(y) + \frac{m\delta_k^2}{\alpha_k^{\mu_N}} + 2 \frac{\delta_k}{\alpha_k^{\mu_N}} \sum_{j=1}^{N-1} \alpha_k^{\mu_j} + \alpha_k^{1-\mu_N} \|y - x^*\|_{\mathbb{R}^n}^2$$

or

$$\varphi_N(x^k) \leq \varphi_N(y) + 2\delta_k + \frac{m\delta_k^2}{\alpha_k^{\mu_N}} + 2 \frac{\delta_k}{\alpha_k^{\mu_N}} \sum_{j=1}^{N-1} \alpha_k^{\mu_j} + \alpha_k^{1-\mu_N} \|y - x^*\|_{\mathbb{R}^n}. \quad (2.7)$$

Since $\delta_k, \delta_k/\alpha_k \rightarrow 0$ as $k \rightarrow \infty$, and φ_N is level-bounded, then $\{x^k\}$ is bounded. Let $\{x^l\} \subset \{x^k\}$ such that $x^l \rightarrow \bar{x}$ as $l \rightarrow \infty$. We shall prove that $\bar{x}$ is a solution of (1.1). Indeed, from (2.6), (2.7) and $\varphi_N(x) \geq 0 \forall x \in \mathbb{R}^n$ it follows

$$0 \leq \|F(x^l)\|_{\mathbb{R}^m}^2 \leq \alpha_l^{\mu_N} \varphi_N(y) + 2\delta_l \alpha_l^{\mu_N} + m\delta_l^2 + 2\delta_l \sum_{j=1}^{N-1} \alpha_l^{\mu_j} + \alpha_l \|y - x^*\|_{\mathbb{R}^n}.$$

Tending $l \rightarrow \infty$ in the last inequality, the continuous property of F gives $F(\bar{x}) = 0$, i.e., $\bar{x} \in S_0$. Now, we prove that $\bar{x} \in S_1$. For any element $y \in S_0$, from (2.5) we also obtain

$$\begin{aligned} \varphi_1(x^l) + \sum_{j=2}^N \alpha_l^{\mu_j - \mu_1} \varphi_j(x^l) + \alpha_l^{1-\mu_1} \|x^l - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \varphi_1(y) + \sum_{j=2}^N \alpha_l^{\mu_j - \mu_1} \varphi_j(y) + \alpha_l^{1-\mu_1} \|y - x^*\|_{\mathbb{R}^n}^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \varphi_1(x^l) + \sum_{j=2}^N \alpha_l^{\mu_j - \mu_1} \varphi_j(x^l) + \alpha_l^{1-\mu_1} \|x^l - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \varphi_1(y) + 2\delta_l + \sum_{j=2}^N \alpha_l^{\mu_j - \mu_1} \varphi_j(y) + \alpha_l^{1-\mu_1} \|y - x^*\|_{\mathbb{R}^n}^2. \end{aligned} \quad (2.8)$$

After passing $l \rightarrow \infty$ in the last inequality we obtain $\varphi_1(\bar{x}) \leq \varphi_1(y) \forall y \in S_0$. It means that $\bar{x} \in S_1$. Further, we prove that $\bar{x} \in S_2$. For any $y \in S_0 \cap S_1$, again from (2.8) and $\varphi_1(x^l) \geq \varphi_1(y)$ we can write

$$\begin{aligned} \varphi_2(x^l) + \sum_{j=3}^N \alpha_l^{\mu_j - \mu_2} \varphi_j(x^l) + \alpha_l^{1-\mu_2} \|x^l - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \frac{2\delta_l \alpha_l^{\mu_1}}{\alpha_l^{\mu_2}} + \varphi_2(y) + \sum_{j=3}^N \alpha_l^{\mu_j - \mu_2} \varphi_j(y) + \alpha_l^{1-\mu_2} \|y - x^*\|_{\mathbb{R}^n}^2. \end{aligned}$$

By the similar argument, we obtain $\bar{x} \in S_2$. Now, suppose that we have proved $\bar{x} \in \bigcap_{j=0}^{p-1} S_j$, and need to show that $\bar{x} \in S_p$. Taking $y \in \bigcap_{j=0}^p S_j$, from (1.6), (2.5) and $\varphi_j(x^l) \geq \varphi_j(y) = 0, j = 1, 2, \dots, p-1$, it is obvious

$$\begin{aligned} \varphi_p(x^l) + \sum_{j=p+1}^N \alpha_l^{\mu_j - \mu_p} \varphi_j(x^l) + \alpha_l^{1-\mu_p} \|x^l - x^*\|_{\mathbb{R}^n}^2 &\leq \\ &\leq \frac{2\delta_l}{\alpha_l^{\mu_p}} \sum_{j=1}^{p-1} \alpha_l^{\mu_j} + \varphi_p(y) + \sum_{j=p+1}^N \alpha_l^{\mu_j - \mu_p} \varphi_j(y) + \alpha_l^{1-\mu_p} \|y - x^*\|_{\mathbb{R}^n}^2. \end{aligned}$$

It means that $\bar{x} \in S_p$. Consequently, $\bar{x} \in S$. In final, we have to prove that $\bar{x}$ is a solution of (1.1). As above, for any $y \in S$, from (2.7) it deduces that $\varphi_N(\bar{x}) \leq \varphi_N(y) \forall y \in S$. Hence, $\varphi_N(\bar{x}) = \min_{y \in S} \varphi_N(y)$. Theorem is proved now.

REFERENCES

1. *Fletcher R., Reeves C.* Function minimization by conjugate gradients // *Comput. J.* 1964. V. 7. P. 149–154.
2. *Polak E., Ribière G.* Note sur la convergence de directions conjuguées // *Rev. Francaise Inform. Rech. Operat.*, 3e Année. 1969. V. 16. P. 35–43.
3. *Поляк Б.Т.* Метод сопряженных градиентов в задачах на экстремум // *Ж. вычисл. матем. и матем. физ.* 1969. Т. 9. № 4. С. 807–821.
4. *Hestenes M.R., Stiefel E.L.* Methods of conjugate gradients for solving linear systems // *Res. Nat. Bur. Standards Sect.* 1952. V. 5. P. 409–436.
5. *Yuan Y.* On the truncated conjugate gradient method // *Math. Program.* 2000. V. 87. P. 561–571.
6. *Yuan Y.* A class of globally convergent conjugate gradient methods // *Science in China.* 2003. V. 46. P. 253–261.
7. *Yuan Y.* A new stepsize for the steepest descent method // *J. Comput. Math.* 2006. V. 24. P. 149–156.
8. *Powell M.J.D.* A new algorithm for unconstrained optimization // *Nonlinear Program.* 31–66. New York: Acad. Press, 1970.
9. *Powell M.J.D.* Convergence properties of a class of minimization problems // *Nonlinear Program.* 2, 1–27. New York: Acad. Press, 1975.
10. *Powell M.J.D.* On the global convergence of trust region algorithm for unconstrained optimization // *Math. Program.* 1984. V. 29. P. 297–303.
11. *Powell M.J.D., Yuan Y.* A trust region algorithm for equality constrained optimization // *Math. Program.* 1991. V. 49. P. 189–211.
12. *Vardi A.* A trust region algorithm for equality constrained minimization: convergence properties and implementation // *SIAM J. Numer. Analys.* 1985. V. 4. P. 575–591.
13. *Yuan Y.* On the convergence of a new trust region algorithm // *Numer. Math.* 1995. V. 70. P. 515–539.
14. *Васильев Ф.П.* Численные методы решения экстремальных задач. М.: Наука, 1988.
15. *Васильев Ф.П.* Методы оптимизации. М.: Факториал Пресс, 2002.
16. *Nguyen Buong.* Regularization for unconstrained vector optimization of convex functionals in Banach spaces // *Ж. вычисл. матем. и матем. физ.* 2006. Т. 46. № 3. С. 372–378.
17. *Nguyen Buong.* Convergence rates and finite-dimensional approximations for a class of ill-posed variational inequalities // *Украинский матем. журнал.* 1996. Т. 48. С. 1–9.
18. *Nguyen Buong, Phan Vam Loi.* On parameter choice and convergence rates in regularization for a class of ill-posed variational inequalities // *Ж. вычисл. матем. и матем. физ.* 2004. Т. 44. № 10. С. 1735–1744.
19. *Kluge R.* Optimal control with minimum problems and variational inequalities // *Lect. Notes Comput. Sci.* 1975. V. 27. P. 375–382.
20. *Рязанцева И.П.* Операторный метод регуляризации для оптимального решения задач с монотонными отображениями // *Сибирский матем. журнал.* 1983. Т. 24. С. 455–456.
21. *Тихонов А.Н., Арсенин В.Я.* Методы решения плохо обусловленных задач. М.: Наука, 1979.

УДК 519.626

ОБ УСТОЙЧИВОСТИ РЕШЕНИЯ ЗАДАЧ О ВНЕШНЕЙ И ВНУТРЕННЕЙ ОЦЕНКЕ ВЫПУКЛОГО КОМПАКТА ШАРОМ¹⁾

© 2007 г. С. И. Дудов, А. С. Дудова

(410012 Саратов, ул. Астраханская, 83, Саратовский гос. ун-т)

e-mail: dudovsi@info.sgu.ru

Поступила в редакцию 03.05.2007 г.

Рассматриваются конечномерные задачи о внешней и внутренней оценке выпуклого компакта шаром некоторой нормы (задачи об описанном и вписанном шаре). Дается общая характеристика устойчивости решения относительно погрешности задания оцениваемого компакта. Получен новый критерий решения задачи о внешней оценке в форме, связывающей ее с задачей о внутренней оценке нижнего лебегова множества функции расстояния до самой дальней точки оцениваемого компакта. При дополнительном предположении о сильной выпуклости компакта дается количественная оценка устойчивости центра вписанного шара. В предположении сильной квазивыпуклости используемой нормы получена количественная оценка устойчивости центра описанного шара. Библ. 24.

Ключевые слова: оценка выпуклого компакта шаром, устойчивость решения, внешняя и внутренняя оценки, сильная выпуклость.

1. ВВЕДЕНИЕ

Интерес математиков к оценке и аппроксимации достаточно сложных множеств множествами простой геометрической структуры возник очень давно (см., например, [1], [2] и библиографии в них) и возобновлялся по мере появления в математике новых средств исследования, позволяющих рассматривать как новые задачи, так и старые задачи, но на более высоком уровне. Ныне это направление поддерживается в рамках негладкого анализа и недифференцируемой оптимизации, которые дают эффективные инструменты для исследования таких задач.

Задачи по оценке множеств, а также многозначных отображений находят обширные приложения в естествознании, в том числе и в самой математике. Известны многочисленные работы, связанные с внешними и внутренними эллипсоидальными оценками множеств и многозначных отображений (см., например, [3], [4]). Для техники особое значение имеют внешние и внутренние оценки множеств ориентированными параллелепипедами (см., например, [5], [6]). Рассматривались аппроксимации выпуклого компакта линейными комбинациями заданного набора выпуклых компактов простого вида (см., например, [7]), а также внутренние и внешние аппроксимации выпуклого компакта многогранниками (см., например, [8], [9]) и другие задачи.

Наряду с эллипсоидом и многогранником, к числу наиболее простых множеств относится шар любой нормы. В данной статье на предмет устойчивости решения рассматриваются две известные задачи по оценке множества шаром некоторой нормы.

Первая из них – задача об описанном шаре или чебышёвском центре множества. Если D – оцениваемый компакт из конечномерного пространства $\mathbb{R}^p$, а функция $n(x)$ – используемая норма на $\mathbb{R}^p$, то задачу можно записать в виде

$$R(x, D) \equiv \max_{y \in D} n(x - y) \longrightarrow \min_{x \in \mathbb{R}^p}. \quad (1.1)$$

Эту задачу будем также называть задачей о внешней оценке компакта D шаром нормы $n(\cdot)$. Она требует построения шара наименьшего радиуса, содержащего оцениваемый компакт D . Поскольку замена компакта D в задаче (1.1) на его выпуклую оболочку не меняет ее решения, то без потери общности будем далее считать компакт D выпуклым.

¹⁾Работа выполнена при финансовой поддержке РФФИ (код проекта 06-01-00003).

Вторая задача – задача о вписанном шаре или задача о внутренней оценке выпуклого компакта D шаром нормы $n(\cdot)$. Ее можно записать в виде

$$\rho(x, \Omega) \equiv \min_{y \in \Omega} n(x - y) \longrightarrow \max_{x \in D}. \quad (1.2)$$

Здесь $\Omega = \overline{\mathbb{R}^p \setminus D}$, а функция $\rho(x, \Omega)$ выражает расстояние от точки x до множества Ω в норме $n(\cdot)$. Задача (1.2) требует построения шара нормы $n(\cdot)$ с максимальным радиусом, который содержится в D .

О некоторых свойствах целевых функций задач (1.1) и (1.2), а также самих решений как для случая евклидовой нормы, так и для произвольной см., например, [1], [10]–[14].

В этой работе исследуем задачи (1.1) и (1.2) на устойчивость решения относительно погрешности задания оцениваемого выпуклого компакта D . Уточним постановку вопроса.

В практических ситуациях информация о компакте D может носить приближенный характер, т.е. вместо D нам может быть известен другой выпуклый компакт D_ε такой, что $h(D, D_\varepsilon) \leq \varepsilon$. Здесь $\varepsilon > 0$ – известная погрешность задания компакта D , а

$$h(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \|a - b\|, \sup_{b \in B} \inf_{a \in A} \|a - b\| \right\} \quad (1.3)$$

есть расстояние Хаусдорфа между множествами A и B в евклидовой норме $\|\cdot\|$. Таким образом, о решениях (1.1) и (1.2) мы можем судить по решению приближенных задач

$$R(x, D_\varepsilon) \equiv \max_{y \in D_\varepsilon} n(x - y) \longrightarrow \min_{x \in \mathbb{R}^p}, \quad (1.4)$$

$$\rho(x, \Omega_\varepsilon) \equiv \min_{y \in \Omega_\varepsilon} n(x - y) \longrightarrow \max_{x \in D_\varepsilon}, \quad (1.5)$$

где $\Omega_\varepsilon = \overline{\mathbb{R}^p \setminus D_\varepsilon}$. Требуется охарактеризовать устойчивость решений задач в целом и по возможности дать количественную оценку устойчивости через ε оптимальных значений целевых функций и самих множеств решений, т.е. радиусов и центров описанных и вписанных шаров.

Отметим, что исследование устойчивости решения задачи выпуклого программирования, как известно (см., например, [15]), может опираться на сильную выпуклость (см. [16, с. 181]) целевой функции, если она имеется. Однако функция $R(x, D)$, являясь выпуклой по x на $\mathbb{R}^p$, ни при каких условиях на D и $n(\cdot)$ не является сильно выпуклой на каком-либо телесном выпуклом множестве. А функция $\rho(x, \Omega)$ является вогнутой по x на выпуклом компакте D , но на любом его выпуклом телесном подмножестве не является сильно вогнутой. Уточняющие комментарии по этому поводу имеются в разд. 2.

Содержание работы таково. В разд. 2 приводятся используемые вспомогательные понятия и факты. Главными из них являются сведения из сильно выпуклого анализа (см. [9]), лежащие в основе получения количественной оценки устойчивости центров описанного и вписанного шаров. В разд. 3 дана общая характеристика устойчивости решения задач (1.1) и (1.2), т.е. не требующая каких-либо дополнительных условий на D и $n(\cdot)$. В разд. 4 получен новый критерий решения задачи (1.1) о внешней оценке в форме, связывающей ее с задачей о внутренней оценке нижнего лебегова множества функции $R(x, D)$. Именно он оказался удобным для оценки устойчивости центра вписанного шара. В разд. 5 при дополнительном предположении о сильной выпуклости компакта D получена количественная оценка устойчивости центра вписанного шара, а в предположении сильной квазивыпуклости используемой нормы $n(\cdot)$ дана оценка устойчивости центра описанного шара.

Авторам неизвестны работы, где бы ранее ставился вопрос об устойчивости центра вписанного или описанного шара. Дело, по-видимому, в том, что, как показывает один из приведенных в разд. 5 примеров, без какой-либо количественной характеристики выпуклости оцениваемого компакта получить количественную оценку устойчивости центра вписанного шара нельзя. Понятие сильно выпуклого множества как раз дает пример количественной характеристики выпуклости множества.

2. ВСПОМОГАТЕЛЬНЫЕ ФАКТЫ

Для общей характеристики и оценки устойчивости решения задач (1.1) и (1.2) нам будут нужны некоторые сведения о свойствах функций $R(x, D)$ и $\rho(x, \Omega)$, а также понятия и факты из сильно выпуклого анализа (см. [8], [9]). Кроме уже введенных будем использовать следующие обозна-

чения: $\bar{A}$, $\text{int}A$, $\text{co}A$ – соответственно, замыкание, внутренность, выпуклая оболочка множества A ; $A \dot{-} B = \{c : c + B \subset A\}$ – разность Понтрягина множеств A и B ;

$$Bn(x, r) = \{y \in \mathbb{R}^p : n(x - y) \leq r\}, \quad Sn(x, r) = \{y \in \mathbb{R}^p : n(x - y) = r\}$$

суть шар и сфера нормы $n(\cdot)$ с центром в точке x и радиусом r ; $B(x, r)$, $S(x, r)$ – шар и сфера евклидовой нормы с центром в точке x и радиусом r ;

$$Q^R(x, D) = \{y \in D : R(x, D) = n(x - y)\}$$

есть множество точек касания компакта D и сферы $Sn(x, R(x))$; $Q^p(x, \Omega) = \{y \in \Omega : \rho(x, \Omega) = n(x - y)\}$ – проекция точки x на множество Ω ; $\langle x, y \rangle$ – скалярное произведение элементов x и $y \in \mathbb{R}^p$; $K(A) = \{v \in \mathbb{R}^p : \exists \alpha \geq 0, a \in A, v = \alpha a\}$, $K^+ = \{w \in \mathbb{R}^p : \langle v, w \rangle \geq 0, \forall v \in K\}$; $K(x, A)$ – конус возможных направлений множества A в точке x (см. [12, с. 31]);

$$n^*(w) = \max_{v: n(v) \leq 1} \langle v, w \rangle$$

есть полярная норма к $n(\cdot)$;

$$G^R(\lambda) = \{x \in \mathbb{R}^p : R(x, D) \leq \lambda\}, \quad \Omega^R(\lambda) = \overline{\mathbb{R}^p \setminus G^R(\lambda)},$$

$$D(x) = \{y \in \mathbb{R}^p : \rho(y, \Omega) \geq \rho(x, \Omega)\}, \quad \Omega(x) = \{y \in \mathbb{R}^p : \rho(y, \Omega) \leq \rho(x, \Omega)\};$$

$$\text{diam } A = \sup_{x, y \in A} \|x - y\|$$

есть диаметр множества A в евклидовой норме, $0_p = (0, \dots, 0) \in \mathbb{R}^p$.

2.1. Известно, что все нормы на конечномерном пространстве $\mathbb{R}^p$ являются эквивалентными и, следовательно, для конкретно выбранной нормы $n(\cdot)$ найдутся положительные константы C_1 и C_2 такие, что

$$C_1 \|x\| \leq n(x) \leq C_2 \|x\| \quad \forall x \in \mathbb{R}^p. \quad (2.1)$$

Далее по тексту, не оговаривая каждый раз специально, будем считать, что константы C_1 и C_2 соответствуют неравенствам (2.1). Кроме того, любая норма является конечной и выпуклой на $\mathbb{R}^p$ функцией. Субдифференциал нормы можно записать в виде (см. [11, с. 161])

$$\partial n(0_p) = \{v \in \mathbb{R}^p : n^*(v) \leq 1\}, \quad (2.2)$$

$$\partial n(x) = \{v \in \mathbb{R}^p : n^*(v) = 1, n(x) = \langle v, x \rangle\}, \quad x \neq 0_p.$$

Теперь напомним некоторые общие свойства функций $R(x, D)$ и $\rho(x, \Omega)$.

Теорема 1 (см. [11, с. 163]). Функция $R(x, D)$ является выпуклой по x на $\mathbb{R}^p$, а ее субдифференциал может быть выражен формулой

$$\underline{\partial} R(x, D) = \text{co}\{\partial n(x - z) : z \in Q^R(x, D)\}. \quad (2.3)$$

Теорема 2 (см. [17]). Функция $\rho(x, \Omega)$ является вогнутой по x на выпуклом множестве D , а ее супердифференциал в точках $x \in \text{int}D$ может быть выражен формулой

$$\overline{\partial} \rho(x, \Omega) = \text{co}\{\partial n(x - z) \cap K^+(z, D) : z \in Q^p(x, \Omega)\}. \quad (2.4)$$

В [18] доказана

Лемма 1. Если точки x и y таковы, что $\rho(y, \Omega) \geq \rho(x, \Omega)$, то

$$\rho(y, \Omega) = \rho(y, \Omega(x)) + \rho(x, \Omega).$$

Из (2.1) легко следует

Лемма 2. Если радиус максимального шара нормы $n(\cdot)$, вложенного в множество D , равен δ , то радиус максимального евклидова шара, вложенного в то же множество D , не более чем δ/C_1 .

2.2. Далее будем использовать некоторые понятия и факты из сильно выпуклого анализа (см. [8], [9]).

Определение 1. Множество $A \subset \mathbb{R}^p$ называется r -сильно выпуклым, если оно представимо в виде пересечения замкнутых евклидовых шаров радиуса r .

Определение 2. Пусть A – ограниченное множество из $\mathbb{R}^p$, а числа $r \geq \rho > 0$ такие, что $B(0_\rho, \rho) \stackrel{*}{\subset} A \neq \emptyset$. Сильно выпуклой оболочкой радиуса r множества A называется множество, получаемое при пересечении всех замкнутых евклидовых шаров радиуса r , которые содержат данное множество. Она обозначается через $\text{strco}_r A$.

Теорема 3 (см. [9, с. 298]). Пусть a_0 и a_1 из $\mathbb{R}^p$ таковы, что $0 < \|a_0 - a_1\| < 2r$. Тогда r -сильно выпуклая оболочка множества, состоящего из точек a_0 и a_1 , выражается следующим образом:

$$\text{strco}_r(\{a_0\} \cup \{a_1\}) = \bigcup_{\alpha \in [0, 1]} B(a(\alpha), r(\alpha)),$$

где

$$a(\alpha) = (1 - \alpha)a_0 + \alpha a_1, \quad r(\alpha) = r - \sqrt{r^2 - \alpha(1 - \alpha)\|a_0 - a_1\|^2}. \quad (2.5)$$

Теорема 4 (см. [9, с. 302]). Компактное множество A является r -сильно выпуклым множеством тогда и только тогда, когда r -сильно выпуклая оболочка произвольного конечного подмножества его точек непуста и содержится в A .

Докажем следующий факт.

Лемма 3. Пусть A есть r -сильно выпуклое множество. Если радиус наибольшего евклидова шара, вложенного в A , не превосходит $\delta > 0$, то

$$\text{diam } A \leq 2\sqrt{\delta(2r - \delta)}.$$

Доказательство. Предположим противное. Тогда существуют точки a_0 и a_1 из A такие, что

$$\|a_0 - a_1\| > 2\sqrt{\delta(2r - \delta)}. \quad (2.6)$$

Так как, по теореме 4, имеет место включение

$$\text{strco}_r(\{a_0\} \cup \{a_1\}) \subset A,$$

то из теоремы 3 вытекает, что

$$B(a(\alpha), r(\alpha)) \subset A \quad \forall \alpha \in [0, 1], \quad (2.7)$$

где $a(\alpha)$, $r(\alpha)$ определены формулами (2.5). Для значения $\alpha = 0.5$, используя (2.6), получаем

$$r(0.5) = r - \sqrt{r^2 - \frac{1}{4}\|a_0 - a_1\|^2} > r - \sqrt{r^2 - \delta(2r - \delta)} = \delta. \quad (2.8)$$

Здесь следует иметь в виду, что максимальный радиус вложенного евклидова шара не может превосходить радиус сильной выпуклости множества, т.е. $\delta \leq r$. Таким образом, соотношения (2.7) и (2.8) говорят о том, что радиус вложенного в A шара с центром в точке $a(0.5)$ больше δ . Это противоречит условию леммы.

2.3. В некоторых случаях будем накладывать на норму специальные дополнительные условия.

Определение 3. Будем говорить, что норма $n(\cdot)$ является строго (r -сильно) квазивыпуклой, если ее шар единичного радиуса является строго (r -сильно) выпуклым множеством.

Очевидно, сильно выпуклое множество всегда является строго выпуклым, а обратное верно не всегда. Соответственно, сильно квазивыпуклая норма всегда является строго квазивыпуклой, но не всегда, как показывают примеры, строго квазивыпуклая норма будет одновременно и сильно квазивыпуклой. Необходимо отметить, что, являясь положительно-однородной функцией, любая норма не может быть строго и тем более сильно выпуклой функцией на любом выпуклом телесном множестве. Это обстоятельство, на наш взгляд, мотивирует использование во введенном понятии приставки “квази” и соотносит его с традиционным понятием квазивыпуклой функции (см., например, [12, с. 51]).

Функции $R(x, D)$ и $\rho(x, \Omega)$ также не могут быть строго или сильно выпуклыми (вогнутыми) функциями на любом выпуклом телесном множестве. Действительно, на луче $x + \lambda(x - z)$, $\lambda > 0$, где z – любая точка из $Q^R(x, D)$, функция $R(x, D)$ ведет себя как аффинная. Аналогично, на отрезке, соединяющем точку $x \in D$ с любой точкой z из проекции $Q^p(x, \Omega)$, аффинно ведет себя функция $\rho(x, \Omega)$. Однако строгая или сильная квазивыпуклость используемой нормы для функции $R(x, D)$ и, соответственно, строгая или сильная выпуклость множества D для функции $\rho(x, \Omega)$ дают возможность сравнивать их поведение на некоторых отрезках с поведением строго (сильно) выпуклых или строго (сильно) вогнутых функций (см. [19], [20]). В частности, имеет место

Лемма 4 (см. [19]). *Если $n(\cdot)$ – строго квазивыпуклая норма, а точки x_1 и x_2 таковы, что $R(x_1, D) \leq R(x_2, D) < R(x_1, D) + n(x_1 - x_2)$, то*

$$R(\alpha x_1 + (1 - \alpha)x_2, D) < \alpha R(x_1, D) + (1 - \alpha)R(x_2, D) \quad \forall \alpha \in (0, 1).$$

2.4. Приведем некоторые свойства лебеговых множеств функций $R(x, D)$ и $\rho(x, \Omega)$.

Лемма 5. *Если $n(\cdot)$ есть r -сильно квазивыпуклая норма и*

$$\lambda \geq \min_{x \in \mathbb{R}^p} R(x, D),$$

то множество $G^R(\lambda)$ является λr -сильно выпуклым.

Доказательство. Действительно, из самого определения функции $R(x, D)$ следует справедливость равенства

$$\begin{aligned} G^R(\lambda) &= \{x \in \mathbb{R}^p : \max_{y \in D} n(x - y) \leq \lambda\} = \\ &= \{x \in \mathbb{R}^p : n(x - y) \leq \lambda, \forall y \in D\} = \bigcap_{y \in D} \text{Bn}(y, \lambda). \end{aligned} \quad (2.9)$$

Поскольку шар $\text{Bn}(0_p, 1)$ является r -сильно выпуклым множеством, то легко видеть, что шар $\text{Bn}(y, \lambda) = y + \lambda \text{Bn}(0_p, 1)$ является $r\lambda$ -сильно выпуклым. Поэтому из (2.9) и следует утверждение леммы.

Лемма 6. *Если $x \neq \Omega$, то справедливо представление*

$$D(x) = D^* \text{Bn}(0_p, \rho(x, \Omega)), \quad (2.10)$$

при этом $D(x)$ является r -сильно выпуклым множеством, если множество D является r -сильно выпуклым.

Доказательство. Представление множества $D(x)$ в виде (2.10) получено в [18]. Вторая часть утверждения вытекает из известного свойства сильно выпуклых множеств ([см. 9, с. 292]).

Относительно конуса возможных направлений нижнего лебегова множества выпуклой функции имеет место (см., например, [12, с. 221–223])

Лемма 7. *Пусть $h(x)$ – выпуклая конечная на $\mathbb{R}^p$ функция, для точки x_0 выполняется условие $h(x_0) = 0$ и $0_p \notin \underline{\partial} h(x_0)$. Тогда для множества $A = \{x \in \mathbb{R}^p : h(x) \leq 0\}$ имеет место формула $K(x_0, A) = -K^+(\underline{\partial} h(x_0))$. Здесь $\underline{\partial} h(x_0)$ – субдифференциал функции $h(\cdot)$ в точке x .*

3. ОБЩАЯ ХАРАКТЕРИСТИКА УСТОЙЧИВОСТИ

Понятно, что свойства решений задач (1.1) и (1.2), в том числе и устойчивость, зависят от свойств оцениваемого компакта и используемой нормы. Приведем факты, касающиеся устойчивости, которые не требуют каких-либо дополнительных условий на D и $n(\cdot)$.

Введем обозначения для радиусов и множеств центров описанных и вписанных шаров в точных и приближенных задачах:

$$\begin{aligned} R^* &= \min_{x \in \mathbb{R}^p} R(x, D), \quad X^R(D) = \{x \in \mathbb{R}^p : R(x, D) = R^*\}, \\ R_\varepsilon &= \min_{x \in \mathbb{R}^p} R(x, D_\varepsilon), \quad X^R(D_\varepsilon) = \{x \in \mathbb{R}^p : R(x, D_\varepsilon) = R_\varepsilon\}, \end{aligned}$$

$$\rho^* = \max_{x \in D} \rho(x, \Omega), \quad X^p(D) = \{x \in D : \rho(x, \Omega) = \rho^*\},$$

$$\rho_\varepsilon = \max_{x \in D_\varepsilon} \rho(x, \Omega_\varepsilon), \quad X^p(D_\varepsilon) = \{x \in D_\varepsilon : \rho(x, \Omega_\varepsilon) = \rho_\varepsilon\}.$$

3.1. Сначала получим оценку устойчивости радиусов вписанных и описанных шаров. Имеет место следующая

Лемма 8. Для любой точки $x \in \mathbb{R}^p$ справедливы неравенства

$$|R(x, D) - R(x, D_\varepsilon)| \leq C_2 \varepsilon, \quad (3.1)$$

$$|\rho(x, \Omega) - \rho(x, \Omega_\varepsilon)| \leq C_2 \varepsilon. \quad (3.2)$$

Доказательство. Для произвольной точки x возьмем точку $y_x \in Q^R(x, D)$, т.е. $y_x \in D$, и при этом

$$R(x, D) = n(x - y_x). \quad (3.3)$$

Поскольку $h(D, D_\varepsilon) \leq \varepsilon$, то найдется точка $y_x^\varepsilon \in D_\varepsilon$ такая, что $\|y_x - y_x^\varepsilon\| \leq \varepsilon$ и, значит, в силу (2.1),

$$n(y_x - y_x^\varepsilon) \leq C_2 \varepsilon. \quad (3.4)$$

Используя (3.4), получаем

$$n(x - y_x) \leq n(x - y_x^\varepsilon) + n(y_x^\varepsilon - y_x) \leq n(x - y_x^\varepsilon) + C_2 \varepsilon. \quad (3.5)$$

С другой стороны, поскольку $y_x^\varepsilon \in D_\varepsilon$, то

$$n(x - y_x^\varepsilon) \leq R(x, D_\varepsilon). \quad (3.6)$$

Теперь из (3.3), (3.5) и (3.6) получаем

$$R(x, D) \leq R(x, D_\varepsilon) + C_2 \varepsilon. \quad (3.7)$$

Проводя аналогичные рассуждения и меняя местами D и D_ε , имеем также

$$R(x, D_\varepsilon) \leq R(x, D) + C_2 \varepsilon. \quad (3.8)$$

Из (3.7) и (3.8) следует (3.1).

Подобные несложные рассуждения дают и неравенство (3.2).

Теорема 5. Справедливы неравенства

$$|R^* - R_\varepsilon| \leq C_2 \varepsilon, \quad (3.9)$$

$$|\rho^* - \rho_\varepsilon| \leq C_2 \varepsilon. \quad (3.10)$$

Доказательство. Для точки $x^* \in X^p(D)$, применяя неравенство (3.2), имеем

$$\rho(x^*, \Omega_\varepsilon) \geq \rho(x^*, \Omega) - C_2 \varepsilon = \rho^* - C_2 \varepsilon. \quad (3.11)$$

Поскольку для точки $x_\varepsilon \in X^p(D_\varepsilon)$ выполняется $\rho_\varepsilon = \rho(x_\varepsilon, \Omega_\varepsilon) \geq \rho(x^*, \Omega_\varepsilon)$, то из (3.11) получаем

$$\rho_\varepsilon \geq \rho^* - C_2 \varepsilon. \quad (3.12)$$

Аналогичные рассуждения приводят к неравенству

$$\rho^* \geq \rho_\varepsilon - C_2 \varepsilon. \quad (3.13)$$

Из (3.12) и (3.13) следует (3.10). Подобным же образом доказывается неравенство (3.9).

Следствие 1. Для точки $x_\varepsilon \in X^p(D_\varepsilon)$ выполняется неравенство

$$|R^* - R(x_\varepsilon, D)| \leq 2C_2 \varepsilon. \quad (3.14)$$

Доказательство. Действительно, используя (3.1) и (3.9), получаем

$$\begin{aligned} |R^* - R(x_\varepsilon, D)| &\leq |R^* - R_\varepsilon| + |R_\varepsilon - R(x_\varepsilon, D)| = \\ &= |R^* - R_\varepsilon| + |R(x_\varepsilon, D_\varepsilon) - R(x_\varepsilon, D)| \leq 2C_2 \varepsilon. \end{aligned}$$

3.2. Теперь рассмотрим вопрос об устойчивости самих множеств решений задач (1.1) и (1.2), т.е. множеств центров описанных и вписанных шаров.

Под $Kv(\mathbb{R}^p)$ будем понимать метрическое пространство всех выпуклых компактов из $\mathbb{R}^p$ с метрикой Хаусдорфа (1.3). Каждому элементу $D \in Kv(\mathbb{R}^p)$ можно сопоставить множество центров описанных шаров $X^R(D)$ и вписанных шаров $X^p(D)$. Поэтому можно рассматривать многозначные отображения

$$X^R(\cdot) : Kv(\mathbb{R}^p) \longrightarrow 2^{\mathbb{R}^p}, \quad X^p(\cdot) : Kv(\mathbb{R}^p) \longrightarrow 2^{\mathbb{R}^p}.$$

Теорема 6. *Многозначные отображения $X^R(\cdot)$ и $X^p(\cdot)$ являются полунепрерывными сверху всюду на $Kv(\mathbb{R}^p)$.*

Доказательство. Сначала покажем, что

$$\lim_{\varepsilon \downarrow 0} (R(x_\varepsilon, D_\varepsilon) - R(x_0, D)) = 0, \quad (3.15)$$

если $x_\varepsilon \rightarrow x_0$ при $\varepsilon \downarrow 0$. В самом деле, запишем очевидное неравенство

$$|R(x_\varepsilon, D_\varepsilon) - R(x_0, D)| \leq |R(x_\varepsilon, D_\varepsilon) - R(x_\varepsilon, D)| + |R(x_\varepsilon, D) - R(x_0, D)|. \quad (3.16)$$

Здесь $|R(x_\varepsilon, D_\varepsilon) - R(x_\varepsilon, D)| \leq C_2 \varepsilon$ в силу леммы 8. Кроме того, функция $R(x, D)$, как выпуклая и конечная по x на $\mathbb{R}^p$ функция, является всюду непрерывной. Поэтому и второе слагаемое в правой части неравенства (3.16) также стремится к нулю при $\varepsilon \downarrow 0$. Тем самым (3.15) доказано.

Пусть $\{D_{\varepsilon_i}\}$, $i = 1, 2, \dots$, – произвольная последовательность элементов из $Kv(\mathbb{R}^p)$ такая, что $h(D, D_{\varepsilon_i}) \leq \varepsilon_i$, где $\varepsilon_i \downarrow 0$ при $i \rightarrow \infty$.

Пусть также последовательность $\{x_{\varepsilon_i}\}$, $i = 1, 2, \dots$, такова, что

$$x_{\varepsilon_i} = X^R(D_{\varepsilon_i}), \quad x_{\varepsilon_i} \rightarrow x_0, \quad i \rightarrow \infty. \quad (3.17)$$

Покажем, что

$$x_0 \in X^R(D). \quad (3.18)$$

Действительно, если это неверно, то для точки $x^* \in X^R(D)$ выполняется неравенство

$$R(x_0, D) - R(x^*, D) = \Delta > 0. \quad (3.19)$$

В соответствии с леммой 8 можно записать

$$|R(x^*, D_{\varepsilon_i}) - R(x^*, D)| \leq C_2 \varepsilon_i. \quad (3.20)$$

Кроме того, ввиду (3.15) имеем

$$|R(x_{\varepsilon_i}, D_{\varepsilon_i}) - R(x_0, D)| = \delta_i \downarrow 0, \quad i \rightarrow \infty. \quad (3.21)$$

Из (3.20) и (3.21) следуют неравенства

$$R(x_{\varepsilon_i}, D_{\varepsilon_i}) \geq R(x_0, D) - \delta_i, \quad R(x^*, D) + C_2 \varepsilon_i \geq R(x^*, D_{\varepsilon_i}),$$

из которых получаем

$$R(x_{\varepsilon_i}, D_{\varepsilon_i}) - R(x^*, D_{\varepsilon_i}) \geq R(x_0, D) - R(x^*, D) - \delta_i - C_2 \varepsilon_i.$$

Отсюда, учитывая (3.19), получаем, что при достаточно больших значениях индекса i , выполняется неравенство $R(x_{\varepsilon_i}, D_{\varepsilon_i}) - R(x^*, D_{\varepsilon_i}) \geq \Delta/2$. Это противоречит тому, что $x_{\varepsilon_i} \in X^R(D_{\varepsilon_i})$. Тем самым включение (3.18) доказано, т.е. предел сходящейся последовательности $\{x_{\varepsilon_i}\}$, $i = 1, 2, \dots$, элементов, выбираемых, соответственно, из $X^R(D_{\varepsilon_i})$, обязательно содержится в $X^R(D)$. Поскольку последовательность $\{x_{\varepsilon_i}\}$, $i = 1, 2, \dots$, обладающая сходимостью в (3.17), выбиралась произвольно, то мы доказали полунепрерывность сверху многозначного отображения $X^R(\cdot)$ на произвольно выбранном элементе $D \in Kv(\mathbb{R}^p)$.

Полунепрерывность сверху многозначного отображения $X^p(\cdot)$ всюду на $Kv(\mathbb{R}^p)$ доказывается аналогично.

3.3. Сделаем несколько замечаний, касающихся теорем 5 и 6.

Замечания. 1. Нетрудно представить примеры, показывающие, что оценки (3.9) и (3.10) являются точными. Кроме того, легко видеть, что если расстояние Хаусдорфа измерять в используемой норме, т.е. заменить евклидову норму в (1.3) на $n(\cdot)$, то следует заменить константу C_2 в (3.9) и (3.10) на единицу.

2. Покажем, что многозначные отображения $X^R(\cdot)$ и $X^p(\cdot)$ могут не обладать полунепрерывностью снизу.

Пример 1. Пусть $p = 3$, $x = (x^{(1)}, x^{(2)}, x^{(3)})$. В качестве оцениваемого выпуклого компакта возьмем отрезок $D = \text{co}\{(0, 1, 0), (0, -1, 0)\}$, а норму $n(\cdot)$ сконструируем следующим образом. Возьмем квадрат и круг в ортогональных плоскостях

$$T_1 = \{x \in \mathbb{R}^3 : \max\{|x^{(1)}|, |x^{(2)}|\} \leq 1, x^{(3)} = 0\},$$

$$T_2 = \{x \in \mathbb{R}^3 : (x^{(2)})^2 + (x^{(3)})^2 \leq 1, x^{(1)} = 0\}.$$

Выпуклая оболочка $M = \text{co}\{T_1, T_2\}$ этих симметричных относительно нулевого элемента множеств образует выпуклое тело, также симметричное относительно 0_3 . В качестве нормы $n(\cdot)$ возьмем функцию Минковского этого тела

$$n(x) = \inf\{\alpha > 0 : x \in \alpha M\}.$$

Теперь берем произвольную последовательность $\delta_i \downarrow 0$ при $i \rightarrow \infty$ и

$$D_{\varepsilon_i} = \text{co}\{(0, 1 - \delta_i, -\sqrt{2\delta_i - \delta_i^2}), (0, -1 + \delta_i, \sqrt{2\delta_i - \delta_i^2})\}.$$

Нетрудно подсчитать, что $h(D, D_{\varepsilon_i}) = \varepsilon_i$, где $\varepsilon_i = \sqrt{2\delta_i}$. При этом можно убедиться в том, что $X^R(D_{\varepsilon_i}) = \{0_3\}$. Однако, как нетрудно видеть, $X^R(D) = \text{co}\{(-1, 0, 0), (1, 0, 0)\}$. Таким образом, любая точка $x^* \neq 0_3$ из $X^R(D)$ не может быть пределом последовательности элементов x_{ε_i} , выбираемых из $X^R(D_{\varepsilon_i})$. Это и означает, что на данном элементе $D \in Kv(\mathbb{R}^p)$ многозначное отображение $X^R(\cdot)$ не является полунепрерывным снизу.

Пример 2. Пусть $p = 2$, $x = (x^{(1)}, x^{(2)})$, $n(x) = \sqrt{(x^{(1)})^2 + (x^{(2)})^2}$ – евклидова норма,

$$D = \{x \in \mathbb{R}^2 : |x^{(1)}| \leq 2, |x^{(2)}| \leq 1\},$$

$$D_\varepsilon = \{x \in \mathbb{R}^2 : |x^{(1)}| \leq 2, (x^{(1)})^2 + (x^{(2)} + r_\varepsilon - 1)^2 \leq r_\varepsilon^2, \\ (x^{(1)})^2 + (x^{(2)} - r_\varepsilon + 1)^2 \leq r_\varepsilon^2\}.$$

Нетрудно видеть, что $h(D, D_\varepsilon) = \varepsilon$ при $r_\varepsilon = (4 - \varepsilon^2)/2\varepsilon$. А поскольку $X^p(D) = \text{co}\{(-1, 0), (1, 0)\}$, $X^p(D_\varepsilon) = \{0_2\}$ при сколь угодно малых $\varepsilon > 0$, то, как и в примере 1, приходим к выводу, что многозначное отображение $X^p(\cdot)$ не является полунепрерывным снизу на данном элементе $D \in Kv(\mathbb{R}^p)$.

Замечание 3. Полунепрерывность сверху многозначного отображения $X^p(\cdot)$ на элементе $D \in Kv(\mathbb{R}^p)$ означает наличие сходящегося процесса

$$\rho(X^p(D_\varepsilon), X^p(D)) = \sup_{x \in X^p(D_\varepsilon)} \inf_{y \in X^p(D)} \|x - y\| \rightarrow 0, \quad \varepsilon \downarrow 0. \quad (3.22)$$

Здесь $\rho(A, B)$ – уклонение множества A от множества B . Поэтому можно поставить вопрос об оценке скорости сходимости данного процесса по ε . Однако пример, приведенный в разд. 5, говорит о том, что без дополнительного предположения относительно компакта D , связанного с количественной характеристикой выпуклости D , получить такую оценку нельзя. Соответственно, для оценки скорости сходимости процесса

$$\rho(X^R(D_\varepsilon), X^R(D)) \rightarrow 0, \quad \varepsilon \downarrow 0, \quad (3.23)$$

по-видимому, требуются дополнительные условия количественного характера на используемую норму.

4. КРИТЕРИЙ РЕШЕНИЯ ЗАДАЧИ О ВНЕШНЕЙ ОЦЕНКЕ

Поскольку функция $R(x, D)$ является выпуклой по x на $\mathbb{R}^p$, то, в соответствии с известным фактом из выпуклого анализа (см., например, [11, с. 142]), для того чтобы точка x^* являлась решением задачи (1.1), т.е. центром описанного шара, необходимо и достаточно, чтобы выполнялось включение

$$0_p \in \partial R(x^*, D), \quad (4.1)$$

где субдифференциал $\partial R(\cdot, D)$ выражен формулой (2.3).

4.1. Приведем другой критерий решения задачи (1.1), связывающий ее с задачей о внутренней оценке нижнего лебегова множества функции $R(x, D)$ шаром нормы $n(\cdot)$. Именно он оказался удобным для количественной оценки устойчивости центра описанного шара.

Теорема 7. Точка x^* является решением задачи (1.1) тогда и только тогда, когда для любого $\lambda > R^*$ она является центром вложенного в множество $G^R(\lambda)$ шара наибольшего радиуса, т.е. решением задачи

$$\rho(x, \Omega^R(\lambda)) \equiv \min_{y \in \Omega^R(\lambda)} n(x - y) \longrightarrow \max_{x \in G^R(\lambda)}. \quad (4.2)$$

При этом радиус вложенного шара есть $\rho(x^*, \Omega^R(\lambda)) = \lambda - R^*$.

Доказательство. Необходимость докажем в четыре этапа. Итак, пусть точка x^* является решением задачи (1.1) и $\lambda > R^*$.

1. Покажем, что справедливо включение

$$\text{Вп}(x^*, \lambda - R^*) \subset G^R(\lambda). \quad (4.3)$$

Действительно, поскольку $D \subset \text{Вп}(x^*, R^*)$, то для любого $y \in D$ выполняется условие $x^* \in \text{Вп}(y, R^*)$. Отсюда вытекает включение

$$\text{Вп}(x^*, \lambda - R^*) \subset \text{Вп}(y, \lambda) \quad \forall y \in D. \quad (4.4)$$

А так как множество $G^R(\lambda)$ представимо в виде (2.9), то из (4.4) следует (4.3).

2. Возьмем произвольную точку

$$z \in Q^R(x^*, D) = \{y \in D: R^* = n(x^* - y)\} \quad (4.5)$$

и сопоставим ей точку

$$z_\lambda = x^* + \frac{\lambda - R^*}{n(x^* - z)}(x^* - z). \quad (4.6)$$

Покажем, что

$$z_\lambda \in Q^p(x^*, \Omega^R(\lambda)) = \{y \in \Omega^R(\lambda): \rho(x^*, \Omega^R(\lambda)) = n(x^* - y)\}, \quad (4.7)$$

т.е. точка z_λ принадлежит проекции точки x^* на множество $\Omega^R(\lambda)$ в норме $n(\cdot)$.

В самом деле, с одной стороны, из (4.5), (4.6) имеем

$$n(z_\lambda - z) = n(x^* - z) + \lambda - R^* = \lambda, \quad (4.8)$$

т.е.

$$z_\lambda \in \text{Sn}(z, \lambda) = \{y \in \mathbb{R}^p: n(z - y) = \lambda\}. \quad (4.9)$$

В то же время из (4.6) вытекает равенство

$$n(z_\lambda - x^*) = \lambda - R^* \quad (4.10)$$

и поэтому в силу (4.3) имеем

$$z_\lambda \in G^R(\lambda). \quad (4.11)$$

Поскольку, в соответствии с (4.5), $z \in D$, то шар $\text{Вп}(z, \lambda)$, в силу представления (2.9), содержит множество $G^R(\lambda)$. Поэтому из (4.9) и (4.11) следует, что z_λ является граничной точкой множества $G^R(\lambda)$, а следовательно, и множества $\Omega^R(\lambda)$. Таким образом, учитывая (4.3) и (4.10), заключаем,

что шар $\text{Вп}(x^*, \lambda - R^*)$ содержится в $G^R(\lambda)$ и касается своей граничной точкой z_λ множества $\Omega^R(\lambda)$. Это и означает справедливость (4.7).

3. Покажем, что имеет место включение

$$K(\partial n(x^* - z_\lambda)) \subset K^+(z_\lambda, G^R(\lambda)), \quad (4.12)$$

где точка z_λ соответствует точке $z \in Q^R(x^*, D)$ согласно (4.6).

Так как $z \in D$, то из (2.9) следует включение

$$G^R(\lambda) \subset \text{Вп}(z, \lambda). \quad (4.13)$$

А поскольку из (4.9) и (4.11) следует, что $z_\lambda \in G^R(\lambda) \cap \text{Вп}(z, \lambda)$, то из (4.13) получаем

$$K(z_\lambda, G^R(\lambda)) \subset K(z_\lambda, \text{Вп}(z, \lambda)).$$

Тогда для сопряженных конусов выполняется обратное включение

$$K^+(z_\lambda, \text{Вп}(z, \lambda)) \subset K^+(z_\lambda, G^R(\lambda)). \quad (4.14)$$

Так как $z_\lambda \neq z$, то из формулы (2.2) следует $0_p \notin \partial n(z - z_\lambda)$. Поэтому, учитывая (4.8) и применяя лемму 7, получаем

$$K(z_\lambda, \text{Вп}(z, \lambda)) = K^+(\partial n(z - z_\lambda)). \quad (4.15)$$

Из (4.6) имеем

$$z_\lambda - z = \left(1 + \frac{\lambda - R^*}{n(x^* - z)}\right)(x^* - z), \quad z_\lambda - x^* = \frac{\lambda - R^*}{n(x^* - z)}(x^* - z),$$

т.е. $x^* - z_\lambda = \beta(z - z_\lambda)$, где $\beta > 0$. Поэтому, в силу (2.2), выполняется условие

$$\partial n(x^* - z_\lambda) = \partial n(z - z_\lambda). \quad (4.16)$$

Отметим также, что поскольку $0_p \notin \partial n(z - z_\lambda)$, то (см. [21, с. 314–316]) имеем

$$K^{++}(\partial n(z - z_\lambda)) = K(\partial n(z - z_\lambda)). \quad (4.17)$$

Теперь из (4.14)–(4.17) следует (4.12).

4. Напомним, что поскольку точку x^* считаем решением задачи (1.1), то, в соответствии с критерием ее решения (4.1) и формулой (2.3), выполняется включение

$$0_p \in \text{co}\{\partial n(x^* - z) : z \in Q^R(x^*, D)\}. \quad (4.18)$$

Множество $G^R(\lambda)$, как нижнее лебегово множество выпуклой функции, является выпуклым. Тогда, по теореме 2, функция $\rho(x, \Omega^R(\lambda))$ вогнута по x на $G^R(\lambda)$. Поэтому, в соответствии с известным фактом выпуклого анализа (см. [11, с. 142]), критерием того, что некоторая точка $x_0 \in \text{int } G^R(\lambda)$ является точкой максимума вогнутой функции $\rho(x, \Omega^R(\lambda))$, т.е. решением задачи (4.2), является соотношение $0_p \in \overline{\partial \rho}(x_0, \Omega^R(\lambda))$. Следовательно, для того чтобы доказать, что точка x^* есть решение задачи (4.2), нам, учитывая формулу супердифференциала функции расстояния (2.4), надо показать справедливость соотношения

$$0_p \in \text{co}\{\partial n(x^* - y) \cap K^+(y, G^R(\lambda)) : y \in Q^p(x^*, \Omega^R(\lambda))\}. \quad (4.19)$$

Действительно, используя (4.12), получаем

$$\begin{aligned} \partial n(x^* - z_\lambda) &= \partial n(x^* - z_\lambda) \cap K(\partial n(x^* - z_\lambda)) \subset \\ &\subset \partial n(x^* - z_\lambda) \cap K^+(z_\lambda, G^R(\lambda)). \end{aligned} \quad (4.20)$$

Из (4.6) следует $x^* - z_\lambda = \gamma(x^* - z)$, где $\gamma = (R^* - \lambda)(n(x^* - z))^{-1} < 0$. Поэтому из (2.2) вытекает

$$\partial n(x^* - z_\lambda) = -\partial n(x^* - z). \quad (4.21)$$

Учитывая (4.21) и используя (4.18) и (4.20), имеем

$$\begin{aligned} 0_p \in \operatorname{co}\{\partial n(x^* - z_\lambda) : z \in Q^R(x^*, D)\} \subset \\ \subset \operatorname{co}\{\partial n(x^* - z_\lambda) \cap K^+(z_\lambda, G^R(\lambda)) : z \in Q^R(x^*, D)\}. \end{aligned} \quad (4.22)$$

Здесь, как было показано в п. 2 доказательства, все точки z_λ , соответствующие точкам $z \in Q^R(x^*, D)$, согласно формуле (4.7) содержатся в $Q^p(x^*, \Omega^R(\lambda))$. Поэтому из (4.22) следует (4.19). Необходимость доказана, причем, как это следует из сказанного выше, радиус максимального шара, вложенного в $G^R(\lambda)$, есть $\rho(x^*, \Omega^R(\lambda)) = \lambda - R^*$.

Достаточность. Пусть точка x^* является центром шара максимального радиуса, вложенного в $G^R(\lambda)$. Радиус этого шара, как было установлено выше, равен $\lambda - R^*$. Следовательно, имеем включение

$$\operatorname{Bn}(x^*, \lambda - R^*) \subset G^R(\lambda). \quad (4.23)$$

Поскольку множество $G^R(\lambda)$ представимо в виде (2.9), то из (4.23) следует

$$\operatorname{Bn}(x^*, \lambda - R^*) \subset \operatorname{Bn}(y, \lambda)$$

для любого $y \in D$. Отсюда получаем $x^* \in \operatorname{Bn}(y, R^*)$, или $n(x^* - y) \leq R^*$ для всех $y \in D$. Это и означает, что

$$R(x^*) \equiv \max_{y \in D} n(x^* - y) \leq R^*,$$

т.е. точка x^* является решением задачи (1.1).

4.2. Непосредственно из теоремы 7 ввиду доказанной эквивалентности задач (1.1) и (4.2) получаем критерий единственности решения.

Следствие 2. Для того чтобы задача (1.1) имела единственное решение, необходимо и достаточно, чтобы при некотором значении $\lambda > R^*$ задача (4.2) имела единственное решение.

Приведем следующее простое достаточное условие единственности решения задачи (1.1).

Теорема 8. Если $n(\cdot)$ является строго квазивыпуклой нормой, то решение задачи (1.1) единственно.

Доказательство. Действительно, предположим, что решение задачи (1.1) неединственно и точки $x_1 \neq x_2$ таковы, что

$$R(x_1, D) = R(x_2, D) = \min_{x \in \mathbb{R}^p} R(x, D). \quad (4.24)$$

Поскольку для точек x_1 и x_2 выполняются условия леммы 4, то, учитывая (4.24), для $\alpha \in (0, 1)$ получаем противоречие:

$$R(\alpha x_1 + (1 - \alpha)x_2, D) < \alpha R(x_1, D) + (1 - \alpha)R(x_2, D) = \min_{x \in \mathbb{R}^p} R(x, D).$$

Отметим, что, как показывают примеры, строгая квазивыпуклость нормы не является необходимым условием единственности решения задачи (1.1). Нетрудно также представить пример, показывающий, что строгая выпуклость оцениваемого компакта D при отсутствии строгой квазивыпуклости нормы $n(\cdot)$ не гарантирует единственности решения.

5. КОЛИЧЕСТВЕННАЯ ХАРАКТЕРИЗАЦИЯ УСТОЙЧИВОСТИ ПРИ ДОПОЛНИТЕЛЬНЫХ ПРЕДПОЛОЖЕНИЯХ

Получим оценку скорости сходимости процессов (3.22) и (3.23) при некоторых дополнительных условиях на используемую норму или оцениваемый компакт.

5.1. Скорость сходимости процесса (3.23) оценим в предположении сильной квазивыпуклости нормы.

Теорема 9. Пусть $n(\cdot)$ является r -сильно квазивыпуклой нормой. Если точка x^* – решение точной задачи (1.1), а точка x_ε – решение приближенной задачи (1.4) для $\varepsilon \in [0, \varepsilon_0]$, где $\varepsilon_0 =$

$= rR^*/(2(rC_2 + C_2/C_1))$, то справедливо неравенство

$$\|x^* - x_\varepsilon\| \leq 4 \sqrt{\frac{C_2}{C_1} r \varepsilon (R^* + 2C_2 \varepsilon)}. \quad (5.1)$$

Доказательство. Напомним, что строгая, и тем более сильная, квазивыпуклость нормы, по теореме 8, влечет единственность решения задач (1.1) и (1.4), т.е. $X^R(D) = \{x^*\}$, $X^R(D_\varepsilon) = \{x_\varepsilon\}$.

Считая $x_\varepsilon \neq x^*$, возьмем в качестве λ функцию $R(x_\varepsilon, D)$. Тогда, по теореме 7, максимальный радиус шара нормы $n(\cdot)$, вложенного в множество $G^R(\lambda)$, равен $R(x_\varepsilon, D) - R^* > 0$ и, в силу (3.14), не превышает величину $\delta = 2C_2\varepsilon$. Поэтому, по лемме 2, радиус евклидова шара, вложенного в $G^R(\lambda)$, не более чем $\delta_0 = 2C_2\varepsilon/C_1$. В соответствии с леммой 5, множество $G^R(\lambda)$ является $rR(x_\varepsilon, D)$ -сильно выпуклым. Используя (3.14), нетрудно сделать вывод, что при $\varepsilon \in [0, \varepsilon_0]$ выполняется неравенство $\delta_0 \leq rR(x_\varepsilon, D)$, которое гарантирует выполнение естественного условия: радиус вложенного евклидова шара не должен превышать радиус сильной выпуклости множества $G^R(\lambda)$. Это позволяет воспользоваться леммой 3, в результате чего получаем

$$\text{diam } G^R(\lambda) \leq 2\sqrt{\delta_0(2rR(x_\varepsilon, D) - \delta_0)} < 2\sqrt{2\delta_0 rR(x_\varepsilon, D)}. \quad (5.2)$$

Поскольку точки x^* и x_ε содержатся в $G^R(\lambda)$, то имеем

$$\|x^* - x_\varepsilon\| \leq \text{diam } G^R(\lambda). \quad (5.3)$$

Из (5.2), (5.3), используя (3.14), получаем (5.1).

5.2. Теперь получим оценку скорости сходимости процесса (3.22) в предположении сильной выпуклости компакта D . Сразу отметим, что, как и для любого строго выпуклого компакта, в этом случае задача (1.2) имеет единственное решение (см. [14]). Кроме того, сильная выпуклость D гарантирует, что $\text{int } D \neq \emptyset$, а из теоремы 6 вытекает включение $X^p(D_\varepsilon) \subset \text{int } D$ при всех достаточно малых значениях $\varepsilon > 0$.

Теорема 10. Пусть D является r -сильно выпуклым множеством, а ε достаточно мало, чтобы $\varepsilon \in [0, \varepsilon_0]$, где $\varepsilon_0 = C_1 r / (2C_2)$, и $X^p(D_\varepsilon) \subset \text{int } D$. Если точка x^* – решение точной задачи (1.2), а x_ε – некоторое решение приближенной задачи (1.5), то справедлива оценка

$$\|x^* - x_\varepsilon\| \leq 4 \sqrt{\frac{C_2}{C_1} \varepsilon \left(r - \frac{C_2}{C_1} \varepsilon \right)}. \quad (5.4)$$

Доказательство. Возьмем произвольную точку $x_\varepsilon \in X^p(D_\varepsilon)$ и произвольную точку $x \in D$ такие, что $\rho(x, \Omega) \geq \rho(x_\varepsilon, \Omega)$. В силу леммы 1 выполняется равенство

$$\rho(x, \Omega) = \rho(x, \Omega(x_\varepsilon)) + \rho(x_\varepsilon, \Omega). \quad (5.5)$$

Отсюда, учитывая, что x^* – решение задачи (1.1), вытекает равенство

$$\rho(x^*, \Omega(x_\varepsilon)) = \max_{x \in D(x_\varepsilon)} \rho(x, \Omega(x_\varepsilon)), \quad (5.6)$$

а также неравенство

$$\rho(x^*, \Omega(x_\varepsilon)) \leq |\rho(x^*, \Omega) - \rho(x_\varepsilon, \Omega_\varepsilon)| + |\rho(x_\varepsilon, \Omega_\varepsilon) - \rho(x_\varepsilon, \Omega)|. \quad (5.7)$$

Используя (3.2) и (3.10), из (5.7) получаем

$$\rho(x^*, \Omega(x_\varepsilon)) \leq 2C_2\varepsilon. \quad (5.8)$$

В силу (5.6) неравенство (5.8) дает нам оценку сверху радиуса шара, вложенного в $D(x_\varepsilon)$. Тогда, применяя лемму 2, мы можем дать оценку сверху и для максимального радиуса евклидова шара, вложенного в $D(x_\varepsilon)$:

$$\max_{x \in D(x_\varepsilon)} \min_{y \in \Omega(x_\varepsilon)} \|x - y\| \leq 2 \frac{C_2}{C_1} \varepsilon. \quad (5.9)$$

Поскольку $X^p(D_\varepsilon) \subset \text{int } D$, то $x_\varepsilon \notin \Omega$ и, по лемме 6, множество $D(x_\varepsilon)$ является r -сильно выпуклым. Значения $\varepsilon \in [0, \varepsilon_0]$ ввиду (5.9) гарантируют нам, что радиус евклидова шара, вложенного в

$D(x_\varepsilon)$, не превышает радиуса сильной выпуклости этого множества. Поскольку точки x^* и x_ε содержатся в $D(x_\varepsilon)$, то, применяя теорему 4, имеем

$$\text{strco}_r(\{x^*\}, \{x_\varepsilon\}) \subset D(x_\varepsilon). \quad (5.10)$$

Из теоремы 3 следует, что максимальный радиус евклидова шара, вложенного в $\text{strco}_r(\{x^*\}, \{x_\varepsilon\})$, равен $r - \sqrt{r^2 - \|x^* - x_\varepsilon\|^2}/4$. Поэтому, учитывая (5.9) и (5.10), получаем неравенство

$$r - \sqrt{r^2 - \frac{1}{4}\|x^* - x_\varepsilon\|^2} \leq 2 \frac{C_2}{C_1} \varepsilon,$$

простое преобразование которого дает (5.4).

5.3. Приведем некоторые комментарии.

1. Покажем, что порядок погрешности в оценке (5.4) не уменьшаем.

Пример 3. Пусть, как и в теореме 3, точки a_0 и a_1 из $\mathbb{R}^p$ таковы, что $0 < \|a_0 - a_1\| < 2r$. Пусть $n(\cdot) = \|\cdot\|$ – евклидова норма, а

$$D = \text{strco}_r(\{a_0\}, \{a_1\}). \quad (5.11)$$

Считая далее значения $\delta > 0$ достаточно малыми, вводим обозначение

$$\hat{D}(\delta) = \text{co} \left\{ D, B \left(a \left(\frac{1}{2} + \delta \right), r \left(\frac{1}{2} + \delta^2 \right) \right) \right\}, \quad (5.12)$$

где $a(\cdot)$ и $r(\cdot)$ определены в (2.5). Нетрудно видеть, что для этих множеств центры и радиусы вписанных шаров выражаются следующими значениями:

$$X^p(D) = \left\{ a \left(\frac{1}{2} \right) \right\}, \quad \rho^* = r \left(\frac{1}{2} \right), \quad (5.13)$$

$$X^p(\hat{D}(\delta)) = \left\{ a \left(\frac{1}{2} + \delta \right) \right\}, \quad \rho(\delta) = r \left(\frac{1}{2} + \delta^2 \right). \quad (5.14)$$

Из (5.11), (5.12) следует, что

$$h(D, \hat{D}(\delta)) \leq \rho \left(B \left(a \left(\frac{1}{2} + \delta \right), r \left(\frac{1}{2} + \delta^2 \right) \right), D \right). \quad (5.15)$$

А поскольку, в силу теоремы 3, имеем $B(a(1/2 + \delta), r(1/2 + \delta)) \subset D$, то из (5.15) получаем

$$\begin{aligned} h(D, \hat{D}(\delta)) &\leq \rho \left(B \left(a \left(\frac{1}{2} + \delta \right), r \left(\frac{1}{2} + \delta^2 \right) \right), B \left(a \left(\frac{1}{2} + \delta \right), r \left(\frac{1}{2} + \delta \right) \right) \right) = \\ &= r \left(\frac{1}{2} + \delta^2 \right) - r \left(\frac{1}{2} + \delta \right) = \sqrt{r^2 - \left(\frac{1}{4} - \delta^2 \right) \|a_0 - a_1\|^2} - \sqrt{r^2 - \frac{1}{4} \|a_0 - a_1\|^2} + \delta^2. \end{aligned} \quad (5.16)$$

Выберем, для определенности, точки a_0 и a_1 так, чтобы выполнялось равенство $\|a_0 - a_1\| = \sqrt{2}r$. Тогда из (5.16) имеем

$$h(D, \hat{D}(\delta)) \leq \delta^2 (\sqrt{2}r + 1). \quad (5.17)$$

Возьмем $D_\varepsilon = \hat{D}(\delta(\varepsilon))$ в качестве $\delta(\varepsilon) = [\varepsilon / (\sqrt{2}r + 1)]^{1/2}$. Тогда, в соответствии с (5.17), выполняется неравенство $h(D, D_\varepsilon) \leq \varepsilon$. Таким образом, учитывая (5.13), (5.14) и (2.5), получаем

$$\|x^* - x_\varepsilon\| = \left\| a \left(\frac{1}{2} \right) - a \left(\frac{1}{2} + \delta(\varepsilon) \right) \right\| = \delta(\varepsilon) \|a_0 - a_1\| = \frac{\sqrt{2}r}{\sqrt{2}r + 1} \sqrt{\varepsilon},$$

что и требовалось показать.

2. Теперь покажем, что скорость сходимости процесса (3.22) невозможно оценить, не имея какой-либо количественной характеристики выпуклости оцениваемого компакта.

Пример 4. Пусть $x_1 \neq x_2$ – точки из $\mathbb{R}^p$, $n(\cdot) = \|\cdot\|$ – евклидова норма, а $r_0 > 0$ такое, что при всех $\varepsilon \in [0, 1]$ и $r \geq r_0$ выполняется

$$D_1(r, \varepsilon) = \text{strco}_r\{B(x_1, 1), B(x_2, 1 - \varepsilon)\} \neq \emptyset,$$

$$D_2(r, \varepsilon) = \text{strco}_r\{B(x_1, 1 - \varepsilon), B(x_2, 1)\} \neq \emptyset.$$

Очевидно, что (см. [9, с. 305])

$$\lim_{r \rightarrow \infty} h(D_1(r, \varepsilon), \text{co}\{B(x_1, 1), B(x_2, 1 - \varepsilon)\}) = 0, \quad (5.18)$$

$$\lim_{r \rightarrow \infty} h(D_2(r, \varepsilon), \text{co}\{B(x_1, 1 - \varepsilon), B(x_2, 1)\}) = 0. \quad (5.19)$$

Поэтому, нетрудно видеть, что каждому значению ε можно сопоставить значение $r = r(\varepsilon)$ так, чтобы выполнялись условия

$$X^p(D_1(r(\varepsilon), \varepsilon)) = \{x_1\}, \quad X^p(D_2(r(\varepsilon), \varepsilon)) = \{x_2\}$$

и при этом $r(\varepsilon) \rightarrow \infty$ для $\varepsilon \downarrow 0$. В силу (5.18) и (5.19), при достаточно малых $\varepsilon > 0$ радиусы вложенных шаров для множеств $D_i(r(\varepsilon), \varepsilon)$ будут равны единице и $h(D_1(r(\varepsilon), \varepsilon), D_2(r(\varepsilon), \varepsilon)) = \varepsilon$.

Таким образом, каждому достаточно малому значению $\varepsilon > 0$ мы сопоставили пару строго выпуклых множеств, отличающихся в метрике Хаусдорфа на ε , с фиксированными одинаковыми радиусами вписанных шаров и фиксированными, но отличными друг от друга центрами этих шаров. Отметим, что множества $D_i(r(\varepsilon), \varepsilon)$, $i = 1, 2, \dots$, являются по построению даже сильно выпуклыми, однако их радиус сильной выпуклости $r(\varepsilon) \rightarrow \infty$ для $\varepsilon \downarrow 0$.

3. Авторы предполагают, что выводы, сделанные выше относительно задачи о внутренней оценке, справедливы и для задачи о внешней оценке. Однако построение соответствующих примеров вызвало затруднения.

4. Приведем пример возможного использования полученных результатов.

Один из способов приближенного решения задачи (1.2) может заключаться в предварительной аппроксимации оцениваемого выпуклого компакта D многогранником. Известны различные способы аппроксимации выпуклого компакта многогранником (см., например, [8], [22]), для которых получена соответствующая оценка погрешности аппроксимации. Имея аппроксимирующий многогранник, мы можем решить задачу о его внутренней оценке, которая сводится к задаче линейного программирования (см. [14]). Далее, используя оценку погрешности аппроксимации, можно, по теореме 5, дать оценку погрешности радиуса вписанного шара, а если D есть r -сильно выпуклое множество, то, по теореме 10, и погрешности центра вписанного шара.

5. В [23] для случая евклидовой нормы получена асимптотическая оценка устойчивости решения задачи наилучшего приближения в метрике Хаусдорфа заданного выпуклого компакта шаром, которая, как известно (см. [7]), эквивалентна задаче о построении шарового слоя минимальной толщины, содержащего границу приближаемого компакта. Эту задачу с задачами о внешней и внутренней оценке, как показано в [24], связывает то обстоятельство, что все они, а также и некоторые другие, являются частными случаями задачи о наилучшем приближении в метрике Хаусдорфа выпуклого компакта шаром фиксированного радиуса в зависимости от величины этого радиуса. Поэтому характеристика устойчивости решения этой последней задачи могла бы дать единое представление об устойчивости целого круга задач.

СПИСОК ЛИТЕРАТУРЫ

1. Боннезен Т., Фенхель В. Теория выпуклых тел. М.: Фазис, 2002.
2. Том Л.Ф. Расположения на плоскости, на сфере и в пространстве. М.: Физматлит, 1958.
3. Черноусько Ф.Л. Оценивание фазового состояния динамических систем: метод эллипсоидов. М.: Наука, 1988.
4. Kurzhanski A.B., Valyi I. Ellipsoidal calculus for estimation and control. Boston: Birkhauser, 1997.
5. Абрамов О.В., Здор В.В., Супоня А.А. Допуски и номиналы систем управления. М.: Наука, 1976.
6. Дудов С.И. О задаче фиксированных допусков // Ж. вычисл. матем и матем. физ. 1997. Т. 37. № 8. С. 937–944.

7. Никольский М.С., Силин Д.Б. О наилучшем приближении выпуклого компакта элементами аддиала // Тр. Матем. ин-та РАН. 1995. Т. 211. С. 338–354.
8. Половинкин Е.С. Сильно выпуклый анализ // Матем. сб. 1996. Т. 187. № 2. С. 102–130.
9. Половинкин Е.С., Балашов М.В. Элементы выпуклого и сильно выпуклого анализа. М.: Физматлит, 2004.
10. Лейхтвейс К. Выпуклые множества. М.: Наука, 1985.
11. Пишеничный Б.Н. Выпуклый анализ и экстремальные задачи. М.: Наука, 1980.
12. Демьянов В.Ф., Васильев Л.В. Недифференцируемая оптимизация. М.: Наука, 1981.
13. Кларк Ф. Оптимизация и негладкий анализ. М.: Наука, 1986.
14. Дудов С.И. Внутренняя оценка выпуклого множества телом нормы // Ж. вычисл. матем. и матем. физ. 1996. Т. 36. № 5. С. 153–159.
15. Карманов В.Г. Математическое программирование. М.: Наука, 1986.
16. Васильев Ф.П. Численные методы решения экстремальных задач. М.: Наука, 1988.
17. Дудов С.И. Субдифференцируемость и супердифференцируемость функции расстояния // Матем. заметки. 1997. Т. 61. № 4. С. 530–542.
18. Дудов С.И. Об обобщенном градиенте функции расстояния // Итоги науки и техн. Сер. Современ. матем. и ее прилож. Тр. междунар. конф., посвящ. 90-летию Л.С. Понтрягина. Т. 2. Негладкий анализ и оптимизация. М., 1999. Т. 61. С. 5–14.
19. Дудов С.И., Златорунская И.В. Равномерная оценка выпуклого компакта шаром произвольной нормы // Матем. сб. 2000. Т. 191. № 10. С. 13–38.
20. Дудова А.С. Свойства функции расстояния до строго и сильно выпуклого множества // Математика. Механика. Саратов: Изд-во Саратовского ун-та, 2006. Вып. 8. С. 44–47.
21. Демьянов В.Ф., Малоземов В.Н. Введение в минимакс. М.: Наука, 1972.
22. Дудова А.С. Об аппроксимации выпуклого компакта многогранником // Математика. Механика. Саратов: Изд-во Саратовского ун-та, 2005. Вып. 7. С. 45–47.
23. Дудова А.С. Об устойчивости решения задачи наилучшего приближения выпуклого компакта шаром // Изв. вузов. Математика. 2006. № 7. С. 25–33.
24. Дудов С.И. Взаимосвязь некоторых задач по оценке выпуклого компакта шаром // Матем. сб. 2007. Т. 198. № 1. С. 43–58.

УДК 519.624.3

ЧИСЛЕННЫЙ АНАЛИЗ СПЕКТРА ЗАДАЧИ ОРРА–ЗОММЕРФЕЛЬДА¹⁾

© 2007 г. С. Л. Скороходов

(119991 Москва, ул. Вавилова, 40, ВЦ РАН)

e-mail: skor@ccas.ru

Поступила в редакцию 29.03.2007 г.

Разработан высокоточный метод вычисления собственных значений λ_n и собственных функций оператора Орра–Зоммерфельда. Метод основан на представлении решения в виде комбинации разложений в степенные ряды и на сшивке этих разложений. Скорость сходимости разложений исследована на основе теории рекуррентных уравнений. Для течений Куэтта и Пуазейля в канале детально исследовано поведение спектра при увеличении числа Рейнольдса R . Показано, что для течения Куэтта собственные значения λ_n , рассматриваемые как функции числа R , имеют счетное множество точек ветвления $R_k > 0$, в которых кратность собственных значений равна двум. Приведены первые 10 этих точек с точностью в 10 дес. зн. ц. Библ. 41. Фиг. 12. Табл. 1.

Ключевые слова: дифференциальное уравнение Орра–Зоммерфельда, численный анализ спектра уравнения Орра–Зоммерфельда, течения Куэтта, Пуазейля, Куэтта–Пуазейля, исследование скорости сходимости.

1. ВВЕДЕНИЕ

1.1. Оператор Орра–Зоммерфельда возникает при исследовании течения вязкой несжимаемой жидкости в канале (см. [1]–[6]). Основными уравнениями здесь являются безразмерные нестационарные уравнения Навье–Стокса с условиями непротекания и прилипания жидкости на границах области.

В настоящей работе рассматривается течение в двумерном канале $\Omega(x, y)$ с продольной координатой $x \in (-\infty, \infty)$ и поперечной координатой $y \in (-1, 1)$. Основное стационарное течение жидкости имеет лишь горизонтальную скорость $U(y)$, для него обычно рассматривают три случая.

Случай I. Течение Куэтта, когда стенки канала движутся в противоположные стороны с одинаковой скоростью. Этому случаю соответствует профиль основного потока со скоростью

$$U(y) = y. \quad (1.1)$$

Случай II. Течение Пуазейля, когда на торцах канала имеется перепад давления, вызывающий движение. Этому случаю соответствует профиль скорости

$$U(y) = 1 - y^2. \quad (1.2)$$

Случай III. Течение Куэтта–Пуазейля, когда скорость основного потока описывается параболой с вещественными коэффициентами

$$U(y) = ay^2 + by + c. \quad (1.3)$$

Рассматривая малые возмущения скорости относительно основного потока $U(y)$, вводя функцию тока $\Psi = \Psi(x, y, t)$ и исследуя периодические по x решения в форме

$$\Psi(x, y, t) = \Phi(y, t)e^{i\alpha x}, \quad \alpha > 0, \quad (1.4)$$

¹⁾ Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-01-00295, 07-01-00503) и Программы № 3 ОМН РАН.

получают линейное уравнение для $\Phi(y, t)$:

$$\left\{ \frac{\partial^2}{\partial y^2} - \alpha^2 \right\} \frac{\partial \Phi}{\partial t} = i\alpha \left\{ \frac{1}{i\alpha R} \left(\frac{\partial^2}{\partial y^2} - \alpha^2 \right)^2 - U \left(\frac{\partial^2}{\partial y^2} - \alpha^2 \right) + \frac{d^2}{dy^2} U \right\} \Phi;$$

здесь α – волновое число в представлении (1.4), а параметр R – число Рейнольдса основного потока. В актуальных задачах течения жидкости наиболее интересен случай больших чисел $R \gg 1$; этому мы уделим особое внимание.

Далее, используя для решения $\Phi(y, t)$ разделение переменных

$$\Phi(y, t) = \varphi(y) e^{-i\alpha\lambda t} \quad (1.5)$$

с неизвестной постоянной λ , для амплитуды $\varphi(y)$ получают знаменитое уравнение Орра–Зоммерфельда четвертого порядка (см. [1]–[3], [5], [6])

$$\frac{1}{i\alpha R} [\varphi^{(IV)}(y) - 2\alpha^2 \varphi''(y) + \alpha^4 \varphi(y)] - (U(y) - \lambda) [\varphi'(y) - \alpha^2 \varphi(y)] + U''(y) \varphi(y) = 0. \quad (1.6)$$

Отметим здесь, что представления (1.4), (1.5) эквивалентны поиску функции тока $\Psi(x, y, t)$ в виде бегущей волны

$$\Psi(x, y, t) = \varphi(y) e^{i\alpha(x - \lambda t)} \quad (1.7)$$

с неизвестной скоростью λ .

Краевыми условиями для (1.6) являются однородные условия

$$\varphi(-1) = \varphi'(-1) = 0, \quad (1.8)$$

$$\varphi(1) = \varphi'(1) = 0. \quad (1.9)$$

Таким образом, поставленная задача (1.6), (1.8), (1.9) состоит в нахождении собственных значений λ (СЗ) и собственных функций $\varphi(y)$ (СФ) на отрезке $y \in [-1, 1]$ при заданной аналитической функции $U(y)$. В [7], [8] показано, что множество СЗ задачи счетно, а соответствующее множество СФ полно.

1.2. Перепишем уравнение (1.6) в операторной форме:

$$L\varphi(y) = -\lambda D\varphi(y), \quad (1.10)$$

где

$$L = \frac{1}{i\alpha R} D^2 - U(y)D + U''(y), \quad (1.11)$$

$$D = \frac{d^2}{dy^2} - \alpha^2.$$

Оператор L задачи (1.10), (1.8), (1.9) является несамосопряженным и включает старшую производную $\frac{d^4}{dy^4}$ с малым коэффициентом $(\alpha R)^{-1} \ll 1$, что значительно осложняет анализ спектра (см. [9]–[12]) и исследование устойчивости малых возмущений скорости относительно основного потока (см. [1]–[6], [13]–[15]).

Из представления (1.7) в предположении $\alpha > 0$ следует, что если хоть одно собственное значение λ_k расположено в верхней полуплоскости $\text{Im}(\lambda_k) > 0$, то малые возмущения функции тока $\Psi(x, y, t)$ будут со временем экспоненциально быстро нарастать; это, в свою очередь, приводит к развитию турбулентности течения.

1.3. Для численного решения задачи (1.6), (1.8), (1.9) используют равномерные и неравномерные разностные сетки [16], [17], разложения по многочленам Чебышёва I рода (см. [18], [19], [6]) и по специальным функциям Чандрасекара (см. [20], [21]), проекционный метод с выбором специального базиса (см. [22]), комбинированные методы (см. [23], [24]), метод расщепления оператора задачи (см. [25]–[28]), метод локализации неустойчивых СЗ (см. [29]), теорию возмущений (см. [30]) и др. Обзоры этих методов можно найти также в [2], [3], [9], [31].

Как было отмечено в большинстве указанных работ, основная трудность задачи заключается в том, что мы имеем дело не с обычной задачей на СЗ, а с соответствующей задачей для пучка операторов. Помимо этого, оператор L из (1.11) является несамосопряженным и включает малый коэффициент $(\alpha R)^{-1}$ при старшей производной. Поэтому разработка эффективного метода решения задачи и исследование ее спектра при числах Рейнольдса $R \gg 1$ остается по-прежнему актуальным.

1.4. План работы следующий. В разд. 2 изложен метод решения на основе построения системы степенных разложений с центрами в граничных точках $y = -1$ и $y = 1$, а также на использовании сшивки разложений в одной точке.

В разд. 3 даны особенности метода в случае наличия четности или нечетности искомого решения. В частности, это относится к течению Пуазейля.

В разд. 4 подробно исследованы асимптотики решений рекуррентных уравнений, которые возникают для расчета коэффициентов разложений. С помощью теории Пуанкаре–Биркгофа показано, что эти коэффициенты очень быстро падают и обеспечивают весьма высокую скорость сходимости разложений.

В разд. 5 подробно исследован спектр задачи λ_n для течения Куэтта. Показано, что СЗ, рассматриваемые как функции числа Рейнольдса R , имеют счетное множество точек ветвления $R_k > 0$, в которых значения $\lambda_n(R_k)$ имеют кратность 2. Разработан эффективный метод вычисления точек ветвления R_k и двойных СЗ, а также даны первые 10 этих точек. Показано, что $\lambda_n(R_k) \rightarrow -i/\sqrt{3}$ при $k \rightarrow \infty$. Исследован также спектр задачи при комплексных волновых числах α .

В разд. 6 представлены спектральные портреты для течения Пуазейля, некоторые собственные функции и вычислены две ветви нейтральной кривой.

2. МЕТОД РЕШЕНИЯ

Уравнение (1.6) будем рассматривать в комплексной плоскости $y \in \mathbb{C}$ для общего случая течения Куэтта–Пуазейля $U(y) = ay^2 + by + c$. Отдельные случаи (1.1) и (1.2) течения Куэтта и Пуазейля изучим особо.

Параметр λ будем находить далее в процессе решения, а пока будем считать его некоторым фиксированным числом.

Уравнение (1.6) является линейным с переменными коэффициентами, поэтому все особые точки его решения определяются его коэффициентами (см. [32]).

Зависимость решения $\varphi(y)$ от параметров задачи a, b, c, α, R и СЗ λ будем обозначать через $\varphi(a, b, c, \alpha, R, \lambda; y)$, иногда для сокращения записи будем писать также $\varphi(y)$.

Легко убеждаемся, что любое решение $\varphi(a, b, c, \alpha, R, \lambda; y)$ уравнения (1.6) является регулярной функцией в любой конечной точке y_0 , а особой будет лишь бесконечно удаленная точка $y = \infty$.

Поскольку мы ищем решение $\varphi(y)$ на отрезке $y \in [-1, 1]$, то будем представлять $\varphi(y)$ в виде соответствующих разложений по степеням $(1+y)^k$ и $(1-y)^k$.

2.1. Окрестность точки $y = -1$

Сначала построим решение уравнения (1.6) в окрестности точки $y = -1$. С учетом краевых условий (1.8) представим регулярное решение $\varphi(a, b, c, \alpha, R, \lambda; y)$ в виде

$$\varphi(a, b, c, \alpha, R, \lambda; y) = \sum_{k=0}^{\infty} d_k (y+1)^{k+2}, \quad |y+1| < \infty, \quad (2.1)$$

где коэффициенты $d_k = d_k(a, b, c, \alpha, R, \lambda)$ зависят от параметров задачи a, b, c, α, R и искомого λ . Ниже будет показана скорость убывания коэффициентов d_k .

Далее для решения уравнения (1.6) нам потребуются первые четыре производные по y , которые мы получим последовательным дифференцированием ряда (2.1):

$$\frac{\partial \varphi}{\partial y} = \sum_{k=0}^{\infty} (k+2) d_k (y+1)^{k+1}, \quad \frac{\partial^2 \varphi}{\partial y^2} = \dots \quad (2.2)$$

Здесь использована запись с частными производными по y , поскольку далее нам потребуется находить также производные решения $\varphi(a, b, c, \alpha, R, \lambda; y)$ по спектральному параметру λ .

Затем, записывая функцию $U(y)$ из (1.3) в виде трехчлена по степеням $(y + 1)^k$ и подставляя разложения (2.1), (2.2) в уравнение (1.6), получаем рекуррентное уравнение для коэффициентов d_k :

$$\begin{aligned} d_k = & \frac{-i\alpha\alpha^3 R}{(k+2)k(k^2-1)}d_{k-6} + \frac{i\alpha^3 R(2a-b)}{(k+2)k(k^2-1)}d_{k-5} + \\ & + \frac{i\alpha R[a(k-4)(k-1) - \alpha^2(a+c-b-\lambda)] - \alpha^4}{(k+2)k(k^2-1)}d_{k-4} + \\ & + \frac{i\alpha R(k-2)(b-2a)}{(k+2)(k+1)k}d_{k-3} + \frac{i\alpha R(a+c-b-\lambda) + 2\alpha^2}{(k+1)(k+2)}d_{k-2}, \\ & k \geq 2. \end{aligned} \quad (2.3)$$

Соотношение (2.3) является линейным разностным уравнением порядка 6 с переменными коэффициентами и служит для последовательного вычисления d_k при возрастании $k = 2, 3, \dots$. При этом значения d_0 и d_1 мы можем выбрать произвольными и дополнительно учесть, что

$$d_k = 0, \quad k = -1, -2, \dots$$

2.2. Два решения в окрестности точки $y = -1$

Теперь построим два линейно независимых решения уравнения (1.6), удовлетворяющих условиям (1.8). Для этого используем разложение (2.1), коэффициенты d_0 и d_1 выберем двумя независимыми способами, а последующие d_k при $k \geq 2$ будем вычислять по (2.3).

Первое из решений $\varphi_1(y)$ определим с помощью задания коэффициентов

$$d_0^{(1)} = 1, \quad d_1^{(1)} = 0, \quad (2.4)$$

а второе решение $\varphi_2(y)$ – с помощью задания

$$d_0^{(2)} = 0, \quad d_1^{(2)} = 1. \quad (2.5)$$

Отметим здесь, что условия (1.8) и (2.4) соответствуют постановке задачи Коши в точке $y = -1$ для уравнения (1.6) с ненулевой лишь второй производной, а условия (1.8) и (2.5) – задаче Коши в точке $y = -1$ с ненулевой лишь третьей производной.

Решения $\varphi_1(y)$, $\varphi_2(y)$ и их производные (2.2) мы можем вычислить с любой наперед заданной точностью.

Теперь запишем общий вид решения уравнения (1.6) с краевыми условиями (1.8):

$$\varphi(y) = p_1\varphi_1(y) + p_2\varphi_2(y), \quad (2.6)$$

где p_1 и p_2 – произвольные комплексные постоянные, которые найдем далее в процессе решения задачи на СЗ.

2.3. Окрестность точки $y = 1$

Построим аналогично предыдущему представление общего решения уравнения (1.6) с краевыми условиями (1.9).

Запишем регулярное решение $\varphi(a, b, c, \alpha, R, \lambda; y)$, удовлетворяющее обоим условиям (1.9), в виде ряда

$$\varphi(a, b, c, \alpha, R, \lambda; y) = \sum_{k=0}^{\infty} e_k(1-y)^{k+2}, \quad |1-y| < \infty, \quad (2.7)$$

где коэффициенты e_k зависят от параметров задачи $a, b, c, \alpha, R, \lambda$. Ниже будет показана скорость убывания коэффициентов e_k .

Дифференцируя ряд (2.7) четыре раза по y , получаем необходимые для уравнения (1.6) производные:

$$\frac{\partial \Phi}{\partial y} = - \sum_{k=0}^{\infty} (k+2) e_k (1-y)^{k+1}, \quad \frac{\partial^2 \Phi}{\partial y^2} = \dots \quad (2.8)$$

Записывая функцию $U(y)$ из (1.3) в виде трехчлена по степеням $(1-y)^k$ и подставляя разложения (2.7), (2.8) в уравнение (1.6), получаем соотношение для коэффициентов e_k :

$$\begin{aligned} e_k = & \frac{-i\alpha^3 R}{(k+2)k(k^2-1)} e_{k-6} + \frac{i\alpha^3 R(2a+b)}{(k+2)k(k^2-1)} e_{k-5} + \\ & + \frac{i\alpha R[a(k-4)(k-1) - \alpha^2(a+b+c-\lambda)] - \alpha^4}{(k+2)k(k^2-1)} e_{k-4} - \\ & - \frac{i\alpha R(k-2)(b+2a)}{(k+2)(k+1)k} e_{k-3} + \frac{i\alpha R(a+b+c-\lambda) + 2\alpha^2}{(k+1)(k+2)} e_{k-2}, \\ & k \geq 2. \end{aligned} \quad (2.9)$$

Это уравнение отличается от (2.3) лишь сменой знака перед параметром b .

Выбирая значения e_0 и e_1 произвольными и учитывая, что

$$e_k = 0, \quad k = -1, -2, \dots,$$

мы вычисляем все коэффициенты e_k при $k = 2, 3, \dots$

2.4. Два решения в окрестности точки $y = 1$

Построим два линейно независимых решения $\Phi_3(y)$ и $\Phi_4(y)$ вида (2.7), удовлетворяющих краевым условиям (1.9). Первое из них, $\Phi_3(y)$, определим с помощью задания коэффициентов

$$e_0^{(1)} = 1, \quad e_1^{(1)} = 0, \quad (2.10)$$

а второе решение, $\Phi_4(y)$, определим с помощью задания

$$e_0^{(2)} = 0, \quad e_1^{(2)} = 1. \quad (2.11)$$

Отметим здесь, что условия (1.9) и (2.10) соответствуют постановке задачи Коши в точке $y = 1$ для уравнения (1.6) с ненулевой лишь второй производной, а условия (1.9) и (2.11) – задаче Коши в точке $y = 1$ с ненулевой лишь третьей производной.

Решения $\Phi_3(y)$ и $\Phi_4(y)$ и их производные (2.8) мы можем вычислить с любой наперед заданной точностью.

Теперь запишем общий вид решения уравнения (1.6) с краевыми условиями (1.9):

$$\Phi(y) = p_3 \Phi_3(y) + p_4 \Phi_4(y), \quad (2.12)$$

где p_3 и p_4 – произвольные комплексные постоянные, которые найдем далее.

2.5. Сшивка решений

Для нахождения решения уравнения (1.6), удовлетворяющего всем краевым условиям (1.8) и (1.9), необходимо гладко сшивать общие решения (2.6), (2.12) и их первые три производные в некоторой точке y_* .

Запишем условия сшивки:

$$\begin{aligned} p_1\Phi_1(y_*) + p_2\Phi_2(y_*) &= p_3\Phi_3(y_*) + p_4\Phi_4(y_*), \\ p_1\Phi_1'(y_*) + p_2\Phi_2'(y_*) &= p_3\Phi_3'(y_*) + p_4\Phi_4'(y_*), \\ p_1\Phi_1''(y_*) + p_2\Phi_2''(y_*) &= p_3\Phi_3''(y_*) + p_4\Phi_4''(y_*), \\ p_1\Phi_1'''(y_*) + p_2\Phi_2'''(y_*) &= p_3\Phi_3'''(y_*) + p_4\Phi_4'''(y_*). \end{aligned} \quad (2.13)$$

Введем вектор $\mathbf{z}$ с компонентами

$$z_1 = p_1, \quad z_2 = p_2, \quad z_3 = -p_3, \quad z_4 = -p_4$$

и перепишем систему (2.13) в векторном виде:

$$A\mathbf{z} = \mathbf{0}, \quad (2.14)$$

где

$$A = \begin{vmatrix} \Phi_1(y_*) & \Phi_2(y_*) & \Phi_3(y_*) & \Phi_4(y_*) \\ \Phi_1'(y_*) & \Phi_2'(y_*) & \Phi_3'(y_*) & \Phi_4'(y_*) \\ \Phi_1''(y_*) & \Phi_2''(y_*) & \Phi_3''(y_*) & \Phi_4''(y_*) \\ \Phi_1'''(y_*) & \Phi_2'''(y_*) & \Phi_3'''(y_*) & \Phi_4'''(y_*) \end{vmatrix} \quad (2.15)$$

с элементами $a_{jk} = \Phi_k^{(j)}(a, b, c, \alpha, R, \lambda; y_*)$. Для нахождения нетривиального решения системы (2.14) необходимо выполнение условия

$$\det(A) = 0,$$

которое эквивалентно равенству нулю вронскиана $\text{Wr}(\Phi_1, \Phi_2, \Phi_3, \Phi_4; \lambda, y_*)$:

$$\text{Wr}(\lambda) = \text{Wr}(\Phi_1, \Phi_2, \Phi_3, \Phi_4; \lambda, y_*) = 0. \quad (2.16)$$

Уравнение (2.16) является основным для нахождения искомых СЗ λ_m .

Вычислив значение λ_m , найдем коэффициенты p_k представлений (2.6), (2.12) решения исходной задачи (1.6), (1.8), (1.9). Поскольку собственные функции $\Phi_m(a, b, c, \alpha, R, \lambda_m; y)$ определены с точностью до мультипликативной константы, то положим $p_4 = 1$ и остальные коэффициенты p_1, p_2, p_3 найдем из любых трех уравнений системы (2.13), например из

$$\begin{aligned} p_1\Phi_1(y_*) + p_2\Phi_2(y_*) - p_3\Phi_3(y_*) &= \Phi_4(y_*), \\ p_1\Phi_1'(y_*) + p_2\Phi_2'(y_*) - p_3\Phi_3'(y_*) &= \Phi_4'(y_*), \\ p_1\Phi_1''(y_*) + p_2\Phi_2''(y_*) - p_3\Phi_3''(y_*) &= \Phi_4''(y_*). \end{aligned}$$

2.6. Итерационный метод вычисления СЗ λ

Для решения уравнения (2.16) был использован итерационный метод Ньютона:

$$\lambda^{(n+1)} = \lambda^{(n)} - \frac{\text{Wr}(\lambda^{(n)})}{\text{Wr}'(\lambda^{(n)})}, \quad n = 0, 1, \dots \quad (2.17)$$

Необходимая в (2.17) производная $\text{Wr}'(\lambda)$ вычислялась явно с помощью дифференцирования по λ элементов матрицы A из (2.15). Например, разложение для значения $\frac{\partial^4 \Phi_1(y_*)}{\partial \lambda \partial y^3}$ имеет вид

$$\frac{\partial^4 \Phi_1(y_*)}{\partial \lambda \partial y^3} = \sum_{k=1}^{\infty} (k+2)(k+1)k \frac{\partial d_k}{\partial \lambda} (y_* + 1)^{k-1}. \quad (2.18)$$

Производные коэффициентов $\frac{\partial d_k}{\partial \lambda}$ здесь вычисляются рекуррентно из соотношения, получаемого дифференцированием по λ уравнения (2.3):

$$\begin{aligned} \frac{\partial d_k}{\partial \lambda} = & \frac{-ia\alpha^3 R}{(k+2)k(k^2-1)} \frac{\partial d_{k-6}}{\partial \lambda} + \frac{i\alpha^3 R(2a-b)}{(k+2)k(k^2-1)} \frac{\partial d_{k-5}}{\partial \lambda} + \\ & + \frac{i\alpha R[a(k-4)(k-1) - \alpha^2(a+c-b-\lambda)] - \alpha^4 \partial d_{k-4}}{(k+2)k(k^2-1)} \frac{\partial d_{k-4}}{\partial \lambda} + \\ & + \frac{i\alpha^3 R}{(k+2)k(k^2-1)} d_{k-4} + \frac{i\alpha R(k-2)(b-2a)}{(k+2)(k+1)k} \frac{\partial d_{k-3}}{\partial \lambda} + \\ & + \frac{i\alpha R(a+c-b-\lambda) + 2\alpha^2 \partial d_{k-2}}{(k+1)(k+2)} \frac{\partial d_{k-2}}{\partial \lambda} - \frac{i\alpha R}{(k+1)(k+2)} d_{k-2}, \\ & k \geq 2. \end{aligned}$$

Значения коэффициентов d_0 и d_1 берутся из (2.4) с учетом $\frac{\partial d_0}{\partial \lambda} = \frac{\partial d_1}{\partial \lambda} = 0$.

Аналогично вычислялись все необходимые слагаемые в выражении для $Wr'(\lambda)$. Легко показать, что таких слагаемых всего 96, точность вычисления каждого из них может быть достигнута любая.

Начальные значения $\lambda^{(0)}$ в итерационном процессе (2.17) брались исходя из метода продолжения по параметру. Так если мы знаем спектр задачи при числе Рейнольдса R , то легко вычислить спектр при малом изменении числа $R_1 = R + \delta$.

Помимо этого эффективным инструментом аппроксимации собственных значений служило обобщение принципа аргумента для регулярной в области $\mathcal{G}$ функции $Wr(\lambda)$:

$$\sum_{k=1}^K \lambda_k^j = \frac{1}{2\pi i} \int_{\partial \mathcal{G}} \lambda^j \frac{Wr'(\lambda)}{Wr(\lambda)} d\lambda, \quad j = 0, 1, \dots, \quad (2.19)$$

где λ_k – нули функции $Wr(\lambda)$ в области $\mathcal{G}$, а K – число нулей в $\mathcal{G}$ с учетом их кратности. При значении $j = 0$ формула (2.19) представляет собой принцип аргумента.

Производная $Wr'(\lambda)$ в (2.19) вычислялась с использованием разложений типа (2.18), численное интегрирование проводилось по формуле Гаусса, а в качестве области $\mathcal{G}$ выбирались круги в плоскости спектрального параметра λ .

Точность получаемых СЗ проверялась сравнением со всеми доступными данными (см. [6], [18], [22], [26], [29]), а также перевычислением результатов с большей точностью вплоть до 100 верных дес. зн. ц. для общего течения Куэтта–Пуазейля с числом Рейнольдса R до значений $R = 10^6$.

3. УЧЕТ ЧЕТНОСТИ РЕШЕНИЙ

Если основной поток $U(y)$ является четной функцией $U(y) = ay^2 + c$, то решения $\varphi(y)$ уравнения (1.6) обладают свойством четности или нечетности. Учтем это явно с помощью вида разложения в точке $y = 0$.

3.1. Четные решения

Представив решение $\varphi(y)$ в виде

$$\varphi(y) = \sum_{k=0}^{\infty} g_k y^{2k}, \quad |y| < \infty, \quad (3.1)$$

найдя производные φ'' , φ^{IV} и подставив эти разложения в (1.6), получим для коэффициентов $g_k =$

$= g_k(a, c, \alpha, R, \lambda)$ рекуррентное уравнение

$$g_k = \frac{2\alpha^2 + i\alpha R(c - \lambda)}{2k(2k - 1)} g_{k-1} + \frac{i\alpha R[2a(k-3)(2k-3) - \alpha^2(c - \lambda)] - \alpha^4}{4k(2k-1)(k-1)(2k-3)} g_{k-2} - \frac{i\alpha^3 a R}{4k(2k-1)(k-1)(2k-3)} g_{k-3}, \quad k \geq 2. \quad (3.2)$$

Далее будет показана скорость падения коэффициентов g_k .

Два независимых четных решения $\varphi_3(a, c, \alpha, R, \lambda; y)$ и $\varphi_4(a, c, \alpha, R, \lambda; y)$ вида (3.1) построим с помощью различного выбора коэффициентов g_0 и g_1 :

$$g_0^{(1)} = 1, \quad g_1^{(1)} = 0, \quad g_0^{(2)} = 0, \quad g_1^{(2)} = 1.$$

Теперь запишем общий вид четного решения уравнения (1.6):

$$\varphi(y) = p_3 \varphi_3(y) + p_4 \varphi_4(y),$$

где константы p_3, p_4 найдем далее в процессе сшивки.

Необходимые в дальнейшем разложения для производных этих решений по аргументу y и параметру λ находятся аналогично предыдущему.

3.2. Нечетные решения

Представив решение $\varphi(y)$ в виде

$$\varphi(y) = \sum_{k=0}^{\infty} h_k y^{2k+1}, \quad |y| < \infty, \quad (3.3)$$

получим для коэффициентов $h_k = h_k(a, c, \alpha, R, \lambda)$ рекуррентное уравнение

$$h_k = \frac{2\alpha^2 + i\alpha R(c - \lambda)}{2k(2k + 1)} h_{k-1} + \frac{i\alpha R[2a(k-1)(2k-5) - \alpha^2(c - \lambda)] - \alpha^4}{4k(2k+1)(2k-1)(k-1)} h_{k-2} - \frac{i\alpha^3 a R}{4k(2k+1)(2k-1)(k-1)} h_{k-3}, \quad k \geq 2. \quad (3.4)$$

Два независимых нечетных решения $\varphi_3(a, c, \alpha, R, \lambda; y)$ и $\varphi_4(a, c, \alpha, R, \lambda; y)$ вида (3.3) построим с помощью различного выбора коэффициентов h_0 и h_1 :

$$h_0^{(1)} = 1, \quad h_1^{(1)} = 0, \quad h_0^{(2)} = 0, \quad h_1^{(2)} = 1.$$

Запишем теперь общий вид нечетного решения уравнения (1.6):

$$\varphi(y) = p_3 \varphi_3(y) + p_4 \varphi_4(y),$$

где константы p_3, p_4 найдем далее в процессе сшивки.

3.3. Сшивка решений

Для построения четного или нечетного решения $\varphi(y)$ задачи (1.6), (1.8), (1.9) необходимо гладко сшить разложение с центром в точке $y = -1$ и разложение с центром в точке $y = 0$; точку сшивки y_* наиболее удобно выбрать на интервале $y_* \in (-1, 0)$.

3.4. Нечетность функции потока $U(y)$

Когда скорость основного потока $U(y)$ является нечетной функцией, вещественной на отрезке $y \in [-1, 1]$:

$$U(y) = by, \quad \text{Im}(b) = 0, \quad (3.5)$$

то собственные значения и функции также имеют определенную симметрию. Так, если имеется СЗ λ_0 и соответствующая ему СФ $\phi_0(y)$, то, используя комплексное сопряжение для уравнения (1.6), убеждаемся, что значение λ_1 и функция $\phi_1(y)$:

$$\lambda_1 = -\lambda_0^*, \quad \phi_1(y) = \phi_0^*(-y), \quad (3.6)$$

также являются собственными. Таким образом, в случае (3.5) СЗ обладают симметрией относительно мнимой оси.

4. АСИМПТОТИКА РЕШЕНИЙ РЕКУРРЕНТНЫХ УРАВНЕНИЙ

Полученные рекуррентные уравнения (2.3), (2.9) и (3.2), (3.4) являются линейными однородными уравнениями шестого и третьего порядков соответственно и могут быть исследованы с помощью теории линейных разностных уравнений. Теория Паункаре–Перрона (см. [33]) в этом случае не может быть применена, поскольку соответствующие характеристические уравнения имеют кратные корни, все равные нулю. Для этого случая используется теория Биркгофа (см. [34]–[36]), являющаяся обобщением теории Паункаре–Перрона.

4.1. Уравнения шестого порядка

Уравнения (2.3), (2.9) имеют шесть линейно независимых решений. Запишем асимптотику решений лишь уравнения (2.3), а для уравнения (2.9) эта асимптотика получается заменой знака перед параметром b .

Удобно записать не сами решения $d_k^{(m)}$, $m = 1, \dots, 6$, а асимптотику отношений $d_k^{(m)}/d_{k-1}^{(m)}$ при $k \rightarrow \infty$. Используя результаты работ [34]–[36], получаем для (2.3), что первые два решения имеют один характер асимптотики, а последние четыре – другой характер:

$$\frac{d_k^{(m)}}{d_{k-1}^{(m)}} = \frac{D_1^{(m)}}{k} + \frac{D_2^{(m)}}{k^2} + \dots, \quad k \rightarrow \infty, \quad m = 1, 2, \quad (4.1)$$

$$\frac{d_k^{(m)}}{d_{k-1}^{(m)}} = \frac{C_1^{(m)}}{\sqrt{k}} + \frac{C_2^{(m)}}{k} + \dots, \quad k \rightarrow \infty, \quad m = 3, 4, 5, 6. \quad (4.2)$$

Приведем значения первых двух коэффициентов D_1, D_2 и C_1, C_2 :

$$D_1^{(m)} = \pm\alpha, \quad D_2^{(m)} = -3D_1^{(m)}, \quad (4.3)$$

$$C_1^{(m)} = (i\alpha Ra)^{1/4}, \quad C_2^{(m)} = \frac{b-2a}{4a}(C_1^{(m)})^2, \quad a \neq 0, \quad (4.4)$$

а все последующие коэффициенты могут быть найдены рекуррентно с помощью символьных вычислений.

Четыре различных ветви корня в равенстве (4.4) задают четыре различных решения для $C_1^{(m)}$ и последующих коэффициентов. Два независимых решения $D_1^{(m)}$ и последующих коэффициентов определяются выбором знака $\pm$ в (4.3).

4.2. Уравнения пятого порядка

В случае $a = 0$ уравнения (2.3), (2.9) не содержат коэффициентов с номером $k - 6$, порядок уравнений равен пяти и они имеют пять линейно независимых решений. Асимптотика отношений $d_k^{(m)}/d_{k-1}^{(m)}$ при $k \rightarrow \infty$ для $m = 1, \dots, 5$ строится также с помощью теории Биркгофа. Получаем, что первые два решения не изменяют характера асимптотики (4.1), а последние три решения имеют отличающееся от (4.2) поведение:

$$\frac{d_k^{(m)}}{d_{k-1}^{(m)}} = \frac{C_1^{(m)}}{k^{2/3}} + \frac{C_2^{(m)}}{k^{4/3}} + \dots, \quad k \rightarrow \infty, \quad m = 3, 4, 5,$$

где для первых коэффициентов $C_n^{(m)}$ было получено

$$C_1^{(m)} = (i\alpha Rb)^{1/3}, \quad C_2^{(m)} = -\frac{\alpha^2}{3C_1^{(m)}}. \quad (4.5)$$

Три различных ветви корня в (4.5) задают три различных решения для $C_1^{(m)}$ и последующих коэффициентов.

4.3. Уравнения третьего порядка

В случае существования четного решения (3.1) коэффициенты g_k удовлетворяют уравнению (3.2) третьего порядка, для асимптотики трех его независимых решений было получено разложение

$$\frac{g_k^{(1)}}{g_{k-1}^{(1)}} = \frac{D_1}{k^2} + \frac{D_2}{k^3} + \dots, \quad k \rightarrow \infty, \quad (4.6)$$

$$\frac{g_k^{(m)}}{g_{k-1}^{(m)}} = \frac{C_1^{(m)}}{k} + \frac{C_2^{(m)}}{k^2} + \dots, \quad k \rightarrow \infty, \quad m = 2, 3. \quad (4.7)$$

Первые два коэффициента этих разложений таковы:

$$D_1 = \frac{\alpha^2}{4}, \quad D_2 = \frac{\alpha^2}{8},$$

$$C_1^{(m)} = \pm \frac{\sqrt{i\alpha R a}}{2}, \quad C_2^{(m)} = \frac{\alpha^2 + i\alpha R(c - \lambda) - 10C_1^{(m)}}{8}, \quad a \neq 0.$$

В случае существования нечетного решения вида (3.3) коэффициенты h_k удовлетворяют уравнению (3.4) третьего порядка, для асимптотики трех его независимых решений получены разложения (4.6), (4.7) с коэффициентами

$$D_1 = \frac{\alpha^2}{4}, \quad D_2 = -\frac{\alpha^2}{8},$$

$$C_1^{(m)} = \pm \frac{\sqrt{i\alpha R a}}{2}, \quad C_2^{(m)} = \frac{\alpha^2 + i\alpha R(c - \lambda) - 14C_1^{(m)}}{8}, \quad a \neq 0.$$

4.4. Уравнения второго порядка в случае $a = 0$

В случае $a = 0$ уравнения (3.2), (3.4) для четных и нечетных решений не содержат коэффициентов с номером $k - 3$, порядок уравнений равен двум и они имеют по два линейно независимых решения. Асимптотика отношений $g_k^{(m)}/g_{k-1}^{(m)}$ и $h_k^{(m)}/h_{k-1}^{(m)}$ при $k \rightarrow \infty$ для $m = 1, 2$ имеет вид

$$\frac{g_k^{(m)}}{g_{k-1}^{(m)}} = \frac{C_1^{(m)}}{k^2} + \frac{C_2^{(m)}}{k^3} + \dots, \quad \frac{h_k^{(m)}}{h_{k-1}^{(m)}} = \frac{D_1^{(m)}}{k^2} + \frac{D_2^{(m)}}{k^3} + \dots,$$

где для первых коэффициентов $C_1^{(m)}, C_2^{(m)}, D_1^{(m)}, D_2^{(m)}$ получены выражения

$$C_1^{(1)} = D_1^{(1)} = \frac{\alpha^2}{4}, \quad C_1^{(2)} = D_1^{(2)} = \frac{\alpha^2 + i\alpha R(c - \lambda)}{4},$$

$$C_2^{(m)} = \frac{\frac{5}{2}\alpha C_1^{(m)}[2\alpha + iR(c - \lambda)] - 8(C_1^{(m)})^2 - \frac{3}{4}\alpha^3[\alpha + iR(c - \lambda)]}{8C_1^{(m)} - 2\alpha^2 - i\alpha R(c - \lambda)}, \quad (4.8)$$

$$D_2^{(m)} = \frac{\frac{3}{2}\alpha D_1^{(m)}[2\alpha + iR(c - \lambda)] - 8(D_1^{(m)})^2 - \frac{1}{4}\alpha^3[\alpha + iR(c - \lambda)]}{8D_1^{(m)} - 2\alpha^2 - i\alpha R(c - \lambda)}; \quad (4.9)$$

в формулах (4.8), (4.9) индекс $m = 1, 2$.

Из полученного асимптотического поведения коэффициентов d_k, e_k, g_k, h_k разложений (2.1), (2.7), (3.1), (3.3), согласно признаку Даламбера сходимости рядов (см. [37]), следует, что все разложения сходятся для любой конечной точки $y \in \mathbb{C}$, причем скорость сходимости весьма высокая.

5. СПЕКТР ЗАДАЧИ ДЛЯ ТЕЧЕНИЯ КУЭТТА

5.1. В [38], [39] с помощью свойств функции Эйри $\text{Ai}(t)$ было доказано, что течение Куэтта устойчиво при всех волновых числах $\alpha > 0$ и числах Рейнольдса $R > 0$. Вместе с тем динамика СЗ при изменении параметров α и R оказалась довольно сложной, что и заставило нас уделить этому особое внимание.

Представим результаты расчета спектра задачи для течения Куэтта с потоком $U(y) = y$. Точность вычислений при этом достигала 100 верных дес. зн. ц., что полностью исключило возможность “артефактов”.

Как было отмечено в п. 3.4, спектр обладает симметрией (3.6) относительно мнимой оси.

Как показано в [11], [12], при больших числах R спектр расположен на нескольких ветвях в полуполосе $\{\lambda; \text{Re}(\lambda) \in (-1, 1), \text{Im}(\lambda) < 0\}$. Две ветви расположены сверху и снизу от отрезка, соединяющего две точки комплексной плоскости с координатами $\lambda = -i/\sqrt{3}$ и $\lambda = 1$; на этих ветвях содержится конечное число точек спектра. Третья и четвертая ветви симметричны первым двум относительно мнимой оси. Последняя ветвь включает чисто мнимые СЗ на луче $\{\text{Im}(\lambda) < -1/\sqrt{3}\}$, их счетное множество с предельной точкой $\lambda = -i\infty$.

На фиг. 1 представлены вычисленные СЗ задачи (отмечены кружочками) для значений параметров $\alpha = 1, R = 4000$; на мнимой оси здесь показаны СЗ с $\text{Im}(\lambda) > -1$, сплошной линией отмечены отрезки $\mathcal{L}_1$ и $\mathcal{L}_2$, соединяющие точки $\lambda = -1$ и $\lambda = 1$ с точкой $\lambda_* = -i/\sqrt{3}$. Это распределение СЗ совпадает с результатами в [11], [12], [6].

Такую картину называют “спектральный галстук”, детальному исследованию этого портрета при больших числах Рейнольдса R посвящены работы [11], [12]. В частности, там выведены асимптотические при $R \rightarrow \infty$ формулы для координат СЗ на всех ветвях. Из этих результатов следует, что первые четыре ветви стремятся к своему предельному положению – отмеченным отрезкам $\mathcal{L}_1$ и $\mathcal{L}_2$, причем расстояние до отрезков имеет асимптотику $\sim \log R/\sqrt{R}$.

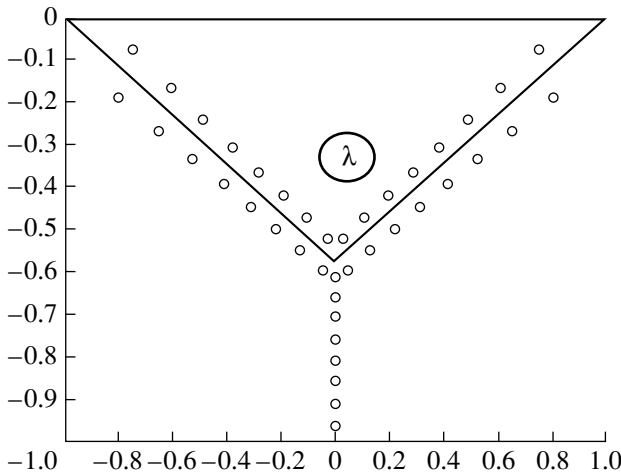
При увеличении числа R количество СЗ и их плотность на первых четырех ветвях возрастают. На фиг. 2 представлен спектр задачи для значений параметров $\alpha = 1, R = 10^4$ и отрезки $\mathcal{L}_1$ и $\mathcal{L}_2$; на мнимой оси показаны СЗ с $\text{Im}(\lambda) > -1$. Над точкой $\lambda_* = -i/\sqrt{3}$ здесь видны два близких СЗ $\lambda = -0.545i$ и $\lambda = -0.554i$.

На фиг. 3 дано сравнение в окрестности отрезка $\mathcal{L}_1$ (показан сплошной линией) точных СЗ (отмечены кружочками) и асимптотик из [11], [12] (отмечены крестиками) для $\alpha = 1, R = 10^4$. Видна высокая точность результатов из [11], [12], причем при увеличении числа R эта точность становится еще выше.

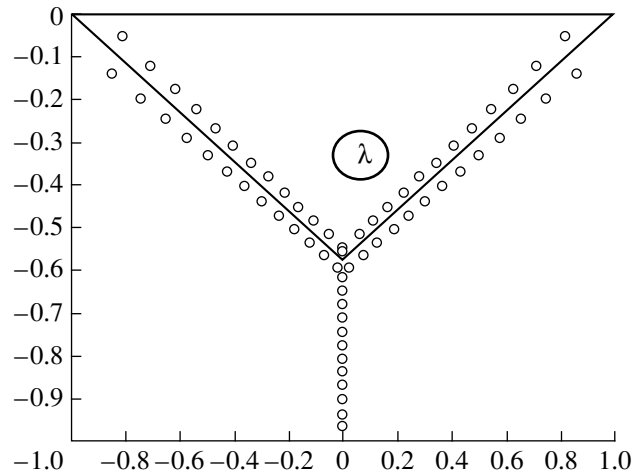
Однако в окрестности узловой точки пересечения ветвей, т.е. точки $\lambda_* = -i/\sqrt{3}$, формулы из [11], [12] перестают работать и здесь необходим детальный численный анализ спектра. Как будет показано далее, именно здесь при увеличении числа R происходит переход точек спектра с нижней ветви на четыре первые ветви. Далее приведем результаты расчета спектра λ_n при фиксированном значении волнового числа $\alpha = 1$ и увеличении числа $R > 0$. Значения λ_n будем рассматривать как функции числа R , т.е. $\lambda_n = \lambda_n(R)$.

При малых значениях R весь спектр λ_n лежит на мнимой отрицательной оси. Приведем с точностью 5 дес. зн. ц. первые 4 значения λ_n для числа $R = 1$:

$$\lambda_1 = -9.31481i, \quad \lambda_2 = -20.57596i, \quad \lambda_3 = -38.94677i, \quad \lambda_4 = -60.05523i.$$



Фиг. 1.



Фиг. 2.

При увеличении числа R до величины $R = 49$ все значения λ_n приближаются к началу координат. Дадим первые 4 значения λ_n при $R = 49$:

$$\lambda_1 = -0.28817i, \quad \lambda_2 = -0.66644i, \quad \lambda_3 = -0.75028i, \quad \lambda_4 = -1.18513i.$$

При дальнейшем увеличении числа $R > 49$ значение $\lambda_1(R)$ останавливается при $R_a = 49.0570$, достигая величины $\lambda_1(R_a) = -0.288165i$, а затем удаляется от начала координат, медленно приближаясь к значению $\lambda_2(R)$.

Параллельно этому процессу при достижении $R = R_b = 54.93702$ значение λ_3 останавливается, достигая величины $\lambda_3(R_b) = -0.715297i$, а затем удаляется от начала координат, приближаясь к значению $\lambda_4(R)$. При числе $R = R_1 = 61.917759$ значения $\lambda_3(R)$ и $\lambda_4(R)$ сталкиваются, образуя первое двойное собственное значение

$$\lambda_3(R_1) = \lambda_4(R_1) = -0.799834979i. \quad (5.1)$$

При дальнейшем увеличении $R > R_1$ это двойное СЗ распадается на два простых, движущихся в комплексной плоскости симметрично относительно мнимой оси и удаляющихся от нее.

Численный расчет показывает, что при комплексификации параметра R в окрестности точки $R = R_1$ два СЗ $\lambda_3(R)$ и $\lambda_4(R)$ имеют поведение

$$\lambda_{3,4}(R) = \pm\sqrt{R - R_1}\Psi(R) + \Phi(R), \quad (5.2)$$

где $\Psi(R)$ и $\Phi(R)$ – регулярные функции в окрестности точки $R = R_1$, т.е. функции $\lambda_3(R)$ и $\lambda_4(R)$ имеют при $R = R_1$ точку ветвления второго порядка.

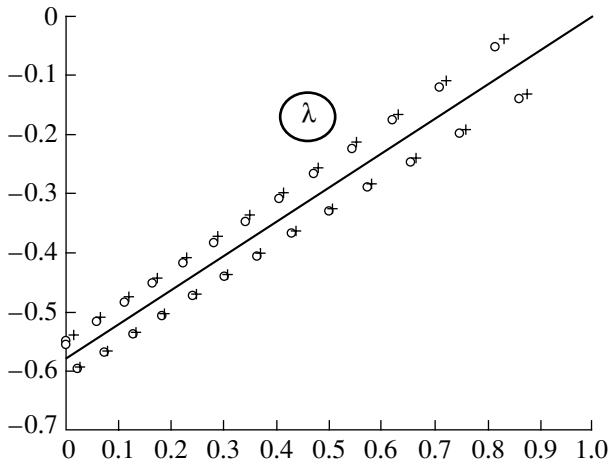
Аналогичное поведение имеет пара $\lambda_1(R)$ и $\lambda_2(R)$ при дальнейшем увеличении числа Рейнольдса R . При $R_2 = 65.52023$ значения $\lambda_1(R)$ и $\lambda_2(R)$ сталкиваются на мнимой оси, образуя второе двойное СЗ

$$\lambda_1(R_2) = \lambda_2(R_2) = -0.38816096i, \quad (5.3)$$

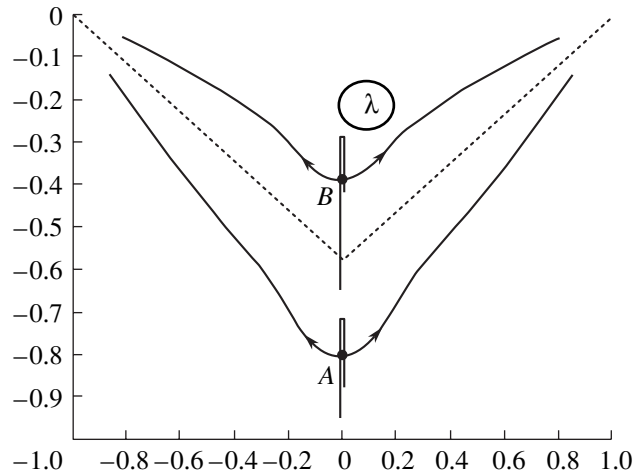
которое при дальнейшем увеличении $R > R_2$ также распадается на два простых, симметричных относительно мнимой оси. Таким образом, точка $R = R_2$ также является точкой ветвления второго порядка для функций $\lambda_1(R)$ и $\lambda_2(R)$.

Траектории СЗ $\lambda_1(R)$, $\lambda_2(R)$ и $\lambda_3(R)$, $\lambda_4(R)$ при увеличении числа $R \in [25, 10^4]$ показаны на фиг. 4. Для наглядности движение СЗ вверх и вниз по мнимой оси здесь представлено в виде двух близких параллельных путей. Точка A соответствует первому двойному СЗ (5.1), точка B – второму двойному СЗ (5.3). Пунктиром показаны отрезки $\mathcal{L}_1$ и $\mathcal{L}_2$, к которым значения $\lambda_1(R)$ и $\lambda_2(R)$ симметрично приближаются сверху, а значения $\lambda_3(R)$ и $\lambda_4(R)$ – снизу.

Последующие значения $\lambda_n(R)$ с номерами $n > 4$ при увеличении R ведут себя более сложным образом и их надо рассматривать четверками.



Фиг. 3.



Фиг. 4.

Значение $\lambda_5(R)$ ведет себя аналогично поведению $\lambda_1(R)$: сначала оно приближается к началу координат по мнимой оси, затем останавливается при $R = 224.300$, потом удаляется от начала координат. Дальнейшую его динамику опишем ниже.

Следующие два значения $\lambda_6(R)$ и $\lambda_7(R)$ вначале ведут себя аналогично поведению $\lambda_1(R)$ и $\lambda_2(R)$: значение $\lambda_6(R)$ приближается к началу координат, затем останавливается, потом движется навстречу $\lambda_7(R)$ и образует третье двойное СЗ при $R = R_3 = 205.77778$:

$$\lambda_6(R_3) = \lambda_7(R_3) = -0.66527009i. \quad (5.4)$$

При числе $R = R_3$ это двойное СЗ распадается на два простых СЗ, движущихся симметрично в комплексной плоскости, сначала удаляясь от мнимой оси на расстояние порядка 0.01, а затем приближаясь к ней. При $R = R_4 = 214.40338$ эти два простых СЗ опять сталкиваются на мнимой оси и образуют следующее, четвертое двойное СЗ:

$$\lambda_6(R_4) = \lambda_7(R_4) = -0.64739767i. \quad (5.5)$$

На фиг. 5 показаны траектории значений $\lambda_6(R)$ и $\lambda_7(R)$ при увеличении числа $R \in [182, 228]$. Точка A здесь соответствует двойному СЗ (5.4), точка B – двойному СЗ (5.5).

При дальнейшем увеличении числа $R > R_4$ это четвертое двойное СЗ распадается на два простых СЗ, движущихся в разные стороны по мнимой оси. Далее одно из этих СЗ, движущихся от начала координат, при числе $R = R_5 = 233.273196$ сталкивается со значением $\lambda_8(R)$ и образует пятое двойное СЗ:

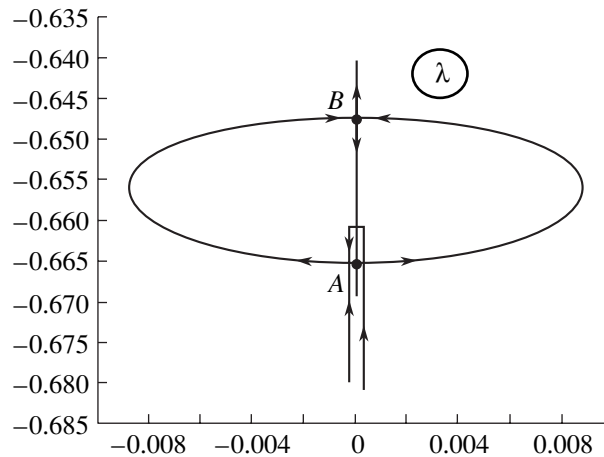
$$\lambda_8(R_5) = \lambda_{(6,7)}(R_5) = -0.70574959i. \quad (5.6)$$

Здесь необходимо отметить, что после распада двойного СЗ на два простых мы не можем однозначно приписать номер каждому простому СЗ – необходимо некоторое правило их нумерации. Поэтому в (5.6) у второго СЗ использован индекс (6, 7), показывающий, что только одно из исходных значений $\lambda_6(R)$ и $\lambda_7(R)$ участвует в образовании двойного СЗ (5.6).

При дальнейшем увеличении числа $R > R_5$ двойное СЗ (5.6) распадается на два простых значения, движущихся симметрично в комплексной плоскости и приближающихся к предельным отрезкам $\mathcal{L}_1$ и $\mathcal{L}_2$ снизу. Траектории этих СЗ аналогичны нижним “усам” на фиг. 4.

Одновременно с описанным последним процессом при увеличении числа $R > R_4$, как было сказано ранее, четвертое двойное СЗ (5.5) распадается на два простых чисто мнимых значения. Одно из этих СЗ $\lambda_{(6,7)}$ движется к началу координат и при числе $R = R_6 = 253.42383$ сталкивается со значением $\lambda_5(R)$, образуя шестое двойное СЗ:

$$\lambda_5(R_6) = \lambda_{(6,7)}(R_6) = -0.458236635i. \quad (5.7)$$



Фиг. 5.

При дальнейшем увеличении числа $R > R_6$ двойное СЗ (5.7) распадается на два простых значения, движущихся симметрично в комплексной плоскости и приближающихся к предельным отрезкам $\mathcal{L}_1$ и $\mathcal{L}_2$ сверху. Траектории этих СЗ аналогичны верхним “усам” на фиг. 4.

На этом заканчивается описание траекторий “квартета” взаимосвязанных СЗ с номерами $n = 5, 6, 7, 8$.

При дальнейшем увеличении числа R все последующие СЗ с номерами $4m + 1, 4m + 2, 4m + 3, 4m + 4$ ($m \geq 2$) ведут себя аналогично этой четверке. Сначала два СЗ $\lambda_{4m+2}(R)$ и $\lambda_{4m+3}(R)$ сталкиваются на мнимой оси, образуя новое двойное СЗ с номером $4m - 1$, затем оно распадается на два простых СЗ, движущихся симметрично в комплексной плоскости по дугам, аналогичным фиг. 5. Далее эти СЗ опять сталкиваются на мнимой оси, образуя следующее двойное СЗ с номером $4m$, которое затем распадается на два простых чисто мнимых СЗ. Далее одно из этих СЗ сталкивается с СЗ $\lambda_{4m+4}(R)$, образуя следующее двойное СЗ с номером $4m + 1$, которое затем распадается на два простых симметричных СЗ, движущихся в плоскости и приближающихся к предельным отрезкам $\mathcal{L}_1$ и $\mathcal{L}_2$ снизу. Далее второе СЗ из пары с номерами $4m + 2$ и $4m + 3$ сталкивается с СЗ $\lambda_{4m+1}(R)$, образуя новое двойное СЗ с номером $4m + 2$; затем оно распадается на два симметричных простых СЗ, приближающихся к отрезкам $\mathcal{L}_1$ и $\mathcal{L}_2$ сверху.

Отличие в поведении каждой следующей четверки состоит в том, что новые двойные СЗ расположены все ближе к узловой точке $\lambda_* = -i/\sqrt{3}$, а траектории между двумя двойными СЗ, аналогичные фиг. 5, все сильнее прижимаются к мнимой оси. Так, СЗ $\lambda_{34}(R)$ и $\lambda_{35}(R)$ при числе $R_{31} = 4454.06$ образуют двойное СЗ $\lambda_{34,35}(R_{31}) = -0.6111i$, которое затем распадается на два простых СЗ, удаляющихся от мнимой оси на расстояние 0.005 и образующих при $R_{32} = 4567.6$ следующее двойное СЗ; $\lambda_{34,35}(R_{32}) = -0.6014i$.

Аналогичная картина образования и распада двойных СЗ имеет место при изменении волнового числа $\alpha > 0$.

Проведенный численный анализ при вещественных $\alpha > 0$ и числах Рейнольдса вплоть до $R = 10^6$ подтвердил, что все СЗ для течения Куэтта остаются в полуполосе $\{\lambda: \operatorname{Re}(\lambda) \in (-1, 1), \operatorname{Im}(\lambda) < 0\}$, т.е. течение устойчиво при $\alpha > 0$. Далее будут представлены результаты при комплексном $\alpha > 0$, говорящие о возникновении неустойчивости.

5.2. Вычисление точек ветвления

При решении уравнения (2.16) методом Ньютона (2.17) трудности возникают лишь для СЗ, близких к двойным значениям. Это связано с тем, что производная $W'(\lambda)$ становится близкой к нулю (а в самом двойном СЗ величина $W'(\lambda) = 0$) и процесс (2.17) оказывается очень медленно сходящимся.

Таблица

m, n	R_k	$\lambda_m(R_k) = \lambda_n(R_k)$
3, 4	61.9177587281	$-0.799834980826i$
1, 2	65.5202290647	$-0.388160961976i$
6, 7	205.7777805692	$-0.665270088736i$
6, 7	214.4033833588	$-0.647397672095i$
8, (6, 7)	233.2731959346	$-0.705749592614i$
5, (6, 7)	253.4238279236	$-0.458236637641i$
10, 11	469.8812933897	$-0.649812853640i$
10, 11	494.5999327932	$-0.628386917108i$
12, (10, 11)	517.2435245325	$-0.670248402414i$
9, (10, 11)	559.0809339956	$-0.488334539640i$

Для преодоления этой трудности был использован модифицированный метод Ньютона, учитывающий вторую производную $W\Gamma''$:

$$\lambda^{(k+1)} = \lambda^{(k)} - T \pm \sqrt{T^2 - \frac{2W\Gamma(\lambda^{(k)})}{W\Gamma''(\lambda^{(k)})}}, \quad T = \frac{W\Gamma'(\lambda^{(k)})}{W\Gamma''(\lambda^{(k)})}, \quad (5.8)$$

где ветвь в (5.8) выбиралась такой, чтобы обеспечить непрерывность соответствующей функции $\lambda(R)$. Необходимые в (5.8) производные $W\Gamma'(\lambda)$ и $W\Gamma''(\lambda)$ вычислялись явно с помощью дифференцирования коэффициентов разложений (2.1) и (2.7).

Скорость сходимости и устойчивость процесса (5.8) была принципиально выше сходимости и устойчивости метода (2.17).

Точки ветвления R_k и соответствующие им двойные СЗ $\lambda_m(R_k) = \lambda_n(R_k)$ вычислялись также на основе представления (5.2). Из него следует, что

$$\Lambda(R) = [\lambda_m(R) - \lambda_n(R)]^2$$

является регулярной функцией в точке ветвления R_k и имеет там простой ноль. Это позволяет вычислить величину R_k с высокой точностью.

Приведем таблицу первых десяти точек ветвления R_k и соответствующих им двойных СЗ $\lambda_m(R_k) = \lambda_n(R_k)$ с номерами m и n для течения Куэтта с волновым числом $\alpha = 1$. Номера в скобках $(n, n+1)$, как и ранее, показывают, что лишь одно СЗ из исходных значений $\lambda_n(R)$ и $\lambda_{n+1}(R)$ участвует в образовании приводимого двойного СЗ. Т.е. запись в первом столбце 8, (6, 7) означает, что в образовании этого двойного СЗ участвует СЗ с номером 8 и одно из СЗ с номерами 6 и 7.

5.3. Спектр в случае комплексных волновых чисел α

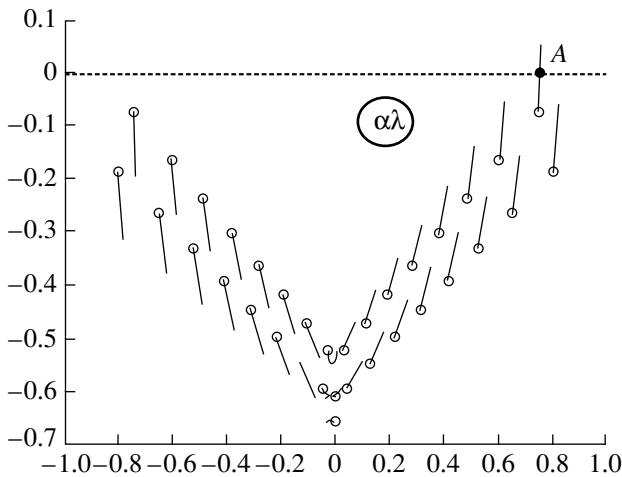
При возникновении возмущений с не вещественным α , т.е. при представлении функции тока $\Psi(x, y, t)$ в виде

$$\Psi(x, y, t) = \varphi(y)e^{i\alpha x - i\alpha\lambda t}$$

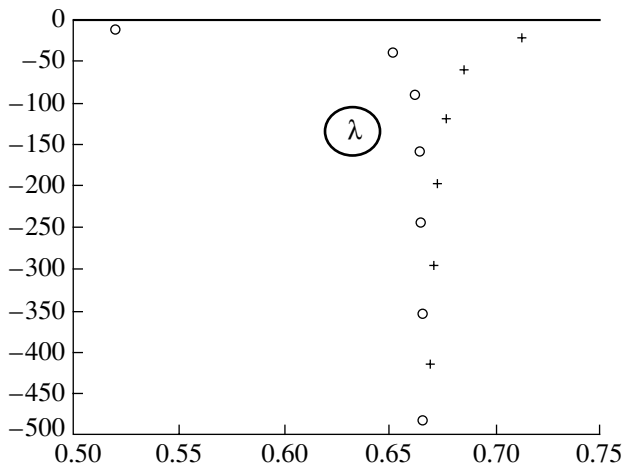
мы получаем, что с ростом времени $t > 0$ неустойчивость будет развиваться, если найдется хоть одно СЗ λ_n , для которого $\text{Im}(\alpha\lambda_n) > 0$.

Приведем результаты расчета величин $\alpha\lambda_n$ при фиксированном числе $R = 4000$ и непрерывном увеличении лишь $\gamma = \text{Im}(\alpha)$. На фиг. 6 даны траектории величин $\alpha\lambda_n$ при увеличении $\gamma \in [0, 0.15]$, когда $\alpha = 1 + i\gamma$. Кружочки – значения $\alpha\lambda_n$ при $\gamma = 0$; эти величины совпадают с СЗ, показанными на фиг. 1. Из фиг. 6 видно, что симметрии относительно мнимой оси теперь нет и уже при сравнительно малом значении $\text{Im}(\alpha) \approx 0.08637$ происходит переход первой величины $\alpha\lambda_1$ в верхнюю полуплоскость (отмечено точкой А).

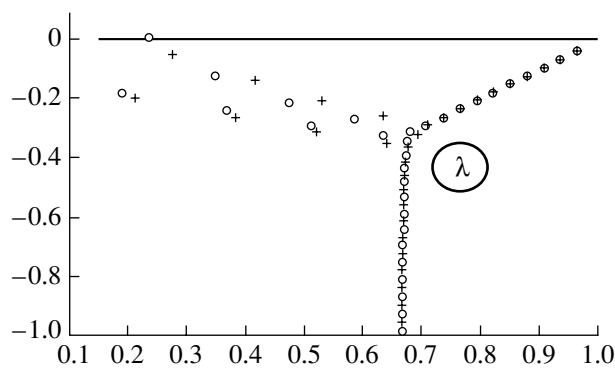
Таким образом, внесение даже затухающих по продольной координате x возмущений приводит для течения Куэтта к развитию неустойчивости во времени.



Фиг. 6.



Фиг. 7.



Фиг. 8.

6. СПЕКТР ЗАДАЧИ ДЛЯ ТЕЧЕНИЯ ПУАЗЕЙЛЯ

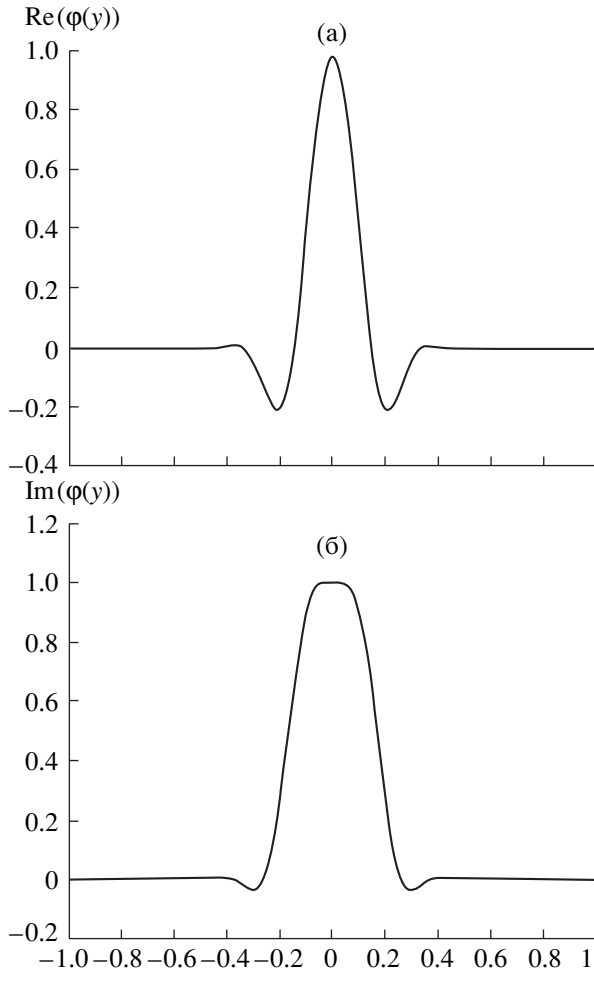
Представим результаты вычисления СЗ в случае течения Пуазейля $U(y) = 1 - y^2$. На фиг. 7 дана часть спектра задачи для чисел $\alpha = 1$ и $R = 1$. Кружочками отмечены СЗ, соответствующие четным СФ, а крестиками – соответствующие нечетным СФ. Значения λ_n с ростом номера n асимптотически приближаются к прямой $\text{Re}(\lambda) = 2/3$, причем $\text{Im}(\lambda_n) \rightarrow -\infty$.

При фиксированном $\alpha = 1$ и непрерывном увеличении R все СЗ сначала приближаются к вещественной оси $\text{Re}(\lambda)$, расходясь на две ветви налево и направо. При значении $R_1 = 5814.828757$ первое СЗ $\lambda_1(R_1) = 0.26123274$, соответствующее четной СФ, пересекает ось $\text{Im}(\lambda) = 0$ и затем движется в верхней полуплоскости. При $R_2 = 31956.451004$ это СЗ $\lambda_1(R_2) = 0.19201999$ опять возвращается в нижнюю полуплоскость, что опять соответствует устойчивости течения. Все остальные СЗ λ_n , $n > 1$, как показывают расчеты, остаются в нижней полуплоскости при всех $R > 0$ и $\alpha > 0$. Это поведение СЗ полностью согласуется с известными результатами (см. [6], [18]).

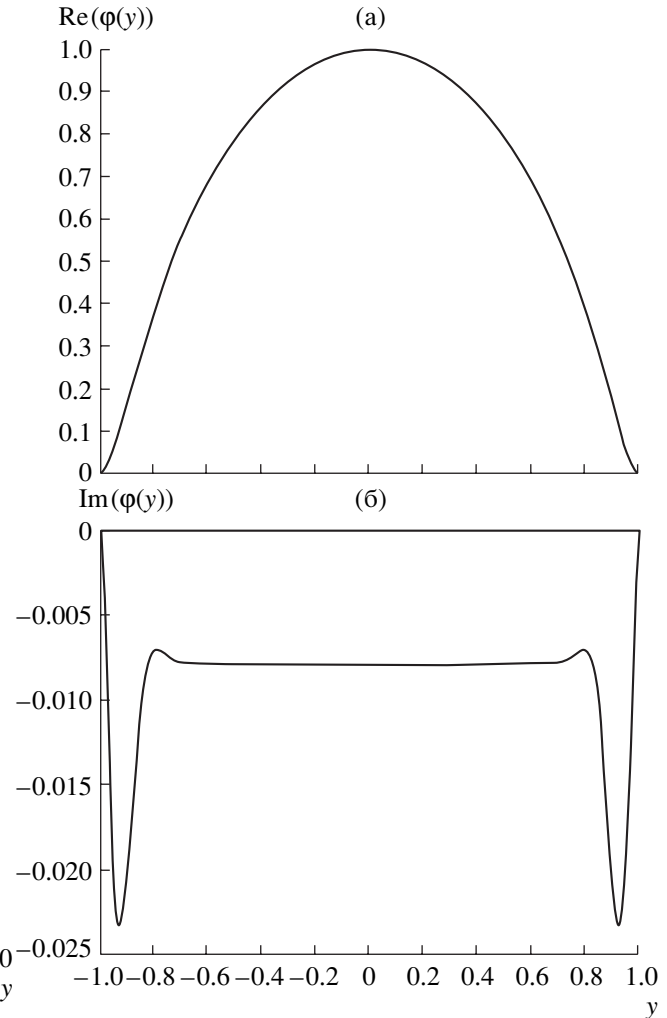
На фиг. 8 дан спектр задачи для параметров $\alpha = 1$ и $R = 10^4$ (кружочки – СЗ для четных СФ, а крестики – для нечетных СФ). Портрет этого спектра приведен также в [6], таблица первых 32 СЗ представлена в [18] с 5–8 дес. зн. ц., а неустойчивое СЗ вычислено также в [26] и [29]. Характерным здесь является чрезвычайно близкое расположение СЗ на правой ветви портрета для четных и нечетных СФ. Так, разность между соответствующими СЗ здесь достигает величины 2×10^{-5} .

Сравнение полученных высокоточных результатов с данными из [18] показало, что в [18] пропущено одно СЗ $\lambda = 0.2127257824 - 0.1993606948i$ (на фиг. 8 это самое левое СЗ для нечетной СФ), а точность некоторых λ_n с номерами $n > 24$ падает до 4 дес. зн. ц.

Сравнение λ_1 с данными из [26] и [29] показало падение точности результатов в [26] до 5 дес. зн. ц., а результаты [29] имели верные все 10 дес. зн. ц.



Фиг. 9.



Фиг. 10.

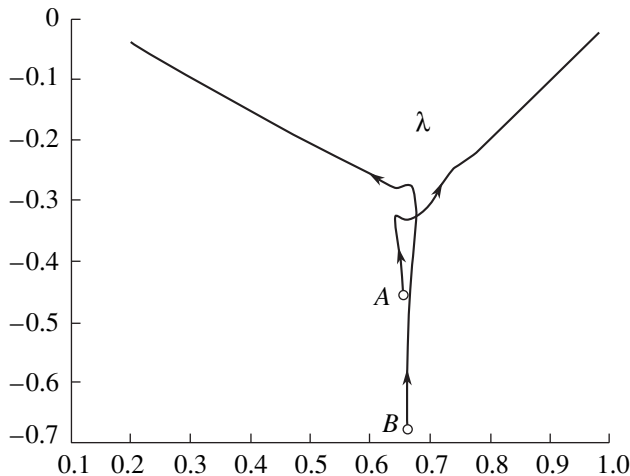
На фиг. 9 представлены вещественная (график (а)) и мнимая (график (б)) части четной СФ $\varphi(y)$, соответствующей неустойчивому СЗ $\lambda = 0.2375264888 + 0.00373967i$ для параметров $\alpha = 1$ и $R = 10^4$.

На фиг. 10 показаны вещественная (график (а)) и мнимая (график (б)) части четной СФ $\varphi(y)$, соответствующей первому устойчивому СЗ $\lambda = 0.96464251 - 0.03518658i$ для параметров $\alpha = 1$ и $R = 10^4$. Здесь ярко виден эффект смены характера решения $\varphi(y)$ в пограничном слое; подобная картина имеет место и для других СФ.

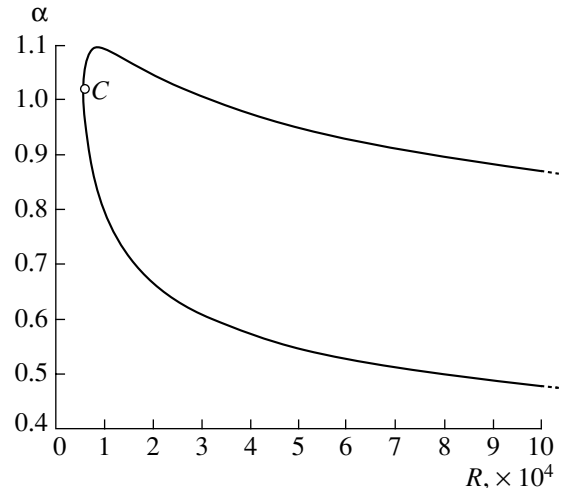
Траектории СЗ $\lambda_n(R)$, рассматриваемые как функции вещественного числа R при фиксированном волновом числе $\alpha > 0$, имеют сложное поведение в окрестности узловой точки $\lambda_* = 2/3 - i/3$. На фиг. 11 показаны траектории двух соседних СЗ $\lambda_n(R)$ при фиксированном $\alpha = 1$ и увеличении R . Траектория, начинающаяся в точке A и уходящая направо от узловой точки, представлена для $R \in [190, 10^5]$, а начинающаяся в точке B и уходящая налево – для $R \in [350, 10^5]$. Подобное сложное движение совершают и другие СЗ: они сначала поднимаются вверх вблизи вертикальной прямой $\text{Re}(\lambda) = 2/3$ до узловой точки, а затем расходятся на две ветви налево и направо.

Однако точек ветвления R_k и соответствующих им двойных СЗ при вещественных значениях $\alpha > 0$ и $R > 0$ обнаружить не удалось: окрестность узловой точки λ_* соседние СЗ проходят при различных числах R . По-видимому, эти особые точки могут быть найдены при комплексных α и R .

При выборе другого вещественного волнового числа $\alpha > 0$ траектории первого СЗ $\lambda_1(R)$ имеют поведение, аналогичное описанному при $\alpha = 1$. Так, в область неустойчивости $\text{Im}(\lambda) > 0$ попадает лишь первое СЗ $\lambda_1(R)$ при изменении R в конечном интервале $R \in (R_l, R_r)$, где $R_l = R_l(\alpha)$ и $R_r = R_r(\alpha)$. Эти кривые $R_l(\alpha)$ и $R_r(\alpha)$ на плоскости (R, α) называют нижней и верхней ветвями ней-



Фиг. 11.



Фиг. 12.

тральной кривой (см. [4], [6]). На фиг. 12 рассчитаны эти ветви для чисел $R \leq 10^5$, они хорошо согласуются с известными результатами из [6]. Было перевычислено минимальное число Рейнольдса R_c , выше которого возникает неустойчивость хотя бы для одного значения α . На фиг. 12 этому соответствует точка C с координатами $R_c = 5772.2218$, $\alpha = 1.02055$, причем $\lambda_1(R_c) = 0.26400056$. Полученное значение R_c совпало с результатом $R_c = 5772.22$ из [18], [6].

7. ЗАКЛЮЧЕНИЕ

1. Разработан высокоточный метод расчета спектра и собственных функций задачи Орра–Зоммерфельда для общего течения Куэтта–Пуазейля при больших числах Рейнольдса. При этом спектр вычислялся при непрерывном изменении функции скорости основного потока $U(y)$ с помощью метода продолжения по параметру; началом такого изменения служил спектр для течения Куэтта либо Пуазейля.

2. Были подробно исследованы спектры для течения Куэтта и течения Пуазейля.

3. Проведенный численный анализ спектра задачи для течения Куэтта выявил сложное поведение СЗ вблизи узловой точки $\lambda_* = -i/\sqrt{3}$, а также существование точек ветвления и двойных СЗ. Это позволяет предположить, что верна

Гипотеза. Собственные значения $\lambda_n(R)$ задачи Орра–Зоммерфельда для течения Куэтта, рассматриваемые как функции числа Рейнольдса R при фиксированном волновом числе $\alpha > 0$, имеют счетное множество точек ветвления второго порядка $R_k > 0$, в которых двойные СЗ $\lambda_n(R_k)$ являются чисто мнимыми отрицательными и справедливо предельное соотношение

$$\lim_{k \rightarrow \infty} \lambda_n(R_k) \rightarrow -\frac{i}{\sqrt{3}}.$$

4. Для течения Пуазейля такие точки ветвления при вещественных значениях $\alpha > 0$ и $R > 0$ обнаружить не удалось. По-видимому, эти особые точки могут быть найдены при комплексных α и R .

Добавим здесь, что подобные точки ветвления и двойные СЗ были также обнаружены и вычислены в [40], [41] при исследовании волнового сфероидального уравнения, где были построены специальные методы их анализа на основе аппроксимаций Паде и их обобщений.

СПИСОК ЛИТЕРАТУРЫ

1. Линь Цзя-Цзяо. Теория гидродинамической устойчивости. М.: Изд-во иностр. лит., 1958.
2. Бетчов Р., Криминале В. Вопросы гидродинамической устойчивости. М.: Мир, 1971.
3. Drazin R.G., Reid W.H. Hydrodynamic stability. Cambridge: Cambridge Univ., 1981.

4. Жук В.И. Волны Толлмина–Шлихтинга и солитоны. М.: Наука, 2001.
5. Дразин Ф. Введение в теорию гидродинамической устойчивости. М.: Физматлит, 2005.
6. Бойко А.В., Грек Г.Р., Довгаль А.В., Козлов В.В. Физические механизмы перехода к турбулентности в открытых течениях. М.: Ин-т компьютер. исследований, 2006.
7. Schensted I.V. Contributions to the theory of hydrodynamic stability: Ph. D. thesis, 1960, Univ. of Michigan.
8. DiPrima R.C., Habetler G.J. A completeness theorem for non-selfadjoint eigenvalue problems in hydrodynamic stability // Arch. Ration. Mech. and Anal. 1969. V. 34. P. 218–227.
9. Reddy S.C., Schmid P.J., Henningson D.S. Pseudospectra of the Orr–Sommerfeld operator // SIAM J. Appl. Math. 1993. V. 53. № 1. P. 15–47.
10. Степин С.А. Несамосопряженные сингулярные возмущения и спектральные свойства задачи Орра–Зоммерфельда // Матем. сб. 1997. Т. 188. С. 129–146.
11. Dyachenko A.V., Shkalikov A.A. On the Orr–Sommerfeld equation with linear profile. arXiv: math.FA/0212127 v3 18 Dec. 2002.
12. Шкаликов А.А. Спектральные портреты оператора Орра–Зоммерфельда при больших числах Рейнольдса // Современная математика. Фундаментальное направление. 2003. Т. 3. С. 89–112.
13. Жук В.И., Проценко И.Г. Асимптотика решений уравнения Орра–Зоммерфельда в окрестностях двух ветвей нейтральной кривой // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 11. С. 1737–1753.
14. Жук В.И., Проценко И.Г. Асимптотическая структура волновых возмущений в теории устойчивости плоского течения Куэтта–Пуазейля // Ж. вычисл. матем. и матем. физ. 2005. Т. 45. № 6. С. 1068–1080.
15. Жук В.И., Проценко И.Г. Асимптотическая модель эволюции возмущений в плоском течении Куэтта–Пуазейля // Докл. РАН. 2006. Т. 411. № 1. С. 20–24.
16. Thomas H.H. The stability of plane Poiseuille flow // Phys. Rev. 1953. V. 91. № 4. P. 780–783.
17. Жарилкасинов А., Лисейкин В.Д., Скобелев Б.Ю., Яненко Н.Н. Применение неравномерной сетки для численного решения задачи Орра–Зоммерфельда // Числ. методы механ. сплошной среды. Новосибирск, 1983. Т. 14. № 5. С. 45–54.
18. Orszag S.A. Accurate solution of the Orr–Sommerfeld equation // J. Fluid Mech. 1971. V. 50. Part 4. P. 689–703.
19. Orszag S.A. Numerical simulation of incompressible flows within simple boundaries. I. Galerkin (Spectral) representations // Study Appl. Math. 1971. V. 50. № 4. P. 293–327.
20. Chandrasekhar S. Hydrodynamic and hydromagnetic stability. Oxford: Oxford Univ. Press, 1961.
21. Gallagher A.P., Mercer A.McD. On the behaviour of small disturbances in plane Couette flow // J. Fluid Mech. 1962. V. 13. P. 91–100.
22. Нейман-заде М.И., Шкаликов А.А. О вычислении собственных значений задачи Орра–Зоммерфельда // Фундаментальная и прикл. матем. 2002. Т. 8. Вып. 1. С. 301–305.
23. Слепцов А.Г. Проекционно-сеточные методы решения задачи Орра–Зоммерфельда // Числ. методы механ. сплошной среды. Новосибирск: ИТПМ СО АН СССР, 1983. Т. 14. № 5. С. 111–126.
24. Нармурадов Ч.Б. Об одном эффективном методе решения уравнения Орра–Зоммерфельда // Матем. моделирование. 2005. Т. 17. № 9. С. 35–42.
25. Абрамов А.А., Альварес Л.М. Метод решения жестких краевых задач, основанный на расщеплении оператора // Ж. вычисл. матем. и матем. физ. 1989. Т. 29. № 12. С. 1800–1810.
26. Альварес Л.М., Диткин В.В. О численном решении уравнения Орра–Зоммерфельда // Ж. вычисл. матем. и матем. физ. 1990. Т. 30. № 4. С. 611–615.
27. Абрамов А.А., Ульянова В.И. Об одном методе решения уравнения типа бигармонического с сингулярно входящим малым параметром // Ж. вычисл. матем. и матем. физ. 1992. Т. 32. № 4. С. 567–575.
28. Ульянова В.И. Об использовании метода расщепления оператора в задачах на собственные значения для уравнения типа Орра–Зоммерфельда // Ж. вычисл. матем. и матем. физ. 1993. Т. 33. № 4. С. 635–639.
29. Курочкин С.В. Метод выявления неустойчивости и поиска неустойчивых собственных значений в задаче Орра–Зоммерфельда // Ж. вычисл. матем. и матем. физ. 2001. Т. 41. № 1. С. 86–94.
30. Дубровский В.В., Кадченко С.И., Кравченко В.Ф., Садовничий В.А. Новый метод приближенного вычисления первых собственных чисел спектральной задачи Орра–Зоммерфельда // Докл. РАН. 2001. Т. 378. № 4. С. 443–446.
31. Trefethen L.N. Pseudospectra of linear operators // SIAM. Rev. 1997. V. 39. № 3. P. 383–406.
32. Голубев В.В. Лекции по аналитической теории дифференциальных уравнений. М.–Л.: Гостехтеориздат, 1950.
33. Гельфонд А.О. Исчисление конечных разностей. М.: Наука, 1967.

34. *Birkhoff G.D.* Formal theory of irregular difference equations // *Acta Math.* 1930. V. 54. P. 205–246.
35. *Birkhoff G.D., Trjitzinsky W.* Analytic theory of singular difference equations // *Acta Math.* 1932. V. 60. P. 1–89.
36. *Wimp J., Zeilberger D.* Resurrecting the asymptotics of linear recurrences // *J. Math. Analys. and Appl.* 1985. V. 111. P. 162–176.
37. *Смирнов В.И.* Курс высшей математики. Т. 1. М.: Гостехтеориздат, 1954.
38. *Романов В.А.* Устойчивость плоскопараллельного течения Куэтта // Докл. АН СССР. 1971. Т. 196. № 5. С. 1049–1051.
39. *Романов В.А.* Устойчивость плоскопараллельного течения Куэтта // Функц. анализ и его прилож. 1973. Т. 7. Вып. 2. С. 62–73.
40. *Скороходов С.Л., Христофоров Д.В.* Вычисление точек ветвления собственных значений, соответствующих волновым сфероидальным функциям // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 7. С. 1195–1210.
41. *Скороходов С.Л., Христофоров Д.В.* Точки ветвления собственных значений, соответствующих сфероидальным волновым функциям // Междунар. конф. “Тихонов и современная матем.” Москва, 19–25 июня 2006 г. Тезисы докл. секц. “Функц. анализ и дифференц. ур-ния”. С. 259–260.

УДК 519.624.2

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ ПЕРЕХОДА К ХАОСУ ДИССИПАТИВНОГО ОСЦИЛЛЯТОРА ДУФФИНГА С ДВУХЧАСТОТНЫМ ВОЗБУЖДЕНИЕМ

© 2007 г. Т. В. Завражина

(03680 Киев, пр-т. акад. Глушкова, 40,

Международ. научно-учебный центр информ.

технологий и систем при НАНУ и МОНУ, Украина)

e-mail: dep@irtc.org.ua

Поступила в редакцию 12.07.2005 г.

Переработанный вариант 10.05.2007 г.

Предлагается методика математического моделирования хаотизации колебаний существенно нелинейного диссипативного осциллятора Дуффинга с двухчастотным возбуждением на инвариантном торе в $\mathbb{R}^2$. Она основана на совместном применении метода продолжения решения по параметру, критериев устойчивости Флоке, теории ветвления, метода высокоточного численного интегрирования Эверхарта. Данный подход использован для численного построения субгармонических решений при переходе рассматриваемого осциллятора к хаосу через последовательность бифуркаций кратного увеличения периода. Подтверждено значение одной из универсальных постоянных, полученных ранее автором при исследовании хаотизации колебаний диссипативных осцилляторов с одночастотным периодическим возбуждением. Библ. 21. Фиг. 4. Табл. 1.

Ключевые слова: динамическая система, осциллятор Дуффинга, периодическое решение на торе, бифуркация, хаос, численный метод Эверхарта, теория Флоке, универсальная постоянная Фейгенбаума.

1. ВВЕДЕНИЕ

Вопросам исследования возможных сценариев перехода от регулярных к хаотическим режимам движения детерминированных динамических систем на фазовых плоскостях с одночастотным периодическим возбуждением посвящена обширная библиография (см., например, [1]–[8]). Качественное исследование отдельных областей фазового пространства систем, в которых имеют место устойчивые и неустойчивые режимы движения, и путей развития неустойчивостей, связанных с возможностью возникновения периодических, субгармонических, квазипериодических и хаотических движений, проведено в [1], [2], [5]–[8] с применением различных приближенных аналитических и численных методов, а также их сочетаний. Следует отметить растущий интерес к исследованиям нелинейных динамических систем, находящихся под действием множественных возмущающих нагрузок различной частоты (см. [9]–[12]). Разнообразные подходы к исследованию таких многочастотных нелинейных систем, которые возникают в различных задачах небесной механики, физики, радиофизики и др., обсуждаются в [9]–[14].

В связи с тем, что с увеличением размерности вектора частот возмущающих нагрузок динамических систем происходит усложнение структуры резонансных поверхностей, наиболее полно изучены одночастотные случаи. Поэтому в настоящей работе предложена методика математического моделирования переходов к хаосу нелинейных осцилляторов с двухчастотным возбуждением через последовательности бифуркаций кратного увеличения периода. Для примера нелинейного диссипативного осциллятора Дуффинга на базе разработанной методики построена универсальная по Фейгенбауму последовательность первых восьми бифуркаций удвоения периода, которая приводит к возникновению хаотических колебаний осциллятора с появлением странного аттрактора.

2. МЕТОДИКА ИССЛЕДОВАНИЯ ПЕРЕХОДОВ К ХАОСУ ДИНАМИЧЕСКИХ СИСТЕМ С ДВУХЧАСТОТНЫМ ВОЗБУЖДЕНИЕМ

Будем исследовать эволюцию установившихся периодических колебаний осциллятора

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, \lambda, \omega_1 t, \omega_2 t), \quad (1)$$

где $\mathbf{x} = (x_1(t), \dots, x_N(t))$ есть N -мерный вектор фазовых координат, λ – управляющий параметр, представляющий собой интенсивность внешнего возмущения с постоянными частотами ω_1, ω_2 ;

$$\mathbf{f}(\mathbf{x}, \lambda, \omega_1 t, \omega_2 t) = (f_1(\mathbf{x}, \lambda, \omega_1 t, \omega_2 t), \dots, f_N(\mathbf{x}, \lambda, \omega_1 t, \omega_2 t))$$

есть нелинейная вектор-функция размерности N , непрерывно дифференцируемая по $\mathbf{x}$ и λ необходимое число раз.

Введем две дополнительные переменные $\varphi_1 = \omega_1 t, \varphi_2 = \omega_2 t$. С учетом этого векторное уравнение (1) заменим системой

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \lambda, \boldsymbol{\varphi}), \quad \frac{d\boldsymbol{\varphi}}{dt} = \boldsymbol{\omega}, \quad (2)$$

где $\boldsymbol{\varphi} = (\varphi_1(t), \varphi_2(t)), \boldsymbol{\omega} = (\omega_1, \omega_2)$.

Для построения резонансных периодических решений системы (2) на инвариантном торе в $\mathbb{R}^2$, удовлетворяющих резонансному соотношению

$$k_1 \omega_1 + k_2 \omega_2 = 0, \quad (3)$$

перейдем в системе (2) к собственному времени $\tau = \Omega t$, при котором будем строить искомые решения периода $T = 2\pi$. Составим (см. [15]) следующие уравнения относительно T :

$$\omega_1 T = 2\pi k_2, \quad \omega_2 T = -2\pi k_1. \quad (4)$$

Тогда на основе (4) частота Ω может быть определена из резонансного соотношения (3) одним из двух эквивалентных равенств

$$\Omega = \frac{\omega_1}{k_2} = -\frac{\omega_2}{k_1}. \quad (5)$$

Здесь k_1, k_2 – целые числа. С учетом указанной замены переменных система (2) примет вид

$$\Omega \frac{d\mathbf{x}}{d\tau} = \mathbf{f}(\mathbf{x}, \lambda, \boldsymbol{\varphi}), \quad \Omega \frac{d\boldsymbol{\varphi}}{d\tau} = \boldsymbol{\omega}. \quad (6)$$

Проведем исследование эволюции периодических решений системы (6) при изменении управляющего параметра λ .

3. МЕТОД ПОСТРОЕНИЯ СУБГАРМОНИЧЕСКИХ ДВИЖЕНИЙ НА ТОРЕ В $\mathbb{R}_2$

Запишем систему (6) в общем виде:

$$d\mathbf{y}/d\tau = \mathbf{g}(\mathbf{y}(\tau), \lambda), \quad (7)$$

где $\mathbf{g}(\mathbf{y}(\tau), \lambda) = (g_1(\mathbf{y}(\tau), \lambda), \dots, g_P(\mathbf{y}(\tau), \lambda))$ – нелинейная вектор-функция размерности $P = N + 2$, непрерывно дифференцируемая по $\mathbf{y}$ и λ необходимое число раз и периодическая по τ с периодом $T = 2\pi$:

$$\mathbf{g}(\mathbf{y}(\tau), \lambda) = \mathbf{g}(\mathbf{y}(\tau + T), \lambda).$$

Здесь $\mathbf{y}(\tau) = (y_1(\tau), \dots, y_P(\tau))$ есть P -мерный вектор.

Исследуем эволюцию установившихся субгармонических движений периода $kT, k \geq 1$, системы (7) при изменении управляющего параметра λ . С этой целью построим непрерывные субгармонические решения $\mathbf{y}(\tau)$ системы (7), удовлетворяющие условиям периодичности вида

$$\mathbf{y}(0) = \mathbf{y}(kT). \quad (8)$$

Методика решения двухточечной краевой задачи (7), (8) базируется на пошаговом алгоритме замены исходной системы (7) системой линейных дифференциальных уравнений с периодиче-

скими коэффициентами, полученной путем линеаризации исходных уравнений движения в окрестности некоторого известного решения.

Построим kT -периодическое решение $\mathbf{y}_s(\tau)$ системы (7), соответствующее заданному значению параметра $\lambda = \lambda_s$. Для этого рассмотрим m -й промежуточный шаг численной реализации метода продолжения решения по параметру. Пусть при значении параметра $\lambda = \lambda_m$ известно в общем случае приближенное субгармоническое решение $\mathbf{y}_m(\tau)$ исходной системы нелинейных дифференциальных уравнений. В предположении непрерывной зависимости $\mathbf{y}(\tau)$ системы (7) от начальных условий $\mathbf{y}(0)$ и параметра λ представим его в окрестности состояния $\mathbf{y}_m(\tau)$, λ_m в виде $\mathbf{y}(\tau) = \mathbf{y}(\mathbf{y}(0), \lambda)$. Тогда условие периодичности рассматриваемого решения запишется в форме

$$\mathbf{y}(0) = \mathbf{y}(\mathbf{y}(0), \lambda, kT). \quad (9)$$

При близком λ_m значении параметра $\lambda_{m+1} = \lambda_m + \delta\lambda_m$ субгармоническое решение $\mathbf{y}_{m+1}(\tau)$ системы (7), удовлетворяющее условиям вида (9), представим следующим образом:

$$\mathbf{y}_{m+1}(\tau) = \mathbf{y}_m(\tau) + \delta\mathbf{y}_m(\tau), \quad (10)$$

где $\delta\mathbf{y}_m(\tau)$ – приращение искомой вектор-функции. Соответствующие решению $\mathbf{y}_{m+1}(\tau)$ начальные условия имеют вид

$$\mathbf{y}_{m+1}(0) = \mathbf{y}_m(0) + \delta\mathbf{y}_m(0), \quad (11)$$

где $\delta\mathbf{y}_m(0)$ – вектор вариаций начальных условий. Подставим искомое решение (10) при $\lambda = \lambda_{m+1}$ в уравнение (7) и получим

$$\frac{d\mathbf{y}_m}{d\tau} + \delta\left(\frac{d\mathbf{y}_m}{d\tau}\right) = \mathbf{g}(\mathbf{y}_m(\tau) + \delta\mathbf{y}_m(\tau), \lambda_m + \delta\lambda_m). \quad (12)$$

Правую часть этого уравнения разложим в ряд Тейлора в окрестности известного состояния $\mathbf{y}(\tau) = \mathbf{y}_m(\tau)$, $\lambda = \lambda_m$. Ограничиваясь членами первого порядка малости по $\delta\mathbf{y}_m(\tau)$, $\delta\lambda_m$, заменим (7) линейной системой вида

$$\frac{d\mathbf{y}_m}{d\tau} + \delta\left(\frac{d\mathbf{y}_m}{d\tau}\right) = \mathbf{g}(\mathbf{y}_m(\tau), \lambda_m) + g_Y(\mathbf{y}_m(\tau), \lambda_m)\delta\mathbf{y}_m(\tau) + \mathbf{g}_\lambda(\mathbf{y}_m(\tau), \lambda_m)\delta\lambda_m,$$

где g_Y и $\mathbf{g}_\lambda$ – соответственно, матрица-функция и вектор-функция производных функции $\mathbf{g}(\mathbf{y}_m, \lambda_m)$ по $\mathbf{y}$ и λ :

$$g_Y = \partial\mathbf{g}/\partial\mathbf{y}, \quad \mathbf{g}_\lambda = \partial\mathbf{g}/\partial\lambda.$$

Тогда неизвестная вектор-функция $\delta\mathbf{y}_m(\tau)$ должна удовлетворять уравнению

$$\delta\left(\frac{d\mathbf{y}_m}{d\tau}\right) = g_Y(\mathbf{y}_m(\tau), \lambda_m)\delta\mathbf{y}_m + \mathbf{g}_\lambda(\mathbf{y}_m(\tau), \lambda_m)\delta\lambda_m \quad (13)$$

и задача построения приближенного субгармонического решения (10) нелинейной системы (7) сводится к отысканию удовлетворяющего условиям периодичности

$$\delta\mathbf{y}_m(0) = \delta\mathbf{y}_m(kT) \quad (14)$$

решения линейного относительно $\delta\mathbf{y}(\tau)$ и $\delta\lambda$ неоднородного уравнения в вариациях (13) с периодическими коэффициентами g_Y и $\mathbf{g}_\lambda$. Согласно (10), (11), вектор вариаций фазовых координат $\delta\mathbf{y}_m(\tau)$ удовлетворяет соотношению

$$\delta\mathbf{y}_m(0) = \mathbf{y}_{m+1}(\mathbf{y}_m(0) + \delta\mathbf{y}_m(0), \lambda_m + \delta\lambda_m) - \mathbf{y}_m(\mathbf{y}_m(0), \lambda_m). \quad (15)$$

Разложим правую часть отношения (15) в ряд Тейлора в окрестности состояния $\mathbf{y}(0) = \mathbf{y}_m(0)$, $\lambda = \lambda_m$:

$$\delta\mathbf{y}_m(\tau) = \frac{\partial\mathbf{y}(\mathbf{y}_m(0), \lambda_m)}{\partial\mathbf{y}(0)}\delta\mathbf{y}_m(0) + \frac{\partial\mathbf{y}(\mathbf{y}_m(0), \lambda_m)}{\partial\lambda}\delta\lambda_m + \mathbf{r}(\mathbf{y}_m(0), \lambda_m). \quad (16)$$

Здесь $\mathbf{r}(\mathbf{y}_m(0), \lambda_m)$ есть P -мерная функция невязки, имеющая порядок отброшенных членов разложения (16).

При заданном $\delta\lambda_m$ вектор вариаций начальных условий $\delta\mathbf{y}_m(0)$ может быть определен из системы линейных алгебраических уравнений

$$[Z_{m+1}(kT) - E]\delta\mathbf{y}_m(0) = -\mathbf{z}_{\lambda_{m+1}}(kT)\delta\lambda_m - \mathbf{R}_m^{k_m}(kT), \quad (17)$$

полученной из соотношения (16) с учетом условий периодичности (8), (14). Невязка $\mathbf{R}_m^{k_m}(kT)$ предыдущего шага учитывается в (17) для увеличения точности при определении $\delta\mathbf{y}_m(0)$. На $(m+1)$ -м шаге варьирования параметра λ матрица монодромии $Z_{m+1}(kT) = \partial\mathbf{y}(kT)/\partial\mathbf{y}(0)$ однородного уравнения в вариациях определяется из нормированной фундаментальной матрицы решений $Z_{m+1}(\tau)$ однородного уравнения в вариациях

$$\frac{dZ_{m+1}}{d\tau} = g_Y Z_{m+1}, \quad (18)$$

где $Z_{m+1}(0) = E$, E – единичная матрица; вектор

$$\mathbf{z}_{\lambda_{m+1}}(kT) = \frac{\partial\mathbf{y}(kT)}{\partial\lambda}$$

определяется из частного решения $\mathbf{z}_{\lambda_{m+1}}(\tau)$ неоднородной системы уравнений в вариациях (13):

$$\frac{d\mathbf{z}_{\lambda_{m+1}}}{d\tau} = g_Y \mathbf{z}_{\lambda_{m+1}} + \mathbf{g}_\lambda, \quad \mathbf{z}_{\lambda_{m+1}}(0) = 0. \quad (19)$$

Дальнейшее построение решения (10) системы (7), удовлетворяющего условиям периодичности вида (8), основывалось на подходе, который при $\lambda_{m+1} = \lambda_m + \delta\lambda_m$ заключается в определении на основании (17) соответствующего ему вектора начальных условий (11), удовлетворяющего условиям периодичности, и решении еще одной, $(P+2)$ -й задачи Коши для исходной нелинейной системы (7) с полученными начальными условиями. При этом построение решений $Z_{m+1}(\tau)$, $\mathbf{z}_{\lambda_{m+1}}(\tau)$ для $(P+1)$ -й задачи Коши (17), (19) и искомого решения дополнительной $(P+2)$ -й задачи Коши выполнялось с использованием высокоточного метода численного интегрирования Эверхарта 11-го порядка (см. [2]). Все вычисления проводились с удвоенной точностью.

Построенное решение $\mathbf{y}_{m+1}(\tau)$ задачи Коши для системы (7) с начальными условиями $\mathbf{y}_{m+1}(0)$ при $\lambda = \lambda_{m+1}$ удовлетворяет условиям kT -периодичности лишь приближенно. Уточнение начальных условий при $\lambda = \lambda_{m+1}$ может быть проведено при помощи представления

$$\mathbf{y}_{m+1}(0) = \mathbf{y}_m(0) + \delta\mathbf{y}_m(0) + \sum_{i=1}^{k_m+1} \delta\mathbf{y}_{R_{m+1}^i}(0) \quad (20)$$

и модифицированной схемы Ньютона

$$\delta\mathbf{y}_{R_{m+1}^i}(0) = -[Z_{m+1}(kT) - E]^{-1} \mathbf{R}_{m+1}^i, \quad i = 1, 2, \dots, k_{m+1}. \quad (21)$$

Здесь k_{m+1} – количество уточняющих итераций на $(m+1)$ -м шаге; $\mathbf{R}_{m+1}^i = \mathbf{y}_{m+1}^i(kT) - \mathbf{y}_{m+1}^i(0)$ – невязка, имеющая порядок отброшенных членов в (14).

Продолжая варьирование параметра λ , можно построить решение $\mathbf{y}_s(\tau)$ системы (7), соответствующее заданному значению параметра $\lambda = \lambda_s$, в том случае, когда на каждом шаге продолжения матрица периодичности $[Z_{m+1}(kT) - E]$ остается невырожденной. В противном случае точка $(\mathbf{y}_b, \lambda_b)$ пространства состояний и параметра, для которой $\det[Z_{m+1}(kT) - E] = 0$, является особой. В ней происходит нарушение условий единственности решений уравнения (7) и возможно разветвление решений. При этом для выполнения следующего шага продолжения решения по параметру необходимо привлечение методов теории ветвления (см. [2]).

Вопрос об исследовании устойчивости невозмущенного движения $\mathbf{y}_{m+1}(\tau)$ динамической системы сводится к изучению устойчивости тривиальных решений уравнений возмущенного движения первого приближения

$$\delta\left(\frac{d\mathbf{y}_{m+1}(\tau)}{d\tau}\right) = g_Y \delta\mathbf{y}_{m+1}(\tau) \quad (22)$$

с периодическими коэффициентами. Этот вопрос решается на основании теорем Флоке (см. [16], [17]) путем анализа мультипликаторов $\rho_j, j = 1, 2, \dots, P$, системы (22), которые представляют собой собственные значения матрицы монодромии $Z_{m+1}(kT)$. Случаи, когда существует хотя бы один мультипликатор ρ_j , удовлетворяющий соотношениям $|\rho_j| = 1, |\rho_j| < 1, i \neq j$, являются критическими. В пространстве состояний им соответствуют точки бифуркаций (y_b, λ_b) , в которых возможно разветвление решений. Характер пересечения мультипликатором ρ_j единичной окружности определяет тип реализующейся бифуркационной точки и свойства ответвляющихся решений. В дальнейшем будем придерживаться следующей классификации точек бифуркации в случае $|\rho_j| = 1$ в зависимости от значений отношения

$$\frac{\arg \rho_j}{2\pi} = \frac{u}{v} < 1$$

(где $v \neq 0, u, v$ – целые числа), соответствующих ответвлению субгармонического решения с периодом vkT ($v \geq 1$), кратным периоду kT исходного решения. Состояние системы (22), в котором $\arg \rho_j = 0$, соответствует в пространстве состояний точке бифуркации рождения (исчезновения) пары субгармонических решений исходного периода kT . Если мультипликатор ρ_j действителен и удовлетворяет соотношению $\arg \rho_j = \pi$, то соответствующую точку пространства состояний будем называть точкой бифуркации удвоения периода. Ответвляющееся решение имеет период $2kT$. Случаи $\arg \rho_j = 2\pi/3$ и $\arg \rho_j = \pi/2$ отвечают бифуркациям утроения и учетверения периода исходного решения. Построение ответвляющихся субгармонических решений может быть проведено только методами теории ветвления (см. [2], [18], [19]).

Отметим, что для системы (17), в соответствии с теоремой Андронова–Витта (см. [18]), матрица периодичности вырождена в каждой точке кривой периодического решения. Для отыскания составляющих вектора $\delta y_m(0)$ из (17) пользуются тем обстоятельством, что на основании теоремы Пуанкаре (см. [20]) для случая одночастотного возбуждения одну из компонент этого вектора можно задавать произвольно. При заданных значениях k_1 и k_2 получаемая указанным образом периодическая траектория лежит на некотором торе в $\mathbb{R}^2$ (см. [15]). Поэтому степень вырожденности матрицы периодичности решения в каждой точке кривой периодического решения повышается на единицу и задавать произвольно можно уже две компоненты вектора $\delta y_m(0)$.

4. ЧИСЛЕННАЯ РЕАЛИЗАЦИЯ ЗАДАЧИ

Будем исследовать установившиеся периодические колебания осциллятора Дуффинга

$$\ddot{x} + \mu \dot{x} + cx + dx^3 = \lambda(\sin \omega_1 t + \sin \omega_2 t) \quad (23)$$

при изменении управляющего параметра λ . Здесь принято $\mu = 0.02, c = -1, d = 1, \omega_1 = 1.8, \omega_2 = 1.2$. Точкой обозначено дифференцирование по времени t . Согласно (2), можно представить неавтономное уравнение (23) в автономной форме

$$\begin{aligned} \ddot{x} + \mu \dot{x} + cx + dx^3 &= \lambda(\sin \varphi_1 + \sin \varphi_2), \\ \dot{\varphi}_1 &= \omega_1, \quad \dot{\varphi}_2 = \omega_2. \end{aligned} \quad (24)$$

С учетом параметров уравнения (23) легко найти значения коэффициентов в резонансных соотношениях (3) соответственно:

$$k_1 = -2, \quad k_2 = 3. \quad (25)$$

Тогда на основании (5) и (25) получим значение частоты $\Omega = 0.6$. В результате перехода к собственному времени τ система (24) примет вид

$$\begin{aligned} \Omega^2 x'' + \mu \Omega x' + cx + dx^3 &= \lambda(\sin \varphi_1 + \sin \varphi_2), \\ \Omega \varphi_1' &= \omega_1, \quad \Omega \varphi_2' = \omega_2, \end{aligned} \quad (26)$$

где штрихом обозначено дифференцирование по собственному времени τ .

Приведем систему (26) к виду (7), для чего воспользуемся представлением $y_1 = x, y_2 = y'_1 = x', \Phi_1 = y_3, \Phi_2 = y_4$:

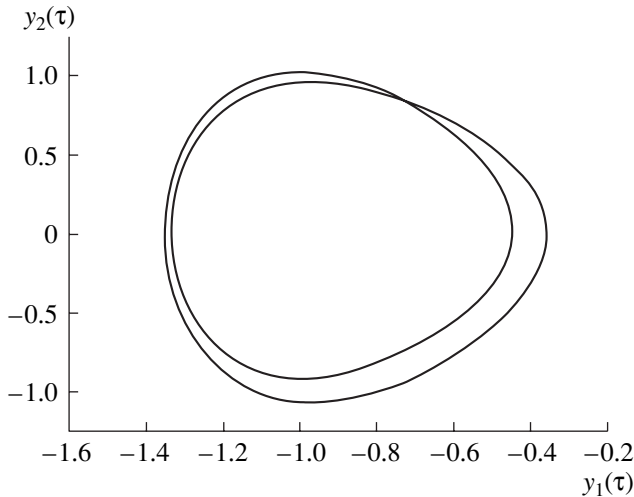
$$\begin{aligned} y'_1 &= y_2, \\ y'_2 &= \frac{1}{\Omega^2}(-\mu\Omega y_2 - cy_1 - dy_1^3 + \lambda(\sin y_3 + \sin y_4)), \\ y'_3 &= \frac{\omega_1}{\Omega}, \quad y'_4 = \frac{\omega_2}{\Omega}. \end{aligned} \quad (27)$$

Для построения решений периода $T = 2\pi$ эволюционной задачи (27) на первом шаге продолжения решения по параметру в качестве порождающего решения $\mathbf{y}_{(0)}(\tau) = (y_{1(0)}(\tau), y_{2(0)}(\tau), y_{3(0)}(\tau), y_{4(0)}(\tau))$ при $\lambda = \lambda_{(0)}$ рассматривалось следующее состояние равновесия системы (27): $\lambda_{(0)} = 0, y_{1(0)}(\tau) = -1, y_{2(0)}(\tau) = y_{3(0)}(\tau) = y_{4(0)}(\tau) = 0$. Для исходного значения параметра $\lambda = \lambda_{(0)}$ оно представляет собой искомое T -периодическое решение.

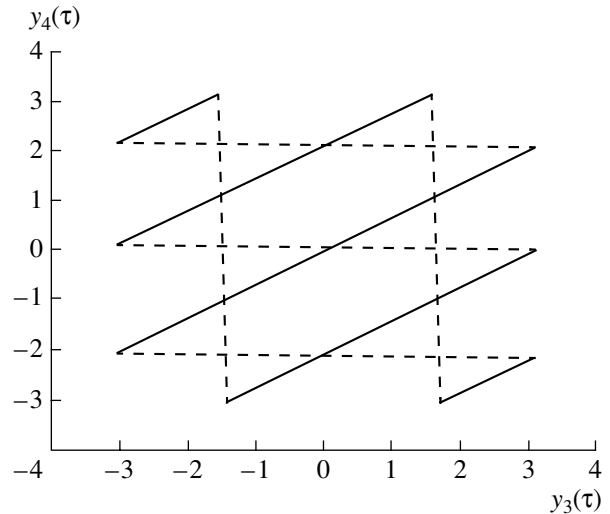
Анализ мультипликаторов $\rho_j, j = 1, 2, 3, 4$, матрицы монодромии проводился на каждом шаге построения ветви решений периода T . Обнаружены два устойчивых участка этой ветви, соответствующие диапазонам изменения параметра $\lambda \in [0, \lambda^1]$ и $\lambda \in (\lambda^2, \lambda_{(0)})$. При $\lambda^1 = 0.09032706045$ и $\lambda^2 = 0.01080031321$ имеют место прямая и обратная бифуркации рождения и исчезновения пары решений исходного периода соответственно, для которых один из мультипликаторов ρ_j становится действительным и выполняются соотношения $|\rho_j| = 1, \arg \rho_j = 0$. В соответствующих точках $(\mathbf{y}^1, \lambda^1), (\mathbf{y}^2, \lambda^2)$ пространства состояний происходит вырождение матрицы периодичности $[Z(T) - E]$, стоящей в левой части системы уравнений (17), и становятся возможными ответвления неустойчивого, а затем устойчивого периодических решений периода T . Регуляризация решений в указанных точках была выполнена путем смены ведущего параметра. При этом в качестве ведущего параметра выбиралась одна из компонент $\delta y_1(0)$ или $\delta y_2(0)$ вектора вариации начальных условий $\delta \mathbf{y}(0) = (\delta y_1(0), \delta y_2(0), \delta y_3(0), \delta y_4(0))$. В результате бифуркации $(\mathbf{y}^1, \lambda^1)$ ветвь T теряет устойчивость, а после бифуркации $(\mathbf{y}^2, \lambda^2)$ вновь ее приобретает и остается устойчивой вплоть до точки бифуркации $(\mathbf{y}^{(0)}, \lambda^{(0)})$ удвоения периода, после которой становится неустойчивой. В бифуркационном состоянии системы, реализующемся при $\lambda = \lambda_{(0)}$, один из мультипликаторов ρ_j действителен и удовлетворяет соотношениям $|\rho_j| = 1, \arg \rho_j = \pi$. Значения параметра $\lambda^{(0)}$ и начальных условий $\mathbf{y}^{(0)}(0)$ в данной точке бифуркации приведены в таблице. Отметим, что $y_4^{(0)}(0) = 0$. На фиг. 1, 2 представлены фазовые портреты системы (27), построенные в указанной точке бифуркации. Легко видеть, что фазовая траектория замыкается в фазовой плоскости $(y_1(\tau), y_2(\tau))$, образуя двойную яйцевидную орбиту (см. фиг. 1), изображение $\tilde{y}_3(\tau) = y_3(\tau) - 2\pi \left[\frac{y_3(\tau)}{2\pi} \right], \tilde{y}_4(\tau) = y_4(\tau) - 2\pi \left[\frac{y_4(\tau)}{2\pi} \right]$ фазовой траектории замыкается после k_1 оборотов по оси $y_3(\tau)$ и k_2 оборотов

Таблица

n	$\lambda^{(n)}$	$y_1^{(n)}(0)$	$y_2^{(n)}(0)$	$y_3^{(n)}(0)$
0	0.0640669998199	-0.8070613821677	-1.0152781354637	0
1	0.0717780885722	-0.5668823068001	-0.8301051607929	-2.243430×10^{-5}
2	0.0727549016335	-0.5307588507032	-0.7805242894483	-2.307746×10^{-5}
3	0.0728838267248	-0.5268845776521	-0.7773235481717	-2.362579×10^{-5}
4	0.0729050272294	-0.5268935961944	-0.7778628866289	-2.356230×10^{-5}
5	0.0729093497767	-0.5269400844770	-0.7780115272255	-2.355764×10^{-5}
6	0.0729102763652	-0.5269499740912	-0.7780488910936	-2.355154×10^{-5}
7	0.0729104749216	-0.5269518367785	-0.7780538643222	-2.355597×10^{-5}



Фиг. 1.



Фиг. 2.

по оси $y_4(\tau)$ на квадратах $-\pi < \tilde{y}_3 < \pi$, $-\pi < \tilde{y}_4 < \pi$ (см. фиг. 2) фазовой плоскости $(y_3(\tau), y_4(\tau))$. Таким образом, согласно теореме из [15], замкнутой фазовой траектории на рассмотренных плоскостях соответствует замкнутая кривая на торе в $\mathbb{R}^2$.

Ответвление в бифуркационной точке $(y^{(0)}(0), \lambda^{(0)})$ субгармонического решения задачи (27) удвоенного периода $2T$ происходит с одновременной потерей устойчивости ветви периода T . Далее исходная ветвь неустойчива. Построение ответвляющегося субгармонического решения периода $2T$ в малой окрестности точки бифуркации проведено на базе уравнений разветвления второго порядка (см. [2]) с учетом $V = 2$, $k = 1$. За пределами бифуркационной окрестности ветвь указанного периода оказалась устойчивой вплоть до следующей точки бифуркации удвоения периода $(y^{(1)}(0), \lambda^{(1)})$ (см. таблицу).

Дальнейшее исследование эволюции субгармонических решений при изменении управляющего параметра и анализ возникающих при этом точек бифуркации позволил построить для задачи (27) последовательность первых восьми бифуркаций удвоения периода. Значения параметра λ^n и начальных условий $y^{(n)}(0)$ в соответствующих точках бифуркаций ($n = 0, \dots, 7$) приведены в таблице. Отметим, что для всех найденных точек бифуркации выполняется $y_4^{(n)}(0) = 0$.

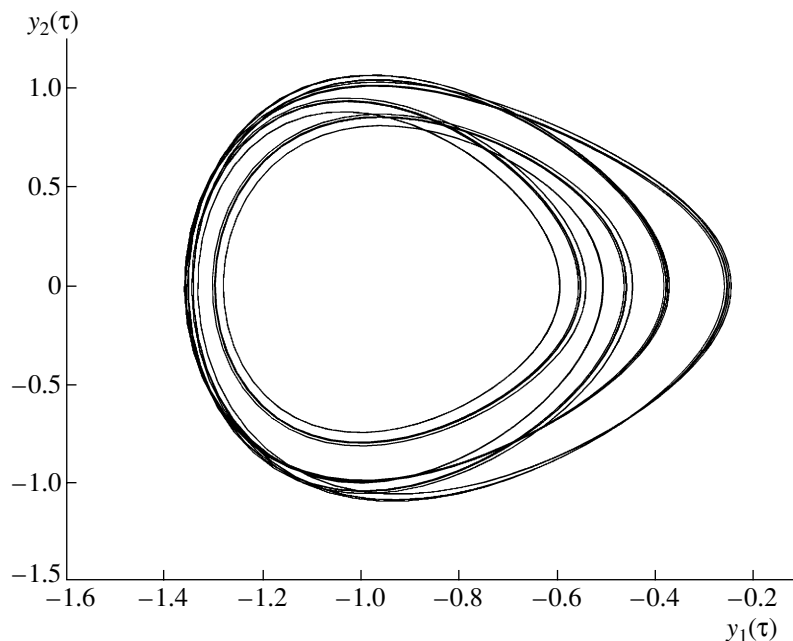
В ходе последовательных бифуркаций удвоения периода происходит качественное преобразование фазовых портретов только в фазовой плоскости $(y_1(\tau), y_2(\tau))$. На фиг. 3 в фазовой плоскости $(y_1(\tau), y_2(\tau))$ представлен фазовый портрет $128T$ -периодической траектории. Очевидно, что она существенно усложняется по сравнению с T -периодической орбитой, изображенной на фиг. 1. В фазовой плоскости $(y_3(\tau), y_4(\tau))$ преобразования фазового портрета в результате последовательных бифуркаций не происходит и он полностью совпадает с соответствующим изображением для T -периодической траектории, представленным на фиг. 2.

Полученная бифуркационная последовательность обладает универсальными свойствами по Фейгенбауму, поскольку процесс изменения масштаба управляющего параметра λ при последовательных бифуркациях для $n \rightarrow \infty$ подчиняется соотношению

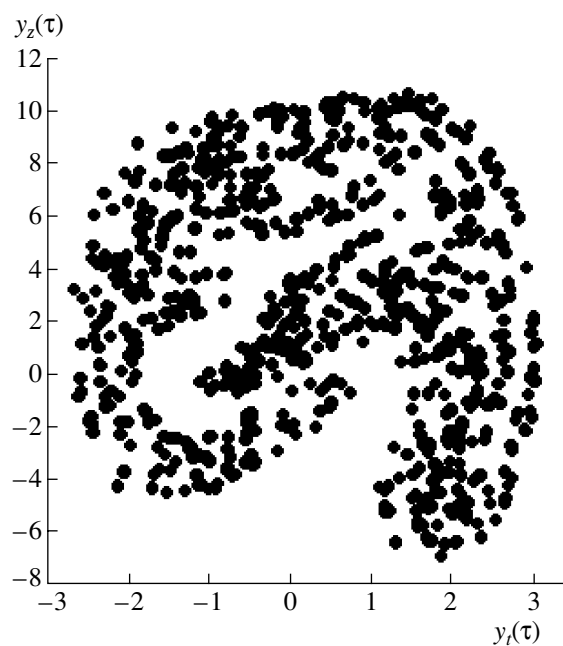
$$\delta^{(n)} = \frac{\lambda^{(n+1)} - \lambda^{(n)}}{\lambda^{(n+2)} - \lambda^{(n+1)}} \rightarrow \delta,$$

где $\delta = 4.666623832$. Для сравнения приведем константу $\delta = 4.669201609\dots$, полученную Фейгенбаумом путем моделирования квадратичного точечного отображения. Найденная нами в результате численных расчетов константа δ позволяет по закону геометрической прогрессии

$$\lambda_{\infty} = \lambda^{(n)} + \frac{(\lambda^{(n+1)} - \lambda^{(n)})\delta}{(\delta - 1)}$$



Фиг. 3.



Фиг. 4.

предсказать предельное значение параметра λ_∞ на пороге хаоса, к которому со скоростью δ сходятся бифуркационные значения параметра $\lambda^{(n)}$. В нашем случае предельное значение указанного параметра составляет $\lambda_\infty = 0.07291052907400$.

Как следует из полученных нами результатов, в пределе $\lambda = \lambda_\infty$ бесконечной последовательности бифуркаций удвоения периода в некоторой области фазового пространства существует бесконечное множество неустойчивых движений периода $2^n T$, $n = 0, 1, \dots, 7$, порождающее при $\lambda > \lambda_\infty$ так называемый хаотический аттрактор Фейгенбаума (см. [2], [21]). Его изображение для задачи

(27) представлено на фиг. 4. Оно получено методом секущей гиперповерхности Пуанкаре в результате численного интегрирования уравнений (27) и одновременного выделения временной выборки $\tau_q = qT + \tau_0$, $q = 0, 1, \dots, 20000$, с периодом T и $\tau_0 = T/2$ (см. фиг. 4) при значении параметра $\lambda = 0.102910529$. Многократное увеличение любого фрагмента на фиг. 4 позволяет обнаружить вложенную слоистую структуру, присущую “странным” аттракторам.

СПИСОК ЛИТЕРАТУРЫ

1. Анищенко В.С. Сложные колебания в простых системах: механизмы возникновения, структура и свойства динамического хаоса в радиофизических системах. М.: Наука, 1990.
2. Гуляев В.И., Завражина Т.В., Завражина Н.М. Универсальные закономерности зарождения хаотических движений нелинейных механических систем. Киев: ВИПОЛ, 1999.
3. Мун Ф. Хаотические колебания. М.: Мир, 1990.
4. Неймарк Ю.И., Ланда П.С. Стохастические и хаотические колебания. М.: Наука, 1987.
5. Huseyin Koncay, Lin Rui. An intrinsic multiple-scale harmonic balance method for nonlinear vibration and bifurcation problems // Internat. J. Non-Linear Mech. 1991. V. 26. № 5. P. 727–740.
6. Pei Qinyuam, Li Li. The chaotic behaviour of a nonlinear oscillator // Appl. Math. and Mech. 1993. V. 14. № 5. P. 377–387.
7. Van Doorin R. On the transition from regular to chaotic behaviour in the Duffing oscillator // J. Sound and Vibration. 1988. V. 123. № 2. P. 327–339.
8. Yagasaki K., Sakata M., Kimura K. Dynamics of a weakly nonlinear system subjected to combined parametric and external excitation // Trans. ASME. J. Appl. Mech. 1990. V. 57. № 1. P. 209–217.
9. Белогородцев А.Б. Теория регулярного и хаотического поведения слабонелинейных осцилляторов с квазипериодическим воздействием // Тезисы докл. II Всес. конф. “Нелинейные колебания механич. систем.”. Горький, 1990. С. 30–31.
10. Brindley J. Noisy periodicity and non chaotic strange attractors in forced nonlinear oscillators // Tagungsber/ Math. Forschungsinst., Oberwolfach. 1991. № 42. P. 5.
11. Wojewoda J., Kapitaniak T. Oscillations of a quasiperiodically forced system with dry friction // J. Sound and Vibration. 1993. V. 163. № 2. P. 379–384.
12. Yagasaki K. The influence of phase locking on chaotic behavior in a two-frequency perturbation of Duffing's equation // Mem. Fac. Engng. Tamagava Univ., 1991. № 26. P. 15–21.
13. Бибииков Ю.Н. Многочастотные нелинейные колебания и их бифуркации. Л.: ЛГУ, 1991.
14. Самойленко А.М. Элементы математической теории многочастотных колебаний. М.: Наука, 1987.
15. Арнольд В.И. Обыкновенные дифференциальные уравнения. М.: Наука, 1971.
16. Демидович Б.П. Лекции по математической теории устойчивости. М.: Наука, 1967.
17. Йосс Ж., Джозеф Д. Элементарная теория устойчивости и бифуркаций. М.: Мир, 1983.
18. Гуляев В.И., Баженов В.А., Гоцуляк Е.А. и др. Устойчивость периодических процессов в нелинейных механических системах. Львів: Вища школа, 1983.
19. Вайнберг М.М., Треногин В.А. Теория ветвления решений нелинейных уравнений. М.: Наука, 1969.
20. Пуанкаре А. Избранные труды. Т. 1–3. М.: Наука, 1971–1974.
21. Feigenbaum M.J. Universal bahaviour in nonlinear systems // Los Alamos Sci. 1980. V. 1. № 1. P. 4–27.

УДК 519.633

НЕЛИНЕЙНЫЙ ИСТОЧНИК В ДИФФУЗИОННЫХ МЕТОДАХ ФИЛЬТРАЦИИ ИЗОБРАЖЕНИЙ¹⁾

© 2007 г. Г. В. Борисенко, А. М. Денисов

(119992 Москва, Ленинские горы, МГУ, ВМК)

e-mail: gborisenko@gmail.com, den@cs.msu.su

Поступила в редакцию 14.05.2007 г.

Предлагается алгоритм диффузионной фильтрации изображений, основанный на решении начально-краевой задачи для двумерного уравнения диффузии со специальным нелинейным источником. Библ. 9. Фиг. 10.

Ключевые слова: уравнение диффузии, нелинейный источник, фильтрация изображений.

Алгоритмы обработки изображений, основанные на решении смешанных задач для уравнения диффузии, имеют широкую область применения. Основная идея подобных алгоритмов состоит в том, что решение задачи Коши для двумерного уравнения диффузии с постоянным коэффициентом эквивалентно свертке начального условия с функцией Гаусса. Таким образом, в результате решения задачи Коши с начальным условием $f(x, y)$, представляющим собой исходное изображение, в некоторый момент T получается сглаженное изображение. Таким же фильтрующим свойством будет обладать и решение смешанной задачи в ограниченной области, на границе которой задано подходящее краевое условие. Так как процесс диффузии локально определяется значением коэффициента диффузии, то, выбирая тот или иной вид этого коэффициента, можно добиться появления у алгоритма свойств, полезных для решения задач, возникающих при обработке изображений. В настоящее время существует большое количество методов диффузионной фильтрации изображений, основанных на выборе различных типов коэффициентов диффузии (см., например, [1]–[7]).

Цель данной работы состоит в том, чтобы показать, что диффузионные методы обработки изображений могут быть модифицированы на основе введения в уравнение диффузии нелинейного источника специального вида.

1. ДИФФУЗИОННЫЙ ФИЛЬТР С НЕЛИНЕЙНЫМ ИСТОЧНИКОМ

Пусть в прямоугольной области $\bar{\Omega}$, где $\Omega = \{(x, y), 0 < x < a, 0 < y < b\}$, задана функция $f(x, y)$, определяющая исходное монохромное изображение. Рассмотрим задачу для уравнения диффузии:

$$u_t = D^2 \Delta u - Q_1(x, y; u)u, \quad (x, y) \in \Omega, \quad t > 0, \quad (1)$$

$$u|_{\Gamma} = 0, \quad t > 0, \quad \Gamma = \partial\Omega, \quad (2)$$

$$u(x, y, 0) = f(x, y), \quad (x, y) \in \bar{\Omega}, \quad (3)$$

где D^2 – постоянный коэффициент диффузии, а функция $Q_1(x, y; u)$ описывает нелинейный источник. Краевое условие (2) соответствует предположению, что вблизи границы области Ω нет элементов полезного изображения. Функция $Q_1(x, y; u)$ определяется следующим образом. Пусть (x, y) – произвольная точка области Ω . Обозначим через $O(x, y)$ некоторую малую окрестность точки (x, y) . В качестве $O(x, y)$ можно взять круг фиксированного радиуса с центром в точке (x, y) или квадрат с фиксированной стороной и ориентацией и центром в (x, y) . Тогда имеем

$$Q_1(x, y; u) = \left(1 + \left(\lambda \int_{O(x, y)} \bar{u}(s, p, t) ds dp \right)^2 \right)^{-1}, \quad \lambda > 0, \quad (4)$$

¹⁾ Работа выполнена при финансовой поддержке РФФИ (код проекта 05-01-00232).

где

$$\bar{u}(s, p, t) = \begin{cases} u(s, p, t), & (s, p) \in \bar{\Omega}, \\ 0, & (s, p) \notin \bar{\Omega}. \end{cases} \quad (5)$$

В качестве отфильтрованного изображения берется решение $u(x, y, T)$ задачи (1)–(3) в некоторый момент времени T .

Опишем область возможного применения метода диффузионной фильтрации, определяемого начально-краевой задачей (1)–(3). Пусть исходное изображение $f(x, y)$ представляет собой некоторое полезное изображение, искаженное высококонтрастным шумом. Если для фильтрации подобного изображения применить простой диффузионный метод, определяемый задачей (1)–(3) с $Q_1(x, y; u) \equiv 0$, то управляющим параметром алгоритма является только величина D^2T . При этом большие значения D^2T , при которых интенсивность шума будет существенно снижена, приводят к искажению полезной информации на изображении. Действие нелинейного источника $Q_1(x, y; u)u$ состоит в следующем. В тех точках (x, y) , в малой окрестности которых содержится много точек полезного изображения, величина интеграла

$$\int_{O(x, y)} \bar{u}(s, p, t) ds dp \quad (6)$$

велика и функция $Q_1(x, y; u)$ близка к нулю. Следовательно, в этих точках действие источника незначительно. В тех же точках (x, y) , в окрестности которых находится только шум, величина интеграла (6) мала и функция $Q_1(x, y; u)$ оказывает существенное влияние на уменьшение величины решения $u(x, y, t)$ в точке (x, y) , действуя как “отрицательный источник”. Положительная постоянная λ является параметром фильтра.

Приведем пример, иллюстрирующий результат работы алгоритма, определяемого задачей (1)–(3). При численной реализации данного метода смешанная задача решается на основе неявной разностной схемы из [8]. Значения $Q_1(x, y; u)$ вычисляются на предыдущем слое по времени. Для вычисления двукратного интеграла (6) строится кубатурная формула второго порядка аппроксимации с использованием 9-точечного шаблона (см. [9]).

На фиг. 1 приведено исходное зашумленное изображение $f(x, y)$, на фиг. 2 – результат фильтрации, полученный при решении линейной задачи (1)–(3) с $Q_1(x, y; u) \equiv 0$, а на фиг. 3 – результат фильтрации, соответствующий решению нелинейной задачи (1)–(3) с $Q_1(x, y; u)$, определяемой по формуле (4).

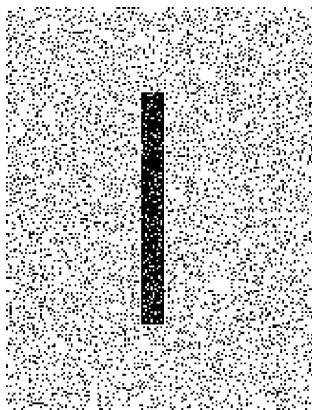
В расчетах предполагалось, что интенсивность изображения меняется от нуля до 255, причем ноль соответствует белому цвету. Величина D^2T в линейном и нелинейном случаях одинакова. Параметр имеет значение $\lambda = 0.01$. В качестве окрестности интегрирования $O(x, y)$ выбран квадрат с центром в точке (x, y) и со стороной в три пикселя. Приведенный пример показывает, что нелинейный источник $Q_1(x, y; u)u$ оказывает существенное положительное влияние на удаление шума.

Рассмотрим другой вариант нелинейного источника. Вследствие того что в приведенной постановке коэффициент $Q_1(x, y; u)$ всегда больше нуля, нелинейный “отрицательный источник” действует в каждой точке изображения. Соответственно, наряду с уменьшением интенсивности высококонтрастного шума предложенный диффузионный фильтр уменьшает и интенсивность полезных объектов. В некоторых случаях при больших значениях T этот эффект может приводить к снижению контрастности изображений. Чтобы избежать потерь интенсивности объектов при фильтрации и увеличить контраст, модифицируем алгоритм, выбрав функцию, определяющую нелинейный источник, в следующем виде:

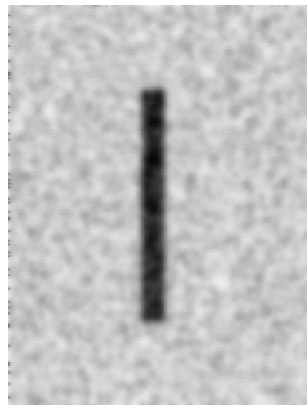
$$Q_2(x, y; u) = R \left(\left(\int_{O(x, y)} \bar{u}(s, p, t) ds dp \right)^2 \right), \quad (7)$$

где $\bar{u}(s, p, t)$ имеет вид (5), а функция $R(z)$ такова, что верно следующее:

$$R(z) > 0, \quad z \in [0, Z_0), \quad R(z) \leq 0, \quad z \in [Z_0, Z_1], \quad R(z) = 0, \quad z \notin [0, Z_1].$$



Фиг. 1.



Фиг. 2.



Фиг. 3.

Величина Z_1 определяется как значение квадрата интеграла (6) по области, в которой все точки имеют максимальную интенсивность. Другими словами,

$$\left(\int_{O(x,y)} \bar{u}(s, p, t) dsdp \right)^2 \leq Z_1 \quad \forall (x, y) \in \Omega.$$

Рассмотрим действие функции $Q_2(x, y; u)$ в алгоритме диффузионной фильтрации изображений. В тех точках (x, y) области Ω , где $0 \leq \left(\int_{O(x,y)} \bar{u}(s, p, t) dsdp \right)^2 < Z_0$, имеем $Q_2(x, y; u) > 0$. Следовательно, в них располагаются “отрицательные источники”. При фильтрации член $Q_2(x, y; u)u$ будет существенно уменьшать интенсивность в таких точках. Величина параметра Z_0 определяет порог отсекающего шума. Варьируя значение этого параметра, можно делать алгоритм более или менее чувствительным к зашумленности фильтруемого изображения $f(x, y)$. В то же время если величина интеграла в окрестности некоторой точки лежит в интервале $Z_0 \leq \left(\int_{O(x,y)} \bar{u}(s, p, t) dsdp \right)^2 \leq Z_1$, то считаем, что эта точка принадлежит полезному объекту. Функция $Q_2(x, y; u)$ в этом случае неположительная, а значит, в рассматриваемой точке либо расположен “положительный источник”, либо $Q_2(x, y; u)u = 0$. Это соответствует тому, что полезные объекты на изображении усиливаются или остаются без существенных изменений. Вид функции $R(z)$ является управляющим параметром данного метода.

Результаты работы фильтров можно проиллюстрировать на следующем примере. На фиг. 4 представлено исходное зашумленное изображение. На фиг. 5 и 6 приводятся результаты фильтрации методами, основанными на решении задачи (1)–(3), с функциями $Q_1(x, y; u)$ и $Q_2(x, y; u)$, определяемыми (4) и (7) соответственно.

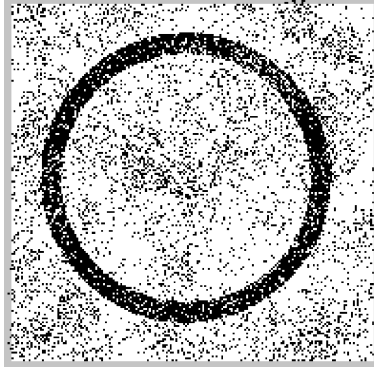
Функция $R(z)$ была выбрана в виде

$$R(z) = A \cos(B \lambda^2 z) (1 + \lambda^2 z), \quad z \in [0, (2288)^2], \quad R(z) = 0, \quad z \notin [0, (2288)^2].$$

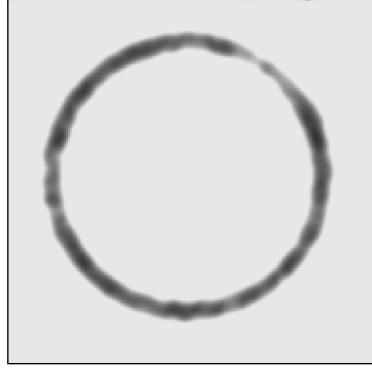
Значение параметра D^2T одинаково. Параметр $\lambda = 0.001$ для обоих методов, $A = 3$, $B = 0.009$.

2. АНИЗОТРОПНЫЙ ДИФFUЗИОННЫЙ ФИЛЬТР С НЕЛИНЕЙНЫМ ИСТОЧНИКОМ

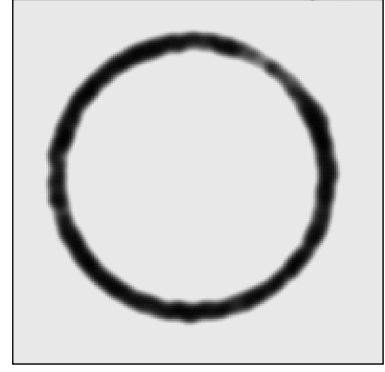
Линейный фильтр (1)–(3) с $Q_1(x, y; u) \equiv 0$ является самым простым диффузионным фильтром. При фильтрации он не учитывает информацию ни об интенсивности объектов, ни о направлении контуров. Существует класс изображений, на которых полезными объектами являются достаточно тонкие и поврежденные линии. Для их фильтрации нужно использовать алгоритмы, учитывающие локальные направления линий. Одним из наиболее эффективных алгоритмов обработки таких изображений является анизотропный диффузионный фильтр, предложенный в [6]. Этот фильтр построен таким образом, чтобы уменьшать диффузию в направлении, перпендикулярном линиям, тем больше, чем сильнее выражена их граница. С другой стороны, диффу-



Фиг. 4.



Фиг. 5.



Фиг. 6.

зия вдоль границы продолжается, фильтруя шум без разрушения контура. Чтобы использовать при фильтрации информацию о направлении линий, в качестве коэффициента диффузии выбирают матрицы специального вида.

Анизотропный алгоритм диффузионной фильтрации описывается следующей задачей:

$$\begin{aligned} u_t &= \operatorname{div}(H(u) \times \nabla u), \quad (x, y) \in \Omega, \quad t > 0, \\ u|_{\Gamma} &= 0, \quad t > 0, \quad \Gamma = \partial\Omega, \\ u(x, y, 0) &= f(x, y), \quad (x, y) \in \bar{\Omega}. \end{aligned} \quad (8)$$

Коэффициент диффузии выбирают в виде матрицы, зависящей от решения

$$H(u) = \frac{1}{\hat{u}_x^2 + \hat{u}_y^2} \begin{pmatrix} g(|\nabla \hat{u}|^2) \hat{u}_x^2 + \hat{u}_y^2 & (g(|\nabla \hat{u}|^2) - 1) \hat{u}_x \hat{u}_y \\ (g(|\nabla \hat{u}|^2) - 1) \hat{u}_x \hat{u}_y & \hat{u}_x^2 + g(|\nabla \hat{u}|^2) \hat{u}_y^2 \end{pmatrix},$$

где

$$g(s) = \frac{1}{1 + s/M}, \quad M > 0,$$

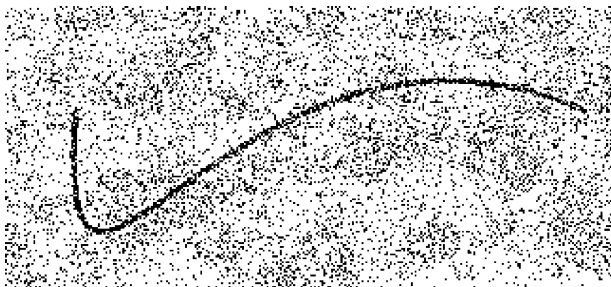
$$\hat{u}(x, y, t) = \frac{1}{\pi \sigma^2} \iint_{\mathbb{R}^2} \bar{u}(s, p, t) \exp\left(-\frac{(x-s)^2 + (y-p)^2}{\sigma^2}\right) ds dp, \quad \sigma > 0,$$

а $\bar{u}(s, p, t)$ определяется формулой (5).

Этот фильтр сам по себе уже является достаточно эффективным, однако добавление в него нелинейного источника помогает дополнительно улучшить его свойства. Несмотря на то, что в анизотропном фильтре размытие границ мало, оно все-таки происходит. При этом с увеличением времени фильтрации из-за нечеткости контура скорость размытия увеличивается. С другой стороны, сглаживание вдоль поврежденного контура усредняет интенсивность на границе, делая тем самым ее менее контрастной. Добавление в алгоритм нелинейного источника, определяемого по формуле (7), позволяет улучшить решение обозначенных проблем. Нелинейный источник, с одной стороны, ускоряет удаление шума с изображения, помогая уменьшить диффузию в направлении, перпендикулярном линии. С другой стороны, расположенные на объекте “положительные источники” увеличивают интенсивность линии, компенсируя снижение контрастности, вызванное диффузией.

Модифицированный анизотропный фильтр с нелинейным источником задается в виде

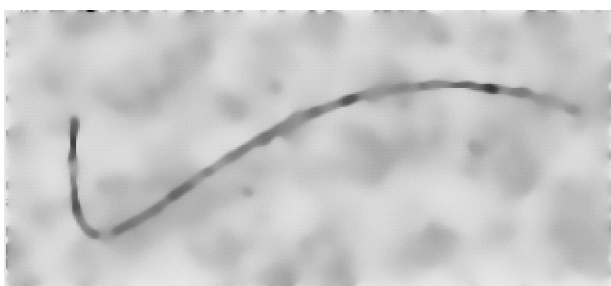
$$\begin{aligned} u_t &= \operatorname{div}(H(u) \times \nabla u) - Q_2(x, y; u)u, \quad (x, y) \in \Omega, \quad t > 0, \\ u|_{\Gamma} &= 0, \quad t > 0, \quad \Gamma = \partial\Omega, \\ u(x, y, 0) &= f(x, y), \quad (x, y) \in \bar{\Omega}. \end{aligned} \quad (9)$$



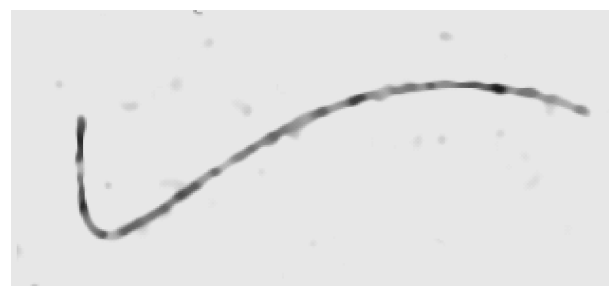
Фиг. 7.



Фиг. 8.



Фиг. 9.



Фиг. 10.

Фиг. 7–10 иллюстрируют работу диффузионного фильтра (9) в сравнении с фильтрами, основанными на решении задач (1)–(3) и (8). На фиг. 7 представлено исходное изображение. На фиг. 8 – результат фильтрации методом, основанным на решении задачи (1)–(3), с коэффициентом (7). Несмотря на то, что этот фильтр был достаточно успешно использован в предыдущем примере, в данном случае малость линии и ее поврежденность не позволили сохранить полезную информацию и удалить шум одновременно. Фиг. 9 иллюстрирует результат работы исходного анизотропного фильтра (8), а фиг. 10 – изображение, отфильтрованное методом, построенным на основе решения задачи (9).

Количество шагов по времени для всех трех отфильтрованных изображений одинаково. Функция $R(z)$ была выбрана в виде

$$R(z) = A \cos(B \lambda^2 z)(1 + \lambda^2 z), \quad z \in [0, (2288)^2], \quad R(z) = 0, \quad z \notin [0, (2288)^2].$$

Параметр $\lambda = 0.001$, $A = 3$, $B = 0.009$, $\sigma = 0.05$, $M = 20$. В качестве окрестности интегрирования $O(x, y)$ выбран квадрат с центром в точке (x, y) и со стороной в три пикселя.

Проведенные расчеты показывают, что использование нелинейного источника позволяет в ряде случаев улучшить диффузионный метод фильтрации изображений.

СПИСОК ЛИТЕРАТУРЫ

1. Perona P., Malik J. Scale space and edge detection using anisotropic diffusion // IEEE Trans. Pattern Analys. and Machine Intelligence. 1990. V. 12. P. 629–639.
2. Weickert J. Foundations and applications of nonlinear anisotropic diffusion filtering // Z. angew. Math. und Mech. 1996. Bd. 76. S. 283–286.
3. Weickert J. A review of nonlinear diffusion filtering // Lect. Notes Comput. Sci. 1997. V. 1252. P. 3–28.
4. Acharya T., Ray A. Image processing. Principles and applications. ISBN 0471719986. Wiley, 2005.
5. Paragios N., Chen Y., Faugeras O. Mathematical models in computer vision: The Handbook. ISBN 0-387-26371-3. Berlin: Springer, 2006.
6. Weickert J. Anisotropic diffusion in image processing. Stuttgart: Teubner, 1998.
7. Борисенко Г.В., Денисов А.М., Крылов А.С. Об одном методе диффузионной фильтрации изображений // Программирование. 2004. № 5. С. 43–47.
8. Самарский А.А., Гулин А.В. Численные методы математической физики. М.: Научн. мир, 2000.
9. Бахвалов Н.С. О приближенном вычислении кратных интегралов // Вестн. МГУ. Сер. Матем., физ., астрон. 1959. № 4.

УДК 519.633

НЕОБХОДИМЫЕ УСЛОВИЯ ДЛЯ ε -РАВНОМЕРНОЙ СХОДИМОСТИ РАЗНОСТНЫХ СХЕМ ДЛЯ ПАРАБОЛИЧЕСКИХ УРАВНЕНИЙ С ДВИЖУЩИМИСЯ ПОГРАНИЧНЫМИ СЛОЯМИ¹⁾

© 2007 г. Г. И. Шишкин

(620219 Екатеринбург, ул. С. Ковалевской, 16, ИММ УрО РАН)

e-mail: shishkin@imm.uran.ru

Поступила в редакцию 20.04.2007 г.

Рассматривается сеточная аппроксимация краевой задачи для сингулярно возмущенного параболического уравнения типа реакции-диффузии в области с границами, движущимися в сторону положительного направления оси x . При малых значениях параметра ε (параметра при старших производных уравнения; $\varepsilon \in (0, 1]$) в окрестности левой боковой границы S_1^L появляется движущийся пограничный слой. В случае стационарных пограничных слоев классические разностные схемы на кусочно-равномерных сетках, сгущающихся в слоях, сходятся ε -равномерно со скоростью $O(N^{-1} \ln N + N_0)$, где величины N и N_0 определяют число узлов сетки по x и t . Для рассматриваемой задачи классические разностные схемы на основе равномерных сеток сходятся лишь при условии $N^{-1} + N_0^{-1} \ll \varepsilon$. Оказывается, что в классе разностных схем на основе прямоугольных сеток, сгущающихся по x и t в окрестности множества S_1^L , сходимость при условии $N^{-1} + N_0^{-1} \leq \varepsilon^{1/2}$ недостижима. Рассмотрение поперечников, аналогичных поперечникам по Колмогорову, позволило установить необходимые и достаточные условия для ε -равномерной сходимости аппроксимаций решения краевой задачи. С использованием этих условий строится схема, сходящаяся ε -равномерно со скоростью $O(N^{-1} \ln N + N_0)$. Библ. 18.

Ключевые слова: краевая задача для параболических уравнений, возмущающий параметр ε , параболическое уравнение реакции-диффузии, разностная аппроксимация, движущийся пограничный слой, поперечники по Колмогорову, ε -равномерная сходимость.

1. ВВЕДЕНИЕ

Для сингулярно возмущенных краевых задач хорошо известна проблема разработки специальных сеточных методов, погрешность решений которых слабо зависит от величины параметра ε , в частности методов, сходящихся ε -равномерно (см. [1]–[4]). В настоящее время достаточно хорошо разработан метод построения ε -равномерно сходящихся схем на специальных сетках – сетках, априорно сгущающихся в пограничных слоях (см., например, [2], [5]–[8] в случае уравнений в частных производных и обыкновенных дифференциальных уравнений (см. [9])). Методы, использующие кусочно-равномерные сетки, сгущающиеся в пограничных слоях, получили достаточно широкое распространение ввиду их простоты и удобства в использовании (см., например, [5]–[8] и библиографию там же).

Заметим, что специальные численные методы для сингулярно возмущенных параболических уравнений практически рассматривались лишь для задач со стационарными пограничными и внутренними слоями. Отметим специфику в построении специальных разностных схем в случае движущихся пограничных и внутренних слоев. Для начальной сингулярно возмущенной задачи с движущимся сосредоточенным источником, в окрестности которого появляется движущийся внутренний слой, в [10]–[12] построены специальные разностные схемы, сходящиеся ε -равно-

¹⁾ Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-01-00729, 04-01-89007-НВО_а), Нидерландской организации научных исследований NWO (проект № 047.016.008), Булевского центра исследований по информатике (BCRI) Национального университета Ирландии г. Корк, а также MACSI – Ассоциации по приложениям математики в науке и технике в Ирландии по математической инициативе Ирландского научного фонда (Mathematics Applications Consortium for Science and Industry in Ireland (MACSI) under the Science Foundation Ireland (SFI) mathematics initiative).

мерно. При построении схем в окрестности траектории движущегося источника использовались сетки, сгущающиеся по оси x и не являющиеся прямоугольными. В связи с достаточной сложностью таких схем возникает интерес к методам построения более простых разностных схем, а также к альтернативным численным методам (методам на апостериорно адаптирующихся сетках), сходящимся ε -равномерно.

В случае сингулярно возмущенных задач с движущимися пограничными и внутренними слоями исследование условий, являющихся необходимыми и достаточными для ε -равномерной сходимости численных методов, представляется актуальной проблемой.

В настоящей работе рассматривается краевая задача для сингулярно возмущенного параболического уравнения типа реакции-диффузии в области с движущимися границами, скорость движения которых направлена в сторону положительного направления оси x . При малых значениях параметра ε в окрестности множества S_1^L – левой части боковой границы – появляется движущийся пограничный слой. Для краевой задачи строится разностная схема, сходящаяся ε -равномерно.

Заметим, что классические разностные схемы на основе равномерных сеток для задач такого типа сходятся лишь при условии $N^{-1} + N_0^{-1} \ll \varepsilon$, где величины N и N_0 определяют число узлов сетки по переменным x и t (см. утверждение теоремы 2). Оказывается, что в классе разностных схем на основе прямоугольных сеток, сгущающихся по x и t в окрестности множества S_1^L , не существует схем, сходящихся уже при условии $P_0^{-1} \approx \varepsilon$, где P_0 – число узлов используемых сеток, $P_0 \approx NN_0$ (см. утверждение теоремы 3).

Для того чтобы найти условия, необходимые для ε -равномерной сходимости схем на адаптирующихся сетках, исследуются поперечники, аналогичные поперечникам по Колмогорову. Рассмотрение поперечников позволило установить условия, накладываемые на триангуляцию области, как необходимые, так и достаточные для ε -равномерной сходимости оптимальных аппроксимаций решений краевой задачи при $P \rightarrow \infty$, где P – число узлов, используемых в триангуляции (см. утверждения теорем 4–10 и их обсуждения в разд. 6, 7). Из условий, необходимых для ε -равномерной сходимости оптимальных аппроксимаций решений в случае криволинейной границы S_1^L следует, что для ε -равномерной сходимости схемы необходимо использовать сетки, являющиеся в окрестности движущегося пограничного слоя существенно анизотропными, согласованными с границей S_1^L в ее окрестности. Такого типа сетки соответствуют прямоугольным сеткам, сгущающимся в окрестности пограничного слоя для краевой задачи, записанной в системе координат, в которой граница S_1^L переходит в отрезок прямой, параллельной оси t . С использованием таких специальных сеток строится разностная схема, сходящаяся ε -равномерно со скоростью $O(N^{-1} \ln N + N_0^{-1})$.

2. ПОСТАНОВКА ЗАДАЧИ. ЦЕЛЬ ИССЛЕДОВАНИЯ

2.1. В области $\bar{G}$ с границей $S = \bar{G} \setminus G$, где

$$G = \{(x, t) : \beta_1(t) < x < \beta_2(t), t \in (0, T]\}, \quad (2.1)$$

рассмотрим краевую задачу для сингулярно возмущенного параболического уравнения

$$\begin{aligned} Lu(x, t) \equiv \left\{ \varepsilon \frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial t} \right\} u(x, t) &= f(x, t), \quad (x, t) \in G, \\ u(x, t) &= \varphi(x, t), \quad (x, t) \in S. \end{aligned} \quad (2.2)$$

Здесь $f(x, t)$, $(x, t) \in \bar{G}$, $\varphi(x, t)$, $(x, t) \in S$, $\beta_i(t)$, $t \in [0, T]$, $i = 1, 2$, – достаточно гладкие функции, удо-

влетворяющие условиям

$$\begin{aligned} |f(x, t)| \leq M, \quad (x, t) \in \bar{G}; \quad |\varphi(x, t)| \leq M, \quad (x, t) \in S; \\ 0 < v_0 \leq (d/dt)\beta_i(t) \equiv v_i(t) \leq v^0, \quad m_1 \leq \beta_2(t) - \beta_1(t) \leq M_1, \\ t \in [0, T], \quad i = 1, 2, \end{aligned} \quad (2.3)$$

причем $\beta_1(0) = 0, \beta_2(0) = d$; ε – параметр, принимающий произвольные значения из полуинтервала $(0, 1]$; производные $\beta'_i(t)$ определяют скорость движения боковых границ. Считаем, что граница S состоит из множеств S^L и S_0 : $S = S_0 \cup S^L$, где S^L – боковая граница, $S^L = S_1^L \cup S_2^L$, S_1^L и S_2^L – соответственно, левая и правая границы множества G , S_0 – нижнее основание множества G , $S_0 = \bar{S}_0$.

Считаем, что на множестве $S^c = S_0 \cup \bar{S}^L$ – в угловых точках $(0, 0)$ и $(d, 0)$ – выполнены условия согласования, обеспечивающие достаточную гладкость решения задачи при фиксированных значениях параметра ε (см. [13]).

При $\varepsilon \rightarrow 0$ в окрестности множества S_1^L появляется движущийся пограничный слой, экспоненциально убывающий при удалении от множества S_1^L – с ростом x и/или уменьшением t (см. оценку (3.7)).

2.2. Ошибки решений разностных схем, построенных на основе классических разностных аппроксимаций задачи (2.2), (2.1), зависят от величины параметра ε и становятся малыми лишь при значениях параметра ε , существенно превосходящих “эффективные” шаги сеток по x и t . Так, в силу оценок (4.7), (4.10), классическая разностная схема (4.4), (4.6) (см. разд. 4) сходится при условии $(N^{-1} + N_0^{-1} \ll \varepsilon)$

$$\varepsilon^{-1} = o(\min[N, N_0]), \quad N, N_0 \rightarrow \infty, \quad (2.4)$$

где величины N, N_0 определяют число узлов сеток по x и по t соответственно. При нарушении этого условия решения разностной схемы не сходятся к решению задачи (2.2), (2.1). Условие (2.4) является более ограничительным, чем условие $(N^{-1} \ll \varepsilon)$

$$\varepsilon^{-1} = o(N), \quad (2.5)$$

а именно условие сходимости классической схемы для задач в областях с неподвижными границами; заметим, что в условии (2.5) ограничения на шаг сетки по t не накладываются.

Таким образом, в связи с отмеченным поведением сеточных решений (аппроксимирующих решение дифференциальной задачи с движущимся пограничным слоем) возникает интерес к построению специальных разностных схем, погрешность решений которых не зависит от величины параметра ε . В частности, представляют интерес разностные схемы, сходящиеся при более слабом условии, чем условие (2.4) – условие сходимости решений разностной схемы (4.4), (4.6).

Несомненный интерес представляют условия, накладываемые на строящиеся сеточные аппроксимации задачи (2.2), (2.1), которые являются необходимыми и достаточными для ε -равномерной (либо близкой к ε -равномерной) сходимости сеточных решений.

Далее нам понадобится ряд определений в случае разностных схем на сетках с достаточно произвольным распределением узлов.

Определения. 1. Пусть на множестве $E_{\bar{N}, \varepsilon} = E_{\bar{N}} \times E_{\varepsilon}$, где $E_{\bar{N}}$ – подмножество из множества пар натуральных чисел N, N_0 , удовлетворяющих условию

$$N, N_0 \geq M_0, \quad E_{\varepsilon} = \{\varepsilon : \varepsilon \in (0, 1]\},$$

определены функции

$$\psi_i(N^{-1}, N_0^{-1}, \varepsilon), \quad i = 1, 2, \quad \psi_i(N^{-1}, N_0^{-1}, \varepsilon) > 0.$$

Запись

$$\psi_1(N^{-1}, N_0^{-1}, \varepsilon) = \hat{o}(\psi_2(N^{-1}, N_0^{-1}, \varepsilon)) \quad \text{на } E_{\bar{N}, \varepsilon}$$

означает, что найдется такая точка $(\tilde{N}_1^{-1}, \tilde{N}_2^{-1}, \tilde{\varepsilon})$, для которой выполняется соотношение

$$\psi_1(N^{-1}, N_0^{-1}, \varepsilon) [\psi_2(N^{-1}, N_0^{-1}, \varepsilon)]^{-1} \longrightarrow 0 \quad \text{при} \quad (N^{-1}, N_0^{-1}, \varepsilon) \longrightarrow (\tilde{N}^{-1}, \tilde{N}_0^{-1}, \tilde{\varepsilon}),$$

$$(N^{-1}, N_0^{-1}, \varepsilon), (\tilde{N}^{-1}, \tilde{N}_0^{-1}, \tilde{\varepsilon}) \in E_{\bar{N}, \varepsilon}.$$

2. Пусть для сеточной функции $z(x, t)$, $(x, t) \in \bar{G}_h$, – решения некоторой разностной схемы – выполняется оценка

$$|u(x, t) - z(x, t)| \leq M\mu(N^{-1}, N_0^{-1}, \varepsilon), \quad (x, t) \in \bar{G}_h.$$

Здесь сетка $\bar{G}_h$ – прямое произведение сеток по x и t ; $N + 1$ и $N + 1$ – число узлов сеток по x и t соответственно. Будем говорить, что эта оценка *неулучшаема* по вхождению величин N, N_0, ε , если оценка

$$|u(x, t) - z(x, t)| \leq M\mu_0(N^{-1}, N_0^{-1}, \varepsilon), \quad (x, t) \in \bar{G}_h,$$

вообще говоря, неверна в том случае, когда $\mu_0(N^{-1}, N_0^{-1}, \varepsilon) = \hat{o}(\mu(N^{-1}, N_0^{-1}, \varepsilon))$ на $E_{\bar{N}, \varepsilon}$.

3. Пусть при $N, N_0 \longrightarrow \infty, \varepsilon \in E_\varepsilon$ решение разностной схемы сходится к решению краевой задачи в случае условия $N^{-1}, N_0^{-1} = o(\varepsilon^\nu)$, $\varepsilon \in E_\varepsilon$, однако сходимость при условии $N^{-1}, N_0^{-1} = O(\varepsilon^\nu)$, вообще говоря, не имеет места. В этом случае будем говорить, что разностная схема *сходится с дефектом ν* относительно параметра ε при $N, N_0 \longrightarrow \infty$ (или, короче, сходится с дефектом ν). При $\nu = 0$ сходимость схемы ε -равномерная.

4. Пусть $\bar{G}_h$ есть некоторая сетка (вообще говоря, не прямоугольная) на множестве $\bar{G}$ и P_0 – число узлов сетки $\bar{G}_h$. Пусть сеточная функция $z(x, t)$, $(x, t) \in \bar{G}_h$, – решение некоторой разностной схемы. Оценка

$$|u(x, t) - z(x, t)| \leq M\mu(P_0^{-1}, \varepsilon), \quad (x, t) \in \bar{G}_h,$$

неулучшаемая по вхождению величин P_0, ε , определяется подобно неулучшаемой оценке по вхождению величин N, N_0, ε .

В том случае, когда при $P_0 \longrightarrow \infty, \varepsilon \in E_\varepsilon$ решение разностной схемы сходится к решению краевой задачи в случае условия $P_0^{-1/2} = o(\varepsilon^\nu)$, $\varepsilon \in E_\varepsilon$, но сходимость при условии $P_0^{-1/2} = O(\varepsilon^\nu)$, вообще говоря, не имеет места, будем говорить, что схема *сходится с дефектом ν* (при $P_0 \longrightarrow \infty$).

В том случае, когда величина ν может быть выбрана сколь угодно малой, будем говорить, что разностная схема, контролируемая величиной ν , *сходится почти ε -равномерно с дефектом ν* (или, короче, *почти ε -равномерно*).

Таким образом, дефект (ε -равномерной сходимости) классической схемы (4.4), (4.6) (относительно величин N, N_0, ε) равен единице.

Наша цель – для краевой задачи (2.2), (2.1) построить разностную схему, сходящуюся ε -равномерно, и, кроме того, найти условия, являющиеся необходимыми для ε -равномерной и почти ε -равномерной сходимости решений разностных схем, строящихся на основе классических аппроксимаций дифференциального уравнения.

Априорные оценки решения задачи (2.2), (2.1) приводятся в разд. 3. Классическая разностная схема на основе сеток, являющихся прямым произведением сеток по x и t , строится в разд. 4; в разд. 5 обсуждаются трудности, возникающие при построении разностных схем, сходящихся ε -равномерно либо с достаточно малым дефектом, при использовании прямоугольных сеток – прямых произведений сеток по x и t . Построение аппроксимаций решений краевой задачи (2.2), (2.1) на основе интерполянтов на триангуляциях множества $\bar{G}$ – аппроксимаций, сходящихся с малым дефектом ν , в частности сходящихся ε -равномерно, – обсуждается в разд. 6, 7. Результаты разделов 6, 7 используются в разд. 8 при построении схемы, сходящейся ε -равномерно. Замечания и обобщения приводятся в разд. 9.

3. АПРИОРНЫЕ ОЦЕНКИ

Приведем априорные оценки решения краевой задачи (2.2), (2.1), используемые при построениях.

На множестве $\bar{G}$ решение задачи представим в виде регулярной и сингулярной компонент

$$u(x, t) = U(x, t) + W(x, t), \quad (x, t) \in \bar{G}. \quad (3.1)$$

В задаче (2.2), (2.1) перейдем к переменным ξ, t , в которых боковые границы являются неподвижными.

Будет удобно вместо переменной x ввести переменную $\xi = \xi(x, t)$, полагая

$$\xi = \xi(x, t) = d[x - \beta_1(t)][\beta_2(t) - \beta_1(t)]^{-1}, \quad (x, t) \in \bar{G}; \quad (3.2a)$$

через $\xi^{-1}(\xi, t) \equiv x(\xi, t)$ обозначим отображение, обратное $\xi(x, t)$. Для функций $v(x, t)$ и $Z(\xi, t)$ и подобластей $G^0 \subseteq \bar{G}$ будем использовать обозначения

$$\begin{aligned} v(x(\xi, t), t) &= v_\xi(\xi, t) = \{v(x, t)\}_\xi = \tilde{v}(\xi, t), \\ Z(\xi(x, t), t) &= Z_{\xi^{-1}}(x, t) = \{Z(\xi, t)\}_{\xi^{-1}}, \end{aligned} \quad (3.2b)$$

$$G_\xi^0 = \{G^0\}_\xi = \xi(G^0) = \{(\xi, t): (x(\xi, t), t) \in G^0\}. \quad (3.2b)$$

Полагаем

$$\tilde{G}_{\xi^{-1}}^0 = \{\tilde{G}^0\}_{\xi^{-1}} = \xi^{-1}(\tilde{G}^0) = \{(x, t): (\xi(x, t), t) \in \tilde{G}^0\},$$

где $\tilde{G}^0$ – некоторое подмножество из множества $\bar{G}$; $\tilde{G} = \bar{G}_\xi = \{\bar{G}\}_\xi$.

Задача (2.2), (2.1) в переменных ξ, t переходит в краевую задачу

$$\begin{aligned} \tilde{L}\tilde{u}(\xi, t) &\equiv \left\{ \varepsilon A(\xi, t) \frac{\partial^2}{\partial \xi^2} + B(\xi, t) \frac{\partial}{\partial \xi} - \frac{\partial}{\partial t} \right\} \tilde{u}(\xi, t) = \tilde{f}(\xi, t), \quad (\xi, t) \in \tilde{G}, \\ \tilde{u}(\xi, t) &= \tilde{\varphi}(\xi, t), \quad (\xi, t) \in \tilde{S}. \end{aligned} \quad (3.3)$$

Здесь

$$A(\xi, t) = \left\{ \left[\frac{\partial}{\partial x} \xi(x, t) \right]^2 \right\}_\xi, \quad B(\xi, t) = - \left\{ \left[\frac{\partial}{\partial t} \xi(x, t) \right] \right\}_\xi, \quad (\xi, t) \in \bar{G}.$$

В силу условия (2.3),

$$B(\xi, t) \geq B_0 > 0, \quad (\xi, t) \in \bar{G}.$$

Задача (3.3) на множестве

$$\bar{\tilde{G}} = \tilde{G} \cup \tilde{S}, \quad \tilde{G} = \tilde{D} \times (0, T], \quad \tilde{D} = \{\xi: 0 < \xi < d\}, \quad (3.4)$$

есть краевая задача для сингулярно возмущенного параболического уравнения конвекции-диффузии в области с неподвижной боковой границей. Пограничный слой появляется в окрестности множества $\tilde{S}_1^L$ – левой части боковой границы $\tilde{S}^L$, на которую направлен конвективный поток. Выполним оценки регулярной и сингулярной компонент в переменных ξ, t (см., например, [14], [15]).

В переменных ξ, t для сингулярной компоненты выполняется оценка

$$\left| \frac{\partial^{k+k_0}}{\partial \xi^k \partial t^{k_0}} \tilde{W}(\xi, t) \right| \leq M \varepsilon^{-k} \exp(-m \varepsilon^{-1} \xi), \quad (\xi, t) \in \bar{\tilde{G}}; \quad k + 2k_0 \leq 4, \quad (3.5)$$

где m – произвольное число из интервала $(0, m^0)$, $m^0 = d^{-1} m_{1(2,3)} v_{0(2,3)}$.

Возвращаясь к переменным x, t , получаем оценки

$$\left| \frac{\partial^{k_1+k_0}}{\partial x^{k_1} \partial t^{k_0}} U(x, t) \right| \leq M, \quad (3.6)$$

$$\left| \frac{\partial^{k_1+k_0}}{\partial x^{k_1} \partial t^{k_0}} W(x, t) \right| \leq M\varepsilon^{-k_1-k_0} \exp\{-m_1\varepsilon^{-1}[x - \beta_1(t)]\}, \quad (x, t) \in \bar{G}, \quad k_1 + 2k_0 \leq K,$$

где m_1 – произвольное число из интервала $(0, m_0)$:

$$m_0 = m_{1(2.3)} M_{1(2.3)}^{-1} \min_{[0, T]} [(d/dt)\beta_1(t)].$$

Для функции $W(x, t)$ выполняется также оценка

$$\left| \frac{\partial^{k_1+k_0}}{\partial x^{k_1} \partial t^{k_0}} W(x, t) \right| \leq M\varepsilon^{-k_1-k_0} \exp[-m\varepsilon^{-1}r((x, t), S_1^L)], \quad (x, t) \in \bar{G}, \quad k_1 + 2k_0 \leq K, \quad (3.7a)$$

где $r((x, t), S_1^L)$ – расстояние от точки (x, t) до множества S_1^L , m – произвольное число из интервала $(0, m^0)$:

$$m^0 = v_0(1 + (v^0)^2)^{-1/2}, \quad v_0 = v_{0(2.3)}, \quad v^0 = v_{(2.3)}^0. \quad (3.7b)$$

Таким образом, в отличие от задач в областях со стационарными границами, производные по обоим переменным x и t сингулярной компоненты $W(x, t)$ в окрестности движущегося пограничного слоя неограниченно растут при $\varepsilon \rightarrow \infty$. Сингулярная компонента $W(x, t)$ и ее производные по x и t экспоненциально убывают с ростом $\varepsilon^{-1}r((x, t), S_1^L)$.

Функция

$$W_0(x, t) = [\varphi(x, t) - U(x, t)] \exp\{-\varepsilon^{-1}\beta_1'(t)[x - \beta_1(t)]\}, \quad (3.8)$$

$$(x, t) \in \bar{G}, \quad r((x, t), S_1^L) \leq m_1,$$

есть главный член сингулярной компоненты $W(x, t)$:

$$|W(x, t) - W_0(x, t)| \leq M\varepsilon \exp[-m\varepsilon^{-1}r((x, t), S_1^L)], \quad (3.9)$$

$$(x, t) \in \bar{G}, \quad r((x, t), S_1^L) \leq m_1,$$

где $m = m_{(3.7)}$.

При выводе оценок для простоты предполагалось выполненным условие

$$\frac{\partial^{k_1+k_0}}{\partial x^{k_1} \partial t^{k_0}} \varphi(x, t) = 0, \quad \frac{\partial^{k_1+k_0}}{\partial x^{k_1} \partial t^{k_0}} f(x, t) = 0, \quad (x, t) \in S^c, \quad k_1 + k_0 \leq l, \quad (3.10)$$

где $l \geq 6$, обеспечивающее условие согласования на множестве S^c (см. [13]).

Теорема 1. Пусть для данных краевой задачи (2.2), (2.1) выполняется условие $f \in C^{l+\alpha}(\bar{G})$, $\varphi \in C^{l+\alpha}(S_0) \cap C^{(l+\alpha)/2}(\bar{S}^L)$, $\beta_i \in C^{l+\alpha}([0, T])$, $l = K + 4$, $K \geq 2$, $\alpha \in (0, 1)$, и пусть выполняются условия (2.3), (3.10). Тогда для компонент из представления (3.1) решения краевой задачи (2.2), (2.1) справедливы оценки (3.6), (3.7), (3.9); в переменных ξ, t для сингулярной компоненты выполняется оценка (3.5).

4. КЛАССИЧЕСКИЕ РАЗНОСТНЫЕ СХЕМЫ

Выпишем классическую разностную схему для задачи (2.2), (2.1) и укажем некоторые трудности, возникающие при численном решении задач при малых значениях параметра ε .

На полосе

$$\bar{G}^\infty = \mathbb{R} \times [0, T]$$

введем прямоугольные базовые сетки, на основе которых будем строить требуемые сеточные множества. Пусть

$$\bar{G}_h = \bar{G}_h^b = \omega_1 \times \bar{\omega}_0, \quad (4.1)$$

где ω_1 и $\bar{\omega}_0$ – сетки на оси x и отрезке $[0, T]$ соответственно; ω_1 и $\bar{\omega}_0$ – сетки с произвольным распределением узлов, удовлетворяющим лишь условию $h \leq MN^{-1}$, $h_t \leq MN_0^{-1}$, где $h = \max_i h^i$, $h^i = x^{i+1} - x^i$, $x^i, x^{i+1} \in \omega_1$, $h_t = \max_j h_t^j$, $h_t^j = t^{j+1} - t^j$, $t^j, t^{j+1} \in \bar{\omega}_0$. Здесь $N+1$ и N_0+1 – максимальное число узлов на отрезке единичной длины на оси x и число узлов сетки $\bar{\omega}_0$ соответственно. Наибольший интерес для нас будут представлять сетки, равномерные по x и t :

$$\bar{G}_h = \bar{G}_h^b, \quad (4.2)$$

т.е. сетки $\bar{G}_{h(4.1)}^b$, где ω_1 и $\bar{\omega}_0$ – равномерные сетки с шагами $h = N^{-1}$ и $h_t = TN_0^{-1}$.

На множестве $\bar{G}$ строим сетку (порождаемую сеткой $\bar{G}_{h(4.1)}$)

$$\bar{G}_h = \bar{G}_h(\bar{G}_{h(4.1)}^b) = G_h \cup S_h. \quad (4.3)$$

Множество G_h есть множество узлов (x^i, t^j) из $G \cap \bar{G}_h^b$, для которых сегмент $x^i \times (t^{j-1}, t^j]$ целиком принадлежит G . Множество S_h образовано пересечениями прямых $t = t^j$, $t^j \in \bar{\omega}_0$, со сторонами S^L (это множество обозначим через S_h^L), а также узлами (x^i, t^0) , принадлежащими S_0 , для которых сегмент $x^i \times [t^0, t^1)$ целиком принадлежит $G \cup S_0$ (это множество обозначим через S_{0h}); узлы $(0, 0)$, $(d, 0)$ считаем принадлежащими множеству S_{0h} ; полагаем $S_h = S_{0h} \cup S_h^L$.

Задачу (2.2), (2.1) аппроксимируем неявной разностной схемой

$$\begin{aligned} \Lambda z(x, t) \equiv \{\varepsilon \delta_{\bar{x}\bar{x}} - \delta_{\bar{t}}\} z(x, t) &= f(x, t), \quad (x, t) \in G_h, \\ z(x, t) &= \varphi(x, t), \quad (x, t) \in S_h. \end{aligned} \quad (4.4)$$

Здесь $\delta_{\bar{x}\bar{x}} z(x, t)$, $\delta_{\bar{t}} z(x, t)$ – вторая и первая разностные производные;

$$\delta_{\bar{x}\bar{x}} z(x, t) = 2(h^i + h^{i-1})^{-1} \{\delta_x - \delta_{\bar{x}}\} z(x, t),$$

$x = x^i$, h^{i-1} и h^i – левое и правое "плечи" трехточечного шаблона на G_h (оператора $\delta_{\bar{x}\bar{x}}$) с центром в узле $(x^i, t^j) \in G_h$.

Для разностной схемы (4.4), (4.3) справедлив принцип максимума (см. [16]).

Для решения разностной схемы (4.4), (4.3) имеем оценку

$$|u(x, t) - z(x, t)| \leq M(\varepsilon + N^{-1} + N_0^{-1})^{-1} [N^{-1} + N_0^{-1}], \quad (x, t) \in \bar{G}_h, \quad (4.5)$$

где $\bar{G}_h = \bar{G}_{h(4.3)}$. На сетке (порождаемой сеткой $\bar{G}_{h(4.2)}$)

$$\bar{G}_h = \bar{G}_h(\bar{G}_{h(4.2)}^b), \quad (4.6)$$

т.е. на сетке $\bar{G}_{h(4.3)}(\bar{G}_{h(4.1)}^b)$, где $\bar{G}_{h(4.1)}^b$ есть $\bar{G}_{h(4.2)}^b$, получается оценка

$$|u(x, t) - z(x, t)| \leq M[(\varepsilon + N^{-1})^{-2} N^{-2} + (\varepsilon + N_0^{-1})^{-1} N_0^{-1}], \quad (x, t) \in \bar{G}_h, \quad (4.7)$$

неулучшаемая по вхождению величин N, N_0, ε .

Схемы (4.4), (4.3) и (4.4), (4.6) сходятся при неулучшаемом условии $(N^{-1}, N_0^{-1} \ll \varepsilon)$

$$\varepsilon^{-1} = o(\min[N, N_0]), \quad N, N_0 \longrightarrow \infty; \quad (4.8)$$

дефект сходимости схемы (4.4) на сетках (4.3), (4.6) равен единице.

На основе сетки $\bar{G}_h$ построим триангуляцию области $\bar{G}$ и по значениям $z(x, t)$, $(x, t) \in \bar{G}_h$ построим интерполянт $\bar{z}(x, t)$, $(x, t) \in \bar{G}$. Область $\bar{G}$ покроем элементарными прямоугольниками, неправильными четырехугольниками (имеющими стороны, не параллельные осям координат) и треугольниками; неправильные четырехугольники и треугольники имеют некоторыми вершинами узлы, принадлежащие множеству S_h , причем одна из сторон принадлежит множеству S^L . Неправильные четырехугольники разобьем на прямоугольники и неправильные треугольники. На прямых $t = t^j$, $t^j \in \bar{\omega}_0$, по значениям $z(x, t)$, $(x, t) \in \bar{G}_h$, построим линейные (по x) интерполянты $\tilde{z}(x, t)$, $(x, t) \in \bar{G}$, $t \in \bar{\omega}_0$. Все прямоугольники разобьем диагональю на треугольные элементы. Эти правильные и неправильные треугольные элементы образуют триангуляцию множества $\bar{G}$. На треугольных элементах построим линейные интерполянты по значениям $\tilde{z}(x, t)$ в вершинах треугольных элементов на множествах $t = t^j$, $t^j \in \bar{\omega}_0$. Интерполянт $\bar{z}(x, t)$, $(x, t) \in \bar{G}$, построен (см., например, [17]).

В случае разностной схемы (4.4), (4.3) для функции $\bar{z}(x, t)$, $(x, t) \in \bar{G}$, выполняется оценка

$$|u(x, t) - \bar{z}(x, t)| \leq M(\varepsilon + N^{-1} + N_0^{-1})^{-1} [N^{-1} + N_0^{-1}], \quad (x, t) \in \bar{G}. \quad (4.9)$$

В случае схемы (4.4), (4.6) имеем оценку

$$|u(x, t) - \bar{z}(x, t)| \leq M[(\varepsilon + N^{-1})^{-2} N^{-2} + (\varepsilon + N_0^{-1})^{-1} N_0^{-1}], \quad (x, t) \in \bar{G}, \quad (4.10)$$

неулучшаемую по вхождению величин N , N_0 , ε .

Оптимальная по порядку относительно числа узлов сетки $\bar{G}_h$ скорость сходимости схемы на сетке (4.6) получается при условии $N^2 \approx \varepsilon^{-1} N_0$; имеем неулучшаемую оценку

$$|u(x, t) - \bar{z}(x, t)| \leq M(\varepsilon^{4/3} + P_0^{-2/3})^{-1} P_0^{-2/3}, \quad (x, t) \in \bar{G}, \quad (4.11)$$

где P_0 – число узлов сетки $\bar{G}_h$; $P_0 \approx NN_0$.

Оптимальная по порядку скорость сходимости схемы (4.3) получается при условии $N \approx N_0$; для такой схемы выполняется неулучшаемая оценка

$$|u(x, t) - \bar{z}(x, t)| \leq M(\varepsilon + P_0^{-1/2})^{-1} P_0^{-1/2}, \quad (x, t) \in \bar{G}, \quad (4.12)$$

более слабая по сравнению с оценкой (4.11).

Условие $(P_0^{-1/2} \ll \varepsilon)$

$$\varepsilon^{-1} = o(P_0^{1/2}), \quad P_0 \rightarrow \infty, \quad (4.13)$$

является необходимым и достаточным для сходимости разностной схемы (4.4) на сетках (4.3), (4.6) при оптимальной по порядку скорости сходимости схемы; дефект сходимости этих схем равен единице.

Теорема 2. Пусть для решения краевой задачи (2.2), (2.1) выполняются априорные оценки (3.6), (3.7) при $K = 4$. Тогда условие (4.8) (условие (4.13)) является необходимым и достаточным для сходимости разностной схемы (4.4) на сетках (4.3), (4.6) (на сетках (4.3), (4.6) при оптимальной по порядку скорости сходимости схемы). Для сеточных решений справедливы оценки (4.5), (4.7), (4.9)–(4.12).

5. О ПОСТРОЕНИИ ПОЧТИ ε -РАВНОМЕРНО СХОДЯЩИХСЯ СХЕМ НА ЛОКАЛЬНО СГУЩАЮЩИХСЯ СЕТКАХ

Заметим, что особенность решения краевой задачи (2.2), (2.1) – движущийся пограничный слой, экспоненциально убывающий при удалении от множества S_1^L (см. оценку (3.7)). Сингулярная компонента при условии $r((x, t), S_1^L) \geq \sigma$, где $r((x, t), S_1^L)$ – расстояние от точки (x, t) до множества S_1^L , не превосходит величины $M\delta$, где δ – достаточно малое число, в том случае, когда $\sigma = m_1^{-1} \varepsilon \ln \delta^{-1}$, m_1 –

произвольное число из интервала $(0, m^0)$, $m^0 = m_{(3.7)}^0$. Невязка разностной схемы на решении краевой задачи велика, однако лишь на этой σ -окрестности множества S_1^L , являющейся достаточно узкой при малых значениях параметра ε .

5.1. Имея в виду, что при аппроксимации краевой задачи возможно использование схем на достаточно произвольных локально сгущающихся сетках, будет удобно ввести сбалансированные сетки – сетки с произвольным распределением узлов, определяемые лишь величиной P_0 – числом узлов сетки на множестве $\bar{G}$. В случае сеток (4.3), (4.6) число P_0 порядка NN_0 . Сбалансированные сетки на $\bar{G}$, вообще говоря, не являются прямым произведением сеток по x и t . Для сбалансированных сеток подстраивание ориентации шаблонов (положения узлов шаблонов) сеточных уравнений к границе S_1^L , вообще говоря, не предполагается.

Величина P_0 определяет объем вычислительной работы для решения краевой задачи в случае экономичных разностных схем (схем, для которых число вычислительных операций, требующихся для решения, порядка узлов сетки, (см. [16])). Величина $P_0^{-1/2}$ есть характерный размер диаметров элементов разбиения при изотропной триангуляции множества $\bar{G}$.

5.2. Рассмотрим класс разностных схем на основе классических аппроксимаций краевой задачи (2.2), (2.1) в случае локально равномерных сеток, т.е. сеток, являющихся прямым произведением сеток, равномерных по x и t , на подобластях из $\bar{G}$.

Рассматриваем сетки, сгущающиеся в окрестности границы S_1^L , являющиеся прямым произведением равномерных сеток по x и t как в ближайшей окрестности границы S_1^L , так и вне несколько большей ее окрестности.

5.2.1. Для простоты полагаем $\beta_1(t) = t$, $\beta_2(t) = d + t$. Пусть на $\bar{G}$ каким-то образом построена сетка

$$\bar{G}_h^* = \bar{G}_h^*(\rho_1), \quad (5.1a)$$

где $\rho_1 > 0$ – параметр, характеризующий распределение узлов сетки (5.1); этот параметр выбирается ниже; пусть P_0 – число узлов сетки $\bar{G}_h^*$. Эта сетка является прямым произведением равномерных сеток по x и t на множествах

$$G_1 = G_1(\rho_1), \quad G_2 = G \setminus \bar{G}_1(M\rho_1), \quad (5.1b)$$

где

$$G_1(\rho_1) = \{(x, t): x \in (\beta_1(t), \beta_1(t) + \rho_1), t \in (0, T)\}$$

есть правая ρ_1 -окрестность множества S_1^L ; $M\rho_1 \leq 2^{-1}d$. Пусть сетки

$$G_{ih} = G_i \cap \bar{G}_{h(5.1)}^*, \quad i = 1, 2, \quad (5.1в)$$

имеют шаги h_i и h_{it} по x и t соответственно, $i = 1, 2$. Считаем для простоты, что шаблоны четырехточечных неявных схем, имеющие центром узлы из G_{ih} , правильные, т.е. их левые, правые и “нижние” плечи равны h_i и h_{it} соответственно.

Рассмотрим фрагменты сеточной задачи из класса разностных схем на сетках (5.1), аппроксимирующих краевую задачу (2.2), (2.1), а именно фрагменты на множествах $\bar{G}_{1h}$ и $\bar{G}_{2h}$. Пусть $z_i(x, t)$, $(x, t) \in \bar{G}_{ih}$, – решение сеточной задачи

$$\begin{aligned} \Lambda_{(4.4)} z_i(x, t) &= f(x, t), \quad (x, t) \in G_{ih}, \\ z_i(x, t) &= u(x, t), \quad (x, t) \in S_{ih}, \quad i = 1, 2, \end{aligned} \quad (5.2)$$

где $u(x, t)$, $(x, t) \in \bar{G}$, – решение задачи (2.2), (2.1).

Для функций $z_i(x, t)$, $(x, t) \in \bar{G}_{ih}$ (в силу оценок теоремы 1), имеем оценки

$$|u(x, t) - z_1(x, t)| \leq M[(\varepsilon + h_1)^{-2}h_1^2 + (\varepsilon + h_{1t})^{-1}h_{1t}], \quad (x, t) \in \bar{G}_{1h}, \quad (5.3a)$$

$$|u(x, t) - z_2(x, t)| \leq M \left\{ [(\varepsilon + h_2)^{-2}h_2^2 + (\varepsilon + h_{2t})^{-1}h_{2t}] \max_{S_{2h}} |W(x, t)| + h_2^2 + h_{2t}^2 \right\}, \quad (x, t) \in \bar{G}_{2h}, \quad (5.3b)$$

где $W(x, t)$ – сингулярная компонента решения краевой задачи из представления (3.1); главный член компоненты $W(x, t)$ есть функция $W_{0(3.8)}(x, t)$. Оценки (5.3a) и (5.3b) неуклучшаемы по вхождению величин h_1, h_{1t}, ε и h_2, h_{2t}, ε соответственно.

Для того чтобы функция $z_2(x, t)$ сходилась ε -равномерно, необходимо, чтобы для величины ρ_1 выполнялось условие ($\rho_1 \gg \varepsilon$)

$$\rho_1^{-1} = o(\varepsilon^{-1}), \quad \varepsilon \rightarrow 0. \quad (5.4)$$

Оценка для функции $z_1(x, t)$, по порядку скорости сходимости одинаковая с оптимальной относительно h_1, h_{1t} при фиксированном и равном P_1 числе узлов сетки $\bar{G}_{1h}$, получается при условии $(\varepsilon + h_{1t})^{-1}h_{1t} \approx (\varepsilon + h_1)^{-2}h_1^2$:

$$|u(x, t) - z_1(x, t)| \leq M\varepsilon^{-4/3} \rho_1^{2/3} P_1^{-2/3} [1 + \varepsilon^{-4/3} \rho_1^{2/3} P_1^{-2/3}]^{-1}, \quad (x, t) \in \bar{G}_{1h}; \quad (5.5)$$

оценка неуклучшаема по вхождению величин ρ_1, P_1, ε . Из оценки (5.5) при условии (5.4) следует, что функция $z_1(x, t)$ не сходится ε -равномерно при $P_1 \rightarrow \infty$.

Таким образом, не существует локально равномерных сеток $\bar{G}_{h(5.1)}^*$, на которых решения задач (5.2) при $i = 1, 2$ сходятся ε -равномерно, как и почти ε -равномерно, к решению задачи (2.2), (2.1) при $P_0 \rightarrow \infty$. Рассматривая фрагменты задач (5.2) на сетках (5.1), подобным образом убеждаемся, что не существует сеток, на которых решения задач (5.2) сходятся уже при условии

$$P_0^{-1} \geq \varepsilon, \quad P_0 \rightarrow \infty, \quad \varepsilon \in (0, 1]. \quad (5.6)$$

5.2.2. Отсюда вытекает, что в случае семейства сеток (5.1), как и семейства локально равномерных сеток

$$\bar{G}_h^*, \quad (5.7)$$

являющихся равномерными (прямым произведением равномерных сеток по x и t) в правой ρ -окрестности множества S_1^L , где $m\varepsilon \leq \rho \leq M\varepsilon$, и сеточных аппроксимаций задачи, являющихся классическими в ρ -окрестности множества S_1^L , не существует схем, сходящихся при условии (5.6).

Теорема 3. Для краевой задачи (2.2), (2.1) в классе разностных схем на локально равномерных сбалансированных сетках (5.7) в том случае, когда сеточные аппроксимации краевой задачи являются классическими на тех подобластях, на которых сетки (5.7) равномерны, не существует схем, сходящихся при условии (5.6).

Замечание 1. В силу соотношения (4.8), дефект сходимости разностной схемы (4.4) на сетках (4.3), (4.6) равен единице. Из соотношения (5.6) следует, что дефект сходимости схемы (относительно величин P_0, ε) в том случае, когда сеточные аппроксимации являются классическими на тех подобластях, на которых сетки равномерны, не ниже, чем 2^{-1} . Из приведенных построений вытекает, что в случае задачи (2.2), (2.1) использование локально сгущающихся сеток (получаемых, например, последовательным измельчением шагов сеток по x и t в подобластях, где ошибки решения достаточно большие) не позволяет существенно ослабить условие (4.8) – условие сходимости классических разностных схем. На достаточно общих локально сгущающихся сетках невозможно понизить дефект сходимости (относительно величин N, N_0, ε) в условии (4.8) больше, чем в 2 раза. Это утверждение согласуется с результатами п. 6.3 (имея в виду соотношение $P_{(6.5)} = NN_0$). Непосредственное использование техники априорно, либо апостериорно адаптирующихся сеток, не учитывающей ориентацию движущегося пограничного слоя, не позволяет существенно улучшить дефект сходимости адаптирующейся схемы по сравнению с классической схемой (4.4) на сетках (4.3) (4.6).

6. О ПОСТРОЕНИИ ε -РАВНОМЕРНЫХ И ПОЧТИ ε -РАВНОМЕРНЫХ АППРОКСИМАЦИЙ РЕШЕНИЙ ЗАДАЧИ (2.2), (2.1)

Рассмотрим некоторые специфические особенности, связанные с триангуляцией области $\bar{G}$, которые возникают при построении ε -равномерных аппроксимаций решений сингулярно возмущенной задачи (2.2), (2.1). При рассмотрении аппроксимаций решений используем аналог поперечников по Колмогорову (см., например, [2], [18] и библиографию там).

6.1. Нас интересует аппроксимация множества $\mathcal{U}$ – множества решений класса краевых задач (2.2), (2.1) (определяемого условиями (2.3)) в пространстве X – множества непрерывных функций с равномерной нормой. Решения задач считаем достаточно гладкими на $\bar{G}$ при фиксированных значениях параметра ε ; для решений и их компонент из представления (3.1) выполняются оценки (3.6), (3.7).

Опишем аппроксимации решений. Пусть $\bar{G}^h$ – множество точек (будем говорить – *сетка*) на $\bar{G}$. Сетки $\bar{G}^h$ могут быть как структурными (порождаемыми некоторым регулярным семейством линий), так и неструктурными. Число узлов сетки $\bar{G}^h$ на $\bar{G}$ обозначим через P ; $\bar{G}^h = \bar{G}^h(P)$. Пусть T_P есть *триангуляция* (разбиение) множества $\bar{G}$, *порождаемая сеткой* $\bar{G}^h$ (см., например, [17]); считаем, что узлы сетки $\bar{G}^h$ являются вершинами треугольных элементов, причем треугольные элементы образованы отрезками прямых, проходящих через узлы $\bar{G}^h$, если хотя бы один из концов стороны треугольника принадлежит $\bar{G}^h$, либо отрезками кривых, если оба конца стороны принадлежат границе $\bar{S}^L$; $G^h = \bar{G}^h \cap G$. На множестве $\bar{G}^h$ пусть определена некоторая сеточная функция $u^h(x, t)$, $(x, t) \in \bar{G}^h$; через $\bar{u}^h(x, t)$, $(x, t) \in \bar{G}$, обозначим ее линейный интерполюнт, строящийся по значениям функции $u^h(x, t)$ в вершинах треугольных элементов. Совокупность таких интерполянтов при фиксированной триангуляции T_P обозначим через U_P^h ; совокупность всевозможных допустимых сеточных множеств $\bar{G}^h$ (однако с числом узлов, равным P на $\bar{G}$) и триангуляций T_P на их основе обозначим через $\mathcal{T}_P$ (будем говорить, что $\mathcal{T}_P$ есть *множество разбиений* области $\bar{G}$). Классы допустимых сеточных множеств и триангуляций на них уточняются ниже. Это множество разбиений $\mathcal{T}_P$ и совокупность интерполянтов U_P^h (для каждой триангуляции из $\mathcal{T}_P$) определяют *пространство* X . Определим величину $d_P(\mathcal{U}, X)$ (*поперечник*) соотношением

$$d_P(\mathcal{U}, X) = \inf_{\mathcal{T}_P} \sup_{u \in \mathcal{U}} \inf_{\bar{u}^h \in U_P^h} \|u - \bar{u}^h\|, \quad (6.1)$$

где $\|\cdot\|$ – норма в C (определение поперечника по Колмогорову см., например, в [18, гл. 3]). Величина $d_P(\mathcal{U}, X)$ есть погрешность при оптимальной точности приближения множества $\mathcal{U}$ в пространстве X (на сетке с P узлами), или, короче, *погрешность оптимальной аппроксимации*.

Определения. 5. Пусть

$$d_P^i(\mathcal{U}, X) = d_P(\mathcal{U}, X; \bar{G}_i^h), \quad i = 1, 2,$$

суть поперечники, порождаемые двумя семействами сеток $\bar{G}_i^h = \bar{G}_i^h(P)$, $i = 1, 2$. В том случае, когда для поперечников $d_P^i(\mathcal{U}, X)$ выполняется оценка равномерно относительно P

$$md_P^1(\mathcal{U}, X) \leq d_P^2(\mathcal{U}, X) \leq Md_P^1(\mathcal{U}, X), \quad P \geq M_1,$$

где M_1 может быть достаточно большим, будем говорить, что поперечники $d_P^1(\mathcal{U}, X)$ и $d_P^2(\mathcal{U}, X)$ *эквивалентны*.

6. Через $\rho_1(T_P^j)$ и $\rho_2(T_P^j)$, где T_P^j – треугольный элемент из разбиения T_P , обозначим радиусы вписанной и описанной окружностей для элемента T_P^j , $j = 1, 2, \dots, J$, $J = J(P)$ – число треугольных

элементов разбиения T_P (считаем выполненным условие $J \approx P$). Триангуляцию T_P назовем *изотропной*, если выполняется условие

$$\rho_1^{-1}(T_P^j) \rho_2(T_P^j) \leq M, \quad j = 1, 2, \dots, J,$$

однако величины $\rho_1^{-1}(T_P^j)$, $\rho_2(T_P^j)$ и величина $\eta(T_P^j) = \rho_1^{-1}(T_P^j) \rho_2(T_P^j)$ – коэффициент анизотропии для элемента T_P^j – могут сильно отличаться от элемента к элементу. Триангуляцию T_P назовем *анизотропной* (с коэффициентом анизотропии разбиения $\eta \geq M_0$, где M_0 может быть достаточно большим), если выполняется условие

$$\eta \equiv \sup_j \eta(T_P^j) \geq M_0, \quad j = 1, 2, \dots, J,$$

постоянная M_0 не зависит от параметров ε, P . Считаем, что совокупность $\mathcal{T}_P$ определяется величиной η ; таким образом, $T_P = T_P(\eta)$, $\mathcal{T}_P = \mathcal{T}_P(\eta)$, $d_P(\mathcal{U}, X) = d_P(\mathcal{U}, X; \eta)$.

7. Подобным образом вводится изотропная и анизотропная *триангуляции на подмножествах* из $\bar{G}$. В том случае, когда поперечники рассматриваются на подмножестве $\bar{G}^0 \subset \bar{G}$ (в этом случае поперечник обозначаем через $d_P(\mathcal{U}, X; \bar{G}^0)$), величина $\|u - \bar{u}^h\|$ в (6.1) вычисляется лишь по треугольным элементам, целиком принадлежащим $\bar{G}^0$.

8. Элементы триангуляции, для которых выполняется условие

$$\eta(T_P^j) \longrightarrow \infty, \quad \text{например при } P \longrightarrow \infty \text{ и/или } \varepsilon \longrightarrow 0,$$

назовем *существенно анизотропными*; триангуляцию T_P , имеющую такие элементы, назовем *существенно анизотропной*.

Величина $d_P(\mathcal{U}, X)$ стремится к нулю при $P \longrightarrow \infty$, однако эта сходимость к нулю не является ε -равномерной. Вообще говоря, приближения сходятся при $P \longrightarrow \infty$ лишь при некоторых соотношениях между величинами P и ε .

9. Пусть для поперечника $d_P(\mathcal{U}, X)$ выполняется оценка

$$d_P(\mathcal{U}, X) \leq M \lambda(\varepsilon^{-\nu} P^{-1/2}), \quad P \longrightarrow \infty, \quad \varepsilon \in (0, 1]; \quad (6.2)$$

$\lambda(\xi) \longrightarrow 0$ при $\xi \longrightarrow 0$ равномерно относительно параметра ε , $\nu \geq 0$; постоянная M , вообще говоря, зависит от ν . При $\nu = 0$ будем говорить, что поперечник (погрешность оптимальной аппроксимации) *сходится ε -равномерно*. При $\nu > 0$ будем говорить, что поперечник (погрешность оптимальной аппроксимации) *сходится с дефектом* (ε -равномерной сходимости) ν (относительно величин P, ε). В том случае, когда величина $\nu > 0$ может быть выбрана сколь угодно малой, причем для поперечника на триангуляциях T_P , контролируемых величиной ν , выполняется оценка (6.2), будем говорить, что поперечник (погрешность оптимальной аппроксимации) *сходится почти ε -равномерно с дефектом ν* (или, короче, *сходится почти ε -равномерно*).

Наибольший интерес представляют приближения множества $\mathcal{U}$, сходящиеся с возможно меньшим дефектом и, в частности, сходящиеся ε -равномерно.

6.2. Рассмотрим оценку поперечника в том случае, когда движущаяся граница S_1^L является отрезком прямой

$$\beta_1(t) = \nu_1 t, \quad t \in [0, T], \quad \nu_1 \in [\nu_0, \nu^0], \quad (6.3)$$

где $\nu_0 = \nu_{0(2.3)}$, $\nu^0 = \nu_{(2.3)}^0$.

Рассмотрение модельных задач, подобное приведенному в разд. 5, показывает, что в случае изотропных разбиений области $\bar{G}$ для поперечника выполняется оценка снизу

$$d_P(\mathcal{U}, X) \geq m(1 + \varepsilon P)^{-1}. \quad (6.4)$$

На множестве

$$\bar{G}^0 = \bar{G}_{1(5.1)}(\rho_1), \quad \rho_1 = M\varepsilon, \quad (6.5)$$

а именно на правой ρ_1 -окрестности множества S_1^L , получается оценка

$$d_p(\mathcal{U}, X; \bar{G}_{(6.5)}^0) \geq m(1 + \varepsilon P)^{-1}, \quad (6.6)$$

неулучшаемая по вхождению величин P, ε .

В случае анизотропных разбиений с учетом явного вида функции $W_{0(3.8)}(x, t)$, где $\beta_1'(t) = v_1$, находим оценку

$$d_p(\mathcal{U}, X) \geq m \min[(\varepsilon\eta + (\varepsilon\eta)^{-1})P^{-1}; 1]. \quad (6.7)$$

На множестве $\bar{G}_{(6.5)}^0$ имеем оценку

$$d_p(\mathcal{U}, X; \bar{G}^0) \geq m \min[(\varepsilon\eta + (\varepsilon\eta)^{-1})P^{-1}; 1]; \quad (6.8)$$

оценка неулучшаема по вхождению величин P, ε, η .

Теорема 4. Пусть для компонент решения краевой задачи (2.2), (2.1) из представления (3.1) выполняются оценки (3.6), (3.7), (3.9), где $K = 2$. Тогда в случае условия (6.3) для поперечников $d_p(\mathcal{U}, X)$, $d_p(\mathcal{U}, X; \bar{G}^0)$ на изотропной триангуляции T_P (анизотропной триангуляции T_P) выполняются оценки (6.4), (6.6) (оценки (6.7), (6.8)); оценки (6.6) и (6.8) неулучшаемы по вхождению величин P, ε и P, ε, η соответственно.

6.3. Пусть выполняется условие (6.3). В силу оценки (6.6), а также оценки (6.8) при условии

$$\eta \geq M, \quad \text{величина } \eta \text{ не зависит от } P \text{ и } \varepsilon \quad (6.9)$$

дефект сходимости погрешности оптимальной аппроксимации на множестве $\bar{G}_{(6.5)}^0$ равен 2^{-1} . Таким образом, дефект сходимости поперечников $d_p(\mathcal{U}, X)$ на изотропной триангуляции, а также анизотропной триангуляции при условии (6.9) не меньше, чем 2^{-1} .

Условие

$$\eta = \eta(\varepsilon, P), \quad \eta = o(\varepsilon^{-1}P), \quad \eta^{-1} = o(\varepsilon P) \quad \text{при } P \rightarrow \infty, \quad \varepsilon \in (0, 1], \quad (6.10)$$

является необходимым для того, чтобы дефект сходимости поперечника на анизотропной триангуляции T_P был меньше, чем 2^{-1} .

Теорема 5. Пусть выполняется условие теоремы 4. В случае условия (6.3) использование существенно анизотропных триангуляций, удовлетворяющих условию (6.10), является необходимым для сходимости поперечника $d_p(\mathcal{U}, X)$ с дефектом меньше чем 2^{-1} , а также для ε -равномерной либо почти ε -равномерной сходимости.

Из условия (6.10) следует, что при

$$\eta = \eta(\varepsilon, P) \equiv \varepsilon^{-1}\eta_0(P) \quad (6.11)$$

дополнительное условие $(\eta_0(P)P, \eta_0^{-1}(P)P \gg 1)$

$$\eta_0(P), \quad \eta_0^{-1}(P) = o(P), \quad P \rightarrow \infty, \quad (6.12)$$

является необходимым для ε -равномерной сходимости поперечника $d_p(\mathcal{U}, X; \bar{G}^0)$ на подмножестве $\bar{G}^0 = \bar{G}_{(6.5)}^0$; при условии (6.11) с учетом априорных оценок теоремы 1 получаем ε -равномерную оценку сверху

$$d_p(\mathcal{U}, X; \bar{G}^0) \leq MP^{-1}[\eta_0(P) + \eta_0^{-1}(P)]; \quad (6.13)$$

оценка неулучшаема по вхождению величины P .

Для поперечника $d_p(\mathcal{U}, X)$ при условии (6.11) получается ε -равномерная оценка

$$d_p(\mathcal{U}, X) \leq M \min\{P^{-1} \ln P[\eta_0(P) + \eta_0^{-1}(P)]; 1\}, \quad (6.14)$$

при дополнительном условии $(P^{-1} \ln P [\eta_0(P) + \eta_0^{-1}(P)] \ll 1)$

$$\eta_0(P), \quad \eta_0^{-1}(P) = o(P \ln^{-1} P), \quad P \rightarrow \infty, \quad (6.15)$$

несколько более сильном, чем условие (6.12), поперечник $d_P(\mathcal{U}, X)$ сходится ε -равномерно.

Теорема 6. Пусть выполняется условие теоремы 4. В случае условия (6.3) условие $\{(6.11), (6.12)\}$ является необходимым, а условие $\{(6.11), (6.15)\}$ – достаточным для ε -равномерной сходимости (при $P \rightarrow \infty$) поперечника $d_P(\mathcal{U}, X)$ при анизотропном разбиении (6.11). Для поперечников $d_P(\mathcal{U}, X; \bar{G}^0)$ и $d_P(\mathcal{U}, X)$ в случае условия (6.11) выполняются оценки (6.13) (неулучшаемая по P) и (6.14) соответственно.

Из неулучшаемой оценки (6.13) и представления (6.11) для величины ν вытекает, что в том случае, когда поперечник на $\bar{G}^0$ сходится с дефектом $\nu < 2^{-1}$, элементы триангуляции на множестве $\bar{G}^0$ являются существенно анизотропными.

6.4. Будем предполагать выполненным условие

$$\text{множество } S_1^L \text{ является отрезком гладкой кривой} \quad (6.16)$$

с ограниченной ненулевой кривизной, и пусть триангуляция T_P является анизотропной, удовлетворяющей условию

$$\text{стороны треугольных элементов триангуляции – отрезки прямых.} \quad (6.17)$$

На триангуляциях $T_P = T_P(\eta)$ с анизотропией η определим поперечник $d_P^*(\mathcal{U}, X)$ соотношением

$$d_P^*(\mathcal{U}, X) = \inf_{\eta} d_P(\mathcal{U}, X; \eta). \quad (6.18)$$

Будем говорить, что поперечник $d_P^*(\mathcal{U}, X)$ является оптимальным относительно анизотропии на триангуляциях $T_P(\eta)$, или, короче, *оптимальным поперечником*.

В случае условий (6.16), (6.17) для поперечника $d_P^*(\mathcal{U}, X)$, рассматриваемого на $\bar{G}_{(6.5)}^0$, с учетом явного вида функции $W_{0(3.8)}(x, t)$ получается неулучшаемая оценка

$$d_P^*(\mathcal{U}, X; \bar{G}^0) \geq m(1 + \varepsilon^{1/2} P)^{-1}. \quad (6.19)$$

Теорема 7. Пусть выполняется условие теоремы 4. Тогда в случае условий (6.16), (6.17) для поперечника $d_{P(6.18)}^*(\mathcal{U}, X; \bar{G}^0)$ выполняется оценка (6.19), неулучшаемая по вхождению величин P, ε .

Замечание 2. В силу (6.19), дефект сходимости поперечника $d_P^*(\mathcal{U}, X; \bar{G}^0)$ в случае криволинейной границы S_1^L есть 4^{-1} . Таким образом, для поперечника $d_P^*(\mathcal{U}, X)$ в случае триангуляций, порождаемых треугольными элементами, удовлетворяющими условию (6.17), и интерполянтов $\bar{u}^h(x, t)$, линейных на элементах T_P^j , дефект сходимости меньше, чем 4^{-1} а в случае условия (6.16) недостижим.

Из приведенных рассмотрений вытекает, что для построения оптимальных аппроксимаций, для поперечников которых дефект сходимости меньше, чем 4^{-1} , необходимо использовать триангуляции T_P с криволинейными треугольными элементами T_P^j и/или интерполянты, строящиеся на T_P^j по значениям функции $u^h(x, t)$, отличные от линейных.

Определения. 10. Треугольные элементы триангуляции T_P назовем *правильными* в том случае, когда стороны треугольных элементов являются отрезками прямых, соединяющих вершины треугольника, если хотя бы одна вершина не принадлежит границе S_1^L , или отрезком границы S_1^L , если обе вершины принадлежат S_1^L . Триангуляцию $\bar{G}$ назовем *триангуляцией на правильных элементах*, если все его треугольные элементы правильные. В этой триангуляции стороны треуголь-

ных элементов “криволинейны” (отрезки границы S_1^L), лишь если оба конца стороны принадлежат границе S_1^L .

11. Треугольные элементы, у которых стороны могут быть отрезками кривых безотносительно к взаимному расположению вершин треугольника и множества S_1^L , отнесем к элементам с криволинейными сторонами (или, короче, к *криволинейным треугольным элементам*). Триангуляцию T_P назовем *триангуляцией на криволинейных элементах*, если среди треугольных элементов есть криволинейные треугольные элементы.

Оценка (6.19) сохраняется и в случае триангуляций на правильных элементах.

Теорема 8. Пусть выполняется условие теоремы 4. В случае условия (6.6) и триангуляций T_P на правильных элементах скорость сходимости поперечника $d_P^*(\mathcal{U}, X)$ с дефектом сходимости меньше, чем 4^{-1} , недостижима.

Замечание 3. Рассмотрение модельных примеров показывает, что в случае условия (6.16) для сходимости поперечника d_P^* (погрешности оптимальной аппроксимации) с дефектом сходимости меньше, чем 4^{-1} , необходимо использовать триангуляции на криволинейных элементах, как и интерполянты, строящиеся по $u^h(x, t)$, отличные от линейных по x и t .

7. АППРОКСИМАЦИИ РЕШЕНИЙ ЗАДАЧИ (2.2), (2.1) В СЛУЧАЕ КРИВОЛИНЕЙНОЙ ГРАНИЦЫ S_1^L

7.1. Обсудим построение триангуляций и интерполянтов, для которых поперечник в случае криволинейной границы S_1^L при $P \rightarrow \infty$ сходится ε -равномерно.

При построении ε -равномерно сходящихся поперечников можно в m -окрестности границы S_1^L перейти к новой системе координат, в которой граница становится отрезком прямой, причем для границы выполняется условие, подобное (6.3), и воспользоваться результатами п. 6.3 (см. теорему 6).

При построении триангуляций T_P на криволинейных элементах T_P^j и подходящих интерполянтов, строящихся на T_P^j по значениям $u^h(x, t)$, удобно использовать такой подход, связанный с заменой координат в задаче (2.2), (2.1). Пусть в новой системе координат $\xi, t, \xi = \xi(x, t)$ граница S_1^L становится отрезком прямой $\tilde{S}_1^L = \{S_1^L\}_\xi$, параллельной оси t (см., например, замену координат (3.2) в разд. 3); множество $\bar{G}$ и решение $u(x, t)$ переходят в множество $\bar{G}^\wedge$ и функцию $u^\wedge(\xi, t)$ соответственно; $\bar{G}^\wedge = \{\bar{G}\}_\xi = \tilde{\bar{G}}, u^\wedge(\xi, t) = \{u(x, t)\}_\xi = \tilde{u}(\xi, t)$.

Пусть

$$d_P(\mathcal{U}^\wedge, X^\wedge) = d_P(\mathcal{U}^\wedge, X^\wedge; \bar{G}^{\wedge h}, \bar{u}^{\wedge h}(\cdot), \eta) \quad (7.1a)$$

есть поперечник, определяемый (подобно поперечнику (6.1)) множеством $\mathcal{U}^\wedge$ решений краевой задачи (3.3), (3.4) и такими компонентами: сеточными множествами $\bar{G}^{\wedge h} \subset \bar{G}^\wedge$, триангуляциями $T_P^\wedge$, строящимися на основе этих множеств, и интерполянтами $\bar{u}^{\wedge h}(\xi, t), (\xi, t) \in T_P^{\wedge j}$, линейными по ξ, t на элементах триангуляции. Эти компоненты определяют множество разбиений $\mathcal{T}_P^\wedge$ и пространство $X^\wedge$. Поперечник $d_P^*(\mathcal{U}^\wedge, X^\wedge)$, где

$$d_P(\mathcal{U}^\wedge, X^\wedge) = d_P^*(\mathcal{U}^\wedge, X^\wedge; \bar{G}^{\wedge h}, \bar{u}^{\wedge h}(\cdot)) \equiv \inf_{\eta} d_P(\mathcal{U}^\wedge, X^\wedge; \bar{G}^{\wedge h}, \bar{u}^{\wedge h}(\cdot), \eta), \quad (7.1b)$$

сходится ε -равномерно.

Пусть поперечник

$$d_P^*(\mathcal{U}, X_{\{\xi^{-1}\}}^\wedge) = d_P^*(\mathcal{U}, X_{\{\xi^{-1}\}}^\wedge; \{\bar{G}^{\wedge h}\}_{\xi^{-1}}, \{\bar{u}^{\wedge h}(\cdot)\}_{\xi^{-1}}) \quad (7.2)$$

определяется множеством $\mathcal{U} = \mathcal{U}_{(6.1)}$ решений краевой задачи (2.2), (2.1) и такими компонентами: сеточными множествами $\{\bar{G}^{\wedge h}\}_{\xi^{-1}} \subset \bar{G}$, триангуляциями на криволинейных элементах $\{T_P^{\wedge}\}_{\xi^{-1}}$ и интерполянтами $\{\bar{u}^{\wedge h}(\xi, t)\}_{\xi^{-1}}$ на элементах триангуляции $\{T_P^{\wedge j}\}_{\xi^{-1}}$ (вообще говоря, не являющимися линейными по x, t). Эти компоненты определяют множество разбиений $\mathcal{T}_{P\{\xi^{-1}\}}^{\wedge}$ и пространство $X_{\{\xi^{-1}\}}^{\wedge}$. Поперечник $d_P^*(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge})$ сходится ε -равномерно.

Пусть поперечник

$$d_P^{1*}(\mathcal{U}^{\wedge}, X^{\wedge}) \quad (7.3a)$$

есть поперечник (7.1б), где множество $\bar{G}^{\wedge h}$ – сетка из $\bar{G}^{\wedge}$:

$$\bar{G}_h^{\wedge} = \bar{\omega}_1^{\wedge} \times \bar{\omega}_0, \quad (7.3б)$$

т.е. прямоугольная сетка на $\bar{G}^{\wedge}$ с произвольным распределением узлов сеток $\bar{\omega}_1^{\wedge}$ и $\bar{\omega}_1$ по ξ и t соответственно.

Пусть

$$d_P^{1*}(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge}) \quad (7.4a)$$

есть поперечник (7.2), где множество $\{\bar{G}^{\wedge h}\}_{\xi^{-1}}$ – сетка из $\bar{G}$:

$$\{\bar{G}_h^{\wedge}\}_{\xi^{-1}} = \{\bar{G}_{h(7.3)}^{\wedge}\}_{\xi^{-1}}. \quad (7.4б)$$

Теорема 9. Пусть выполняется условие теоремы 4. Тогда поперечники $d_{P(7.1)}^*(\mathcal{U}^{\wedge}, X^{\wedge})$ и $d_{P(7.2)}^*(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge})$ эквивалентны поперечникам $d_{P(7.3)}^{1*}(\mathcal{U}^{\wedge}, X^{\wedge})$ и $d_{P(7.4)}^{1*}(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge})$ соответственно.

Замечание 4. Рассмотренный подход позволяет на основе поперечников $d_{P(7.3)}^{1*}(\mathcal{U}^{\wedge}, X^{\wedge})$, строящихся на основе прямоугольных сеток для краевой задачи (3.3), (3.4), построить поперечники $d_{P(7.4)}^{1*}(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge})$ для краевой задачи (2.2), (2.1), сходящиеся ε -равномерно. Равномерная по параметру ε сходимость поперечников $d_P^{1*}(\mathcal{U}^{\wedge}, X^{\wedge}; \bar{G}_{h(7.3)}^{\wedge}, \bar{u}^{\wedge h}(\cdot))$ и $d_P^{1*}(\mathcal{U}, X_{\{\xi^{-1}\}}^{\wedge}; \{\bar{G}_{h(7.3)}^{\wedge}\}_{\xi^{-1}}, \{\bar{u}^{\wedge h}(\cdot)\}_{\xi^{-1}})$ является необходимой для сходимости решений классических разностных схем, строящихся в переменных ξ, t , а также решений этих схем, записанных в переменных x, t , к решениям задач (3.3), (3.4) и (2.2), (2.1) соответственно.

7.2. Рассмотрим поперечник $d_{P(7.3)}^{1*}$ в случае сеток $\bar{G}_{h(7.3)}^{\wedge}$, являющихся кусочно-равномерными по ξ :

$$\bar{G}_h^{\wedge} = \bar{G}_h^{\wedge S} \equiv \bar{\omega}_1^{\wedge S} \times \bar{\omega}_0, \quad (7.5)$$

где $\bar{\omega}_1^{\wedge S}$ – кусочно-равномерная сетка, равномерная на множествах $[0, \sigma^{\wedge}]$ и $[\sigma^{\wedge}, d]$. В этом случае для поперечников получается оценка

$$d_P^{1*}(\mathcal{U}^{\wedge}, X^{\wedge}; \bar{G}_{h(7.5)}^{\wedge}) \leq \begin{cases} MP^{-1} \ln P & \text{при } \varepsilon \ln P \leq m_1, \\ M\varepsilon^{-1} P^{-1} & \text{при } \varepsilon \ln P > m_1, \end{cases} \quad (7.6a)$$

неулучшаемая по P, ε , и ε -равномерная (неулучшаемая) оценка

$$d_P^{1*}(\mathcal{U}^{\wedge}, X^{\wedge}; \bar{G}_{h(7.5)}^{\wedge}) \leq MP^{-1} \ln P, \quad (7.6б)$$

т.е. поперечник сходится ε -равномерно.

Для параметров сетки (7.5), на которой реализуется поперечник (будем говорить: *оптимальной сетки*), выполняются оценки

$$\begin{aligned} N^\wedge &\leq \begin{cases} MP^{1/2} \ln^{1/2} P & \text{при } \varepsilon \ln P \leq m_1, \\ M\varepsilon^{-1/2} P^{1/2} & \text{при } \varepsilon \ln P > m_1, \end{cases} \\ N_0^\wedge &\leq \begin{cases} MP^{1/2} \ln^{-1/2} P & \text{при } \varepsilon \ln P \leq m_1, \\ M\varepsilon^{1/2} P^{1/2} & \text{при } \varepsilon \ln P > m_1, \end{cases} \\ \eta(\varepsilon, P) &\leq M\varepsilon^{-1}, \end{aligned} \quad (7.7a)$$

где $N^\wedge + 1$ и $N_0^\wedge + 1$ – число узлов сеток $\bar{\omega}_1^\wedge$ и $\bar{\omega}_0$ соответственно. Для величины $N_\sigma^\wedge$, где $N_\sigma^\wedge + 1$ – число узлов на множестве $[\sigma^\wedge, d]$, выполняется оценка

$$N_\sigma^\wedge \leq \begin{cases} MP^{1/2} \ln^{-1/2} P & \text{при } \varepsilon \ln P \leq m_1, \\ M\varepsilon^{1/2} P^{1/2} & \text{при } \varepsilon \ln P > m_1, \end{cases} \quad (7.7б)$$

причем

$$m^1 \varepsilon \ln P \leq \sigma^\wedge = \sigma^\wedge(\varepsilon, P) \leq M^1 \varepsilon \ln P \quad \text{при } \varepsilon \ln P \leq m_1. \quad (7.7в)$$

Оценки (7.7а, б) неулучшаемы по P, ε .

Для поперечника d_p^{1*} в случае сеток

$$\bar{G}_h^\wedge = \bar{G}_h^{\wedge u} \equiv \bar{\omega}_1^{\wedge u} \times \bar{\omega}_0, \quad (7.8)$$

где $\bar{\omega}_1^{\wedge u}$ – равномерная сетка, получается оценка

$$d_p^{1*}(\mathcal{U}^\wedge, X^\wedge; \bar{G}_{h(7.8)}^\wedge) \leq M\varepsilon^{-1} P^{-1}. \quad (7.9)$$

Для параметров оптимальной сетки выполняются оценки

$$N^\wedge \leq M\varepsilon^{-1/2} P^{1/2}, \quad N_0^\wedge \leq M\varepsilon^{1/2} P^{1/2}, \quad \eta \leq M\varepsilon^{-1}. \quad (7.10)$$

Оценки (7.9), (7.10) неулучшаемы.

Теорема 10. Пусть выполняется условие теоремы 4. Тогда для поперечников $d_p^{1*}(\mathcal{U}^\wedge, X^\wedge)$ на сетках (7.5) и (7.8) выполняются оценки (7.6) и (7.9) соответственно; для параметров оптимальных сеток выполняются оценки (7.7), (7.10).

8. РАЗНОСТНАЯ СХЕМА НА СЕТКЕ, АДАПТИРУЮЩЕЙСЯ В ДВИЖУЩЕМСЯ ПОГРАНИЧНОМ СЛОЕ

Нас интересуют разностные схемы, для которых интерполянты сеточных решений, строящиеся на элементах триангуляции, порождаемых сеточными узлами, сходятся ε -равномерно. Из результатов разд. 6 вытекает, что оптимальные аппроксимации краевой задачи (2.2), (2.1), строящиеся на основе правильных анизотропных триангуляций и линейных интерполянтов, сходятся с дефектом не ниже, чем 4^{-1} , в случае криволинейной границы S_1^L . Отсюда следует, что для решений разностной схемы (4.4) на сетках, порождаемых прямоугольными (базовыми) сетками, дефект сходимости интерполянтов сеточных решений ниже, чем 4^{-1} , оказывается недостижимым. Таким образом, в случае криволинейной границы S_1^L для сходимости интерполянтов сеточных решений с дефектом ниже, чем 4^{-1} (либо ε -равномерно), необходимо использовать сетки, являющиеся согласованными с границей S_1^L , порождающие существенно анизотропные элементы триангуляции.

При построении специальной схемы, сходящейся ε -равномерно, воспользуемся подходом, приведенным выше. При переходе к переменным $\xi, t, \xi = \xi(x, t)$, задача (2.2), (2.1) переходит в задачу (3.3), (3.4) в области с неподвижными боковыми границами. Для этой задачи оптимальные аппроксимации ее решения в случае триангуляций, строящихся на основе кусочно-равномерных сеток по ξ и равномерных по t , и интерполянтов, линейных на элементах триангуляции, сходятся ε -равномерно. Построив ε -равномерно сходящуюся схему для задачи (3.3), (3.4) (в такой “стандартной” разностной схеме линейный интерполянт на основе ее сеточных решений сходится ε -равномерно) и вернувшись к исходным переменным, получим ε -равномерно сходящуюся схему для задачи (2.2), (2.1), интерполянт которой выписывается через сеточные решения, но уже не является линейным в переменных x, t на элементах триангуляции.

В исходных переменных получающиеся сетки уже не будут прямоугольными (распределение узлов сетки согласовано с движением границы S_1^L), что, вообще говоря, влечет некоторые неудобства при построении сеточных областей и численном решении задачи. Однако подобного типа схему можно строить лишь в достаточно малой окрестности границы S_1^L ; вне этой окрестности можно использовать прямоугольные сетки (в исходных переменных) и классические сеточные аппроксимации задачи.

Следуя указанному подходу, от задачи (2.2), (2.1) перейдем к задаче (3.3), (3.4), для которой построим схему на априорно сгущающихся сетках. На множестве $\bar{G}$ введем прямоугольную сетку

$$\bar{G}_h = \bar{\omega}_1 \times \bar{\omega}_0, \quad (8.1)$$

где $\bar{\omega}_1$ и $\bar{\omega}_0$ – сетки на отрезках $\bar{D} = [0, d]$ и $[0, T]$ соответственно; $\bar{\omega}_0 = \bar{\omega}_{0(4.2)}$, $\bar{\omega}_1$ – сетка с произвольным распределением узлов, удовлетворяющим условию $h_\xi \leq MN^{-1}$, где $h_\xi = \max_i h_\xi^i$, $h_\xi^i = \xi^{i+1} - \xi^i$, $\xi^i, \xi^{i+1} \in \bar{\omega}_1$; $N + 1$ – число узлов сетки $\bar{\omega}_1$.

Для решения задачи (3.3), (3.4) используем разностную схему

$$\begin{aligned} \tilde{A}Z(\xi, t) &\equiv \{\varepsilon A(\xi, t)\delta_{\xi\xi} + B(\xi, t)\delta_\xi - \delta_t\}Z(\xi, t) = \tilde{f}(\xi, t), \quad (\xi, t) \in \tilde{G}_h, \\ Z(\xi, t) &= \tilde{\varphi}(\xi, t), \quad (\xi, t) \in \tilde{S}_h, \end{aligned} \quad (8.2)$$

где $\tilde{G}_h = \bar{G} \cap \bar{G}_h$, $\tilde{S}_h = \tilde{S} \cap \bar{G}_h$, $\delta_{\xi\xi}Z(\xi, t)$ и $\delta_\xi Z(\xi, t)$, $\delta_t Z(\xi, t)$ – вторая и первые разностные производные.

На множестве $\bar{G}$ введем сетку, сгущающуюся в окрестности пограничного слоя:

$$\bar{G}_h^S = \bar{\omega}_1^S \times \bar{\omega}_0, \quad (8.3)$$

где $\bar{\omega}_0 = \bar{\omega}_{0(8.1)}$, $\bar{\omega}_1^S = \bar{\omega}_1^S(\sigma)$ – кусочно-равномерная сетка. Шаги сетки $\bar{\omega}_1^S$ постоянны на множествах $[0, \sigma]$ и $[\sigma, d]$ и равны $h_{(1)} = 2\sigma N^{-1}$, $h_{(2)} = 2[d - \sigma]N^{-1}$ соответственно; величина σ выбирается удовлетворяющей условию

$$\sigma = \sigma(\varepsilon, N) = \min[2^{-1}d, M\varepsilon \ln N],$$

где $M = m_{(3.5)}^{-1}$.

Схема (8.2), (8.3) сходится ε -равномерно с оценкой

$$|\tilde{u}(\xi, t) - Z(\xi, t)| \leq M[N^{-1} \ln N + N_0^{-1}], \quad (\xi, t) \in \bar{G}_h.$$

Для интерполянта $\bar{Z}(\xi, t)$, $(\xi, t) \in \bar{G}$, линейного (по ξ, t) на треугольных элементах разбиения, порождаемых сеткой $\bar{G}_h$, выполняется оценка

$$|\tilde{u}(\xi, t) - \bar{Z}(\xi, t)| \leq M[N^{-1} \ln N + N_0^{-1}], \quad (\xi, t) \in \bar{G}.$$

В переменных x, t сетка

$$\bar{G}_{h\xi^{-1}} = \{\bar{G}_{h(8.3)}\}_{\xi^{-1}} \quad (8.4)$$

не является прямым произведением сеток по x и t . Эта сетка равномерна по t и при $t = t^j, t^j \in \bar{\omega}_0$, кусочно-равномерна по x . Переходя к переменным x, t в схеме (8.2), приходим к схеме

$$\begin{aligned} \tilde{\Lambda}_{\xi^{-1}} Z_{\xi^{-1}}(x, t) &\equiv \varepsilon A_{\xi^{-1}}(x, t) \{\delta_{\xi\xi} Z(\xi, t)\}_{\xi^{-1}} + B_{\xi^{-1}}(x, t) \{\delta_{\xi} Z(\xi, t)\}_{\xi^{-1}} - \{\delta_t Z(\xi, t)\}_{\xi^{-1}} = f(x, t), \\ (x, t) &\in \tilde{G}_{h\xi^{-1}}, \\ Z_{\xi^{-1}}(x, t) &= \varphi(x, t), \quad (x, t) \in \tilde{S}_{h\xi^{-1}}. \end{aligned} \quad (8.5)$$

Для функции $Z^*(x, t) = Z_{\xi^{-1}}(x, t), (x, t) \in \bar{G}_h^*$, где $\bar{G}_h^* = \bar{G}_{h\xi^{-1}}$, – решения разностной схемы (8.5), (8.4), выполняется оценка

$$|u(x, t) - Z^*(x, t)| \leq M[N^{-1} \ln N + N_0^{-1}], \quad (x, t) \in \bar{G}_h^*; \quad (8.6a)$$

для интерполянта $\bar{Z}^*(x, t) = \{\bar{Z}(\xi, t)\}_{\xi^{-1}}, (x, t) \in \bar{G}$, получающегося из функции $\bar{Z}(\xi, t)$ переходом к переменным x, t , имеем подобную оценку:

$$|u(x, t) - \bar{Z}^*(x, t)| \leq M[N^{-1} \ln N + N_0^{-1}], \quad (x, t) \in \bar{G}. \quad (8.6b)$$

Теорема 11. Пусть выполняется условие теоремы 2. Тогда разностная схема (8.5), (8.4), аппроксимирующая краевую задачу (2.2), (2.1), сходится ε -равномерно. Для сеточного решения $Z^*(x, t), (x, t) \in \bar{G}_h^*$, и интерполянта $\bar{Z}^*(x, t), (x, t) \in \bar{G}$, справедливы оценки (8.6).

9. ЗАМЕЧАНИЯ И ОБОБЩЕНИЯ

9.1. Подобным образом строятся ε -равномерно сходящиеся разностные схемы для задач с движущимися пограничными слоями в случае параболических уравнений реакции-диффузии с переменными коэффициентами.

9.2. Подход к построению специальных схем на областях с подвижными границами применим и в случае квазилинейных параболических уравнений. На множестве $\bar{G}_{(2.1)}$ рассмотрим краевую задачу для квазилинейного параболического уравнения:

$$\begin{aligned} L_{(2.2)} u(x, t) &= g(x, t, u(x, t)), \quad (x, t) \in G, \\ u(x, t) &= \varphi(x, t), \quad (x, t) \in S. \end{aligned} \quad (9.1)$$

Функция $g(x, t, u)$ достаточно гладкая на множестве $\bar{G} \times R$, причем

$$-M \leq \frac{\partial}{\partial u} g(x, t, u), \quad (x, t, u) \in \bar{G} \times R;$$

коэффициенты уравнения и функция $\varphi(x, t)$ удовлетворяют условиям, указанным в постановке задачи (2.2), (2.1).

Для решения задачи (9.1) справедливы априорные оценки, подобные оценкам в случае задачи (2.2), (2.1). Задаче (9.1) сопоставим разностную схему

$$\begin{aligned} \tilde{\Lambda}_{\xi^{-1}(8.5)} Z_{\xi^{-1}}(x, t) &= g(x, t, Z_{\xi^{-1}}(x, t)), \quad (x, t) \in \tilde{G}_{h\xi^{-1}}, \\ Z_{\xi^{-1}}(x, t) &= \varphi(x, t), \quad (x, t) \in \tilde{S}_{h\xi^{-1}}, \quad \bar{G}_{h\xi^{-1}} = \bar{G}_{h\xi^{-1}(8.4)}. \end{aligned} \quad (9.2)$$

Решение разностной схемы (9.2), (8.4) сходится к решению краевой задачи (9.1), (2.1) ε -равномерно. Для решений этой схемы справедливо утверждение, подобное утверждению теоремы 11.

9.3. Для решения разностной схемы (8.5), (8.4) имеем оценку (см. (8.6))

$$|u(x, t) - Z^*(x, t)| \leq M[N^{-1} \ln N + NP^{-1}], \quad (x, t) \in \bar{G}_h^*,$$

где $P = NN_0$. При условии

$$N \approx N_0 \quad (9.3)$$

получается неулучшаемая ε -равномерная оценка

$$|u(x, t) - Z^*(x, t)| \leq MP^{-1/2} \ln P, \quad (x, t) \in \bar{G}_h^*. \quad (9.4)$$

9.4. Рассмотрим разностную схему (8.5) на сетке, соответствующей оптимальной сетке (7.7) для поперечника (7.3), (7.5). На множестве $\bar{G}$ строим сетку, являющуюся кусочно-равномерной по ξ :

$$\bar{G}_h^S = \bar{G}_h^S \equiv \bar{\omega}_1^S \times \bar{\omega}_0, \quad (9.5a)$$

где $\bar{\omega}_1^S$ и $\bar{\omega}_0$ – кусочно-равномерная и равномерная сетки; $N^\wedge + 1$ и $N_0^\wedge + 1$ – число узлов сеток $\bar{\omega}_1^S$ и $\bar{\omega}_0$ соответственно, причем $N^\wedge N_0^\wedge = P$. Величины $N^\wedge$ и $N_0^\wedge$ определяются соотношениями (см. (7.7a))

$$\begin{aligned} N^\wedge &\leq \begin{cases} P^{1/2} \ln^{1/2} P & \text{при } \varepsilon \ln P \leq m_1, \\ \varepsilon^{-1/2} P^{1/2} & \text{при } \varepsilon \ln P > m_1, \end{cases} \\ N_0^\wedge &\leq \begin{cases} P^{1/2} \ln^{-1/2} P & \text{при } \varepsilon \ln P \leq m_1, \\ \varepsilon^{1/2} P^{1/2} & \text{при } \varepsilon \ln P > m_1; \end{cases} \end{aligned} \quad (9.5b)$$

постоянная m_1 уточняется ниже. Шаги сетки $\bar{\omega}_1^S$ постоянны на множествах $[0, \sigma]$ и $[\sigma, d]$ и равны $h_{(1)} = \sigma N_{(1)}^{-1}$ и $h_{(2)} = (d - \sigma) N_{(2)}^{-1}$ соответственно, где $N_{(1)} + 1$ и $N_{(2)} + 1$ – число узлов сетки на $[0, \sigma]$ и $[\sigma, d]$, $\sigma \in (0, d]$; $h_{(2)} \approx (N_0^\wedge)^{-1}$, причем $N_{(1)} = N^\wedge$ при $\sigma = d$.

Величина σ удовлетворяет условию

$$\sigma = \sigma(\varepsilon, P) = \min[d, M\varepsilon \ln P], \quad (9.5b)$$

где $M = 2^{-1} m_{(3.5)}^{-1}$. Полагаем $m_1 = M^{-1}d$. Сетка $\bar{G}_h^S$ построена. Заметим, что при $\sigma = d$ (т.е. при $\varepsilon \ln P \geq m_1$) сетка $\bar{G}_h^S$ равномерная.

Для решения разностной схемы (8.5) на сетке

$$\bar{G}_{h\xi^{-1}} = \{\bar{G}_{h(9.5)}\}_{\xi^{-1}} \quad (9.6)$$

выполняется ε -равномерная оценка

$$|u(x, t) - Z^*(x, t)| \leq MP^{-1/2} \ln^{1/2} P, \quad (x, t) \in \bar{G}_h^*. \quad (9.7)$$

Из сравнения оценок (9.4) и (9.7) следует, что схема (8.5), (9.6) – схема на основе сетки (9.5), соответствующей оптимальной кусочно-равномерной сетке (7.7) для поперечника (7.3), (7.5), имеет несколько лучшую сходимость по сравнению со схемой (8.5), (8.4), (9.3) – схемой на основе кусочно-равномерной (стандартной) сетки (8.3) при условии (9.3).

СПИСОК ЛИТЕРАТУРЫ

1. Allen D. N., Southwell R. V. Relaxation methods applied to determine the motion, in two dimensions, of viscous fluid past a fixed cylinder // Quart. J. Mech. and Appl. Math. 1955. V. 8. P. 129–145.
2. Бахвалов Н.С. К оптимизации методов решения краевых задач при наличии пограничного слоя // Ж. вычисл. матем. и матем. физ. 1969. Т. 9. № 4. С. 841–859.
3. Ильин А.М. Разностная схема для дифференциального уравнения с малым параметром при старшей производной // Матем. заметки. 1969. Т. 6. Вып. 2. С. 237–248.
4. Дулан Э., Миллер Дж., Шилдерс У. Равномерные численные методы решения задач с пограничным слоем. М.: Мир, 1983.

5. *Шишкин Г.И.* Сеточные аппроксимации сингулярно возмущенных эллиптических и параболических уравнений. Екатеринбург: УрО РАН, 1992.
6. *Miller J.J.H., O'Riordan E., Shishkin G.I.* Fitted numerical methods for singular perturbation problems. Singapore: World Scient., 1996.
7. *Farrell P.A., Hegarty A.F., Miller J.J.H., O'Riordan E., Shishkin G.I.* Robust computational techniques for boundary layers. Boca Raton: Chapman & Hall / CRC Press, Boca Raton, 2000.
8. *Roos H.-G., Stynes M., Tobiska L.* Numerical methods for singularly perturbed differential equations. Convection-diffusion and flow problems. Berlin: Springer, 1996.
9. *Liseikin V. D.* Grid generation methods. Berlin: Springer, 1999.
10. *Shishkin G.I.* On numerical methods on adaptive meshes for a singularly perturbed reaction-diffusion equation with a moving concentrated source // Finite Difference Schemes: Theory and Applic. Proc. Conf. FDS'2000. Vilnius: Inst. Maths & Inf., 2000. P. 205–214.
11. *Шишкин Г.И.* Сеточная аппроксимация сингулярно возмущенного параболического уравнения на составной области в случае сосредоточенного источника на движущейся границе раздела // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 12. С. 1806–1824.
12. *Shishkin G.I., Shishkina L.P., Hemker P.W.* A class of singularly perturbed convection-diffusion problems with a moving interior layer. An a posteriori adaptive mesh technique // Comput. Meth. Appl. Math. 2004. V. 4. № 1. P. 105–127.
13. *Ладыженская О. А., Солонников В. А., Уральцева Н. Н.* Линейные и квазилинейные уравнения параболического типа. М.: Наука, 1967.
14. *Hemker P.W., Shishkin G.I., Shishkina L.P.* High-order time-accuracy schemes for parabolic singular perturbation problems with convection // Rus. J. Numer. Analys. and Math. Modelling. 2002. V. 17. № 1. P. 1–24.
15. *Hemker P.W., Shishkin G.I., Shishkina L.P.* Novel defect-correction high-order, in space and time, accurate schemes for parabolic singularly perturbed convection-diffusion problems // Comput. Meth. Appl. Math. 2003. V. 3. № 3. P. 387–404.
16. *Самарский А.А.* Теория разностных схем. М.: Наука, 1989.
17. *Марчук Г.И.* Методы вычислительной математики. М.: Наука, 1989.
18. *Бабенко К.И.* Основы численного анализа. М.: Наука, 1986.

УДК 519.632.6

SPECTRAL DISCRETIZATIONS OF 3-d ELLIPTIC PROBLEMS AND FAST DOMAIN DECOMPOSITION METHODS¹⁾

© 2007 г. V. Korneev*,**, A. Rytov**

(*St. Petersburg State University, Russia;

**St. Petersburg State Polytechnical University, Russia)

Korneev@amd.stu.neva.ru; Rytov@mail.ru

Received February 15, 2007; in final form May 22, 2007

Спектральная дискретизация 3-мерных эллиптических задач и быстрые методы декомпозиции области. Важный для приложений класс *hp*-дискретизаций эллиптических уравнений 2-го порядка составляют дискретизации посредством спектральных конечных элементов. Получение для них быстрых алгоритмов метода декомпозиции области сдерживалось отсутствием быстрых солверов для основных компонент метода – локальных внутренних задач на подобластях декомпозиции и их гранях. Как было недавно установлено авторами, такие солверы могут быть получены посредством специальных факторизованных предобусловливателей. Факторизованные предобусловливатели строятся, в свою очередь, с использованием некоторой существенной аналогии между матрицами жесткости спектральных и иерархических базисных *hp*-элементов (координатные функции последних определяются как тензорные произведения интегрированных полиномов Лежандра). Эта аналогия позволяет получить для матриц спектральных элементов быстрые методы решения, которые схожи с быстрыми методами решения для матриц иерархических элементов. Опираясь на эти факты и предыдущие результаты по предобусловливанию других компонент, мы получаем быстрые алгоритмы решения методом декомпозиции области для дискретизаций спектральными элементами. Библ. 39.

Ключевые слова: метод декомпозиции области, дискретизации спектральными элементами, быстрые алгоритмы, предобусловливание.

1. INTRODUCTION

In DD (domain decomposition) methods, the main contribution to the computational work arises from two major components. They are related to the solution of local Dirichlet problems on subdomains of decomposition and problems related to their faces. In the case of *hp* methods, a good choice for subdomains of decomposition is the domains occupied by finite elements of the discretization. For this reason, under the conditions of the shape regularity of finite elements of a FE (finite element) assemblage, optimization of these components with respect to the computational work is reduced to obtaining fast preconditioners-solvers for the stiffness matrix of the *p* reference element and the Schur complement related to its boundary.

For some period, analysis of the DD preconditioning for *hp* discretizations was primarily concerned with different substructuring techniques providing almost optimal relative condition numbers and parallelization of computations. In relation to the DD of Dirichlet-Dirichlet type, we refer to the papers [1]–[8] and to the recent book [9] in which a much more complete bibliography is found. In spite of the well developed general theory, the very possibility of optimization, especially in the 3-d case, of main components of DD solvers with respect to the total computational work was confirmed only recently. Starting points for deriving such optimized components for the two major types of *hp* discretizations were primarily the finite-difference preconditioners, suggested in [4], see also [5], and [10] for *hierarchical* and *spectral* reference elements stiffness matrices, respectively. In this paper, we term *hierarchical* the reference element with the form functions, produced by tensor products of integrated Legendre's polynomials. For hierarchical reference elements, especially in 2-d, a number of fast preconditioners-solvers for internal stiffness matrices have been justified theoretically and tested numerically. For instance, the DD type solver (secondary to the primal DD for the EE discretization) of [11], [12], the multilevel solver of [13] and the multilevel solver of [14], which

¹⁾ Research is supported by Russian Fund of Basic Research, grant 05-01-00779-a.

The first author has partially been supported by the Austrian Science Fund (FWF) through the special research programme (SFB) F013, project 16.

is based on multiresolution wavelet decompositions and is also optimal for 3-d. Fast solvers for faces were derived in [15], [16] by application of the multiresolution wavelet solver to mass matrices of 2-d elements and by the method of K-interpolation. The results of computer testing some of these solvers and other components of DD algorithms can be found in [17]–[20] for a wide range of p , reaching as a rule several hundreds. The situation with spectral elements is different, and to the best of the authors' knowledge, until recently only the multilevel solver [21] was known for spectral discretizations. It demonstrated high efficiency in numerical testing, but its theoretical justification has been done only for the 1-d case.

At first glance hierarchical and spectral elements seem rather different. However, lately in [22], [23] it was established the interrelation between the three most popular types of the cubic p reference elements, including the mentioned two, showing that in computations they can be treated with a great extent of similarity. On one hand, it has been established by a diagonal transformation of the known basic finite-difference preconditioners, neglecting some minor terms and easy estimates, and as a consequence its use is computationally cheap. On the other hand, it allows to adapt all solvers, known for reference elements of one type, into the solvers for stiffness matrices of other types reference elements. It is important that solvers obtained by such an adaptation have computational cost the same in the order of p as their initial prototypes. For one of the instances, the fast multilevel solver for spectral elements was justified in [22], which is of the same type as the one suggested in [13] for 2-d hierarchical p elements. In this paper, first of all we obtain fast solvers for internal finite element and face subproblems, arising in DD algorithms for 3-d discretizations by spectral elements. For internal Dirichlet problems on spectral elements, the fast preconditioner-solver is based on multiresolution wavelet preconditioners-solvers for 1-d stiffness and mass matrices, which are similar to those used in [14] in the case of hierarchical reference elements. The set of admissible wavelets satisfies even easier conditions in comparison with the conditions arising for hierarchical elements. The fast preconditioner-solver for a typical face is basically designed by means of the K-interpolation.

Inefficient prolongations from the interface boundary can also compromise optimality of the DD algorithm. In the paper, we present prolongations that are almost optimal in terms of the computational cost. Such prolongations are constructed by means of an inexact iterative solver for inner problems with the multiresolution wavelet preconditioner mentioned above. With these three main fast DD components in hands, it remains only to look for a good preconditioner for the wire basket subproblem, which has though a relatively small dimension $O(\mathcal{R}p)$, where $\mathcal{R}$ is the number of finite elements. We use exactly the one studied for h discretizations in [24] and in [25] and expanded to spectral discretizations in [3] and [6]. The wire basket solver is described in these papers up to the solver for some $O(\mathcal{R}) \times O(\mathcal{R})$ subsystem. Without study of special algorithms, we make a mild assumption on the existence of such algorithm for solving this small subsystem which do not compromise the optimality of DD algorithm in a whole. The concluding result of the paper is that the DD preconditioner-solver, based on the described above components, provides the relative condition number $O((1 + \log p)^2)$, while solving the system of algebraic equations with the DD preconditioner for the matrix requires $O(N(1 + \log p))$ arithmetic operations, where $N = O(\mathcal{R}p^3)$ is the order of the system of the FE algebraic equations.

The proofs of these bounds require analysis of efficiency of all the suggested components. Apart from this, we naturally attract the results on the wire basket preconditioning of the papers cited above. Among them the most important is the reduction of the face subproblem preconditioning to good matrix representations of the norms $_{00}|\cdot|_{1/2, F_i}$ for separate faces F_i . In this and other relations, we should also refer to [26]–[29], where some basic ideas were developed. Also, as it was noted in [6], a number of bounds and techniques related to DD preconditioning for h -version can be almost directly used in the analysis of DD preconditioners-solvers for the discretizations by spectral elements.

The functional space on cubic reference elements of all types, considered in the paper, is the space $\mathcal{Q}_{p, \mathbf{x}}$ of polynomials of order $p \geq 1$ in each variable of $\mathbf{x} = (x_1, x_2, \dots, x_d)$. For the reason of their similarity in our context, the Lagrange elements with the GLL and GLC nodes are included in the paper in one type and termed *spectral* for convenience. By GLL and GLC nodes are implied the nodes of the Gauss–Lobatto–Legendre and Gauss–Lobatto–Chebyshev quadrature formulas, respectively. In the literature the term *spectral* is commonly applied only to the elements of the latter subtype. Elements of the other type, referred to as hierarchical, have tensor products of the integrated Legendre polynomials for coordinate polynomials. DD solvers similar to presented in this paper, but derived for the 3-d discretizations by means of the latter elements, can be found in [15], [16]. In this paper, they play a subsidiary role providing guidelines and better understanding of common features of the two types of hp discretizations.

The paper is organized as follows. In Section 2, we briefly describe the reference elements under consideration and the preconditioners of finite-difference and the first-order finite element types for their stiffness matrices. Although this material is well presented in the literature, we repeat some definitions crucial

for the main content of the paper. Factored preconditioners of a new type for spectral elements are derived in Section 3. In the same section, this result is used to design a fast multiresolution wavelet 3-d solver and a face solver for spectral elements. Section 4 presents a fast DD solver for discretizations of 3-d elliptic equations by spectral elements and results of the analysis of its relative condition and numerical complexity.

Let us list some notations. Signs $<, >, \asymp$ are used for inequalities and equivalence held up to positive absolute constants: A^+ – pseudo-inverse of a matrix A ; $A < B$ with nonnegative matrices A, B implies $\mathbf{v}^T A \mathbf{v} < \mathbf{v}^T B \mathbf{v}$ for any vector $\mathbf{v}$ and similarly for signs $>, \asymp, <, >$; for a symmetric nonnegative matrix A and $\forall \mathbf{v}$ it is assumed $\|\mathbf{v}\|_A^2 := \mathbf{v}^T A \mathbf{v}$; $\tau_0 = (-1, 1)^d$ is the reference cube of dimension d . Notations $|\cdot|_{k, \Omega}, \|\cdot\|_{k, \Omega}$ stand for the semi-norm and the norm in the Sobolev space $H^k(\Omega)$, i.e.,

$$|\mathbf{v}|_{k, \Omega}^2 = \sum_{|q|=k} \int_{\Omega} (D_x^q \mathbf{v})^2 d\mathbf{x}, \quad \|\mathbf{v}\|_{k, \Omega}^2 = \|\mathbf{v}\|_{0, \Omega}^2 + \sum_{l=1}^k |\mathbf{v}|_{l, \Omega}^2,$$

where

$$D_x^q \mathbf{v} := \partial^{|\mathbf{q}|} \mathbf{v} / \partial x_1^{q_1} \partial x_2^{q_2} \dots \partial x_d^{q_d}, \quad \mathbf{q} = (q_1, q_2, \dots, q_d), \quad q_k \geq 0, \quad |\mathbf{q}| = q_1 + q_2 + \dots + q_d;$$

$\mathring{H}^1(\Omega) := \{\mathbf{v} \in H^1(\Omega) : \mathbf{v}|_{\partial\Omega} = 0\}$ is the subspace of functions from $H^1(\Omega)$ vanishing on the boundary $\partial\Omega$. $P_k(x)$ is the Legendre's polynomial of degree $k \geq 1$ for the interval $(-1, 1)$. Relationship $\mathbf{v} \longleftrightarrow \mathbf{v}$ implies that $\mathbf{v}$ is the vector representation of a finite element function $\mathbf{v}$ in a chosen basis.

2. SPECTRAL AND HIERARCHICAL p ELEMENTS AND PRECONDITIONERS FOR STIFFNESS AND MASS MATRICES

2.1. Preconditioners for spectral elements

For the reference elements introduced in this section, we use notation $\mathcal{E}_{sp}$. Coordinates η_i of the GLL nodes on the segment $[-1, 1]$ are defined as roots of the polynomial $(1-s^2)P'_p(s)$, i.e.,

$$(1-\eta_i^2)P'_p(\eta_i) = 0, \quad i = 0, 1, \dots, p. \quad (2.1)$$

Coordinates of the GLC nodes are extremal points of the Chebyshev polynomials

$$\eta_i = \cos \frac{\pi(p-i)}{p}, \quad i = 0, 1, \dots, p. \quad (2.2)$$

Orthogonal meshes with the nodes

$$\mathbf{x} = \boldsymbol{\eta}_{\boldsymbol{\alpha}} = (\eta_{\alpha_1}, \eta_{\alpha_2}, \dots, \eta_{\alpha_d}), \quad \boldsymbol{\alpha} \in \omega = \{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_d) : 0 \leq \alpha_1, \alpha_2, \dots, \alpha_d \leq p\},$$

having coordinates (2.1) or (2.2), will be termed *Gaussian* for brevity. For coordinate polynomials of the spectral reference elements, we use notation $J_{\boldsymbol{\alpha}}(\mathbf{x}) = \mathcal{J}_{\alpha_1}(x_1)\mathcal{J}_{\alpha_2}(x_2)\dots\mathcal{J}_{\alpha_d}(x_d)$, where $\mathcal{J}_i(s)$ is the 1-d polynomial of order p , satisfying equalities $\mathcal{J}_i(\eta_j) = \delta_{i,j}$, $0 \leq j \leq p$, where $\delta_{i,j}$ – Kronecker's delta.

Without loss of generality, it is convenient to assume here $p = 2N$. For $i \leq N$, the steps $\hbar_i := \eta_i - \eta_{i-1}$ of both Gaussian meshes have the same asymptotic behavior $\hbar_i \asymp i/p^2$. One can define a more general class of meshes, which on the segment $[-1, 0]$ satisfy the relationships

$$\eta_0 = -1, \quad \eta_i = \eta_{i-1} + \hbar_i, \quad \eta_N = 0, \quad c_1 \frac{i^\gamma}{N} \leq \hbar_i \leq c_2 \frac{i^\gamma}{N}, \quad \mathbb{N} = \sum_{i=1}^N i^\gamma, \quad (2.3)$$

with some fixed $c_k > 0$ and $\gamma \geq 0$ and are reproduced on $[0, 1]$ by symmetry. For $\gamma = 0$, we have a quasiuniform mesh with $\mathbb{N} = N$ and for $\gamma = 1$ – the mesh, which will be termed *pseudospectral*, with $\mathbb{N} = N(N+1)/2$. In the particular case of $c_1 = c_2 = 1$, one has $\hbar_i = i/\mathbb{N} = 2i/(N^2 + N) = \beta i/p^2$ for the steps of the pseudospectral

mesh, where $\beta \in [4, 8]$. For stiffness matrices of the reference elements with the Gaussian and pseudospectral nodes, induced by the Dirichlet integral

$$a_{\tau_0}(u, v) = \int_{\tau_0} \nabla u \cdot \nabla v \, d\mathbf{x},$$

we introduce notations A_{sp} , $A_{p/s}$, respectively.

Consider an orthogonal mesh specified by the coordinates η_i on the cube τ_0 . Let $\mathcal{H}(\tau_0)$ be the space of functions continuous on $\bar{\tau}_0$ and belonging to $\mathcal{D}_{1,x}$ on each cell. The introduced stiffness matrices can be efficiently preconditioned by the FE matrices $\mathcal{A}_{sp}$, $\mathcal{A}_{p/s}$, generated by the quadratic form a_{τ_0} on the spaces $\mathcal{H}(\tau_0)$ corresponding to the Gaussian and the pseudospectral meshes, respectively. For another preconditioner, it can be considered the simpler matrix

$$\mathbb{A}_h = \Delta_h \otimes \mathbb{D}_h \otimes \mathbb{D}_h + \mathbb{D}_h \otimes \Delta_h \otimes \mathbb{D}_h + \mathbb{D}_h \otimes \mathbb{D}_h \otimes \Delta_h, \quad (2.4)$$

constructed in terms of matrices Δ_h , $\mathbb{D}_h$ defined for the 1-d case as follows. Namely, $\mathbb{D}_h$ is the diagonal matrix

$$\mathbb{D}_h = \text{diag} \left[\tilde{h}_i = \frac{1}{2}(\tilde{h}_i + \tilde{h}_{i+1}) \right]_{i=0}^p, \quad \tilde{h}_i = 0 \quad \text{for } i = 0, p+1, \quad (2.5)$$

and Δ_h is the FE matrix, induced by the bilinear form $(v', w')_{(-1,1)}$ on the space $\mathcal{H}(-1, 1)$ of continuous and piece-wise linear on the mesh η_i :

$$\begin{aligned} (\Delta_h \mathbf{u})|_{i=0} &= -\frac{1}{\tilde{h}_1}(u_1 - u_0), \quad (\Delta_h \mathbf{u})|_{i=p} = \frac{1}{\tilde{h}_p}(u_p - u_{p-1}), \\ (\Delta_h \mathbf{u})|_i &= -\frac{1}{\tilde{h}_i}u_{i-1} + \left(\frac{1}{\tilde{h}_i} + \frac{1}{\tilde{h}_{i+1}} \right) u_i - \frac{1}{\tilde{h}_{i+1}}u_{i+1}, \quad i = 1, 2, \dots, p-1. \end{aligned} \quad (2.6)$$

Lemma 1. For the same p , let the matrices $\mathcal{A}_{sp}$ and $\mathcal{A}_{p/s}$ be obtained on the Gaussian mesh and on the pseudospectral mesh at $\gamma = 1$, respectively, whereas $\mathbb{A}_h$ be obtained on either of these meshes. Then they are spectrally equivalent to the stiffness matrix A_{sp} of the reference element $\mathcal{E}_{sp}$ uniformly in p , i.e.,

$$\mathbb{A}_h, \mathcal{A}_{p/s}, \mathcal{A}_{sp} < A_{sp} < \mathcal{A}_{sp}, \mathcal{A}_{p/s}, \mathbb{A}_h. \quad (2.7)$$

Under the same conditions similar inequalities

$$\mathbb{M}_h, \mathcal{M}_{p/s}, \mathcal{M}_{sp} < M_{sp} < \mathcal{M}_{sp}, \mathcal{M}_{p/s}, \mathbb{M}_h \quad (2.8)$$

hold for the mass matrix M_{sp} of the spectral element, its FE preconditioners, $\mathcal{M}_{p/s}$, $\mathcal{M}_{sp}$ obtained with the use of the space $\mathcal{H}(\tau_0)$, and $\mathbb{M}_h = \mathbb{D}_h \otimes \mathbb{D}_h \otimes \mathbb{D}_h$.

Proof. As it is well known, the inequalities (2.7) for $\mathcal{A}_{sp}$, in 1-d belong to the authors of [30], [31], who analyzed $L_2(-1, 1)$ - and $H^1(-1, 1)$ -stability of the Lagrange interpolation over the set of Gaussian nodes. The step to a greater dimension may be found in [32] and [6]. The spectral equivalence inequalities for $\mathcal{A}_{p/s}$ are a consequence of the identical asymptotic behavior of the Gaussian and the pseudospectral meshes, see for the proof Lemma 6.1 in [17] or [18]. In order to get the inequalities with $\mathbb{A}_h$, it is sufficient to take additionally into account the spectral equivalence of the mass matrix of the 1-d linear element to its diagonal. Similar facts for mass matrices of the spectral and pseudospectral elements result in (2.8).

2.2. Preconditioners for hierarchical p elements

We introduce the set $\mathcal{M}_{1,p} = \{\mathcal{L}_i(s), i = 0, 1, \dots, p\}$ of polynomials on the interval $(-1, 1)$

$$\begin{aligned} \mathcal{L}_0(s) &= \frac{1}{2}(1+s), \quad \mathcal{L}_1(s) = \frac{1}{2}(1-s), \\ \mathcal{L}_i(s) &:= \beta_i \int_{-1}^s P_{i-1}(t) dt = \gamma_i [P_i(s) - P_{i-2}(s)], \quad i \geq 2, \end{aligned} \quad (2.9)$$

where P_i are Legendre's polynomials and

$$\beta_i = \frac{1}{2}\sqrt{(2j-3)(2j-1)(2j+1)}, \quad \gamma_i = 0.5\sqrt{(2i-3)(2i+1)/(2i-1)}.$$

Hence, $\mathcal{L}_i$ for $i = 0, 1$ are "nodal" linear coordinate functions and for $i \geq 2$ are integrated Legendre's polynomials with β_i providing equalities $\|\mathcal{L}_i(s)\|_{0,(-1,1)} = 1$ for the norms in the space $L_2(-1, 1)$. The choice of such factors was made in Korneev/Jensen [5] with the purpose to obtain the finite-difference like preconditioner (2.12) described below.

The hierarchical reference element $\mathcal{E}_{hi}$ is cube τ_0 with coordinate polynomials forming the set

$$\mathcal{M}_{d,p} = \{L_{\alpha}(\mathbf{x}) = \mathcal{L}_{\alpha_1}(x_1)\mathcal{L}_{\alpha_2}(x_2)\dots\mathcal{L}_{\alpha_d}(x_d), \alpha \in \Omega\},$$

which is the basis in $\mathcal{Q}_{p,\mathbf{x}}$. The reference element stiffness matrix A is induced by the set $\mathcal{M}_{d,p}$ and the Dirichlet integral $a_{\tau_0}(\cdot, \cdot)$, whereas by A_I we understand the *internal* stiffness matrix, corresponding to the sub-

set $\mathring{\mathcal{M}}_{d,p} = \{L_{\alpha}, 2 \leq \alpha_k \leq p, k = 1, 2, \dots, d\}$ of coordinate polynomials vanishing on $\partial\tau_0$. If the set $\mathring{\mathcal{M}}_{d,p}$ is reordered in a specific way, see [5], the matrix A_I and the reference element internal mass matrix M_I take the block diagonal forms, which for $d = 3$ are

$$A_I = \text{diag}[A_{eee}, A_{eeo}, \dots, A_{ooo}, A_{ooo}], \quad M_I = \text{diag}[M_{eee}, M_{eeo}, \dots, M_{ooo}, M_{ooo}]. \quad (2.10)$$

In the case $p = 2N + 1$, each of the 8 independent blocks in A_I and M_I is a $N^3 \times N^3$ matrix. Entries of the block A_{eee} are integrals $a_{\tau_0}(L_{\alpha}, L_{\kappa})$ for $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and $\kappa = (\kappa_1, \kappa_2, \kappa_3)$ with even $\alpha_i, \kappa_j, i, j = 1, 2, 3$; entries of the block A_{eoe} correspond to even $\alpha_1, \alpha_3, \kappa_1, \kappa_3$ and odd α_2, κ_2 etc. In other words, e or o at k -th place correspond to even and odd powers α_k, κ_k , respectively, of polynomials $\mathcal{L}_{\alpha_k}(x_k), \mathcal{L}_{\kappa_k}(x_k)$ entering coordinate functions L_{α}, L_{κ} in the integrals $a_{\tau_0}(L_{\alpha}, L_{\kappa})$. Clearly, these integrals define entries $a_{\alpha,\kappa}$ of matrix A_I ordered with the use of multiple indices α, κ according to tensor product structure of A_I . For what follows, it is important that blocks of matrices (2.10) are represented by the sums of the Kronecker's products

$$\begin{aligned} A_{abc} &= \mathbb{K}_{1,a} \otimes \mathbb{K}_{0,b} \otimes \mathbb{K}_{0,c} + \mathbb{K}_{0,a} \otimes \mathbb{K}_{1,b} \otimes \mathbb{K}_{0,c} + \mathbb{K}_{0,a} \otimes \mathbb{K}_{0,b} \otimes \mathbb{K}_{1,c}, \\ M_{abc} &= \mathbb{K}_{0,a} \otimes \mathbb{K}_{0,b} \otimes \mathbb{K}_{0,c}, \quad a, b, c = e, o, \end{aligned} \quad (2.11)$$

of the $N \times N$ matrices $\mathbb{K}_{1,a}$ and $\mathbb{K}_{0,b}$, which, respectively, may be preconditioned by the simple matrices

$$\mathcal{D} = \text{diag}[4i^2]_{i=1}^N, \quad \Delta = \frac{1}{2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & 0 \\ & & \dots & \dots & \\ 0 & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

The diagonal matrices $\mathbb{K}_{1,a}$ and tridiagonal matrices $\mathbb{K}_{0,b}$ are even/odd blocks of the stiffness and mass matrices

$$\mathbb{K}_1 = \text{diag}[\mathbb{K}_{1,e}, \mathbb{K}_{1,o}], \quad \mathbb{K}_0 = \text{diag}[\mathbb{K}_{0,e}, \mathbb{K}_{0,o}],$$

respectively, of the 1-d reference element, i.e., generated by bilinear forms $(v', w')_{(-1,1)}$ and $(v, w)_{(-1,1)}$.

Lemma 2. For the preconditioners $\mathcal{D}, \Delta$ and

$$\Lambda_e = \mathcal{D} \otimes \Delta \otimes \Delta + \Delta \otimes \mathcal{D} \otimes \Delta + \Delta \otimes \Delta \otimes \mathcal{D}, \quad \mathcal{M} = \Delta \otimes \Delta \otimes \Delta, \quad (2.12)$$

there hold the inequalities

$$\Delta < \mathbb{K}_{0,a} < \Delta, \quad \mathcal{D} < \mathbb{K}_{1,a} < \mathcal{D}, \quad \Lambda_e < A_{abc} < \Lambda_e, \quad \mathcal{M} < M_{abc} < \mathcal{M}, \quad (2.13)$$

with A_{abc} and M_{abc} defined in (2.11).

Proof. Inequalities for A_{abc}, M_{abc} follow from the inequalities for $\mathbb{K}_{0,a}$ and $\mathbb{K}_{1,a}$, representing the case of 1-d, and (2.11), (2.12). The inequalities for $\mathbb{K}_{0,a}$, with Δ replaced by the preconditioner $\Delta + \mathcal{D}^{-1}$ close to

it, and the inequalities for $\mathbb{K}_{1,a}$ were proved in [4], see also [5]. In the form (2.13) the first pair of the inequalities was obtained in [16], see there (4.10).

Derivation of fast preconditioners-solvers for hierarchical and spectral p elements is motivated, firstly, by the relatively simple finite-difference interpretation of the preconditioners, introduced above, and, secondly, by the properties of the corresponding differential operators. For this reason, it is useful to present these differential operators. Let $\pi_1 = (0, 1)^d$. In the case of $d = 2$ the preconditioner

$$\Lambda_e = \mathcal{D} \otimes \Delta + \Delta \otimes \mathcal{D}$$

is the finite-difference approximation of the differential operator

$$Lu \equiv -2 \left(x_1^2 \frac{\partial^2 u}{\partial x_2^2} + x_2^2 \frac{\partial^2 u}{\partial x_1^2} \right), \quad \mathbf{x} \in \pi_1, \quad u|_{\partial\pi_1} = 0, \quad (2.14)$$

on the uniform square mesh of size $h = 1/(N+1)$. The form and even the order of the differential operator depends on d , and, e.g., for $d = 3$ the matrix $h^{-2}\Lambda_e$ is the finite-difference approximation of the 4-th order differential operator

$$Lu \equiv x_1^2 u_{,2,2,3,3} + x_2^2 u_{,1,1,3,3} + x_3^2 u_{,1,1,2,2}, \quad \mathbf{x} \in \pi_1, \quad u|_{\partial\pi_1} = 0, \quad (2.15)$$

on the uniform square mesh of size $h = 1/(N+1)$. Here, e.g., $u_{,1,1,2,2} = \partial^4 u / \partial x_1^2 \partial x_2^2$.

The finite element preconditioners, introduced by the first order elements, are more adapted to the analysis of fast solving procedures. They are easily derived, if one takes into account the form of the finite-difference preconditioners and the corresponding differential operators. Let $d = 3$ and $\mathring{\mathcal{V}}(\pi_1)$ be the space of continuous on $\bar{\pi}_1$ and trilinear on each cubic cell of the mesh functions, vanishing on $\partial\pi_1$. Consider bilinear form

$$b_{\pi_1}(u, v) = \int_{\pi_1} [x_1^2 u_{,2,3} v_{,2,3} + x_2^2 u_{,1,3} v_{,1,3} + x_3^2 u_{,1,2} v_{,1,2}] d\mathbf{x}, \quad (2.16)$$

and define matrix $\Lambda_{e, \text{fem}}$ as the matrix of this bilinear form on the space $\mathring{\mathcal{V}}(\pi_1)$ with the nodal basis.

Lemma 3. *The matrix $\frac{1}{h} \Lambda_{e, \text{fem}}$ is spectrally equivalent to the matrices A_{abc} , Λ_e uniformly in p .*

Proof. In view of Lemma 2, it is sufficient to establish the spectral equivalence $\frac{1}{h} \Lambda_{e, \text{fem}} \asymp \Lambda_e$. Let $\mathring{\mathcal{U}}(0, 1)$ be the space of piece-wise linear on the uniform mesh of size h continuous functions, which vanish at the ends of the interval $(0, 1)$. By $\mathcal{D}_{\text{fem}}$ we denote the matrix generated by the bilinear form $b_{(0,1)}(v, u) = \int_0^1 s^2 v(s) u(s) ds$ on the space $\mathring{\mathcal{U}}(0, 1)$. Direct calculation results in the expression

$$\Lambda_{e, \text{fem}} = \frac{4}{h^2} (\mathcal{D}_{\text{fem}} \otimes \Delta \otimes \Delta + \Delta \otimes \mathcal{D}_{\text{fem}} \otimes \Delta + \Delta \otimes \Delta \otimes \mathcal{D}_{\text{fem}}).$$

Comparing with (2.12), we conclude that the proof is reduced to the proof of the equivalence $\frac{1}{h^3} \mathcal{D}_{\text{fem}} \asymp \mathcal{D}$, which is easily done and omitted.

In 2-d, one can use the FE space $\mathring{\mathcal{V}}_{\Delta}(\pi_1)$ of continuous and vanishing on $\partial\pi_1$ piece-wise linear functions on the triangulation, which is obtained by the subdivision of each square cell of the mesh in two triangles. The preconditioner $\Lambda_{e, \text{fem}}$ is defined as the matrix of the bilinear form

$$b_{\pi_1}(u, v) = \int_{\pi_1} (x_1^2 u_{,2} v_{,2} + x_2^2 u_{,1} v_{,1}) d\mathbf{x} \quad (2.17)$$

on the space $\mathring{\mathcal{V}}_{\Delta}(\pi_1)$. We have $\Lambda_{e, \text{fem}} \asymp h^2 A_{abc}$, $h^2 \Lambda_e$.

The finite element preconditioner may be simplified in particular by the replacement of coefficients in (2.14)–(2.16) by piece-wise constant coefficients. Such coefficients may be defined by the constants different for each cell of the mesh or for each cell of the specially designed coarse (decomposition) mesh, see [12].

It is worth emphasizing one essential difference between preconditioners for the spectral and hierarchical p elements. In the preconditioner Λ_e (2.12), matrices Δ and $\mathcal{D}$ are preconditioners for the “even” blocks $\mathbb{K}_{0,e}$ and $\mathbb{K}_{1,e}$ of the mass and the stiffness matrices, respectively, of 1-d reference element $\mathcal{E}_{hi}$. On the contrary, in Λ_h similar matrices Δ_h and $\mathbb{D}_h$ are preconditioners for the stiffness and the mass matrices of 1-d reference element $\mathcal{E}_{sp}$.

3. FAST PRECONDITIONERS-SOLVERS FOR SPECTRAL ELEMENTS

The preconditioners introduced so far will be termed *source* preconditioners for the two reasons. Firstly, they allow simple finite-difference/FE interpretations (Section 2.2) or are finite-difference/FE by definition (Section 2.1) and, secondly, they have been serving as a starting point for deriving modified preconditioners, more adapted to specific fast solvers. Presently, there are several fast multilevel algorithms applicable to the systems with the matrix Λ_e , such as, e.g., presented in [13] and [14], which are not applicable to the systems with the preconditioners for the spectral elements, appearing in Lemma 1. In this Section, we establish a rather important fact. The matrix $\tilde{\Lambda}_{I,h}$ can be transformed into a matrix, from which another matrix $\Lambda_{I,sp}$ can be obtained by neglecting some secondary terms. The latter matrix, though being quite different from Λ_e , has some basic properties with it in common, which allow to adapt to $\Lambda_{I,sp}$ fast solvers known for Λ_e . It is important that the transformation is cheap and indeed is defined by a diagonal matrix. Below, for a simplification of notations, we sometimes omit index I in the notations of matrices related to the internal unknowns. For instance, we imply by $\mathbb{D}_h$, Δ_h the blocks, which are obtained from matrices (2.5), (2.6) by deleting the first and the last rows and columns.

3.1. Factored preconditioners for Dirichlet problems on spectral elements

The change of variables $\tilde{\mathbf{v}} = \mathbf{C}\mathbf{v}$ by means of the diagonal matrix $\mathbf{C} = p^{-4}\mathbb{D}_h^{-1/2} \otimes \mathbb{D}_h^{-1/2} \otimes \mathbb{D}_h^{-1/2}$ (for 2-d $\mathbf{C} = p^{-2}\mathbb{D}_h^{-1/2} \otimes \mathbb{D}_h^{-1/2}$) transforms $\Lambda_{I,h}$ (2.4) as the matrix of a quadratic form into the matrix

$$\begin{aligned}\tilde{\Lambda}_{I,h} &:= \mathbf{C}^{-1}\Lambda_{I,h}\mathbf{C}^{-1} = p^8\mathbb{D}_h^{1/2} \otimes \mathbb{D}_h^{1/2} \otimes \mathbb{D}_h^{1/2} \Lambda_{I,h} \mathbb{D}_h^{1/2} \otimes \mathbb{D}_h^{1/2} \otimes \mathbb{D}_h^{1/2} = \\ &= p^8(\tilde{\Delta}_h \otimes \mathbb{D}_h^2 \otimes \mathbb{D}_h^2 + \mathbb{D}_h^2 \otimes \tilde{\Delta}_h \otimes \mathbb{D}_h^2 + \mathbb{D}_h^2 \otimes \mathbb{D}_h^2 \otimes \tilde{\Delta}_h), \\ \tilde{\Delta}_h &= \mathbb{D}_h^{1/2} \Delta_h \mathbb{D}_h^{1/2}.\end{aligned}\quad (3.1)$$

Let us introduce also matrices

$$\begin{aligned}\Delta_{sp} &= \text{tridiag}[-1, 2, -1], \quad \mathcal{D}_{sp} = \text{diag}[1, 4, \dots, N^2, (N-1)^2, (N-2)^2, \dots, 4, 1], \\ \tilde{\Lambda}_{I,sp} &= (\Delta_{sp} + \mathcal{D}_{sp}^{-1}) \otimes \mathcal{D}_{sp} \otimes \mathcal{D}_{sp} + \mathcal{D}_{sp} \otimes (\Delta_{sp} + \mathcal{D}_{sp}^{-1}) \otimes \mathcal{D}_{sp} + \mathcal{D}_{sp} \otimes \mathcal{D}_{sp} \otimes (\Delta_{sp} + \mathcal{D}_{sp}^{-1}), \\ \Lambda_{I,sp} &= \Delta_{sp} \otimes \mathcal{D}_{sp} \otimes \mathcal{D}_{sp} + \mathcal{D}_{sp} \otimes \Delta_{sp} \otimes \mathcal{D}_{sp} + \mathcal{D}_{sp} \otimes \mathcal{D}_{sp} \otimes \Delta_{sp},\end{aligned}\quad (3.2)$$

from which two first have the dimension $(p-1) \times (p-1)$ and two others $(p-1)^3 \times (p-1)^3$.

Theorem 1. *Let the matrices $\tilde{\Lambda}_{I,h}$, $\Lambda_{I,sp}$, $\tilde{\Lambda}_{I,sp}$ be obtained on the Gaussian mesh or the pseudospectral mesh at $\tilde{h}_i \asymp ip^2$ for $1 \leq i \leq N$. Then these matrices are spectrally equivalent uniformly in p .*

Proof. It is sufficient to consider the case of the pseudospectral mesh with the steps $\tilde{h}_i = \beta i/p^2$ and to prove the spectral equivalence of the matrix $\tilde{\Delta}_h$ to $(\Delta_{sp} + \mathcal{D}_{sp}^{-1})$, the matrix $p^4\mathbb{D}_h^2$ to $\mathcal{D}_{sp}$ and the matrix $(\Delta_{sp} + \mathcal{D}_{sp}^{-1})$ to Δ_{sp} . In the proof we can use $\tilde{h}_i = ip^2$, since $4 \leq \beta \leq 8$ and we do not pay attention to absolute constants in asymptotic bounds. Taking this and the relation $\tilde{h}_i = \frac{1}{2}(\tilde{h}_i + \tilde{h}_{i+1}) = \frac{1}{2p^2}(2i+1)$ into account, we immedi-

ately conclude that

$$\frac{i^2}{2p^4} \leq \tilde{h}_i^2 \leq \frac{5i^2}{2p^4}, \quad i = 1, 2, \dots, N,$$

and we get the equivalence $\mathcal{D}_{\text{sp}} \asymp p^4 \mathbb{D}_{\hbar}^2$. The matrix $\tilde{\Delta}_{\hbar} = \text{tridiag}[k_{i,i-1}, k_{i,i}, k_{i,i+1}]$ is tridiagonal. Taking into account explicit expressions for $k_{i,j}$ for $i = 1, 2, \dots, N$, we obtain

$$k_{i,i} = \frac{1}{2}(\hbar_i + \hbar_{i+1}) \left(\frac{1}{\hbar_i} + \frac{1}{\hbar_{i+1}} \right) = 2 + \frac{1}{2i(i+1)},$$

$$1 \geq |k_{i,i-1}| = \frac{1}{2\hbar_i} \sqrt{(\hbar_{i-1} + \hbar_i)(\hbar_i + \hbar_{i+1})} = \frac{1}{2i} \sqrt{4i^2 - 1} = \sqrt{1 - \frac{1}{4i^2}} \geq 1 - \frac{1}{4i^2},$$

$$1 \geq |k_{i,i+1}| \geq 1 - \frac{1}{4(i+1)^2}.$$

It is easy to see that the matrix $\tilde{\Delta}_{\hbar}$ is represented by the sum

$$\tilde{\Delta}_{\hbar} = K_{1,I} + D,$$

where

$$K_{1,I} = \text{tridiag}[k_{i,i-1}, (k_{i,i-1} + k_{i,i+1}), k_{i,i+1}], \quad D = \text{diag}[k_{i,i} - k_{i,i-1} - k_{i,i+1}].$$

From the bounds for $k_{i,j}$ obtained above, we have

$$\frac{3}{4} \Delta_{\text{sp}} \leq \tilde{K}_{1,I} \leq \Delta_{\text{sp}}, \quad \frac{1}{4} \mathcal{D}_{\text{sp}}^{-1} \leq D \leq \mathcal{D}_{\text{sp}}^{-1},$$

by means of which one comes to the inequalities

$$\frac{1}{4} \tilde{\Lambda}_{I,\text{sp}} \leq \tilde{\Delta}_{I,\hbar} \leq \tilde{\Lambda}_{I,\text{sp}}. \quad (3.3)$$

For completing the proof of Theorem, we have only to establish the spectral equivalence $(\Delta_{\text{sp}} + \mathcal{D}_{\text{sp}}^{-1}) \asymp \Delta_{\text{sp}}$, which, in turn, requires the proof of the inequality

$$\mathcal{D}_{\text{sp}}^{-1} \leq \Delta_{\text{sp}}. \quad (3.4)$$

It may be completed with the use of Hardy's inequality. Indeed, for $k = 1, 2, \dots, \infty$, $s \neq k$, and $\mathcal{F}$, defined by

$$\mathcal{F}(x) = \begin{cases} \int_x^{\infty} f(t) dt, & s > 1/k, \\ 0, & s = 1/k, \\ \int_0^x f(t) dt, & s < 1/k, \end{cases}$$

according to Hardy's inequality we have

$$\|x^{-s} \mathcal{F}\|_{L_k(0,\infty)} \leq \frac{1}{|s - 1/k|} \|x^{1-s} f\|_{L_k(0,\infty)}.$$

For any $v \in H_0^1(-1, 1)$, the inequalities

$$\|(1+x)^{-1} v\|_{0,(-1,1)}^2 \leq 8 \|v\|_{1,(-1,1)}^2, \quad \|(1-x)^{-1} v\|_{0,(-1,1)}^2 \leq 8 \|v\|_{1,(-1,1)}^2$$

are a direct consequence of Hardy's inequality. From them, we conclude that for the weight function

$$\phi(x) = \begin{cases} 1+x, & x \in [-1, 0], \\ 1-x, & x \in [0, 1], \end{cases}$$

it holds

$$\|\phi^{-1}v\|_{0,(-1,1)}^2 \leq 8|v|_{1,(-1,1)}^2 \quad \forall v \in H_0^1(-1,1). \quad (3.5)$$

Let $\mathcal{V}(-1,1)$ be the space of functions continuous on $[-1,1]$ which are piece-wise linear on the uniform grid of size $h = 1/N$. In (3.5), we can substitute any $v \in \mathcal{V}(-1,1)$ vanishing at the ends of the interval $(-1,1)$.

Then (3.5) becomes the “matrix” inequality $\mathcal{D}_\phi^{-1} < \Delta_\phi$, in which $\mathcal{D}_\phi$ and Δ_ϕ are the matrices of the quadratic forms in the left and the right parts of (3.5). By comparison of $\mathcal{D}_\phi$ with $\mathcal{D}_{sp}$ and Δ_ϕ with Δ_{sp} , we come to (3.4). The comparison can be produced similarly to [16] by means of the auxiliary matrices defined like $\mathcal{D}_\phi$, Δ_ϕ , but for a piece wise constant functions $\hat{\phi}$ and $\check{\phi}$, which are majorant on $[0,1]$ and minorant on (η_1, η_{p-1}) functions for ϕ , see in [16] the proof of (4.10).

The following consequence of Theorem 1 is obvious.

Corollary 1. The matrices $\Lambda_{I,C} := C\Lambda_{I,sp}C$ and $\tilde{\Lambda}_{I,C} := C\tilde{\Lambda}_{I,sp}C$, defined on the pseudospectral at $\gamma = 1$ and the Gaussian meshes, are equivalent in the spectrum to the internal stiffness matrix $A_{I,sp}$ of the spectral reference element $\mathcal{E}_{sp}$ uniformly in p , i.e.,

$$\Lambda_{I,C}, \tilde{\Lambda}_{I,C} < A_{I,sp} < \Lambda_{I,C}, \tilde{\Lambda}_{I,C}. \quad (3.6)$$

The matrix C is diagonal, and thus the arithmetical costs of solving systems with the preconditioners $\Lambda_{I,C}$, $\tilde{\Lambda}_{I,C}$ coincides in the order of p with the arithmetical costs of solving systems with the preconditioners $\Lambda_{I,sp}$, $\tilde{\Lambda}_{I,sp}$, respectively.

The matrix $\Lambda_{I,sp}$ can be viewed as the matrix of 7-point finite difference approximation of the differential operator on τ_0

$$L_{sp}u = -[\phi^2(x_2)\phi^2(x_3)u_{,1,1} + \phi^2(x_1)\phi^2(x_3)u_{,2,2} + \phi^2(x_1)\phi^2(x_2)u_{,3,3}], \quad u|_{\partial\tau_0} = 0. \quad (3.7)$$

Indeed, let us introduce the uniform cubic mesh of size $h = 2/p = 1/N$ and notations $\phi_i = \phi(-1 + ih)$ and $\mathbf{u} = (u_i)_{i_1, i_2, i_3=1}^{p-1}$, assuming the vector $\mathbf{u}$ expanded by zero in all boundary nodes.

Then one can write

$$\Lambda_{I,sp}\mathbf{u}|_{\mathbf{i}} = -\frac{1}{h^2} \sum_{k=1,2,3} \phi_{i_1}^2 \phi_{i_2}^2 [u_{\mathbf{i}-\mathbf{e}_k} - 2u_{\mathbf{i}} + u_{\mathbf{i}+\mathbf{e}_k}], \quad 1 \leq i_1, i_2, i_3 \leq (p-1), \quad (3.8)$$

where $\mathbf{i} = (i_1, i_2, i_3)$, all numbers $k, l, j \in (1, 2, 3)$ are different, $\mathbf{e}_k = (\delta_{k,l})_{l=1}^3$ is the unit vector, which components are the Kronecker's symbols. For $d = 2$ expressions (3.7), (3.8) simplify:

$$\begin{aligned} L_{sp}u &= -[\phi^2(x_2)u_{,1,1} + \phi^2(x_1)u_{,2,2}], \quad u|_{\partial\tau_0} = 0, \\ \Lambda_{I,sp}\mathbf{u}|_{\mathbf{i}} &= -\sum_{k=1,2} \phi_{i_3-k}^2 [u_{\mathbf{i}-\mathbf{e}_k} - 2u_{\mathbf{i}} + u_{\mathbf{i}+\mathbf{e}_k}], \quad \mathbf{i} = (i_1, i_2). \end{aligned} \quad (3.9)$$

Now we can look for common features of the matrices $\Lambda_e, \Lambda_{I,sp}$. At $d = 2$, they and the related differential operators L, L_{sp} are similar. Namely, in each quarter of the square τ_0 , the differential expression of the operator L_{sp} is the same as of the operator L up to a constant factor and rotation and translation of the axes. The same is true for finite-difference operators $\Lambda_e, \Lambda_{I,sp}$ and FE matrices $\Lambda_{e,fem}, B_{I,sp}$, see the definition of the last ones below. At $d = 3$, the differential and finite-difference operators, related to the preconditioners for hierarchical and spectral elements, are different even in the order: L is the differential operator of the 4-th order, whereas L_{sp} of the 2-nd. However, factors $\Delta, \mathcal{D}$ and respectively $\mathcal{D}_{sp}, \Delta_{sp}$ in the representations (2.12), (3.2) of the matrices $\Lambda_e, \Lambda_{I,sp}$ by sums of Kronecker's products are similar. Due to this, all known fast solvers for the systems with the matrices Λ_e (see, e.g., [11], [12], [17], [13], [14]) can be adapted to the systems with the matrices $\Lambda_{I,sp}$ with the same asymptotic arithmetic cost.

As it was noted earlier, the analysis of fast algorithms simplifies, if the preconditioner is a FE matrix. Instead of $\Lambda_{I,sp}, \tilde{\Lambda}_{I,sp}$, it is possible to use spectrally equivalent FE matrices, generated with the use of the

1-st order elements. Let $d = 2$. We divide each square cell of size $\hbar$ on τ_0 in two triangles. On such a triangulation of τ_0 , we introduce the space $\mathring{\mathcal{V}}_\Delta(\tau_0)$ of continuous piece-wise linear functions vanishing on $\partial\tau_0$. The FE preconditioner $B_{I, \text{sp}}$ may be defined as the matrix of the bilinear form

$$b_{\tau_0}(u, v) = \sum_{k=1}^3 \int_{\tau_0} \phi_{3-k}^2 u_{,k} v_{,k} d\mathbf{x} \quad (3.10)$$

on the space $\mathring{\mathcal{V}}_\Delta(\tau_0)$. At $d = 3$, the operator L_{sp} (3.7) is associated with the bilinear form

$$b_{\tau_0}(u, v) = \sum_{k=1}^3 \int_{\tau_0} \phi_{k+1}^2 \phi_{k+2}^2 u_{,k} v_{,k} d\mathbf{x} \quad (3.11)$$

with indices $k+1, k+2$ understood modulo 3. This form is defined on the space $\mathring{\mathcal{V}}(\tau_0)$ of continuous functions, which are trilinear on each cubic cell of the mesh and vanish on $\partial\tau_0$. Then $B_{I, \text{sp}}$ can be defined as the matrix of the bilinear form (3.11) on this space. In a way similar to that used in Lemma 3, it may be shown that

$$B_{I, \text{sp}} \asymp \hbar^{4-d} \Lambda_{I, \text{sp}}. \quad (3.12)$$

Let us note that instead of functions $\varphi(x) = x^2$, $x \in [0, 1]$, and $\phi(x) = \min(x+1, x-1)$, $s \in [-1, 1]$, functions $\varphi(x) = \max(\hbar^2, x^2)$ and $\delta(x) = \max(\hbar, \min(x+1, x-1))$ may be used.

3.2. Multiresolution wavelet solver for Dirichlet problems on 3-d spectral elements

According to Section 3.1, in order to obtain fast preconditioners-solvers for the internal stiffness matrices $A_{I, \text{sp}}$ of spectral elements, it is sufficient to design fast preconditioners-solvers for the preconditioner $\Lambda_{I, \text{sp}}$ or $\tilde{\Lambda}_{I, \text{sp}}$. In this section, we consider a multilevel preconditioner-solver for the matrix $\Lambda_{I, \text{sp}}$, based on the multiresolution wavelet analysis. This solver is similar to the solver, designed by Beuchler/Schneider/Schwab [14] for the preconditioners $\Lambda_{e, \text{fem}}$ of the internal stiffness matrices of the hierarchical p elements.

Taking into account the interpretation of $\Lambda_{I, \text{sp}}$ as the finite-difference operator (3.8), we consider cube τ_0 subdivided by the cubic mesh of size $\hbar = 1/p$. Only for convenience and without loss of generality, it is assumed that $p = 2N$, $N = 2^{\ell_0-1}$. For each $l = 1, 2, \dots, \ell_0$, one can introduce the uniform mesh x_i^l , $i = 0, 1, \dots, 2N_l$, $N_l = 2^{l-1}$, $x_0 = -1$, $x_{2N_l} = 1$ of size $\hbar_l = 2^{1-l}$ and the space $\mathcal{V}_l(-1, 1)$ of continuous on $(-1, 1)$ piece-wise linear functions, vanishing at the ends of this interval. The dimension of $\mathcal{V}_l(-1, 1)$ is $N_l = p_l - 1 = 2^l - 1$ with $p_{\ell_0} = p$. Let $\phi_i^l \in \mathcal{V}_l(-1, 1)$ be the nodal basis function for the node x_i^l , so that $\phi_i^l(x_j^l) = \delta_{i,j}$ and $\mathcal{V}_l(-1, 1) = \text{span}\{\phi_i^l\}_{i=1}^{p_l-1}$. This basis induces the Gram matrices

$$\Delta_l = \hbar_l (\langle (\phi_i^l)', (\phi_j^l)' \rangle_{\omega=1})_{i,j=1}^{p_l-1}, \quad \mathcal{M}_l = \hbar_l^{-1} (\langle \phi_i^l, \phi_j^l \rangle_{\omega=\phi})_{i,j=1}^{p_l-1} \quad (3.13)$$

where

$$\langle v, u \rangle_\omega := \int_{-1}^1 \omega^2 v u d\mathbf{x},$$

and ϕ is the function introduced in Section 3.1. Obviously,

$$\Delta_{\ell_0} = \Delta_{\text{sp}}, \quad \mathcal{M} := \mathcal{M}_{\ell_0} \asymp \mathcal{D}_{\text{sp}}. \quad (3.14)$$

The representation of each $\mathcal{V}_l$ by the direct sum $\mathcal{V}_l = \mathcal{V}_{l-1} \oplus \mathcal{W}_l$ results in the decomposition

$$\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \dots \oplus \mathcal{W}_{\ell_0}$$

with the notations $\mathcal{V} = \mathcal{V}_{\ell_0}$ and $\mathcal{W}_1 = \mathcal{V}_1$. Let $\{\psi_k^l\}_{k,l=1}^{p_{l-1}, \ell_0}$ denote the *multiscale wavelet basis*, composed of some *single scale bases* $\{\psi_k^l\}_{k=1}^{p_{l-1}}$ in the spaces $\mathcal{W}_l$, so that $\mathcal{W}_l = \text{span}\{\psi_k^l\}_{k=1}^{p_{l-1}}$. The multiscale wavelet basis in $\mathcal{V}$, if it is stable in the norms induced by the scalar products (3.13), allows to define 1-d multilevel preconditioners, which in turn lead to the multidimensional tensor product multilevel preconditioner. Before formulating the result, we introduce additional notations. The basis $\{\psi_k^l\}_{k,l=1}^{p_{l-1}, \ell_0}$ induces the matrices

$$\begin{aligned} \Delta_{\text{wlet}} &= (\langle (\psi_i^k)', (\psi_j^l)' \rangle_1)_{i,j=1; k,l=1}^{p_{l-1}, \ell_0}, & \mathcal{M}_{\text{wlet}} &= (\langle \psi_i^k, \psi_j^l \rangle_\phi)_{i,j=1; k,l=1}^{p_{l-1}, \ell_0}, \\ \mathbb{D}_1 &= \text{diag}[\langle (\psi_i^l)', (\psi_i^l)' \rangle_1]_{i,l=1}^{p_{l-1}, \ell_0}, & \mathbb{D}_0 &= \text{diag}[\langle \psi_i^l, \psi_i^l \rangle_\phi]_{i,l=1}^{p_{l-1}, \ell_0}, \end{aligned} \quad (3.15)$$

where, obviously, $\mathbb{D}_1$ and $\mathbb{D}_0$ are the diagonal matrices having the same main diagonals with Δ_{wlet} and $\mathcal{M}_{\text{wlet}}$, respectively. By Q is denoted the transformation matrix from the multiscale wavelet basis to the finite element basis $\{\phi_k^{l_0}\}_{k=1}^{p-1}$. If $\mathbf{v}_{\text{wlet}}$ and $\mathbf{v}$ are the vectors of coefficients of a function from $\mathcal{V}(0, 1)$ in these two bases, respectively, then $\mathbf{v} = Q^T \mathbf{v}_{\text{wlet}}$.

Theorem 2. *There exist such multiscale wavelet bases $\{\psi_k^l\}_{k,l=1}^{p_{l-1}, \ell_0}$ that the matrices $\Delta_{\ell_0}^{-1}$ and $\mathcal{M}_{\ell_0}^{-1}$ are simultaneously spectrally equivalent to the matrices $Q^T \mathbb{D}_1^{-1} Q$ and $Q^T \mathbb{D}_0^{-1} Q$, respectively, uniformly in p . Besides, matrix-vector multiplications $Q \mathbf{v}_{\text{wlet}}$ and $Q^T \mathbf{v}$ require $O(p)$ arithmetic operations.*

Proof. Note that the equivalences of Theorem follow from the equivalences $\Delta_{\text{wlet}} \asymp \mathbb{D}_1$ and $\mathcal{M}_{\text{wlet}} \asymp \mathbb{D}_0$ and the definitions of these matrices and the matrices Δ_{ℓ_0} , $\mathcal{M}_{\ell_0}$, Q . The proof of Theorem is simpler than the proof of corresponding results of Beuchler/Schneider/Schwab [14], see there, e.g., Theorem 3.3, because in our case the weight ϕ is symmetric on the interval $(-1, 1)$ under consideration. The cited authors justified existence of multiscale wavelet bases with the required properties in the more difficult case of the space $\mathcal{V}(0, 1) := \{v \in \mathcal{V}(-1, 1) | v(x) = 0 \text{ at } x \notin (-1, 0)\}$ and the same weight ϕ , which is not symmetric on a half of $(-1, 1)$. Practically the same types of wavelets, reproduced on $(0, 1)$ by the symmetry and easily adjusted in the vicinity of the regular point $x = 0$, can be used.

Theorem 3. *Let $\mathcal{B}_{I, \text{sp}} = C \mathbb{B}_{I, \text{sp}} C$ and*

$$\mathbb{B}_{I, \text{sp}}^{-1} = \begin{cases} (Q^T \otimes Q^T) [\mathbb{D}_0 \otimes \mathbb{D}_1 + \mathbb{D}_1 \otimes \mathbb{D}_0]^{-1} (Q \otimes Q), & d = 2, \\ (Q^T \otimes Q^T \otimes Q^T) [\mathbb{D}_0 \otimes \mathbb{D}_0 \otimes \mathbb{D}_1 + \mathbb{D}_0 \otimes \mathbb{D}_1 \otimes \mathbb{D}_0 + \\ \mathbb{D}_0 \otimes \mathbb{D}_0 \otimes \mathbb{D}_1]^{-1} (Q \otimes Q \otimes Q), & d = 3, \end{cases} \quad (3.16)$$

then $\mathcal{B}_{I, \text{sp}} \asymp A_{I, \text{sp}}$ and therefore

$$\text{cond}[\mathcal{B}_{I, \text{sp}}^{-1} A_{I, \text{sp}}] < 1. \quad (3.17)$$

The arithmetical cost of the operation $\mathcal{B}_{I, \text{sp}}^{-1} \mathbf{v}$ for any $\mathbf{v} \in U_I$ is $O(p^d)$.

Proof. In view of Lemma 1 and Theorem 1, it is sufficient to prove the equivalence

$$\text{cond}[\mathbb{B}_{I, \text{sp}}^{-1} A_{I, \text{sp}}] \asymp 1.$$

The proof of this bound is the consequence of (3.14), Theorem 2 and properties of the Kronecker product.

3.3. Multiresolution wavelet solver for faces

Another important problem in optimization of DD algorithms for spectral discretizations is the development of fast solvers for the internal problems on faces. As it is known (see e.g. [3] and [16]), in the wire basket algorithms it is reduced to the preconditioning of the matrix of the quadratic form $_{00}|\cdot|_{1/2, F_0}^2$ on the subspace of polynomials $\overset{\circ}{\mathcal{Q}}_{p, \mathbf{x}}$ of two variables $\mathbf{x} = (x_1, x_2)$, vanishing on the boundary of F_0 . This quadratic

form is the square of the norm in the space $H_{00}^{1/2}(F_0)$, whereas $F_0 = (-1, 1) \times (-1, 1)$ represents a typical face of the 3-d reference cube. As shown in [33] and [34], one of the characterizations of this norm is

$${}_{00}|\mathbf{v}|_{1/2, F_0}^2 = |\mathbf{v}|_{1/2, F_0}^2 + \int_{F_0} \frac{|\mathbf{v}(\mathbf{x})|^2}{\text{dist}[\mathbf{x}, \partial F_0]} d\mathbf{x}.$$

Theorem 4. Let $d_{0,i}, d_{1,i}$ be the diagonal entries of matrices $\mathbb{D}_0, \mathbb{D}_1$, respectively, and $\mathbb{D}_{1/2}$ be the diagonal $(p-1)^2 \times (p-1)^2$ matrix with the entries on the main diagonal

$$d_{i,j}^{(1/2)} = d_{0,i} d_{0,j} \sqrt{\frac{d_{1,i}}{d_{0,i}} + \frac{d_{1,j}}{d_{0,j}}}.$$

Let also

$$\mathcal{S}_0 = \mathbf{C} \mathbb{S}_0 \mathbf{C}, \quad \mathbb{S}_0^{-1} = (\mathbf{Q}^T \otimes \mathbf{Q}^T) \mathbb{D}_{1/2}^{-1} (\mathbf{Q} \otimes \mathbf{Q}).$$

Then for all $\mathbf{v} \in \mathring{\mathcal{Q}}_{p,\mathbf{x}}$ and vectors $\mathbf{v}$, representing $\mathbf{v}$ in the basis $\mathring{\mathcal{M}}_{2,p}$, the norms ${}_{00}|\mathbf{v}|_{1/2, F_0}$ and $\|\mathbf{v}\|_{\mathcal{S}_0}$ are equivalent uniformly in p .

Proof. The proof is alike that of a similar statement for the case of the hierarchical p element $\mathcal{E}_{\text{hi}}$ (see [16, Theorem 4.4]). First of all, we note that another characterization of the norm ${}_{00}|\mathbf{v}|_{1/2, F_0}$ can be obtained by the Peetre's K-method of interpolation between the norms in spaces $L_2(F_0)$ and $\mathring{H}^1(F_0)$ (see, e.g., [34, p. 66–69, 98–99]). At the same time from Theorem 2 and considerations preceding it, it follows that the matrices

$$\mathbb{D}_0 = \mathbb{D}_0 \otimes \mathbb{D}_0, \quad \mathbb{D}_1 = \mathbb{D}_1 \otimes \mathbb{D}_1 + \mathbb{D}_0 \otimes \mathbb{D}_0$$

are spectrally equivalent to the internal mass and stiffness matrices of the 2-d spectral reference element in a special basis. It is the multiscale wavelet basis modified according to the transformation matrix $\mathbf{C}$. Indeed, if we introduce matrices

$$\mathcal{M} = \mathbf{C}^{-1}(\mathbf{Q}^T \times \mathbf{Q}^T), \quad \mathbf{M}_{\text{wlet}} = \mathcal{M}^T \mathbf{M}_{\text{sp}} \mathcal{M}, \quad \mathbf{A}_{\text{wlet}} = \mathcal{M}^T \mathbf{A}_{\text{sp}} \mathcal{M},$$

then by the chain of equivalences reflected in Lemma 1, Theorems 1, 2, and (3.14), we have

$$\mathbf{M}_{\text{wlet}} \asymp \mathbb{D}_0, \quad \mathbf{A}_{\text{wlet}} \asymp \mathbb{D}_1.$$

Let $\mathbf{v} \in \mathring{\mathcal{Q}}_{p,\mathbf{x}}$, $\mathbf{v}$ be the vector representation of $\mathbf{v}$ in the basis polynomials of the reference element, and $\mathbf{v} = \mathcal{M} \mathbf{v}_{\text{wlet}}$. The above relationships may be also rewritten in the form

$$\|\mathbf{v}_{\text{wlet}}\|_{\mathbb{D}_0} \asymp \|\mathbf{v}\|_{0, \tau_0}, \quad \|\mathbf{v}_{\text{wlet}}\|_{\mathbb{D}_1} \asymp |\mathbf{v}|_{1, \tau_0}, \quad \tau_0 = (-1, 1) \times (-1, 1), \quad (3.18)$$

and by the Peetre's K-method of interpolation we have

$$\|\mathbf{v}_{\text{wlet}}\|_{\mathbb{D}_{1/2}} \asymp {}_{00}|\mathbf{v}|_{1/2, \tau_0}. \quad (3.19)$$

In view of the relation between $\mathbf{v}$ and $\mathbf{v}_{\text{wlet}}$ this completes the proof, i.e.,

$$\|\mathbf{v}_F\|_{\mathcal{S}_0} \asymp {}_{00}|\mathbf{v}|_{1/2, F_0}, \quad (3.20)$$

where now vectors $\mathbf{v}_F \longleftrightarrow \mathbf{v}$ represent 2-d polynomials $\mathbf{v} \in \mathring{\mathcal{Q}}_{p,\mathbf{x}}$ in the basis $L_{\boldsymbol{\alpha}}(\mathbf{x})$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, $2 \leq \alpha_1, \alpha_2 \leq p$, on the representative face F_0 of the spectral reference element.

The presented in Sections 3.2 and 3.3 fast solvers for internal and face problems can be easily generalized on “orthotropic” spectral elements with shape polynomials, having different orders along different axes.

4. DOMAIN DECOMPOSITION ALGORITHM FOR DISCRETIZATIONS BY SPECTRAL ELEMENTS

Now we describe main components of the DD algorithm and formulate the result on its arithmetic cost.

As a model, we consider the Dirichlet problem: find $u \in \mathring{H}^1(\Omega)$ satisfying the identity

$$a_\Omega(u, v) := \int_\Omega \varrho(\mathbf{x}) \nabla u \cdot \nabla v d\mathbf{x} = (f, v)_\Omega \quad \forall v \in \mathring{H}^1(\Omega), \quad (4.1)$$

where $\mathring{H}^1(\Omega)$ is the subspace of functions from $H^1(\Omega)$, vanishing on the boundary $\partial\Omega$. For simplicity, it is assumed that Ω coincides with the computational domain, i.e., it is the domain of an assemblage of compatible and in general curvilinear finite elements occupying domains τ_r , i.e.,

$$\bar{\Omega} = \bigcup_{r=1}^{\mathcal{R}} \bar{\tau}_r.$$

Finite elements and their domains τ_r are specified by nondegenerate mappings $\mathbf{x} = \mathbb{X}^{(r)}(\mathbf{y}) : \bar{\tau}_0 \rightarrow \bar{\tau}_r$ with positive Jacobian's, and it is required that these mappings satisfy the conditions, called the *generalized conditions of the angular (shape) quasiuniformity*. If the mappings are trilinear, i.e., elements have straight edges, these conditions are equivalent to the well known conditions of shape regularity (see, e.g., [35]). In a more general case, they are equivalent to the following ones, see [36]. Suppose, each mapping is represented as a superposition of two nondegenerate mappings $\mathbb{X}^{(r)}(\mathbf{y}) = \tilde{\mathbb{X}}^{(r)}(\mathcal{L}^{(r)}(\mathbf{y}))$, where $x = \tilde{\mathbb{X}}^{(r)}(\mathbf{z}) : \bar{\tau}'_r \rightarrow \bar{\tau}_r$ is a nonlinear and $\mathbf{z} = \mathcal{L}^{(r)}(\mathbf{y}) : \bar{\tau}_0 \rightarrow \bar{\tau}'_r$ is an affine or trilinear mapping (e.g., with the vertices of τ'_r coinciding with those of τ_r). Then τ'_r must be shape regular, and for the nonlinear mappings and their inverses the Jacobians and their components must be uniformly bounded.

The positive coefficient ϱ is accepted to be piece-wise constant and namely such that $\varrho(\mathbf{x}) = \varrho_r = \text{const}$ for $\mathbf{x} \in \tau_r$.

The assemblage of spectral finite elements, associated with a single reference element $\mathcal{E}_{\text{sp}}$ by mappings $\mathbb{X}^{(r)}$, defines the FE space

$$\mathbb{V}(\Omega) = \{v : v \in C(\bar{\Omega}), v(\mathbb{X}^{(r)}(\mathbf{y}))|_{\mathbf{y} \in \tau_0} \in \mathcal{Q}_{p,\mathbf{x}} \text{ for } r = 1, 2, \dots, \mathcal{R}\}, \quad \mathbb{V}(\Omega) \subset H^1(\Omega),$$

and its subspace $\mathring{\mathbb{V}}(\Omega) = \mathbb{V}(\Omega) \cap \mathring{H}^1(\Omega)$. We write the system of FE algebraic equations for the problem (4.1), obtained by means of the subspace $\mathring{\mathbb{V}}(\Omega)$, in the form

$$\mathbf{K}\mathbf{u} = \mathbf{f}. \quad (4.2)$$

At designing the DD solver for (4.2), each p -element is treated as a subdomain of decomposition, typically for many other papers as well.

It is natural to distinguish *internal*, *face*, *edge* and *vertex* degrees of freedom in the FE assemblage and decompose the vector space V of the unknowns into the direct sum of the corresponding subspaces

$$V = V_I \oplus V_F \oplus V_E \oplus V_V.$$

DD solvers or their parts are also often based on the decompositions

$$V = V_I \oplus V_F \oplus V_W, \quad V = V_I \oplus V_B,$$

where $V_B = V_F \oplus V_E \oplus V_V$ and $V_W = V_E \oplus V_V$ are the subspaces of the *interelement boundary* and *wire basket* degrees of freedom. According to these subspaces, the finite element stiffness matrix may be represented in the block forms

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_I & \mathbf{K}_{IB} \\ \mathbf{K}_{BI} & \mathbf{K}_B \end{pmatrix} = \begin{pmatrix} \mathbf{K}_I & \mathbf{K}_{IF} & \mathbf{K}_{IW} \\ \mathbf{K}_{FI} & \mathbf{K}_F & \mathbf{K}_{FW} \\ \mathbf{K}_{WI} & \mathbf{K}_{WF} & \mathbf{K}_{WW} \end{pmatrix} = \begin{pmatrix} \mathbf{K}_I & \mathbf{K}_{IF} & \mathbf{K}_{IE} & \mathbf{K}_{IV} \\ \mathbf{K}_{FI} & \mathbf{K}_F & \mathbf{K}_{FE} & \mathbf{K}_{FV} \\ \mathbf{K}_{EI} & \mathbf{K}_{EF} & \mathbf{K}_E & \mathbf{K}_{EV} \\ \mathbf{K}_{VI} & \mathbf{K}_{VF} & \mathbf{K}_{VE} & \mathbf{K}_V \end{pmatrix}. \quad (4.3)$$

For the corresponding spaces of the FE functions, we use similar notations with V replaced by $\mathbb{V}$. Restrictions of the introduced above spaces to finite elements τ_r are supplied with the additional upper index r , e.g., $\mathbb{V}_B^{(r)}$ denotes the subspace spanned by boundary coordinate functions of a finite element τ_r with $r = 0$ reserved for the reference cube. Similarly, $K_F^{(r)}$ is the block of the stiffness matrix of an element τ_r , generated by face coordinate functions. The spaces $V^{(r)}$ and $\mathbb{V}^{(r)}$ for the reference element will be denoted U and $\mathcal{U} = \mathcal{Q}_{p,\mathbf{x}}$, respectively, with the same indexation for subspaces.

For the FE matrix K , we will design a DD preconditioner-solver $\mathcal{H}$ of the Dirichlet–Ditichlet type, having the form

$$\begin{aligned}\mathcal{H}^{-1} &= \overline{\mathcal{H}}_I^+ + \mathbf{P}_{V_B \rightarrow V} \mathcal{S}_B^{-1} \mathbf{P}_{V_B \rightarrow V}^T, \\ \mathcal{S}_B^{-1} &= \overline{\mathcal{S}}_F^+ + \mathbf{P}_{V_W \rightarrow V_B} (\mathcal{S}_W^B)^{-1} \mathbf{P}_{V_W \rightarrow V_B}^T,\end{aligned}\quad (4.4)$$

and defined by means of three preconditioners-solvers and two prolongation matrices. First we describe the former three.

(i) The block diagonal preconditioner-solver for the internal Dirichlet problems on finite elements has the form

$$\overline{\mathcal{H}}_I^+ := \begin{pmatrix} \mathcal{H}_I^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{where} \quad \mathcal{H}_I = \text{diag}[h_1 \mathfrak{Q}_1 \mathcal{B}_{I,\text{sp}}, h_2 \mathfrak{Q}_2 \mathcal{B}_{I,\text{sp}}, \dots, h_{\mathcal{R}} \mathfrak{Q}_{\mathcal{R}} \mathcal{B}_{I,\text{sp}}]$$

and $\mathcal{B}_{I,\text{sp}}$ is the multiresolution wavelet preconditioner-solver, appearing in Theorem 3. The value h_r is the characteristic size of an element, entering the generalized conditions of angular quasiuniformity. It can be set equal to the arithmetic mean of the inscribed and circumscribed spheres for τ_r . Each block $h_r \mathfrak{Q}_r \mathcal{B}_{I,\text{sp}}$ corresponds to one block $K_I^{(r)}$ in the block K_I for internal unknowns

$$K_I = \text{diag}[K_I^{(1)}, K_I^{(2)}, \dots, K_I^{(\mathcal{R})}]$$

of the FE stiffness matrix K .

(ii) The block diagonal preconditioner-solver for the internal problems on faces of finite elements

$$\overline{\mathcal{S}}_F^+ = \begin{pmatrix} \mathcal{S}_F^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{where} \quad \mathcal{S}_F = \text{diag}[\kappa_1 \mathcal{S}_0, \kappa_2 \mathcal{S}_0, \dots, \kappa_Q \mathcal{S}_0], \quad (4.5)$$

Q is the number of different faces $F_k \subset \Omega$ of the FE discretization, κ_k are factors, $\mathcal{S}_0^{-1}$ is the multiresolution wavelet preconditioner-solver for one face, appearing in Theorem 4. For a face F_k of the discretization, let $r_1(k)$ and $r_2(k)$ be the numbers of two elements $\bar{\tau}_{r_1(k)}$ and $\bar{\tau}_{r_2(k)}$, sharing the face F_k . Then

$$\kappa_k = (h_{r_1(k)} \mathfrak{Q}_{r_1(k)} + h_{r_2(k)} \mathfrak{Q}_{r_2(k)}). \quad (4.6)$$

The matrix $\mathcal{S}_F$ is the preconditioner for the $Q(p-1)^2 \times Q(p-1)^2$ block S_F of the Schur complement

$$S_B = K_B - K_{BI} K_I^{-1} K_{IB}, \quad S_B = \begin{pmatrix} S_F & S_{FW} \\ S_{WF} & S_W \end{pmatrix},$$

which for the reference element is denoted by S_B . Although S_F is not a block diagonal matrix, obviously, it can be represented in the block form with each $(p-1)^2 \times (p-1)^2$ block on the diagonal related to one face. Such a representation is assumed in (4.5).

(iii) The preconditioner-solver $\mathcal{S}_W^B$ related to the wire basket subproblem. It can be interpreted as an efficient preconditioner for the second stage Schur complement $S_W^B = S_W - S_{WF} S_F^{-1} S_{FW}$ (see [15, Section 3]). Below, we briefly describe $\mathcal{S}_W^B$, as well as the prolongation matrix $\mathbf{P}_{V_W \rightarrow V_B}$, following [3] and [6].

The prolongation operations include:

(iv) the prolongation $P_{V_B \rightarrow V}$ from the interelement boundary onto the whole computational domain $\bar{\Omega}$.

(v) the prolongation $P_{V_W \rightarrow V_B}$ from the wire basket onto the interelement boundary.

Since (i) and (ii) have been completely defined, it remains to define (iii)–(v), among which the prolongation $P_{V_B \rightarrow V}$ brings usually the major contribution to the computational cost. For the reason that we have an efficient preconditioner-solver for internal problems on finite elements, this prolongation can be efficiently completed by means of inexact iterative procedures applied element wise in parallel. We present the simplest variant of such procedure, resulting in the extra factor $\log p$ in the estimate of the computational work.

The prolongation matrix $P_{V_B \rightarrow V}$ is defined in such a way, that its restriction to each element is $P_{V_B \rightarrow V}^{(r)} = \mathbb{P}_{U_B \rightarrow U}$, and $\mathbb{P}_{U_B \rightarrow U}$ is a master prolongation matrix, which is used for the prolongation inside any finite element of the discretization. The master prolongation matrix is obtained for the reference element in the following way. Let $\mathcal{A} = \mathcal{A}_{sp}$ or $\mathcal{A} = \mathcal{A}_{p/s}$, $\mathbb{B}_I$ is another preconditioner for the internal block A_I of the reference element stiffness matrix, and $\mathbb{B}_I$ possess a fast solver. To any $\mathbf{v}_B \in U_B$, one can relate vectors $\bar{\mathbf{v}}_B, \tilde{\mathbf{v}}_B \in U_B$, such that $\tilde{\mathbf{v}}_B + \bar{\mathbf{v}}_B = \mathbf{v}_B$ and entries of $\bar{\mathbf{v}}_B$ are equal to the mean value of the corresponding finite element function $\mathbf{v}_B \longleftrightarrow \mathbf{v}_B$ on the boundary $\partial\tau_0$. The prolongation of $\bar{\mathbf{v}}_B$ by the constant is denoted by $\bar{\mathbf{v}} \in U$. The prolongation $\mathbf{u} = \mathbb{P}_{U_B \rightarrow U} \mathbf{v}_B$ is found as the sum of two vectors

$$\mathbf{u} = \bar{\mathbf{v}} + \tilde{\mathbf{u}}, \quad \text{where} \quad \tilde{\mathbf{u}} = (\tilde{\mathbf{u}}_I^T, \tilde{\mathbf{v}}_B^T)^T,$$

in which subvector $\tilde{\mathbf{u}}_I = \mathbf{w}_I^{k_0}$ is produced for some fixed number k_0 of the iterations

$$\mathbf{w}_I^{k+1} = \mathbf{w}_I^k - \sigma_{k+1} \mathbb{B}_I^{-1} (\mathcal{A}_I \mathbf{w}_I^k - \mathcal{A}_{IB} \tilde{\mathbf{v}}_B), \quad \mathbf{w}_I^0 = 0, \quad (4.7)$$

with Chebyshev iteration parameters σ_k . Obviously, the order of the sufficient number k_0 of iterations will not increase, if $\mathcal{A}_I, \mathcal{A}_{IB}$ are replaced by the respective blocks of the reference element stiffness matrix A or any other preconditioner $\mathbb{A}_h$ spectrally equivalent to A . However, in general, multiplications by A_I, A_{IB} can be much more expensive, than by $\mathcal{A}_I, \mathcal{A}_{IB}$. Note, that $\tilde{\mathbf{v}}_B$ can be calculated by means of the quadratures, related to the reference element.

Lemma 4. Suppose $\gamma_I \mathbb{B}_I \leq \mathcal{A}_I \leq \bar{\gamma}_I \mathbb{B}_I$ with positive γ_k . Then at $k_0 \geq c(1 + \log p)/(\log \rho^{-1})$, where $\rho = (1 - \theta)/(1 + \theta)$, $\theta = \sqrt{\gamma_I/\bar{\gamma}_I}$, the inequality

$$\|\mathbb{P}_{U_B \rightarrow U} \mathbf{v}_B\|_A \leq c_{\mathbb{P}, 0} \|\mathbf{v}_B\|_{\mathbb{S}_B} \quad (4.8)$$

holds with the constant $c_{\mathbb{P}}$ independent of p and $\mathbb{S}_B = A_B - A_{B,I} A_I^{-1} A_{I,B}$.

Let us note that the inequality (4.8) is equivalent to

$$|u|_{1, \tau_0} < c_{\mathbb{P}, 0} |\mathbf{v}_B|_{1/2, \partial\tau_0}, \quad (4.9)$$

where $u \longleftrightarrow \mathbf{u}$ and $\mathbf{v}_B \longleftrightarrow \mathbf{v}_B$. If $\mathbf{v}_B$ is a constant vector, then $\tilde{\mathbf{v}}_B = 0$, $A_{IB} \tilde{\mathbf{v}}_B = 0$, and, therefore, $\mathbf{w}_I^k = 0$ for $k \geq 1$, $\tilde{\mathbf{u}} = 0$, and $\mathbb{P}_{U_B \rightarrow U} \mathbf{v}_B = \bar{\mathbf{v}}$.

Proof. Since inequalities (4.8), (4.9) do not change for all $\mathbf{v}_B = \bar{\mathbf{v}}_B + \tilde{\mathbf{v}}_B$ with the same $\tilde{\mathbf{v}}_B$, it is sufficient to prove lemma for the vectors $\mathbf{v}_B = \tilde{\mathbf{v}}_B$. Under this condition, the proof follows the lines of the proof of Proposition 5.1 in [16]. Let vector $\boldsymbol{\varphi}_I$ be the solution of the system $\mathcal{A}_I \boldsymbol{\varphi}_I = \mathcal{A}_{IB} \tilde{\mathbf{v}}_B$, and $\boldsymbol{\varphi}$ has $\boldsymbol{\varphi}_I$ and $\mathbf{v}_B$ for the internal and boundary subvectors. We have the convergence estimate

$$\|\mathbf{u}_I^k - \boldsymbol{\varphi}_I\|_{\mathcal{A}_I} \leq \rho^k \|\boldsymbol{\varphi}_I\|_{\mathcal{A}_I}, \quad (4.10)$$

whence for $\mathbf{u}^k = \mathbf{u}_I^k + \mathbf{v}_B$ it follows

$$\|\mathbf{u}^k - \boldsymbol{\phi}\|_A \leq \rho^k (\|\boldsymbol{\phi}\|_A + \|\mathbf{v}_B\|_{A_B}), \quad (4.11)$$

if we take into account the spectral equivalence of the matrices $\mathcal{A}$ and A , see Lemma 1, and the triangular inequality. In (4.11), the vector $\mathbf{v}_B$ is considered as an element of U with zero entries for the internal degrees of freedom. In order to estimate second summand in the right part, we introduce the polynomial $\mathbf{v}_B \in \mathcal{U}_B(\tau_0)$, $\mathbf{v}_B \longleftrightarrow \mathbf{v}_B$, and the piece-wise bilinear function $\hat{\mathbf{v}}_B \in \mathcal{H}(\tau_0)$ coinciding with $\mathbf{v}_B$ at the nodes. Below, the squares of the norms $\|\cdot\|_{k, \partial\tau_0}$, $|\cdot|_{k, \partial\tau_0}$, $k = 1, 2$, are understood as the sums of squares of the same norms for faces. Equivalence of these norms for the functions $\mathbf{v}_B$ and $\hat{\mathbf{v}}_B$, as follows from Lemma 1, and direct estimation of $|\hat{\mathbf{v}}_B|_{1, \tau_0}$ lead to the estimates

$$\|\mathbf{v}_B\|_{A_B}^2 \equiv |\mathbf{v}_B|_{1, \tau_0}^2 < |\hat{\mathbf{v}}_B|_{1, \tau_0}^2 < p^2 \|\hat{\mathbf{v}}_B\|_{1, \partial\tau_0}^2 < p^2 \|\mathbf{v}_B\|_{1, \partial\tau_0}^2. \quad (4.12)$$

Then Markov's inequality applied face wise, the definition of the norm $\|\mathbf{v}_B\|_{1/2, \partial\tau_0}$, and the factor space argument allow us to obtain

$$\|\mathbf{v}_B\|_{A_B} < p^2 |\mathbf{v}_B|_{1/2, \partial\tau_0}. \quad (4.13)$$

Finally, by means of the trace and continuation theorems in polynomial spaces equipped with Sobolev's norms, see [37], [38], and the definition of $\mathbb{S}_B$, we come to

$$\|\mathbf{v}_B\|_{A_B} < p^2 \|\mathbf{v}_B\|_{\mathbb{S}_B} \quad (4.14)$$

and

$$\|\boldsymbol{\phi}\|_A \equiv \|\mathbf{v}_B\|_{\mathbb{S}_B} < |\mathbf{v}_B|_{1/2, \partial\tau_0}. \quad (4.15)$$

Combining (4.11), (4.14) and (4.15), we get

$$\|\mathbf{u}^k\|_A < (1 + \rho^k (1 + p^2)) \|\mathbf{v}_B\|_{\mathbb{S}_B},$$

hence the statement of Lemma immediately follows.

According to Lemma 1 and Theorem 3, the value of ρ is an absolute constant. Therefore, Lemma 4 and the conditions of the shape quasiuniformity, allow to make the following conclusion.

Corollary 2. If $\mathbb{B}_I = \mathcal{B}_{I, \text{sp}}$, and $k_0 \asymp (1 + \log p)$. Then

$$\|\mathbf{P}_{V_B \rightarrow V} \mathbf{v}_B\|_K \leq c_{\mathbb{P}} \|\mathbf{v}_B\|_{\mathbb{S}_B}, \quad c_{\mathbb{P}} = c c_{\mathbb{P}, 0}, \quad (4.16)$$

with the constant c , depending only on the conditions of the shape regularity.

The wire basket preconditioner $\mathcal{S}_W^B$ and the prolongation matrix $\mathbf{P}_{V_W \rightarrow V_B}$, appearing in (iii) and (v), respectively, were discussed in many papers (see, e.g., [24], [25], [6], [3]), where the choice of these components has been thoroughly analyzed. In their definition, we follow [6] more closely.

Matrices $\mathcal{S}_W^B$, $\mathbf{P}_{V_W \rightarrow V_B}$ are assembled of the scaled standard matrices defined for the reference element. For more definiteness, we assume that the nodes $\boldsymbol{\eta}_\alpha$ of the reference element are the GLL nodes. We denote by $\omega_W \in \omega$ the subset of α , corresponding to the nodes on the wire basket W_0 of τ_0 , by χ_α – the weights of the quadrature, which is assembled of GLL quadratures applied to each edge, and by $\mathbb{S}_{W_0}$ – the matrix of the quadratic form

$$\mathbf{v}_W^T \mathbb{S}_{W_0} \mathbf{v}_W = \inf_c \sum_{\alpha \in \omega_W} \chi_\alpha (\mathbf{v}_\alpha - c)^2, \quad (4.17)$$

where $\mathbf{v}_\alpha$ are the entries of $\mathbf{v}_W$. If D_0 is the diagonal matrix of the quadrature in (4.17) and $\mathbf{z}_1$ contains unity

for all entries, then

$$\mathbb{S}_{W_0} = \mathbb{D}_0 - \frac{\mathbb{D}_0 \mathbf{z}_1 (\mathbb{D}_0 \mathbf{z}_1)^T}{\mathbf{z}_1^T \mathbb{D}_0 \mathbf{z}_1} = \mathbb{D}_0 - \frac{1}{24} \mathbb{D}_0 \mathbf{z}_1 (\mathbb{D}_0 \mathbf{z}_1)^T \quad (4.18)$$

and $\mathcal{J}_w^B$ is assembled of the matrices

$$\mathcal{J}_w^{(r)} = h_r (1 + \log p) \mathbb{Q}_r \mathbb{S}_{W_0}. \quad (4.19)$$

For details of the solving procedure for the systems with the preconditioner $\mathcal{J}_w^B$ of the dimension $O(\mathcal{R}p) \times O(\mathcal{R}p)$, we refer to [3] and [6], where it is described up to the solver for the $O(\mathcal{R}) \times O(\mathcal{R})$ subsystem. According to (4.17), (4.19), each unknown in this subsystem is the constant $c = c_r$ corresponding to one element and is coupled only with the next neighboring elements. If the number of elements $\mathcal{R}$ is not fixed, it is sufficient for our purposes to assume that the arithmetical cost of the pointed out procedure does not exceed $O(\mathcal{R}p^3)$.

Let F_0 be the representative face for the reference element and $\mathbf{v}_{\partial F_0}$ be the vector with entries related to the nodes on ∂F_0 . By definition, the vector $\mathbf{1}_{F_0}$ contains 1-s for all internal nodes of F_0 and 0-s for all nodes of its boundary ∂F_0 , whereas the vector $\mathbf{v}_{F_0}$ with entries for all nodes of $\bar{F}_0$ is the result of the continuation of $\mathbf{v}_{\partial F_0}$ by zero entries to all internal nodes of the face. Let also ∇ be the mean value on ∂F_0 of the finite element function $\mathbf{v}_{\partial F_0} \longleftrightarrow \mathbf{v}_{\partial F_0}$, which, e.g., can be calculated by quadratures. Then the standard matrix $\mathbb{P}_{\partial F_0 \rightarrow \bar{F}_0}$ for the prolongation from the boundary ∂F_0 on the whole face $\bar{F}_0$ is defined as

$$\mathbb{P}_{\partial F_0 \rightarrow \bar{F}_0} \mathbf{v}_{\partial F_0} = \mathbf{v}_{F_0} + \nabla \mathbf{1}_{F_0}. \quad (4.20)$$

A slightly different prolongation is obtained, if ∇ is the mean value over ∂F_0 of the piece-wise linear function with the entries of $\mathbf{v}_{\partial F_0}$ for the nodal values. The prolongation matrix $\mathbb{P}_{V_w \rightarrow V_B}$ is uniquely defined by the condition that its restriction to each face $F_k \subset \Omega$ is $\mathbb{P}_{\partial F_0 \rightarrow \bar{F}_0}$.

We have completely defined the DD preconditioner-solver (4.4).

Theorem 5. *Under the stated assumptions, the DD preconditioner-solver $\mathcal{H}$ provides the condition number $\text{cond}[\mathcal{H}^{-1}\mathbf{K}] \leq c(1 + \log p)^2$, whereas the arithmetical cost of the operation $\mathcal{H}^{-1}\mathbf{f}$ for any $\mathbf{f}$ is $O(p^3(1 + \log p)\mathcal{R})$.*

Proof. By Theorem 3 and conditions of the generalized shape regularity,

$$\underline{\beta}_I \mathcal{H}_I \leq \mathbf{K}_I \leq \bar{\beta}_I \mathcal{H}_I \quad (4.21)$$

with positive constants $\underline{\beta}_I, \bar{\beta}_I$, depending only on the mentioned conditions. Therefore, taking into account Corollary 2 and general results on relative condition numbers of DD preconditioners, we reduce the estimate of $\text{cond}[\mathcal{H}^{-1}\mathbf{K}]$ to deriving the bounds for $\underline{\gamma}_B, \bar{\gamma}_B > 0$ in the inequalities

$$\underline{\gamma}_B \mathcal{J}_B \leq \mathbf{S}_B \leq \bar{\gamma}_B \mathcal{J}_B. \quad (4.22)$$

Indeed, according to (4.21), (4.22), Corollary 2 and, e.g., Theorem 8 of [39], we obtain

$$\underline{\beta} \mathcal{H} \leq \mathbf{K} \leq \bar{\beta} \mathcal{H},$$

with

$$\begin{aligned} \underline{\beta} &= \min\{\underline{\beta}_I, \underline{\gamma}_B\} (1 - \sqrt{1 - c_{\mathbb{P}}^{-2}}) > c_{\mathbb{P}}^{-2} \min\{\underline{\beta}_I, \underline{\gamma}_B\}, \\ \bar{\beta} &= \max\{\bar{\beta}_I, \bar{\gamma}_B\} (1 + \sqrt{1 - c_{\mathbb{P}}^{-2}}) < 2 \max\{\bar{\beta}_I, \bar{\gamma}_B\}. \end{aligned} \quad (4.23)$$

The proof of the bound for the relative condition number is completed by the use of the bounds for $\underline{\gamma}_B$, $\bar{\gamma}_B$ given in Theorem 6 below, which allow us to write

$$\underline{\beta} \geq c_{\mathbb{P}}^{-2} \underline{c} \frac{1}{(1 + \log p)^2}, \quad \bar{\beta} \leq 2 \max\{\bar{\beta}_I, \bar{c}\}. \quad (4.24)$$

The operation $\mathcal{H}^{-1}\mathbf{f}$ involves operations with the preconditioners and prolongations i)–v) with the following arithmetic costs:

- (i) – $O(p^3\mathcal{R})$ according to Theorem 3 and the definition of $\mathcal{H}_I$,
- (ii) – $O(p^2\mathcal{R})$ according to Theorem 4 and (4.5),
- (iii) – $O(p^3\mathcal{R})$ according to the above assumption on the cost of the wire basket preconditioner,
- (iv) – $O(p^3(1 + \log p)\mathcal{R})$ according to Theorem 3 and Corollary 2,
- (v) – $O(p^2\mathcal{R})$ according to the definition of the prolongation $P_{V_W \rightarrow V_B}$ by means of (4.20).

Taking into account these contributions, we conclude that the estimate of theorem for the arithmetic cost of solving the system $\mathcal{H}\mathbf{v} = \mathbf{f}$ with any $\mathbf{f}$, holds as well.

Theorem 6. *Let the preconditioner-solver $\mathcal{S}_B$ be defined according to (4.4)–(4.6) with $\mathcal{S}_0$ defined as in Theorem 4. Then*

$$\underline{\gamma}_B \geq \underline{c} \frac{1}{(1 + \log p)^2}, \quad \bar{\gamma}_B \leq \bar{c} \quad (4.25)$$

with positive constants $\underline{c}$, $\bar{c}$ depending only on the generalized shape regularity conditions.

Proof. Bounds (4.25) remain true for any Schur complement preconditioner $\mathcal{S}_B$, defined as in (4.4)–(4.6), if in (4.5) $\mathcal{S}_0$ is any matrix for which (3.2) holds uniformly in p . This result is found in [3], [6]. Therefore, the proof follows from Theorem 4 of this paper and, e.g., from [6, Theorem 2].

Compared to hierarchical elements, spectral elements are less flexible for adaptive computations. With the help of the subproblem solvers presented in the paper, DD preconditioners-solvers can be designed without losses of efficiency for a range of admissible adaptive spectral discretizations. The adjustment of the analysis of their condition numbers and computational cost can be completed along the lines of [16].

REFERENCE

1. Babuška I., Craig A., Mandel J., Pitkäranta J. Efficient preconditioning for the p -version finite element method in two dimensions // SIAM J. Numer. Anal. 1991. V. 28. № 3. P. 624–661.
2. Ainsworth M. A preconditioner based on domain decomposition for $h - p$ finite element approximation on quasi-uniform meshes // SIAM J. Numer. Anal. 1996. V. 33. P. 1358–1376.
3. Pavarino L.E., Widlund O.B. Polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions // SIAM J. Numer. Anal. 1996. V. 37. № 4. P. 1303–1335.
4. Ivanov S.A., Korneev V.G. Preconditioning in the domain decomposition methods for the p -version with the hierarchical bases // Matematicheskoe modelirovanie (Math. Modelling). 1996. V. 8. № 9. P. 63–73.
5. Korneev V.G., Jensen S. Preconditioning of the p -version of the finite element method // Computer Meth. Appl. Mech. Engrg. 1997. V. 150. № 1–4. P. 215–238.
6. Casarin M.A. Quasi-optimal Schwarz methods for the conforming spectral element discretization // SIAM J. Numer. Anal. 1997. V. 34. № 6. P. 2482–2502.
7. Oden J.T., Patra A., Feng Y.S. Parallel Domain decomposition solver for adaptive hp finite element methods // SIAM J. Numer. Anal. 1997. V. 34. P. 2090–2118.
8. Korneev V.G., Flaherty J., Oden T., Fish J. Hp-version additive Schwarz algorithms on triangular meshes // Matematicheskoe modelirovanie (Math. Modelling). 2002. V. 14. № 2. P. 61–94.
9. Toselli A., Widlund O. Domain decomposition methods – algorithms and theory. Berlin etc.: Springer, 2005.
10. Orzag A. Spectral methods for problems in complex geometries // J. Comput. Phys. 1980. V. 37. P. 70–92.
11. Korneev V.G. Almost optimal method for solving Dirichlet problems on subdomains of decomposition of hierarchical hp -version // Differentsial'nye uravnenia (Different. Equations). 2001. V. 37. № 7. P. 1–10.
12. Korneev V.G. Local Dirichlet problems on subdomains of decomposition in hp discretizations, and optimal algorithms for their solution // Matematicheskoe modelirovanie (Math. Modelling). 2002. V. 14. № 5. P. 51–74.
13. Beuchler S. Multigrid solver for the inner problem in domain decomposition methods for p -FEM // SIAM J. Numer. Anal. 2002. V. 40. № 4. P. 928–944.

14. *Beuchler S., Schneider R., Schwab C.* Multiresolution weighted norm equivalences and applications // *Numer. Math.* 2004. V. 98. № 1. P. 67–97.
15. *Korneev V., Langer U., Xanthis L.* On fast domain decomposition solving procedures for *hp*-discretizations of 3-d elliptic problems // *Comput. Meth. Appl. Math.* 2003. V. 3. № 4. P. 536–559.
16. *Korneev V., Langer U., Xanthis L.* Fast adaptive domain decomposition algorithms for *hp*-discretizations of 2-d and 3-d elliptic equations: recent advances // *Hermis-μπ: Internat. J. Comput. Math. and its Applic.* 2003. V. 4. P. 27–44.
17. *Korneev V., Xanthis L., Anoufrieu I.* Solving finite element *hp* discretizations of elliptic problems by fast DD algorithms // *Tr. SPbGPU (Trans.: St. Petersburg State Polytechn. Univ., 2002. № 485. P. 126–153.)*
18. *Korneev V.G., Xanthis L., Anufriev I.E.* Hierarchical and Lagrange *hp* discretizations and fast domain decomposition solvers for them // *SFB Rept 02-18. Linz: Johannes Kepler Univ. Linz, 2002.*
19. *Anoufrieu I., Korneev V., Rytov A.* Computer testing of fast domain decomposition algorithms for *hp* discretizations of 2-nd order elliptic equations // *Proc. Hellenic-European Conf. Computer Math. and its Applic. HERCMA 2003. Athens: LEA Pubs., 2003. V. 1. P. 99–108.*
20. *Anoufrieu I., Korneev V.* Fast domain decomposition solver for degenerating PDE of 4th order in 3D domain // *Matematika Balkanika.* 2006. V. 6. № 1. P. 101–112.
21. *Shen J., Wang F., Xu J.* A finite element multigrid preconditioner for Chebyshev-collocation methods // *Appl. Numer. Math.* 2000. V. 33. P. 471–477.
22. *Korneev V., Rytov A.* On the interrelation between fast solvers for spectral and hierarchical *p* elements // *Hermis-μπ: Internat. J. Comput. Math. and its Applic.* 2005. V. 6. P. 99–113. <http://www.aueb.gr/pympe/hermis/hermis-volume-6/>.
23. *Korneev V., Rytov A.* On existence of the essential interrelation between spectral and hierarchical elements // *Mesh Methods for Boundary Value Problems and Applic.: Proc. 6-th All-Russian Seminar (Materialy 6-ogo vserossiiskogo seminar).* Kazan: Kazan State Univ., 2005. P. 141–150.
24. *Smith B.F.* A domain decomposition algorithm for elliptic problems in three dimentions // *Numer. Math.* 1991. V. 60. № 4. P. 219–234.
25. *Dryja M., Smith B.F., Widlund O.B.* Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions // *SIAM J. Numer. Analys.* 1994. V. 31. № 6. P. 1662–1694.
26. *Bramble J.H., Pasciak J.E., Schatz A.H.* The construction of preconditioners for elliptic problems by substructuring. I // *Math. Comput.* 1986. V. 47. № 175. P. 103–134.
27. *Bramble J.H., Pasciak J.E., Schatz A.H.* The constraction of preconditioners for elliptic problems by substructuring. II // *Math. Comput.* 1987. V. 49. № 17. P. 1–16.
28. *Bramble J.H., Pasciak J.E., Schatz A.H.* The constraction of preconditioners for elliptic problems by substructuring. III // *Math. Comput.* 1988. V. 51. № 184. P. 415–430.
29. *Bramble J.H., Pasciak J.E., Schatz A.H.* The constraction of preconditioners for elliptic problems by substructuring. IV // *Math. Comput.* 1989. V. 53. № 187. P. 1–24.
30. *Bernardi Ch., Maday Y.* Polynomial interpolation results in Sobolev spaces // *J. Comput. and Appl. Math.* 1992. V. 43. P. 53–80.
31. *Bernardi Ch., Maday Y.* Approximations spectrales de problèmes aux limites elliptiques // *Math. Appl. Vol. 10. Paris: Springer, 1992.*
32. *Canuto C.* Stabilizaiton of spectral methods by finite element bubble functions // *Comput. Meth. Appl. Mech. Engrg.* 1994. V. 116. P. 13–26.
33. *Grisvard P.* Elliptic problems in nonsmooth domains. Boston: Pitman Publishing, 1985.
34. *Lions J.-L., Magenes E.M.* Nonhomogeneous boundary value problems and applications. Vol. 1. Berlin etc.: Springer, 1972.
35. *Ciarlet P.* The finite element method for elliptic problems. Amsterdam etc.: North-Holland Publishing Co., 1978.
36. *Korneev V.G.* O postroienii variačionno-raznostnyh shem vysokogo poriadka točnosti // *Vestn. Leningradskogo Univ.* 1970. V. 25. № 19. P. 28–40 (in Russian). (Engl. trans.: On the construction of variational-difference schemes of higher orders of accuracy // *Vestnik Leningrad Univ. Math.* 1976. V. 3. P. 301–316.)
37. *Bernardi Ch., Dauge M., Maday Y.* Relèvements de traces préservant les polinômes // *C. R. acad. Sci. Paris, Sér. I. Math.* 1992. V. 315. P. 333–338.
38. *Bernardi Ch., Dauge M., Maday Y.* Polymonials in weighted Sobolev spaces: Basics and trace liftings // *Publications du Laboratoire d'Analyse Numerique, R92039. Paris: Univ. Pierre et Marie Curie, 1993.*
39. *Korneev V., Langer U.* Domain decomposition methods and preconditioning // *Encyclopedia Comput. Mech. Vol. 1. John Wiley & Sons, Ltd., 2004. P. 617–647.*

УДК 519.634

АСИМПТОТИКА РЕШЕНИЯ ВРЕМЕННОГО УРАВНЕНИЯ ШРЁДИНГЕРА С МАЛОЙ КОНСТАНТОЙ ПЛАНКА

© 2007 г. А. С. Омуралиев

(720044 Бишкек, пр-т. Мира, 56, Кыргызско-Турецкий ун-т “Манас”)

e-mail: asan.omuraliev@manas.kg

Поступила в редакцию 20.12.2006 г.
Переработанный вариант 14.03.2007 г.

Строится регуляризованная асимптотика решения временного уравнения Шрёдингера, когда малая константа Планка стоит перед пространственной производной. Показано, что асимптотика решения содержит погранслойную функцию, имеющую быстро осциллирующий характер изменения. Библ. 6.

Ключевые слова: сингулярно возмущенное временное уравнение Шрёдингера, регуляризованная асимптотика решения.

ВВЕДЕНИЕ

В данной статье строится регуляризованная асимптотика решения первой краевой задачи для временного уравнения Шрёдингера:

$$L_\varepsilon u(x, t, \varepsilon) \equiv i\partial_t u(x, t, \varepsilon) + \varepsilon^2 a(x) \partial_x^2 u(x, t, \varepsilon) - b(x, t) u(x, t, \varepsilon) = f(x, t), \quad (1)$$

$$(x, t) \in \Omega, \quad u(x, t, \varepsilon)|_{t=0} = h(x), \quad u(x, t, \varepsilon)|_{x=0} = u(x, t, \varepsilon)|_{x=1} = 0,$$

где i – мнимая единица, в наших обозначениях $\varepsilon > 0$ – достаточно малая постоянная Планка, $u(x, t, \varepsilon)$ – волновая функция, которая определяет состояние нерелятивистской системы в момент времени t в двумерном пространстве-времени при квантово-механическом изучении (см. [1]):

$$\partial_t = \frac{\partial}{\partial t}, \quad \partial_x^2 = \frac{\partial^2}{\partial x^2}, \quad \Omega = \{(x, t) : x \in (0, 1), t \in (0, T]\}.$$

Временное уравнение Шрёдингера с малым параметром при всех производных изучено в [2], где построена асимптотика решения задачи Коши с быстро осциллирующими начальными данными для такого уравнения. Не вникая в физический смысл, мы рассматриваем случай, когда малый параметр отсутствует при временной производной. Сделанное упрощение позволило построить регуляризованную асимптотику решения (см. [3]) поставленной задачи (1).

Относительно заданных функций предполагается, что они являются достаточно гладкими и функция $a(x) > 0 \forall x \in [0, 1]$, а также выполнены $h(0) = h(1) = 0$.

1. РЕГУЛЯРИЗАЦИЯ ЗАДАЧИ

Для регуляризации задачи введем регуляризующие переменные (см. [3]) по формулам

$$\xi_j = \frac{\varphi_j(x)}{\varepsilon} \equiv \frac{(-1)^{j-1}}{\varepsilon} \int_{j-1}^x \frac{ds}{\sqrt{a(s)}}, \quad j = 1, 2, \quad (2)$$

и вместо искомой функции $u(x, t, \varepsilon)$ будем изучать расширенную функцию $\tilde{u}(x, t, \xi, \varepsilon)$, $\xi = (\xi_1, \xi_2)$, такую, чтобы сужение посредством регуляризующих функций совпадало с искомой функцией

$$\tilde{u}(x, t, \xi, \varepsilon)|_{\xi = \varphi(x)/\varepsilon} \equiv u(x, t, \varepsilon), \quad \varphi(x) = (\varphi_1(x), \varphi_2(x)). \quad (3)$$

Отсюда с учетом (2) найдем производные по t и x , тогда для расширенной функции $\tilde{u}(x, t, \xi, \varepsilon)$, на основании (1), (3), естественно поставить расширенную задачу

$$\tilde{L}_\varepsilon \tilde{u}(x, t, \xi, \varepsilon) \equiv i\partial_t \tilde{u}(x, t, \xi, \varepsilon) + D_\xi \tilde{u}(x, t, \xi, \varepsilon) + \varepsilon L_\xi \tilde{u}(x, t, \xi, \varepsilon) + \varepsilon^2 L_x \tilde{u}(x, t, \xi, \varepsilon) = f(x, t),$$

$$(x, t, \xi) \in Q, \quad (4)$$

$$\tilde{u}(x, t, \xi, \varepsilon)|_{t=0} = h(x), \quad \tilde{u}(x, t, \xi, \varepsilon)|_{x=0, \xi_1=0} = 0, \quad \tilde{u}(x, t, \xi, \varepsilon)|_{x=1, \xi_2=0} = 0,$$

где

$$Q = \{(x, t, \xi) : (x, t) \in \Omega, \xi \in (0, \infty)^2\}, \quad D_\xi \equiv \sum_{j=1}^2 \partial_{\xi_j}^2 - b(x, t),$$

$$L_\xi \equiv a(x) \sum_{j=1}^2 L_{\xi_j}, \quad L_x \equiv a(x) \partial_x^2, \quad L_{\xi_j} \equiv 2\varphi_j'(x) \partial_{x, \xi_j}^2 + \varphi_j''(x) \partial_{\xi_j}.$$

Решение расширенной задачи (4) будем искать в виде ряда

$$\tilde{u}(x, t, \xi, \varepsilon) = \sum_{k=0}^{\infty} \varepsilon^k u_k(x, t, \xi); \quad (5)$$

тогда для коэффициентов этого ряда получим следующие итерационные задачи:

$$\begin{aligned} Tu_0(x, t, \xi) &= f(x, t), \quad u_0(x, t, \xi)|_{t=0} = h(x), \\ u_0(x, t, \xi)|_{x=0, \xi_1=0} &= 0, \quad u_0(x, t, \xi)|_{x=1, \xi_2=0} = 0, \\ Tu_k(x, t, \xi) &= -L_\xi u_{k-1}(x, t, \xi) - L_x u_{k-2}(x, t, \xi), \\ Tu &\equiv i\partial_t u + D_\xi u, \\ u_k|_{t=0} &= u_k|_{x=0, \xi_1=0} = u_k|_{x=1, \xi_2=0} = 0, \quad k \geq 1. \end{aligned} \quad (6)$$

2. РЕШЕНИЕ ИТЕРАЦИОННЫХ ЗАДАЧ

Итерационные задачи (6) будем решать в классе функций

$$U = \left\{ u(x, t, \xi) : u = v(x, t) + \sum_{j=1}^2 c_j(x, t) \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right), \right.$$

$$\left. v(x, t), c_j(x, t) \in C^\infty(\bar{\Omega}), \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right) = \frac{2}{\sqrt{\pi}} \int_{\frac{\xi_j}{2\sqrt{it}}}^{\infty} \exp(-s^2) ds \right\}.$$

В силу однородности, уравнение (6) при $k=0$ разрешимо в классе U и его решение представимо в виде

$$u_0(x, t, \xi) = v_0(x, t) + \sum_{j=1}^2 c_{0,j}(x, t) \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right),$$

если функции $v_0(x, t)$, $c_{0,j}(x, t)$ будут выбраны как решения уравнений

$$\begin{aligned} i\partial_t v_0(x, t) &= b(x, t) v_0(x, t) + f(x, t), \\ i\partial_t c_{0,j}(x, t) &= b(x, t) c_{0,j}(x, t). \end{aligned} \quad (7)$$

Начальные условия для этих уравнений определяются из следующих соотношений:

$$\begin{aligned} v_0(x, t)|_{t=0} &= h(x), \quad c_{0,j}(x, t)|_{t=0} = c_{0,j}^0(x), \quad j = 1, 2, \\ c_{0,1}(x, t)|_{x=0} + c_{0,2}(x, t)|_{x=0} \operatorname{erfc}\left(\frac{\varphi_2(0)}{2\sqrt{it}}\right) + v_0(0, t) &= 0, \\ c_{0,1}(x, t)|_{x=1} \operatorname{erfc}\left(\frac{\varphi_1(1)}{2\varepsilon\sqrt{it}}\right) + c_{0,2}(x, t)|_{x=1} + v_0(1, t) &= 0. \end{aligned} \quad (8)$$

Так как

$$\lim_{t \rightarrow 0} \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right) = 0,$$

то для функций $c_{0,j}(x, t)$, $j = 1, 2$, начальные условия при $t = 0$ можно задать произвольно, т.е. в виде $c_{0,j}(x, t)|_{t=0} = c_{0,j}^0(x)$, где $c_{0,j}^0(x)$, $j = 1, 2$, – произвольные функции. Эти произвольные функции обеспечат нам разрешимость следующего ($k = 1$) итерационного уравнения в классе U . Обеспечивая разрешимость, мы полагаем

$$L_{\xi,j} c_{0,j}(x, t) \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right) = 0, \quad j = 1, 2. \quad (9)$$

Этот член, входящий в правую часть следующего итерационного уравнения, содержит производную $\partial_{\xi_j} \operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right)$, которая приведет к появлению в решении секулярных членов, особенности которых будут расти с ростом номера итерации.

Решив уравнения (7) при соответствующих начальных условиях из (8), найдем $c_{0,j}(x, t)$ и подставим его в соотношение (9). Выполнение соотношения (9) можно обеспечить за счет выбора произвольной начальной функции $c_{0,j}^0(x)$, которую выберем как решение уравнения

$$q_j^1(x) \frac{dc_{0,j}^0(x)}{dx} + \left[q_j^1(x) \frac{\partial B(x, t)}{\partial x} + q_j^2(x) \right] c_{0,j}^0(x) = 0, \quad j = 1, 2,$$

$$B(x, t) = \frac{1}{i} \int_0^t b(x, s) ds, \quad q_j^1(x) = 2\varphi_j'(x), \quad q_j^2(x) = \varphi_j''(x),$$

где t – параметр.

Решая это уравнение при начальном условии

$$c_{0,j}^0(x)|_{x=j-1} = p_0(j-1),$$

находим

$$c_{0,j}^0(x) = p_0(j-1) \exp[-B_{1,j}(x, t)],$$

$$B_{1,j}(x, t) = \int_0^x \left[\frac{\partial B(s, t)}{\partial s} + \frac{q_j^2(s)}{q_j^1(s)} \right] ds$$

и для функции $c_{0,j}(x, t)$ получаем выражение

$$c_{0,j}(x, t) = p_0(j-1) \exp[B(x, t) - B_{1,j}(x, t)], \quad j = 1, 2.$$

Далее найденные значения $c_{0,j}(x, t)$ и $v_0(x, t)$ подставляем во второе и в третье уравнения из (8),

затем, сокращая первое соотношение на $\exp[B(0, t)]$, а второе на $\exp[B(1, t)]$, получаем

$$\begin{aligned}\beta_{1,1}(0, t)p_0(0) + \beta_{1,2}(0, t)p_0(1)\operatorname{erfc}\left(\frac{\varphi_2(0)}{2\varepsilon\sqrt{it}}\right) &= H_1(0, t), \\ \beta_{1,1}(1, t)p_0(0)\operatorname{erfc}\left(\frac{\varphi_1(1)}{2\varepsilon\sqrt{it}}\right) + \beta_{1,2}(1, t)p_0(1) &= H_1(1, t),\end{aligned}\quad (10)$$

где

$$\begin{aligned}\beta_{1,j}(x, t) &= \exp[-B_{1,j}(x, t)], \\ H_1(x, t) &= h(x) + \frac{1}{i} \int_0^t f(x, s) \exp[-B(x, s)] ds.\end{aligned}$$

Определитель этой системы

$$\Delta = \begin{vmatrix} \beta_{1,1}(0, t) & \beta_{1,2}(0, t)\operatorname{erfc}\left(\frac{\varphi_2(0)}{2\varepsilon\sqrt{it}}\right) \\ \beta_{1,1}(1, t)\operatorname{erfc}\left(\frac{\varphi_1(1)}{2\varepsilon\sqrt{it}}\right) & \beta_{1,2}(1, t) \end{vmatrix}$$

отличен от нуля при достаточно малых $\varepsilon > 0$ и $\forall t \in [0, T]$, поэтому система однозначно разрешима относительно $p_0(0)$, $p_0(1)$. Таким образом, мы однозначным образом определили главный член асимптотики.

Из асимптотического представления функции (см. [4, с. 443])

$$\operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right) \cong \left(\cos\frac{\xi_j^2}{4t} + i\sin\frac{\xi_j^2}{4t}\right)\left(\frac{a_1}{\xi_j/\sqrt{t}} + \frac{a_2}{\xi_j/\sqrt{t}} + \dots\right),$$

которая описывает пограничный слой, заключаем, что он является произведением двух типов функций. Одна из них имеет быстро осциллирующий, а вторая – степенной характер изменения при $\xi_j \rightarrow \infty$.

Рассмотрим теперь итерационное уравнение (6) при $k = 1$; на основании (9) это уравнение запишется в виде

$$Tu_1(x, t, \xi) = 0.$$

В силу однородности, оно разрешимо в U , и его решение представимо в виде

$$u_1(x, t, \xi) = v_1(x, t) + \sum_{j=1}^2 c_{1,j}(x, t)\operatorname{erfc}\left(\frac{\xi_j}{2\sqrt{it}}\right),$$

если функции $v_1(x, t)$, $c_{1,j}(x, t)$ будут решениями уравнений

$$\begin{aligned}i\partial_t v_1(x, t) &= b(x, t)v_1(x, t), \\ i\partial_t c_{1,j}(x, t) &= b(x, t)c_{1,j}(x, t), \quad j = 1, 2,\end{aligned}$$

при начальных условиях

$$v_1(x, t)|_{t=0} = 0, \quad c_{1,j}(x, t)|_{t=0} = c_{1,j}^0(x), \quad j = 1, 2.$$

Первое уравнение имеет тривиальное решение $v_1(x, t) = 0$, а решение второго уравнения запишется в виде

$$c_{1,j}(x, t) = c_{1,j}^0(x) \exp[B(x, t)]. \quad (11)$$

Обеспечивая разрешимость следующего итерационного уравнения из (6) при $k = 2$, из тех же соображений, что и при получении уравнения (9), относительно $c_{1,j}^0(x)$ получаем уравнение

$$q_j^1(x) \frac{dc_{1,j}^0(x)}{dx} + [q_j^1(x) \partial_x(B(x, t)) + q_j^2(x)] c_{1,j}^0(x) = 0, \quad j = 1, 2.$$

Решив это уравнение при начальном условии

$$c_{1,j}^0(x)|_{x=j-1} = p_1(j-1),$$

найдем функцию $c_{1,j}^0(x)$, затем ее подставим в (11). Далее полученное при этом значение функции $c_{1,j}(x, t)$ подставим в систему, которая получается удовлетворением функции $u_1(x, t, \xi)$ граничным условиям. Эта система (10) является однородной относительно $\{p_1(0), p_1(1)\}$. В силу того, что определитель этой системы отличен от нуля, она имеет нулевое решение $p_1(0) = 0, p_1(1) = 0$, тогда и функция $c_{1,j}^0(x)$ обратится в нуль, а потому и $c_{1,j}(x, t) = 0, j = 1, 2$. В результате получаем, что $u_1(x, t, \xi) = 0$.

Далее, повторяя описанный выше процесс, можно определить все коэффициенты частичной суммы ряда (5), причем коэффициенты с нечетными индексами обратятся в нуль.

3. ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА

Чтобы получить решение исходной задачи, в полученной частичной сумме ряда (5) произведем сужение посредством регуляризующих функций:

$$\begin{aligned} u_{\varepsilon,n}\left(x, t, \frac{\varphi(x)}{\varepsilon}\right) &= \sum_{k=0}^n \varepsilon^{2k} \left[v_{2k}(x, t) + \sum_{j=1}^2 c_{2k,j}(x, t) \operatorname{erfc}\left(\frac{\varphi_j(x)}{2\varepsilon\sqrt{it}}\right) \right] = \\ &= \sum_{k=0}^n \varepsilon^{2k} \left[v_{2k}(x, t) + \frac{2}{\sqrt{\pi}} \sum_{j=1}^2 c_{2k,j}(x, t) \int_{\frac{\varphi_j(x)}{2\varepsilon\sqrt{it}}}^{\infty} \exp(-s^2) ds \right]. \end{aligned}$$

Производя сужение в расширенной задаче (4) и учитывая, что

$$(\tilde{L}_\varepsilon \tilde{u}(x, t, \xi, \varepsilon))|_{\xi = \varphi(x)/\varepsilon} \equiv L_\varepsilon u(x, t, \varepsilon),$$

относительно остаточного члена

$$R_\varepsilon(x, t) = u(x, t, \varepsilon) - u_{\varepsilon,n}\left(x, t, \frac{\varphi(x)}{\varepsilon}\right)$$

получаем задачу

$$\begin{aligned} L_\varepsilon R_\varepsilon(x, t) &= \varepsilon^{2n+2} g_n(x, t, \varepsilon), \quad (x, t) \in \Omega, \\ R_\varepsilon(x, t)|_{t=0} &= R_\varepsilon(x, t)|_{x=0} = R_\varepsilon(x, t)|_{x=1} = 0. \end{aligned}$$

На основании приведенных построений заключаем, что известная функция $g_n(x, t, \varepsilon)$ равномерно ограничена по ε и непрерывна в области $\bar{\Omega}$ для любого номера $n = 0, 1, \dots$. Справедлива

Теорема. Пусть заданные функции достаточно гладкие и $a(x) > 0 \forall x \in [0, 1], h(0) = h(1) = 0$. Тогда $u_{\varepsilon,n}(x, t, \varphi(x)/\varepsilon)$ является асимптотическим решением исходной задачи, т.е. для доста-

точно малых $\varepsilon > 0 \forall (x, t) \in \bar{\Omega}, n = 0, 1, \dots$, справедлива оценка

$$|u(x, t, \varepsilon) - u_{\varepsilon, n}(x, t, \varphi(x)/\varepsilon)| < c\varepsilon^{2n+2}.$$

Доказательство данной теоремы основано на принципе максимума, доказанного в [5], и проводится совершенно аналогично теореме об оценке остаточного члена из [6].

СПИСОК ЛИТЕРАТУРЫ

1. Миллер У. Симметрия и разделения переменных. М.: Мир, 1981.
2. Маслов В.П. Асимптотические методы и теория возмущений. М.: Наука, 1988.
3. Ломов С.А. Введение в общую теорию сингулярных возмущений. М.: Наука, 1981.
4. Лаврентьев М.А., Шабат Б.В. Методы теории функций комплексного переменного. М.: Наука, 1987.
5. Ладыженская О.А., Солонников В.А., Уральцева Н.Н. Линейные и квазилинейные уравнения параболического типа. М.: Наука, 1967.
6. Омуралиев А.С. Регуляризация двумерной сингулярно возмущенной параболической задачи // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 8. С. 1423–1432.

УДК 519.634

КОНСЕРВАТИВНАЯ РАЗНОСТНАЯ СХЕМА ДЛЯ ЗАДАЧИ РАСПРОСТРАНЕНИЯ ФЕМТОСЕКУНДНОГО ЛАЗЕРНОГО ИМПУЛЬСА С АКСИАЛЬНО-СИММЕТРИЧНЫМ ПРОФИЛЕМ В СРЕДЕ С КУБИЧНОЙ НЕЛИНЕЙНОСТЬЮ¹⁾

© 2007 г. А. Г. Волков, В. А. Трофимов

(119992 Москва, Ленинские горы, МГУ, ф-т ВММК)

e-mail: agvolkov@cs.msu.su; vatro@cs.msu.su

Поступила в редакцию 27.03.2007 г.

Построены консервативные разностные схемы для задачи распространения фемтосекундного лазерного импульса в кубично-нелинейной среде в аксиально-симметричном случае с учетом временной дисперсии нелинейного отклика среды. Данный процесс описывается нелинейным уравнением Шрёдингера, содержащим производную по времени от нелинейного слагаемого. Записаны инварианты дифференциальной задачи. Показано, что разностные аналоги этих инвариантов сохраняются на решении предложенных схем разностной задачи. В качестве примера приведены результаты компьютерного моделирования самофокусировки светового пучка фемтосекундной длительности. Библ. 24. Фиг. 1.

Ключевые слова: фемтосекундный лазерный импульс, комбинированное нелинейное уравнение Шрёдингера, аксиально-симметричный случай, самофокусировка, кубичная нелинейность.

1. ВВЕДЕНИЕ

Как известно, в настоящее время все большее применение в различных областях науки находят фемтосекундные лазерные импульсы (см. [1]–[5]). Это обусловлено их свойствами: сверхкороткой длительностью, позволяющей исследовать быстро протекающие процессы, и высокой интенсивностью, которая может существенно превосходить интенсивность внутриатомного поля. Широкое применение такие импульсы находят и в задачах передачи информации по оптическим волокнам (см. [5]).

При математическом моделировании распространения фемтосекундных импульсов в нелинейной среде описывается так называемым комбинированным нелинейным уравнением Шрёдингера (КНУШ, см. [5]), которое отличается от обычного нелинейного уравнения Шрёдингера наличием производной по времени от нелинейного отклика среды, если интенсивность импульса не превосходит некоторой величины, которая соответствует пробойю вещества, т.е. его ионизации. Несмотря на широкое применение данного уравнения в литературе, разностные схемы для его решения широко не обсуждались. Следует отметить, что ранее (см. [6]) предложена консервативная разностная схема для случая распространения светового импульса в оптических волокнах (координаты (z, t)). На ее решении сохраняются разностные аналоги имеющихся инвариантов (см. [7]–[9]). Заметим, что для ее построения использовалось предложенное в [9] преобразование. Проведенное в [10] сравнение консервативной разностной схемы и широко используемого в литературе метода расщепления продемонстрировало ее преимущества. В отличие от перечисленных выше работ в настоящей статье рассматривается задача распространения фемтосекундного импульса с учетом его пространственного распределения (координаты (z, r, t)) в среде, обладающей перечисленными выше нелинейными свойствами. Для КНУШ записаны и доказаны инварианты дифференциальной задачи, а также их разностные аналоги. Построенная разностная схема обладает свойством консервативности, что показано ниже.

Разностная схема реализована в виде программы, с помощью которой выполнено компьютерное моделирование нестационарной самофокусировки аксиально-симметричного гауссова пучка. Как известно (см. [11]–[21]), она проявляется во многих задачах взаимодействия лазерного излучения с веществом и является основным препятствием, в частности, для повышения эффективности преобразования световой энергии мощных оптических пучков. Характерной осо-

¹⁾ Работа выполнена при частичной финансовой поддержке РФФИ (код проекта 05-01-00507).

бенностью этой задачи является неограниченный рост интенсивности в нелинейном фокусе. Для его ограничения в литературе предложено учитывать непараксиальный подход (см. [15]), насыщение нелинейности либо нелинейное поглощение (см. [13]), специальным образом выполненное пространственное распределение показателя преломления (см. [21]), генерацию встречной волны вблизи нелинейного фокуса (см. [19]). В настоящей работе изучено влияние дисперсии нелинейного отклика среды на самофокусировку оптического излучения. Показано, что при определенных условиях имеет место ограничение интенсивности в нелинейном фокусе либо изменении продольной координаты его формирования.

2. ОСНОВНЫЕ УРАВНЕНИЯ И ИНВАРИАНТЫ

Распространение оптического излучения фемтосекундной длительности с учетом его пространственного распределения в аксиально-симметричном случае описывается следующим безразмерным комбинированным уравнением Шрёдингера:

$$\begin{aligned} \mathcal{E} &= [A \exp(i\omega t - ikz) + \text{к.с.}]/2, \\ \frac{\partial A}{\partial z} + iD_r \Delta_r A + iD \frac{\partial^2 A}{\partial t^2} + i\alpha |A|^2 A + \alpha \gamma \frac{\partial}{\partial t}(|A|^2 A) &= 0, \quad z > 0, \quad 0 < t < L_t, \quad 0 < r < R, \end{aligned} \quad (1)$$

с начальными и граничными условиями

$$A|_{z=0} = A_0(r, t), \quad r \frac{\partial A}{\partial r} \Big|_{r \rightarrow 0} = A|_{r=R} = A|_{t=0, L_t} = 0. \quad (2)$$

Здесь к.с. – комплексное сопряжение, $\mathcal{E}$ – безразмерная напряженность электрического поля, ω и k – соответственно, безразмерная частота и волновое число светового импульса, $A(z, r, t)$ – нормированная на корень из максимальной интенсивности светового импульса $A_{0m} = \sqrt{I_0}$ на входе в нелинейную среду комплексная амплитуда импульса, распространяющегося вдоль координаты z , которая измеряется в единицах дифракционной длины $l_{\text{dif}} = 2ka^2$, r – поперечная координата, нормированная на характерный радиус a начального профиля пучка на входе в нелинейную среду, R – ее максимальное значение, $\Delta_r = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \right)$ – поперечный оператор Лапласа, записанный в полярных координатах в аксиально-симметричном случае, t – нормированное на длительность τ_p падающего на нелинейную среду импульса время в сопровождающей его системе координат, L_t – безразмерный временной интервал, в пределах которого анализируется процесс распространения импульса.

Заметим, что в выбранной нормировке коэффициент D_r равен единице. Однако он оставлен в уравнении (1) для удобства дальнейшего анализа. Коэффициент D равен отношению дифракционной длины к $l_{\text{dis}} = 2\tau_p^2 \left| \frac{\partial^2 k}{\partial \omega^2} \right|^{-1} \Big|_{\omega=\omega_0}$, т.е. к дисперсионной длине: $D = l_{\text{dif}}/l_{\text{dis}}$. Параметр нелинейности определяется по формуле

$$\alpha = \frac{3\pi k^2 a^2 \chi^{(3)} I_0}{n^2} = l_{\text{dif}}/l_{\text{nl}}$$

как отношение дифракционной длины к характерной нелинейной длине, $\chi^{(3)}$ – коэффициент, описывающий кубическую восприимчивость среды, n – ее показатель преломления, $\gamma = \frac{2}{\omega_0 \tau_p}$ – параметр, обратно пропорциональный произведению длительности импульса на его частоту. Как следует из (1), при его записи пренебрегалось эффектами установления нелинейного отклика, изучение которых выходит за рамки данной работы.

Из-за финитности распределения по времени и пространству на границах $t = 0$, $t = L_t$ и $r = R$ можно ввести дополнительные условия на значение производной по времени от функции A вида

$$\frac{\partial A}{\partial t} \Big|_{t=0, L_t} = 0, \quad \frac{\partial A}{\partial r} \Big|_{r=R} = 0. \quad (3)$$

Они необходимы ниже для записи некоторых инвариантов распространения светового импульса.

Прежде всего отметим, что в процессе взаимодействия лазерного импульса с веществом сохраняется его энергия:

$$I_A(z) = \int_0^{RL_t} \int_0^t |A(z, r, t)|^2 dt \cdot r dr = \text{const.} \quad (4)$$

Для записи других инвариантов нелинейного распространения лазерного импульса введем новую функцию

$$E(z, r, t) = \int_0^t A(z, r, \eta) e^{\frac{i(\eta-t)}{\gamma}} d\eta, \quad (5)$$

которая удовлетворяет релаксационному уравнению

$$\frac{\partial E}{\partial t} + \frac{i}{\gamma} E = A. \quad (6)$$

В новых переменных уравнение (1) преобразуется к виду

$$\frac{\partial E}{\partial z} + iD_r \Delta_r E + iD \frac{\partial^2 E}{\partial t^2} + \alpha \gamma |A|^2 A = 0, \quad z > 0, \quad 0 < t < L_t, \quad 0 < r < R. \quad (7)$$

На входе в нелинейную среду вместо начального распределения относительно функции $A(z, r, t)$ задается условие

$$E(0, r, t) = E_0(r, t) = \int_0^t A_0(r, \eta) e^{\frac{i(\eta-t)}{\gamma}} d\eta. \quad (8)$$

Из способа введения функций $E(z, r, t)$ (см. (5)) следует, что для нее имеют место краевые условия

$$r \frac{\partial E}{\partial r} \Big|_{r \rightarrow 0} = E|_{r=R} = 0, \quad E|_{t=0} = \frac{\partial E}{\partial t} \Big|_{t=0} = \left(\frac{\partial E}{\partial t} + \frac{i}{\gamma} E \right) \Big|_{t=L_t} = 0. \quad (9a)$$

Из уравнения (7) в точке $t = L_t$ следует еще одно условие:

$$\left(\frac{\partial E}{\partial z} - \frac{iD}{\gamma^2} E + iD_r \Delta_r E \right) \Big|_{t=L_t} = 0. \quad (9b)$$

Оно используется для построения консервативной разностной схемы. Используя (6), также можно записать условия

$$\frac{\partial A}{\partial r} \Big|_{r=R} = \frac{\partial}{\partial r} \left(\frac{\partial E}{\partial t} + \frac{i}{\gamma} E \right) \Big|_{r=R} = 0, \quad (9b)$$

а условиям (3) соответствуют условия

$$\frac{\partial E}{\partial r} \Big|_{r=R} = 0, \quad \frac{\partial}{\partial t} \left(\frac{\partial E}{\partial t} + \frac{i}{\gamma} E \right) \Big|_{0, L_t} = 0. \quad (9c)$$

В результате процесс распространения фемтосекундного импульса описывается уравнениями (6), (7) с начальными и граничными условиями (8), (9a).

Уравнение (7), а, следовательно, и (1) обладают следующими инвариантами:

$$I_1(z) = \int_0^{RL_t} \int_0^t \left(|E|^2 - \gamma \text{Im} \left(E \frac{\partial E^*}{\partial t} \right) \right) dt \cdot r dr = \text{const}, \quad (10)$$

$$I_2(z) = \int_0^R \int_0^{L_t} 2\operatorname{Re}(EA^*) dt \cdot r dr = \int_0^R \int_0^{L_t} \frac{\partial |E|^2}{\partial t} dt \cdot r dr = \text{const}, \quad (11)$$

$$I_3(z) = \int_0^R \int_0^{L_t} \left(E \frac{\partial A^*}{\partial t} - E^* \frac{\partial A}{\partial t} \right) dt \cdot r dr = \text{const}, \quad (12)$$

$$I_E(z) = \int_0^R \int_0^{L_t} \left(\left| \frac{\partial E}{\partial t} \right|^2 - \frac{1}{\gamma^2} |E|^2 \right) dt \cdot r dr = \text{const}. \quad (13)$$

Имеет место также спектральный инвариант:

$$I_{\text{SP}}(z) = \int_0^R \int_0^{L_t} A(z, r, t) e^{it/\gamma} dt \cdot r dr = e^{iDz/\gamma^2} \int_0^R \int_0^{L_t} A_0(r, t) e^{\frac{it}{\gamma}} dt \cdot r dr, \quad I_{\text{SP}}(z) = I_{\text{SP}}(0) e^{iDz/\gamma^2}, \quad (14)$$

который в терминах функции $E(z, r, t)$ запишется в виде

$$\int_0^R \int_0^{L_t} E(z, r, L_t) r dr = e^{iDz/\gamma^2} \int_0^R \int_0^{L_t} E_0(r, L_t) r dr. \quad (15)$$

Покажем существование таких инвариантов.

Доказательство существования инварианта (4). Умножим уравнение (1) на rA^* , а сопряженное к нему – на rA , сложим два полученных уравнения и проинтегрируем по t от 0 до L_t и по r от 0 до R . В результате получим

$$\int_0^R \int_0^{L_t} \left[\frac{\partial |A|^2}{\partial z} + iD_r(A^* \Delta_r A - A \Delta_r A^*) + iD \left(\frac{\partial^2 A}{\partial t^2} A^* - \frac{\partial^2 A^*}{\partial t^2} A \right) + \frac{3\alpha\gamma}{2} \frac{\partial |A|^4}{\partial t} \right] dt \cdot r dr = 0. \quad (16)$$

Интегрируя в (16) по частям выражения, содержащие вторые производные, и учитывая граничные условия (2), получаем соотношение

$$\frac{\partial}{\partial z} \int_0^R \int_0^{L_t} |A|^2 dt \cdot r dr = 0,$$

которое означает наличие инварианта (4).

Доказательство существования инварианта (10). Для записи инварианта $I_1(z)$ умножим (7) на rE^* , а сопряженное к нему – на rE . Сложим полученные уравнения и проинтегрируем результат по t от 0 до L_t , по r от 0 до R :

$$\begin{aligned} & \int_0^R \int_0^{L_t} \left(E^* \frac{\partial E}{\partial z} + E \frac{\partial E^*}{\partial z} + iD_r(E^* \Delta_r E + E \Delta_r E^*) + \right. \\ & \left. + iD \left(E^* \frac{\partial^2 E}{\partial t^2} - E \frac{\partial^2 E^*}{\partial t^2} \right) + \alpha\gamma |A|^2 (AE^* + A^*E) \right) dt \cdot r dr = 0. \end{aligned} \quad (17)$$

Учитывая граничные условия (9а), слагаемые, содержащие вторые производные по времени,

преобразуем к виду

$$\begin{aligned} iD \int_0^R \int_0^{L_t} \left(E^* \frac{\partial^2 E}{\partial t^2} - \frac{\partial^2 E^*}{\partial t^2} E \right) dt \cdot r dr &= iD \int_0^R \left[\frac{\partial E}{\partial t} E^* \Big|_{L_t} - \frac{\partial E^*}{\partial t} E \Big|_{L_t} \right] r dr = \\ &= -iD \int_0^R \frac{2i}{\gamma} |E|^2 \Big|_{L_t} r dr = \frac{2D}{\gamma} \int_0^R \int_0^{L_t} \frac{\partial |E|^2}{\partial t} dt \cdot r dr. \end{aligned}$$

Заметим, что интеграл, содержащий оператор Лапласа по координате r , равен нулю, а интеграл в (17) от последней скобки запишется следующим образом:

$$\alpha \gamma \int_0^R \int_0^{L_t} |A|^2 (AE^* + A^*E) dt \cdot r dr = \alpha \gamma \int_0^R \int_0^{L_t} |A|^2 \left(\frac{\partial E}{\partial t} E^* + \frac{\partial E^*}{\partial t} E \right) dt \cdot r dr = \alpha \gamma \int_0^R \int_0^{L_t} |A|^2 \frac{\partial |E|^2}{\partial t} dt \cdot r dr.$$

Таким образом, из (17) следует интеграл вида

$$\int_0^R \int_0^{L_t} \left(\frac{\partial |E|^2}{\partial z} + \frac{\partial |E|^2}{\partial t} \left(\alpha \gamma |A|^2 + \frac{2D}{\gamma} \right) \right) dt \cdot r dr = 0. \quad (18)$$

Другое соотношение запишем, умножив уравнение (7) на $r \frac{\partial E^*}{\partial t}$, а сопряженное к нему – на $r \frac{\partial E}{\partial t}$, вычтя второе из первого и проинтегрировав по t и r :

$$\begin{aligned} \int_0^R \int_0^{L_t} \left[\frac{\partial E^*}{\partial t} \frac{\partial E}{\partial z} - \frac{\partial E}{\partial t} \frac{\partial E^*}{\partial z} + iD_r \left(\frac{\partial E^*}{\partial t} \Delta_r E + \frac{\partial E}{\partial t} \Delta_r E^* \right) + \right. \\ \left. + iD \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial t} \right|^2 + \alpha \gamma |A|^2 \left(A \frac{\partial E^*}{\partial t} - A^* \frac{\partial E}{\partial t} \right) \right] dt \cdot r dr = 0. \end{aligned} \quad (19)$$

Принимая во внимание условия (9а) и (9в), преобразуем интеграл от слагаемого, содержащего iD_r , следующим образом:

$$\int_0^R \int_0^{L_t} iD_r \left(\frac{\partial E^*}{\partial t} \Delta_r E + \frac{\partial E}{\partial t} \Delta_r E^* \right) dt \cdot r dr = \int_0^{L_t} iD_r r \left(\frac{\partial E^*}{\partial r} \frac{\partial E}{\partial t} + \frac{\partial E}{\partial r} \frac{\partial E^*}{\partial t} \right) \Big|_0^R dt - \int_0^R \int_0^{L_t} iD_r \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial r} \right|^2 dt \cdot r dr.$$

Интеграл, содержащий производные функции E по продольной координате, преобразуется следующим образом:

$$\begin{aligned} \int_0^R \int_0^{L_t} \left(\frac{\partial E^*}{\partial t} \frac{\partial E}{\partial z} - \frac{\partial E}{\partial t} \frac{\partial E^*}{\partial z} \pm \frac{1}{2} \left(E \frac{\partial^2 E^*}{\partial z \partial t} + E^* \frac{\partial^2 E}{\partial z \partial t} \right) \right) dt \cdot r dr = \\ = \frac{\partial}{\partial z} \int_0^R \int_0^{L_t} i \operatorname{Im} \left(E \frac{\partial E^*}{\partial t} \right) dt \cdot r dr - \int_0^R \int_0^{L_t} i \operatorname{Im} \frac{\partial}{\partial t} \left(E \frac{\partial E^*}{\partial z} \right) dt \cdot r dr. \end{aligned}$$

В результате из (19) получим интеграл

$$\int_0^R \int_0^{L_t} \left[i \operatorname{Im} \left(\frac{\partial}{\partial z} \left(E \frac{\partial E^*}{\partial t} \right) - \frac{\partial}{\partial t} \left(E \frac{\partial E^*}{\partial z} \right) \right) - iD_r \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial r} \right|^2 + iD \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial t} \right|^2 + \alpha \gamma |A|^2 \left(A \frac{\partial E^*}{\partial t} - A^* \frac{\partial E}{\partial t} \right) \right] dt \cdot r dr = 0. \quad (20)$$

Далее вычислим интеграл от второго слагаемого в (20). Для этого умножим (9б) на rE^* , а сопряженное к (9б) – на rE , вычтем из первого второе и, учтя граничные условия (9а), запишем равенство

$$\begin{aligned} & \int_0^R \left\{ \left(E^* \frac{\partial E}{\partial z} - E \frac{\partial E^*}{\partial z} \right) - 2i \frac{D}{\gamma^2} |E|^2 + i D_r (E^* \Delta_r E + E \Delta_r E^*) \right\} \Big|_{t=L_t} r dr = \\ & = \int_0^R \left\{ \left(E^* \frac{\partial E}{\partial z} - E \frac{\partial E^*}{\partial z} \right) - 2i \left(\frac{D}{\gamma^2} |E|^2 + D_r \left| \frac{\partial E}{\partial r} \right|^2 \right) \right\} \Big|_{t=L_t} r dr = 0, \end{aligned}$$

откуда следует

$$\int_0^R \left\{ \operatorname{Im} \left(E \frac{\partial E^*}{\partial z} \right) \Big|_{t=L_t} \right\} r dr = - \iint_{0 0}^{R L_t} \left(\frac{D}{\gamma^2} \frac{\partial |E|^2}{\partial t} + D_r \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial r} \right|^2 \right) dt \cdot r dr. \quad (21)$$

Подставляя из (21) выражение левого интеграла в (20), получаем соотношение

$$\iint_{0 0}^{R L_t} i \left\{ \operatorname{Im} \frac{\partial}{\partial z} \left(E \frac{\partial E^*}{\partial t} \right) + \left(\alpha |A|^2 + \frac{D}{\gamma^2} \right) \frac{\partial |E|^2}{\partial t} + D \frac{\partial}{\partial t} \left| \frac{\partial E}{\partial t} \right|^2 \right\} dt \cdot r dr = 0. \quad (22)$$

Сложим выражения (18), умножая их на i , и (22), умноженное на $-\gamma$. В результате получим

$$\int_0^R \left[\int_0^{L_t} i \left(\frac{\partial |E|^2}{\partial z} - \gamma \operatorname{Im} \frac{\partial}{\partial z} \left(E \frac{\partial E^*}{\partial t} \right) \right) dt + \frac{i D}{\gamma} \left(|E|^2 - \gamma^2 \left| \frac{\partial E}{\partial t} \right|^2 \right) \Big|_{t=L_t} \right] r dr = 0. \quad (23)$$

Учитывая, что из (9а) следует

$$\left(|E|^2 - \gamma^2 \left| \frac{\partial E}{\partial t} \right|^2 \right) \Big|_{t=L_t} = 0, \quad (24)$$

получаем требуемый инвариант (10).

Доказательство существования инварианта (11). Умножим (9б) на rE^* , сопряженное к нему – на rE , сложим полученный результат и проинтегрируем его по r от 0 до R . В результате получим

$$\int_0^R \left(E^* \frac{\partial E}{\partial z} + E \frac{\partial E^*}{\partial z} + i D_r (E^* \Delta_\perp E - E \Delta_\perp E^*) \right) \Big|_{t=L_t} r dr = 0. \quad (25)$$

Заметим, что интеграл в (25) от слагаемого, содержащего коэффициент D_r , равен нулю. Следовательно, в процессе взаимодействия светового импульса со средой имеет место инвариант

$$\int_0^R |E|^2 \Big|_{t=L_t} r dr = \text{const}, \quad (26)$$

или

$$\iint_{0 0}^{R L_t} \frac{\partial |E|^2}{\partial t} dt \cdot r dr = \text{const}. \quad (27)$$

Далее, умножая релаксационное уравнение (6) на rE^* , сопряженное к нему – на rE и интегрируя по r и t в соответствующих пределах, складываем их и получаем

$$\iint_{0 0}^{R L_t} \frac{\partial |E|^2}{\partial t} dt \cdot r dr = \iint_{0 0}^{R L_t} (E^* A + E A^*) dt \cdot r dr. \quad (28)$$

Из этого равенства с учетом (27) следует, что инвариант (11) имеет место.

Доказательство существования инварианта (12). Умножим релаксационное уравнение (6) на rA^* , а сопряженное к нему – на rA , вычтем из первого второе и проинтегрируем по t от 0 до L_t и по r от 0 до R . В результате получим

$$\int_0^R \int_0^{L_t} \left(A^* \frac{\partial E}{\partial t} - A \frac{\partial E^*}{\partial t} \right) dt \cdot r dr = -\frac{i}{\gamma} \int_0^R \int_0^{L_t} (A^* E + A E^*) dt \cdot r dr = I. \quad (29)$$

Так как имеет место инвариант (11), то интеграл (29) равен $I = \text{const}$. Интегрируя по времени в левой части (29) по частям и учитывая граничные условия (3), получаем, что инвариант (12) имеет место.

Доказательство существования инварианта (13). Инвариант (13) следует из инварианта энергии. Однако доказать его существование можно непосредственно на основе преобразованного уравнения. Для этого умножим уравнение (7) на $r \frac{\partial^2 E^*}{\partial t^2}$, а сопряженное к нему – на $r \frac{\partial^2 E}{\partial t^2}$, сложим оба уравнения и проинтегрируем по t от 0 до L_t и по r от 0 до R , учтя граничные условия (9а), в результате получим соотношение

$$\begin{aligned} \int_0^R \left\{ \int_0^{L_t} \left[-\frac{\partial}{\partial z} \left| \frac{\partial E}{\partial t} \right|^2 + i D_r \left(\frac{\partial^2 E^*}{\partial t^2} \Delta_r E - \frac{\partial^2 E}{\partial t^2} \Delta_r E^* \right) + \alpha \gamma |A|^2 \left(A \frac{\partial^2 E^*}{\partial t^2} + A^* \frac{\partial^2 E}{\partial t^2} \right) \right] dt + \right. \\ \left. + \left(\frac{\partial E^*}{\partial z} \frac{\partial E}{\partial t} + \frac{\partial E}{\partial z} \frac{\partial E^*}{\partial t} \right) \right|_{t=L_t} \right\} r dr = 0. \end{aligned} \quad (30)$$

Выражая вторые производные от функции E из релаксационного уравнения (6), равенство (30) можно преобразовать к виду

$$\begin{aligned} \int_0^R \left\{ \int_0^{L_t} \left[-\frac{\partial}{\partial z} \left| \frac{\partial E}{\partial t} \right|^2 + i D_r \left(\frac{\partial^2 E^*}{\partial t^2} \Delta_r E - \frac{\partial^2 E}{\partial t^2} \Delta_r E^* \right) - \frac{\alpha}{\gamma} |A|^2 (E^* A + E A^*) \right] dt + \right. \\ \left. + \left(\frac{\partial E^*}{\partial z} \frac{\partial E}{\partial t} + \frac{\partial E}{\partial z} \frac{\partial E^*}{\partial t} \right) \right|_{t=L_t} \right\} r dr = 0. \end{aligned} \quad (31)$$

Интегрируя по частям в (31) выражение с множителем $i D_r$ с учетом граничных условий при $t = L_t$, получаем

$$\begin{aligned} i D_r \int_0^R \int_0^{L_t} \left(\frac{\partial^2 E^*}{\partial t^2} \Delta_r E - \frac{\partial^2 E}{\partial t^2} \Delta_r E^* \right) dt \cdot r dr = \\ = i D_r \int_0^R \left[\left(\frac{\partial E^*}{\partial t} \Delta_r E - \frac{\partial E}{\partial t} \Delta_r E^* \right) \right]_{t=L_t} - \int_0^{L_t} \left(\frac{\partial E^*}{\partial t} \frac{\partial \Delta_r E}{\partial t} - \frac{\partial E}{\partial t} \frac{\partial \Delta_r E^*}{\partial t} \right) dt \right] r dr. \end{aligned} \quad (32)$$

Заметим, что последний интеграл в (32) равен нулю. Это легко показать, интегрируя его по частям по координате r и учитывая граничные условия (9а) на производную функции E по r . Замена соответствующий интеграл в (31) правой частью выражения (32) и используя уравнение (9б), из (31) получаем следующее соотношение:

$$\int_0^R \left\{ \int_0^{L_t} \left[-\frac{\partial}{\partial z} \left| \frac{\partial E}{\partial t} \right|^2 - \frac{\alpha}{\gamma} |A|^2 (E^* A + E A^*) \right] dt - \frac{i D}{\gamma^2} \left(E^* \frac{\partial E}{\partial t} - E \frac{\partial E^*}{\partial t} \right) \right|_{t=L_t} \right\} r dr = 0. \quad (33)$$

Далее, умножим уравнение (7) на rE^* , а сопряженное к нему – на rE . Сложив результат и проинтегрировав его по t и по r , нетрудно получить следующее соотношение:

$$\int_0^R \left\{ \int_0^{L_t} \left[\frac{\partial}{\partial z} |E|^2 + \alpha \gamma |A|^2 (E^* A + EA^*) \right] dt + iD \left(E^* \frac{\partial E}{\partial t} - E \frac{\partial E^*}{\partial t} \right) \Big|_{t=L_t} \right\} r dr = 0. \quad (34)$$

Окончательно умножив (34) на $1/\gamma^2$ и сложив с (33), получим требуемый инвариант (13).

Доказательство существования инварианта (15). Умножим уравнение (9б) на r и проинтегрируем от 0 до R :

$$\int_0^R \left[\frac{\partial E}{\partial z} - \frac{iD}{\gamma^2} E + iD_r \Delta_r E \right] \Big|_{t=L_t} r dr = 0. \quad (35)$$

Учитывая краевое условие (9г), из (35) получаем уравнение

$$\frac{\partial}{\partial z} \int_0^R E(z, r, L_t) r dr = \frac{iD}{\gamma^2} \int_0^R E(z, r, L_t) r dr, \quad (36)$$

решая которое с начальным условием (8), получаем спектральный инвариант (14).

Следует подчеркнуть, что приведенные выше инварианты используются для построения консервативных разностных схем.

3. РАЗНОСТНЫЕ СХЕМЫ

Для задачи (6)–(9а) построим консервативную разностную схему. Для этого введем в области $\Omega = (0, L_z) \times (0, R) \times (0, L_t)$ сетку $\omega = \omega_z \times \omega_r \times \omega_t$, где

$$\begin{aligned} \omega_z &= \{z_m = mh_z, m = 0, 1, \dots, N_z, h_z = L_z/N_z\}, \\ \omega_r &= \{r_k = (k + 0.5)h_r, k = 0, 1, \dots, N_r, h_r = R/(N_r + 0.5)\}, \\ \omega_t &= \{t_n = n\tau, n = 0, 1, \dots, N_t, \tau = L_t/N_t\}. \end{aligned} \quad (37)$$

Определим сеточные функции A и E на ω и введем также следующие безиндексные обозначения (см. [22]):

$$\begin{aligned} A &= A_{k,n} = A(z_m, r_k, t_n), \quad \hat{A} = A(z_{m+1}, r_k, t_n), \quad A_{\pm 1} = A(z_m, r_k, t_{n \pm 1}), \quad \overset{0.5}{A} = 0.5(\hat{A} + A), \\ E &= E_{k,n} = E(z_m, r_k, t_n), \quad \hat{E} = E(z_{m+1}, r_k, t_n), \quad E_{\pm 1} = E(z_m, r_k, t_{n \pm 1}), \quad \overset{0.5}{E} = 0.5(\hat{E} + E), \\ |A|^2 &= 0.5(|\hat{A}|^2 + |A|^2). \end{aligned} \quad (38)$$

Оператор Лапласа по времени t и по пространственной координате r аппроксимируется следующим образом (см. [23]):

$$\Lambda_{tt} u = \frac{u(z_m, r_k, t_{n+1}) - 2u(z_m, r_k, t_n) + u(z_m, r_k, t_{n-1}))}{\tau^2}, \quad n = 1, 2, \dots, N_t - 1, \quad (39)$$

$$\Lambda_{rr} u = \frac{1}{r_k h_r} \left(r_{k+0.5h_r} \frac{u(z_m, r_{k+0.5h_r}, t_n) - u(z_m, r_k, t_n)}{h_r} - r_{k-0.5h_r} \frac{u(z_m, r_k, t_n) - u(z_m, r_{k-0.5h_r}, t_n)}{h_r} \right), \quad (40)$$

$$r_{k \pm 0.5} = r_k \pm 0.5h_r, \quad k = 1, 2, \dots, N_r - 1.$$

С учетом введенных выше обозначений разностная схема для уравнений (6), (7) запишется в виде

$$\frac{\hat{E} - E}{h_z} + iD_r \Lambda_{rr} \overset{0.5}{E} + iD \Lambda_{tt} \overset{0.5}{E} + \alpha \gamma |A|^2 \overset{0.5}{A} = 0, \quad (41)$$

$$\frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} + \frac{i}{\gamma} \overset{0.5}{E} = \overset{0.5}{A}. \quad (42)$$

Заметим, что в силу линейности разностного уравнения (42) его можно записать также в виде

$$\frac{\hat{E}_{+1} - \hat{E}_{-1}}{2\tau} + \frac{i}{\gamma} \hat{E} = \hat{A}, \quad n = 1, 2, \dots, N_t - 1. \quad (43)$$

Уравнения (41), (42) необходимо дополнить разностными соотношениями в соответствующих граничных точках:

$$E_{k,0} = E_{N_r,n} = 0, \quad k = 0, 1, \dots, N_r, \quad n = 0, 1, \dots, N_t$$

$$iD_r \frac{E_{1,n}^{0.5} - E_{0,n}^{0.5}}{0.5h_r^2} = - \left(\frac{\hat{E}_{0,n} - E_{0,n}}{h_z} + iD\Lambda_{\bar{t}t}^{0.5} E_{0,n} + \alpha\gamma |A_{0,n}|^2 A_{0,n}^{0.5} \right), \quad n = 1, 2, \dots, N_t - 1, \quad (44)$$

$$\frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^{0.5}}{\tau} + \frac{i}{\gamma} E_{k,N_t}^{0.5} - \tau \frac{E_{k,N_t}^{0.5}}{\gamma^2} = 0, \quad (45)$$

$$\frac{\hat{E}_{k,N_t} - E_{k,N_t}}{h_z} = \frac{iD^{0.5}}{\gamma^2} E_{k,N_t} - iD_r \Lambda_{\bar{r}r}^{0.5} E_{k,N_t}, \quad (46a)$$

$$\frac{\hat{E}_{0,N_t} - E_{0,N_t}}{h_z} - \frac{iD^{0.5}}{\gamma^2} E_{0,N_t} + iD_r \frac{E_{1,N_t}^{0.5} - E_{0,N_t}^{0.5}}{0.5h_r^2} = 0. \quad (46b)$$

Условия (45), (46) аппроксимируют, соответственно, последнее условие в (9a) и (9б). Как следствие из них в граничной точке $n = N_t$ сетки по времени получим разностные уравнения

$$\frac{\hat{E}_{k,N_t} - E_{k,N_t}}{h_z} - \frac{iD}{\tau} \left(\frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^{0.5}}{\tau} + \frac{i}{\gamma} E_{k,N_t}^{0.5} \right) + iD_r \Lambda_{\bar{r}r}^{0.5} E_{k,N_t} = 0, \quad k = 1, 2, \dots, N_r - 1, \quad (47a)$$

$$\frac{\hat{E}_{0,N_t} - E_{0,N_t}}{h_z} - \frac{iD}{\tau} \left(\frac{E_{0,N_t}^{0.5} - E_{0,N_t-1}^{0.5}}{\tau} + \frac{i}{\gamma} E_{0,N_t}^{0.5} \right) + iD_r \frac{E_{1,N_t}^{0.5} - E_{0,N_t}^{0.5}}{0.5h_r^2} = 0, \quad (47b)$$

которые используются при построении консервативной разностной схемы. Начальное условие для сеточной функции E задается в виде

$$E(0, r_k, t_n) = E_0(r_k, t_n), \quad k = 0, 1, \dots, N_r, \quad n = 0, 1, \dots, N_t. \quad (48)$$

Так как построенная разностная схема является нелинейной, то для ее разрешения воспользуемся методом простой итерации:

$$\frac{\hat{E}^{(s+1)} - E^{(s)}}{h} + iD_r \Lambda_{\bar{r}r}^{0.5} E^{(s+1)} + iD \Lambda_{\bar{t}t}^{0.5} E^{(s+1)} + \alpha\gamma |A|^{0.5,0.5} A^{(s)} = 0, \quad (49)$$

$$\frac{E_{+1}^{(s)} - E_{-1}^{(s)}}{2\tau} + \frac{i}{\gamma} E^{(s)} = A^{(s)}, \quad s = 0, 1, \dots \quad (50)$$

В граничных точках итерационный процесс запишется в виде

$$\hat{E}_{k,0}^{(s+1)} = \hat{E}_{N_r,n}^{(s+1)} = 0, \quad k = 0, 1, \dots, N_r, \quad n = 0, 1, \dots, N_t,$$

$$\begin{aligned}
 iD_r \frac{\hat{E}_{1,n}^{(s+1)} - E_{0,n}^{(s+1)}}{0.5h_r^2} &= - \left(\frac{\hat{E}_{0,n}^{(s+1)} - E_{0,n}^{(s+1)}}{h_z} + iD\Lambda_{\bar{t}t}^{(s+1)} E_{0,n}^{(s+1)} + \alpha \gamma |A_{0,n}|^{2^{0.5}} A_{0,n}^{(s)} \right), \quad n = 1, 2, \dots, N_t - 1, \\
 \frac{\hat{E}_{k,N_t}^{(s+1)} - E_{k,N_t}^{(s+1)}}{h_z} - \frac{iD}{\tau} \left(\frac{E_{k,N_t}^{(s+1)} - E_{k,N_t-1}^{(s+1)}}{\tau} + \frac{i}{\gamma} E_{k,N_t}^{(s+1)} \right) &+ iD_r \Lambda_{\bar{r}r}^{(s+1)} E_{k,N_t}^{(s+1)} = 0, \quad k = 1, 2, \dots, N_r - 1 \\
 \frac{\hat{E}_{0,N_t}^{(s+1)} - E_{0,N_t}^{(s+1)}}{h_z} - \frac{iD}{\tau} \left(\frac{E_{0,N_t}^{(s+1)} - E_{0,N_t-1}^{(s+1)}}{\tau} + \frac{i}{\gamma} E_{0,N_t}^{(s+1)} \right) &+ iD_r \frac{E_{1,N_t}^{(s+1)} - E_{0,N_t}^{(s+1)}}{0.5h_r^2} = 0.
 \end{aligned} \tag{51}$$

Значения функций на нулевой итерации ($s = 0$) берутся с предыдущего слоя по z :

$$\hat{E}^{(s=0)} = E.$$

Итерации прекращаются, если выполнено условие

$$\max_{t_n, r_k} |\hat{E}^{(s+1)} - \hat{E}^{(s)}| \leq \varepsilon \max_{t_n, r_k} |\hat{E}^{(s)}| + \delta, \quad \varepsilon, \delta = \text{const} > 0. \tag{52}$$

Нетрудно показать, что разностная схема (41)–(47) имеет следующий порядок аппроксимации: $\Psi = O(h_z^2 + h_r^2/r + \tau)$. Действительно, разностные уравнения (46) и (47) аппроксимируют исходные граничные условия (9а) и (9б) в точке $(z_m + 0.5h, r_k, L_t)$ с погрешностью $\Psi = O(h_z^2 + h_r^2/r + \tau)$. В остальных узлах сетки соответствующие разностные уравнения аппроксимируют дифференциальную задачу в точке $(z_m + 0.5h, r_k, t_n)$ с погрешностью $\Psi = O(h_z^2 + h_r^2/r + \tau^2)$.

Так как при реализации схемы (51), (52) имеются трудности с применением спектрального метода для обращения разностного оператора Лапласа по времени из-за сложного краевого условия, то построим также разностную схему для задачи (1), (2), которая эквивалентна схеме (41)–(48) (схема сохраняет инварианты, записанные выше). Она с учетом введенных обозначений (37)–(40) запишется в виде

$$\frac{\hat{A} - A}{h_z} + iD_r \Lambda_{\bar{r}r}^{0.5} \hat{A} + iD \Lambda_{\bar{t}t}^{0.5} \hat{A} + i\alpha |A|^{2^{0.5}} \hat{A} + \alpha \gamma \frac{|A_{+1}|^{2^{0.5}} A_{+1}^{0.5} - |A_{-1}|^{2^{0.5}} A_{-1}^{0.5}}{2\tau} = 0. \tag{53}$$

Уравнения (53) необходимо дополнить разностными соотношениями в соответствующих граничных точках:

$$A_{k,0} = A_{k,N_t} = A_{N_r,n} = 0, \quad k = 0, 1, \dots, N_r, \quad n = 0, 1, \dots, N_t,$$

$$\begin{aligned}
 iD_r \frac{A_{1,n}^{0.5} - A_{0,n}^{0.5}}{0.5h_r^2} &= - \left(\frac{\hat{A}_{0,n} - A_{0,n}}{h_z} + iD\Lambda_{\bar{t}t}^{0.5} A_{0,n}^{0.5} + i\alpha |A_{0,n}|^{2^{0.5}} A_{0,n}^{0.5} + \alpha \gamma \frac{|A_{0,n+1}|^{2^{0.5}} A_{0,n+1}^{0.5} - |A_{0,n-1}|^{2^{0.5}} A_{0,n-1}^{0.5}}{2\tau} \right), \\
 n &= 1, 2, \dots, N_t - 1.
 \end{aligned} \tag{54}$$

Начальное условие для сеточной функции A задается в виде

$$A(0, r_k, t_n) = A_0(r_k, t_n), \quad n = 0, 1, \dots, N_t. \tag{55}$$

Разностная схема (53)–(55) имеет порядок аппроксимации $\Psi = O(h_r^2/r + h_z^2 + \tau^2)$ в норме C относительно точек $(z_m + 0.5h_z, r_{k+0.5}, t_n)$, а в норме L_2 имеем $\Psi = O(h_z^2 + h_r^2 \sqrt{\ln(1/h_r)} + \tau^2)$.

Так как построенная разностная схема является нелинейной, то для ее разрешения воспользуемся методом простой итерации:

$$\begin{aligned} \frac{\hat{A} - A}{h_z} + iD_r \Lambda_{\bar{r}r} A + iD \Lambda_{\bar{t}t} A + F &= 0, \quad F = i\alpha |A|^{0.5} A + \alpha \gamma \frac{|A_{+1}|^{0.5} A_{+1} - |A_{-1}|^{0.5} A_{-1}}{2\tau}, \\ A_{k,0} &= A_{k,N_t} = A_{N_r,n} = 0, \quad k = 0, 1, \dots, N_r, \quad n = 0, 1, \dots, N_t, \\ iD_r \frac{A_{1,n} - A_{0,n}}{0.5h_r^2} &= - \left(\frac{\hat{A}_{0,n} - A_{0,n}}{h_z} + iD \Lambda_{\bar{t}t} A_{0,n} + i\alpha |A_{0,n}|^{0.5} A_{0,n} + \alpha \gamma \frac{|A_{0,n+1}|^{0.5} A_{0,n+1} - |A_{0,n-1}|^{0.5} A_{0,n-1}}{2\tau} \right), \\ s &= 0, 1, \dots, \quad n = 1, 2, \dots, N_t - 1. \end{aligned} \quad (56)$$

Значения функций на нулевой итерации ($s = 0$) берутся с предыдущего слоя по z :

$$\hat{A}^{(s=0)} = A.$$

Итерации прекращаются, если выполнено условие

$$\max_{t_n, r_k} |\hat{A} - A| \leq \varepsilon \max_{t_n, r_k} |\hat{A}| + \delta, \quad \varepsilon, \delta = \text{const} > 0. \quad (57)$$

Для обращения разностного оператора по координате t на каждой итерации используется псевдоспектральный метод (см. [24]). Для этого, перейдя в фурье-пространство, вместо (56) получим уравнение

$$\frac{(A_\omega - A_\omega)}{h_z} + iD_r \Lambda_{\bar{r}r} \left(\frac{A_\omega + A_\omega}{2} \right) - iD \lambda_j^2 \left(\frac{A_\omega + A_\omega}{2} \right) + F_\omega = 0. \quad (58)$$

Начальное условие (55) запишется в виде

$$A_\omega(0, r_k, \omega_j) = \sum_{n=0}^{N_t} A(0, r_k, t_n) e^{-i\lambda_j t_n}, \quad \lambda_j = \frac{2\pi j}{L_t}, \quad j = 0, 1, \dots, N_t, \quad k = 0, 1, \dots, N_r.$$

Краевые условия (54) запишутся в виде

$$\begin{aligned} A_{\omega k,0} &= A_{\omega k,N_t} = A_{\omega N_r,j} = 0, \quad k = 0, 1, \dots, N_r, \quad j = 0, 1, \dots, N_t, \\ iD_r \frac{A_{\omega 1,j} - A_{\omega 0,j}}{0.5h_r^2} &= - \left(\frac{\hat{A}_{\omega 0,j} - A_{\omega 0,j}}{h_z} - iD \lambda_j^2 A_{\omega 0,j} + F_{\omega 0,j} \right), \quad j = 0, 1, \dots, N_t. \end{aligned}$$

Выше введены следующие обозначения:

$$\begin{aligned} A_\omega &= \sum_{n=0}^{N_t} A(z_m, r_k, t_n) e^{-i\lambda_j t_n}, \quad A(z_m, r_k, t_n) = \frac{1}{(N_t + 1)} \sum_{j=0}^{N_t} A_\omega(z_m, r_k, \omega_j) e^{i\lambda_j t_n}, \quad \lambda_j = \frac{2\pi j}{L_t}, \\ F_\omega &= \sum_{n=0}^{N_t} F(z_m, r_k, t_n) e^{-i\lambda_j t_n}, \quad F(z_m, r_k, t_n) = \frac{1}{(N_t + 1)} \sum_{j=0}^{N_t} F_\omega(z_m, r_k, \omega_j) e^{i\lambda_j t_n}, \quad n = 0, 1, \dots, N_t. \end{aligned}$$

Для решения трехточечных уравнений (58) используется метод прогонки по координате r . Найдя решение на верхнем слое и $(s + 1)$ -й итерации, применим к нему обратное преобразование Фурье. Итерации прекращались при выполнении условия (57).

4. ДОКАЗАТЕЛЬСТВО КОНСЕРВАТИВНОСТИ СХЕМ

Покажем, что разностная схема (41)–(48) является консервативной, т.е. сохраняет следующие разностные аналоги инвариантов:

$$I_1(z_m) = \sum_{k=0}^{N_r-1} r_k h_r \tau \left\{ \sum_{n=1}^{N_t-1} \left[|E|^2 - \gamma \operatorname{Im} \left(E \frac{E_{+1}^* - E_{-1}^*}{2\tau} \right) \right] \right\} + \frac{1}{2} |E_{k, N_t}|^2 - \gamma \operatorname{Im} \left(E_{k, N_t} \frac{E_{k, N_t}^* - E_{k, N_t-1}^*}{2\tau} \right) = \text{const}, \quad (59)$$

$$I_2(z_m) = 2 \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \operatorname{Re}(AE^*) = \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{E_{+1}^* - E_{-1}^*}{2\tau} E + \frac{E_{+1} - E_{-1}}{2\tau} E^* \right) = \text{const}, \quad (60)$$

$$I_3(z_m) = \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{A_{+1}^* - A_{-1}^*}{2\tau} E - \frac{A_{+1} - A_{-1}}{2\tau} E^* \right) = \text{const}, \quad (61)$$

$$I_{\text{SP}}(z_{m+1}) = \sum_{k=0}^{N_r-1} r_k h_r E(z_m, r_k, N_t) = \sum_{k=0}^{N_r-1} r h_r \left(\frac{a_{\text{SP}}}{b_{\text{SP}}} \right)^{n+1} E_0(r_k, N_t), \quad (62)$$

где

$$a_{\text{SP}} = 1 - \left(\frac{Dh}{2\gamma^2} \right)^2 + \frac{iDh}{\gamma^2}, \quad b_{\text{SP}} = 1 + \left(\frac{Dh}{2\gamma^2} \right)^2.$$

Другие инварианты также контролировались при компьютерном моделировании.

Консервативность схемы по инварианту (59). Для записи разностного аналога инварианта $I_1(z_m)$ умножим уравнение (41) на E^* , а сопряженное к нему – на E , сложим и умножим их на $\tau r_k h_r$. Просуммируем по n от 1 до $N_t - 1$ и по k от 1 до $N_r - 1$:

$$\sum_{k=1}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left[E^* \frac{\hat{E} - E}{h_z} + E \frac{\hat{E}^* - E^*}{h_z} + iD(E^* \Lambda_{it}^{0.5} E - E \Lambda_{it}^{0.5} E^*) + iD_r(E^* \Lambda_{rr}^{0.5} E - E \Lambda_{rr}^{0.5} E^*) + \right. \\ \left. + \alpha \gamma |A|^2 (E^* A + E A^*) \right] = 0. \quad (63)$$

Учитывая вторую разностную формулу Грина (см. [22]) и граничные условия (44), преобразуем слагаемые, содержащие разностные операторы Лапласа, к виду

$$\sum_{k=1}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau (E^* \Lambda_{it}^{0.5} E - E \Lambda_{it}^{0.5} E^*) = \sum_{k=1}^{N_r-1} r_k h_r \left(\frac{E_{k, N_t}^{0.5} - E_{k, N_t-1}^{0.5}}{\tau} E_{k, N_t}^* - \frac{E_{k, N_t}^* - E_{k, N_t-1}^*}{\tau} E_{k, N_t}^{0.5} \right), \\ \sum_{n=1}^{N_t-1} \tau \sum_{k=1}^{N_r-1} r_k h_r (E^* \Lambda_{rr}^{0.5} E - E \Lambda_{rr}^{0.5} E^*) = - \sum_{n=1}^{N_t-1} \tau h_r \left(E_{0, n}^* \frac{E_{1, n}^{0.5} - E_{0, n}^{0.5}}{h_r} - E_{0, n}^{0.5} \frac{E_{1, n}^* - E_{0, n}^*}{h_r} \right).$$

Получим следующее выражение:

$$\sum_{k=1}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-1} \tau \left[\frac{|\hat{E}|^2 - |E|^2}{h_z} + \alpha \gamma |A|^2 \left(\frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} E^* + \frac{E_{+1}^* - E_{-1}^*}{2\tau} E \right) \right] + \right. \\ \left. + iD \left(\frac{E_{k, N_t}^{0.5} - E_{k, N_t-1}^{0.5}}{\tau} E_{k, N_t}^* - \frac{E_{k, N_t}^* - E_{k, N_t-1}^*}{\tau} E_{k, N_t}^{0.5} \right) \right\} - iD_r \sum_{n=1}^{N_t-1} \tau h_r \left(E_{0, n}^* \frac{E_{1, n}^{0.5} - E_{0, n}^{0.5}}{h_r} - E_{0, n}^{0.5} \frac{E_{1, n}^* - E_{0, n}^*}{h_r} \right) = 0. \quad (64)$$

Преобразуем (64) таким образом, чтобы в него входили слагаемые в граничных узлах. Для этого умножим краевое условие (44) на $\overset{0.5}{E}_{0,n}^*$, а сопряженное к нему – на $\overset{0.5}{E}_{0,n}$. Сложим их, а результат просуммируем по n от 1 до N_t , предварительно умножив его на $r_0 h_r \tau$. В результате получим

$$iD_r \sum_{n=1}^{N_t-1} \tau h_r \left(\overset{0.5}{E}_{0,n}^* \frac{\overset{0.5}{E}_{1,n} - \overset{0.5}{E}_{0,n}}{h_r} - \overset{0.5}{E}_{0,n} \frac{\overset{0.5}{E}_{1,n}^* - \overset{0.5}{E}_{0,n}^*}{h_r} \right) = - \sum_{n=1}^{N_t-1} r_0 h_r \tau \left[\frac{|\hat{E}_{0,n}|^2 - |E_{0,n}|^2}{h_z} + \right. \\ \left. + \alpha \gamma |A|^2 \left(\frac{\overset{0.5}{E}_{0,n+1} - \overset{0.5}{E}_{0,n-1}}{2\tau} \overset{0.5}{E}_{0,n}^* + \frac{\overset{0.5}{E}_{0,n+1}^* - \overset{0.5}{E}_{0,n-1}^*}{2\tau} \overset{0.5}{E}_{0,n} \right) \right] + iD \left(\frac{\overset{0.5}{E}_{0,N_t} - \overset{0.5}{E}_{0,N_t-1}}{\tau} \overset{0.5}{E}_{0,N_t}^* - \frac{\overset{0.5}{E}_{0,N_t}^* - \overset{0.5}{E}_{0,N_t-1}^*}{\tau} \overset{0.5}{E}_{0,N_t} \right). \quad (65)$$

Выражая последнюю сумму в (64) из (65), преобразуем соотношение (64) к виду

$$\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left[\frac{|\hat{E}|^2 - |E|^2}{h_z} + \alpha \gamma |A|^2 \left(\frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \overset{0.5}{E}^* + \frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \overset{0.5}{E} \right) \right] + \\ + iD \left(\frac{\overset{0.5}{E}_{k,N_t} - \overset{0.5}{E}_{k,N_t-1}}{\tau} \overset{0.5}{E}_{k,N_t}^* - \frac{\overset{0.5}{E}_{k,N_t}^* - \overset{0.5}{E}_{k,N_t-1}^*}{\tau} \overset{0.5}{E}_{k,N_t} \right) = 0. \quad (66)$$

Таким образом, (66) содержит слагаемые при $k=0$. Однако оно не содержит разностной производной по координате z от модуля функции E в точке N_t . Для ее записи умножим (47) на $0.5\tau r_k h_r \overset{0.5}{E}_{k,N_t}^*$, а сопряженное к (47) – на $0.5\tau r_k h_r \overset{0.5}{E}_{k,N_t}$, сложим их и затем просуммируем по k от 1 до N_r-1 . В результате получим

$$\sum_{k=1}^{N_r-1} r_k h_r 0.5\tau \left[\overset{0.5}{E}_{k,N_t}^* \frac{\hat{E}_{k,N_t} - E_{k,N_t}}{h_z} + \overset{0.5}{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - E_{k,N_t}^*}{h_z} - \frac{iD}{\tau} \left(\frac{\overset{0.5}{E}_{k,N_t} - \overset{0.5}{E}_{k,N_t-1}}{\tau} \overset{0.5}{E}_{k,N_t}^* - \frac{\overset{0.5}{E}_{k,N_t}^* - \overset{0.5}{E}_{k,N_t-1}^*}{\tau} \overset{0.5}{E}_{k,N_t} \right) \right] + \\ + 2 \frac{D}{\gamma \tau} |\overset{0.5}{E}_{k,N_t}|^2 + iD_r (\overset{0.5}{E}_{k,N_t}^* \Lambda_{\bar{r}r} \overset{0.5}{E}_{k,N_t} - \overset{0.5}{E}_{k,N_t} \Lambda_{\bar{r}r} \overset{0.5}{E}_{k,N_t}^*) = 0. \quad (67)$$

Применяя вторую разностную формулу Грина к слагаемым с разностным оператором Лапласа по координате r в (67), учитывая краевое условие (47б) в точке ($r_0 = 0.5h_r$, $t = L_t$), преобразуем их к виду

$$\sum_{k=1}^{N_r-1} r_k h_r \frac{\tau}{2} [(\overset{0.5}{E}_{k,N_t}^* \Lambda_{\bar{r}r} \overset{0.5}{E}_{k,N_t} - \overset{0.5}{E}_{k,N_t} \Lambda_{\bar{r}r} \overset{0.5}{E}_{k,N_t}^*)] = -\tau r_0 \left(\overset{0.5}{E}_{0,N_t}^* \frac{\overset{0.5}{E}_{1,N_t} - \overset{0.5}{E}_{0,N_t}}{h_r} - \overset{0.5}{E}_{0,N_t} \frac{\overset{0.5}{E}_{1,N_t}^* - \overset{0.5}{E}_{0,N_t}^*}{h_r} \right).$$

Тогда (67) запишется следующим образом:

$$\sum_{k=1}^{N_r-1} r_k h_r 0.5\tau \left[\overset{0.5}{E}_{k,N_t}^* \frac{\hat{E}_{k,N_t} - E_{k,N_t}}{h_z} + \overset{0.5}{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - E_{k,N_t}^*}{h_z} - \frac{iD}{\tau} \left(\frac{\overset{0.5}{E}_{k,N_t} - \overset{0.5}{E}_{k,N_t-1}}{\tau} \overset{0.5}{E}_{k,N_t}^* - \frac{\overset{0.5}{E}_{k,N_t}^* - \overset{0.5}{E}_{k,N_t-1}^*}{\tau} \overset{0.5}{E}_{k,N_t} \right) \right] + \\ + 2 \frac{D}{\gamma \tau} |\overset{0.5}{E}_{k,N_t}|^2 - iD_r \tau r_0 \left(\overset{0.5}{E}_{0,N_t}^* \frac{\overset{0.5}{E}_{1,N_t} - \overset{0.5}{E}_{0,N_t}}{h_r} - \overset{0.5}{E}_{0,N_t} \frac{\overset{0.5}{E}_{1,N_t}^* - \overset{0.5}{E}_{0,N_t}^*}{h_r} \right) = 0. \quad (68)$$

Для исключения разностной производной по координате r в (68) домножим краевое условие

(44) на $0.5\tau h_r r_0 \overset{0.5}{E}_{0, N_t}^*$, а сопряженное к нему – на $0.5\tau h_r r_0 \overset{0.5}{E}_{0, N_t}$ и сложим их. В результате получим

$$iD_r \tau r_0 \left(\overset{0.5}{E}_{0, N_t}^* \frac{\overset{0.5}{E}_{1, N_t} - \overset{0.5}{E}_{0, N_t}}{h_r} - \overset{0.5}{E}_{0, N_t} \frac{\overset{0.5}{E}_{1, N_t}^* - \overset{0.5}{E}_{0, N_t}^*}{h_r} \right) =$$

$$= -r_0 h_r \tau \left[\frac{|\hat{E}_{0, N_t}|^2 - |E_{0, N_t}|^2}{h_z} + iD \left(\frac{\overset{0.5}{E}_{0, N_t} - \overset{0.5}{E}_{0, N_t-1}}{\tau} \overset{0.5}{E}_{0, N_t}^* - \frac{\overset{0.5}{E}_{0, N_t}^* - \overset{0.5}{E}_{0, N_t-1}^*}{\tau} \overset{0.5}{E}_{0, N_t} \right) + 2 \frac{D}{\gamma} |\overset{0.5}{E}_{0, N_t}| \right].$$

Используя это соотношение, (68) преобразуем к виду

$$\sum_{k=0}^{N_r-1} r_k h_r \left[0.5\tau \frac{|\hat{E}_{k, N_t}|^2 - |E_{k, N_t}|^2}{h_z} - \frac{iD}{2} \left(\frac{\overset{0.5}{E}_{k, N_t} - \overset{0.5}{E}_{k, N_t-1}}{\tau} \overset{0.5}{E}_{k, N_t}^* - \frac{\overset{0.5}{E}_{k, N_t}^* - \overset{0.5}{E}_{k, N_t-1}^*}{\tau} \overset{0.5}{E}_{k, N_t} \right) + \frac{D}{\gamma} |\overset{0.5}{E}_{k, N_t}| \right] = 0. \quad (69)$$

Теперь сложим (66) и (69) и получим

$$\sum_{k=0}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-1} \tau \left[\frac{|\hat{E}|^2 - |E|^2}{h_z} + \alpha \gamma |A|^2 \left(\frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \overset{0.5}{E}^* + \frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \overset{0.5}{E} \right) \right] + \right.$$

$$\left. + \frac{iD}{2} \left(\frac{\overset{0.5}{E}_{k, N_t} - \overset{0.5}{E}_{k, N_t-1}}{\tau} \overset{0.5}{E}_{k, N_t}^* - \frac{\overset{0.5}{E}_{k, N_t}^* - \overset{0.5}{E}_{k, N_t-1}^*}{\tau} \overset{0.5}{E}_{k, N_t} \right) + \frac{\tau}{2} \frac{|\hat{E}_{k, N_t}|^2 - |E_{k, N_t}|^2}{h_z} + \frac{D}{\gamma} |\overset{0.5}{E}_{k, N_t}| \right\} = 0. \quad (70)$$

Таким образом, в соотношение (70) разностная производная по координате z входит во всех узлах сетки, где она отлична от нуля.

С целью преобразования (70) к инварианту (59) запишем еще одно разностное соотношение, умножив для этого уравнение (41) на $r_k h_r (\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*)/(2\tau)$, а сопряженное к нему – на $r_k h_r (\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1})/(2\tau)$. Вычтем из первого произведения второе и полученное соотношение просуммируем по n от 1 до $N_t - 1$ и по k от 1 до $N_r - 1$, предварительно умножив его на шаг по времени τ . В результате получим

$$\sum_{k=1}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left[\frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \frac{\hat{E} - E}{h_z} - \frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \frac{\hat{E}^* - E^*}{h_z} + iD \left(\frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \Lambda_{\bar{r}t} \overset{0.5}{E} - \frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \Lambda_{\bar{r}t} \overset{0.5}{E}^* \right) + \right.$$

$$\left. + iD_r \left(\frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \Lambda_{\bar{r}r} \overset{0.5}{E} - \frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \Lambda_{\bar{r}r} \overset{0.5}{E}^* \right) + \alpha \gamma |A|^2 \left(\frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} A + \frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} A^* \right) \right] = 0. \quad (71)$$

Как и при записи соотношения (70), в (71) необходимо добавить слагаемые, содержащие производную по координате z от функций в узлах сетки при $k = 0$, а также $n = N_t$, и исключить разностную производную по координате r . Для этого преобразуем слагаемое, содержащее разностный оператор Лапласа по координате r в (71), учитывая вторую разностную формулу Грина и граничные условия (46), к виду

$$iD_r \sum_{n=1}^{N_t-1} \tau \sum_{k=1}^{N_r-1} r_k h_r \left(\frac{\overset{0.5}{E}_{+1}^* - \overset{0.5}{E}_{-1}^*}{2\tau} \Lambda_{\bar{r}r} \overset{0.5}{E} - \frac{\overset{0.5}{E}_{+1} - \overset{0.5}{E}_{-1}}{2\tau} \Lambda_{\bar{r}r} \overset{0.5}{E}^* \right) =$$

$$= -iD_r \sum_{k=0}^{N_r-1} r_k h_r \frac{r_{k+0.5}}{2r_k} \left(\frac{\overset{0.5}{E}_{k+1, N_t}^* - \overset{0.5}{E}_{k, N_t}^*}{h_r} \frac{\overset{0.5}{E}_{k+1, N_t-1} - \overset{0.5}{E}_{k, N_t-1}}{h_r} + \frac{\overset{0.5}{E}_{k+1, N_t} - \overset{0.5}{E}_{k, N_t}}{h_r} \frac{\overset{0.5}{E}_{k+1, N_t-1}^* - \overset{0.5}{E}_{k, N_t-1}^*}{h_r} \right) -$$

$$- iD_r \sum_{n=1}^{N_t-1} \tau \overset{0.5}{E}_{0, n+1} \left(\frac{\overset{0.5}{E}_{0, n+1}^* - \overset{0.5}{E}_{0, n-1}^*}{2\tau} \frac{\overset{0.5}{E}_{1, n} - \overset{0.5}{E}_{0, n}}{h_r} + \frac{\overset{0.5}{E}_{0, n+1} - \overset{0.5}{E}_{0, n-1}}{2\tau} \frac{\overset{0.5}{E}_{1, n}^* - \overset{0.5}{E}_{0, n}^*}{h_r} \right).$$

Преобразуем последнюю сумму в (72). Для этого умножим краевое условие (47) на $r_k h_r \frac{E_{0,n+1}^{0.5*} - E_{0,n-1}^{0.5*}}{2}$, сопряженное к нему – на $r_k h_r \frac{E_{0,n+1}^{0.5} - E_{0,n-1}^{0.5}}{2}$. Вычтем из первого произведения второе и полученное соотношение просуммируем от 1 до $N_t - 1$. В результате получим

$$\begin{aligned} iD_r \sum_{n=1}^{N_t-1} \tau r_{0.5} \left(\frac{E_{0,n+1}^{0.5*} - E_{0,n-1}^{0.5*}}{2\tau} \frac{E_{1,n}^{0.5} - E_{0,n}^{0.5}}{h_r} - \frac{E_{0,n+1}^{0.5} - E_{0,n-1}^{0.5}}{2\tau} \frac{E_{1,n}^{0.5*} - E_{0,n}^{0.5*}}{h_r} \right) = \\ = - \sum_{n=0}^{N_t-1} \tau \left[\frac{E_{0,n+1}^{0.5*} - E_{0,n-1}^{0.5*}}{2\tau} \frac{\hat{E}_{0,n} - E_{0,n}}{h_z} - \frac{E_{0,n+1}^{0.5} - E_{0,n-1}^{0.5}}{2\tau} \frac{\hat{E}_{0,n}^* - E_{0,n}^*}{h_z} + \right. \\ \left. + iD \left(\frac{E_{0,n+1}^{0.5*} - E_{0,n-1}^{0.5*}}{2\tau} \Lambda_{it} E_{0,n}^{0.5} - \frac{E_{0,n+1}^{0.5} - E_{0,n-1}^{0.5}}{2\tau} \Lambda_{it} E_{0,n}^{0.5*} \right) + \right. \\ \left. + \alpha \gamma |A_{0,n}|^2 \left(\frac{E_{0,n+1}^{0.5*} - E_{0,n-1}^{0.5*}}{2\tau} A_{0,n}^{0.5} + \frac{E_{0,n+1}^{0.5} - E_{0,n-1}^{0.5}}{2\tau} A_{0,n}^{0.5*} \right) \right]. \end{aligned} \quad (73)$$

Подставим (73) в (72), а затем полученное выражение – в (71). Тогда оно запишется в виде

$$\begin{aligned} \sum_{k=0}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-1} \tau \left[\frac{E_{+1}^{0.5*} - E_{-1}^{0.5*}}{2\tau} \frac{\hat{E} - E}{h_z} - \frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} \frac{\hat{E}^* - E^*}{h_z} + iD \left(\frac{E_{+1}^{0.5*} - E_{-1}^{0.5*}}{2\tau} \Lambda_{it} E^{0.5} - \frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} \Lambda_{it} E^{0.5*} \right) + \right. \right. \\ \left. \left. + \alpha \gamma |A|^2 \left(\frac{E_{+1}^{0.5*} - E_{-1}^{0.5*}}{2\tau} A^{0.5} + \frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} A^{0.5*} \right) \right] - iD_r \frac{1}{2} r_{k+0.5} \left(\frac{E_{k+1,N_t}^{0.5*} - E_{k,N_t}^{0.5*}}{h_r} \frac{E_{k+1,N_t-1}^{0.5} - E_{k,N_t-1}^{0.5}}{h_r} + \right. \right. \\ \left. \left. + \frac{E_{k+1,N_t}^{0.5*} - E_{k,N_t}^{0.5*}}{h_r} \frac{E_{k+1,N_t-1}^{0.5} - E_{k,N_t-1}^{0.5}}{h_r} \right) \right\} = 0. \end{aligned} \quad (74)$$

Таким образом, последнее выражение содержит слагаемые в узлах сетки $k = 0$.

Преобразуем слагаемые в (74), содержащие разностную производную по продольной координате, следующим образом:

$$\begin{aligned} \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{E_{+1}^{0.5*} - E_{-1}^{0.5*}}{2\tau} \frac{\hat{E} - E}{h_z} - \frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} \frac{\hat{E}^* - E^*}{h_z} \right) = \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{\hat{E}_{+1} - \hat{E}_{-1}}{2h_z} E^{0.5} - \frac{E_{+1}^* - E_{-1}^*}{2h_z} E^{0.5*} \right) - \\ - \frac{1}{2} \left(\frac{E_{k,N_t-1}^{0.5}}{E_{k,N_t-1}^{0.5*}} \frac{\hat{E}_{k,N_t} - E_{k,N_t}^*}{h_z} + E_{k,N_t}^{0.5} \frac{\hat{E}_{k,N_t-1}^* - E_{k,N_t-1}^*}{h_z} \right). \end{aligned} \quad (75)$$

Учитывая вторую разностную формулу Грина и граничные условия (44), преобразуем слагаемое, содержащее разностный оператор Лапласа по координате t в (74), к виду

$$\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{E_{+1}^{0.5*} - E_{-1}^{0.5*}}{2\tau} \Lambda_{it} E^{0.5} - \frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2\tau} \Lambda_{it} E^{0.5*} \right) = \sum_{k=0}^{N_r-1} r_k h_r \left| \frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^{0.5}}{\tau} \right|. \quad (76)$$

Подставив в (74) выражения (75) вместо (74), получим соотношение

$$\begin{aligned} \sum_{k=0}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-1} \left[\left(\hat{E}_{n+1}^* - \hat{E}_n^* - E_{n+1}^* + E_n^* \right) + i\alpha |A|^{0.5} \left(\frac{E_{n+1}^* - E_n^*}{2} \hat{E}_n^* + \frac{E_n^* - E_{n-1}^*}{2} \hat{E}_{n+1}^* \right) \right] - \right. \\ \left. - \frac{1}{2} \left(\frac{E_{k,N_t-1}^* - E_{k,N_t}^*}{h_z} + \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{h_z} \right) - \right. \\ \left. - \frac{iD_r}{2} r_{k+0.5} \left(\frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^*}{h_r} + \frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^*}{h_r} \right) + \right. \\ \left. + iD \left| \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{\tau} \right|^2 \right\} = 0. \end{aligned} \quad (77)$$

Теперь необходимо исключить из (77) разностные производные по r . Для этого умножим краевое условие (47) на $r_k h_r (E_{k,N_t}^* - E_{k,N_t-1}^*)/\tau$, а сопряженное к нему – на $r_k h_r (E_{k,N_t}^* - E_{k,N_t-1}^*)/\tau$, вычтем из первого произведения второе, умножим результат на $\tau/2$ и просуммируем полученное выражение по k от 1 до $N_r - 1$. В результате получим

$$\begin{aligned} \sum_{k=1}^{N_r-1} r_k h_r 0.5 \left\{ (E_{k,N_t}^* - E_{k,N_t-1}^*) \frac{E_{k,N_t}^* - E_{k,N_t}^*}{h_z} - (E_{k,N_t}^* - E_{k,N_t-1}^*) \frac{E_{k,N_t}^* - E_{k,N_t}^*}{h_z} - i2D \left| \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{\tau} \right|^2 + \right. \\ \left. + 2 \frac{D}{\gamma} \left(\frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{\tau} E_{k,N_t}^* - \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{\tau} E_{k,N_t}^* \right) + \right. \\ \left. + iD_r [(E_{k,N_t}^* - E_{k,N_t-1}^*) \Lambda_{rr} E_{k,N_t}^* - (E_{k,N_t}^* - E_{k,N_t-1}^*) \Lambda_{rr} E_{k,N_t}^*] \right\} = 0. \end{aligned} \quad (78)$$

Применяя вторую разностную формулу Грина к слагаемым с разностным оператором Лапласа по координате r в (78), учитывая краевое условие (47а) в точке $(r_0 = 0.5h_r, t = L_t)$, преобразуем их к виду

$$\begin{aligned} iD_r \frac{1}{2} \sum_{k=1}^{N_r-1} r_k h_r [(E_{k,N_t}^* - E_{k,N_t-1}^*) \Lambda_{rr} E_{k,N_t}^* + (E_{k,N_t}^* - E_{k,N_t-1}^*) \Lambda_{rr} E_{k,N_t}^*] = \\ = -\frac{iD_r}{2} \sum_{k=0}^{N_r-1} r_{k+0.5} h_r \left[\left(\frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} - \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^*}{h_r} \right) \frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} + \right. \\ \left. + \left(\frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} - \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^*}{h_r} \right) \frac{E_{k+1,N_t}^* - E_{k,N_t}^*}{h_r} \right] + \\ + \frac{iD_r}{2} h_r \left[(E_{0,N_t}^* - E_{0,N_t-1}^*) \frac{E_{1,N_t}^* - E_{0,N_t}^*}{h_r} + (E_{0,N_t}^* - E_{0,N_t-1}^*) \frac{E_{1,N_t}^* - E_{0,N_t}^*}{h_r} \right]. \end{aligned} \quad (79)$$

Для исключения разностной производной по координате r в узлах сетки $k=0$ в (79) домножим краевое условие (44) на $0.5h_r r_0 (E_{0,N_t}^* - E_{0,N_t}^*)$, а сопряженное к нему – на $0.5h_r r_0 (E_{0,N_t}^* - E_{0,N_t}^*)$ и вычтем их друг из друга. В результате получим

$$iD_r 0.5 r_0 \left[(E_{0,N_t}^* - E_{0,N_t}^{*0.5}) \frac{E_{1,N_t}^{0.5} - E_{0,N_t}^{0.5}}{h_r} + (E_{0,N_t}^{0.5} - E_{0,N_t}^*) \frac{E_{1,N_t}^* - E_{0,N_t}^*}{h_r} \right] = - \left[\frac{(E_{0,N_t}^* - E_{0,N_t}^{*0.5}) \hat{E}_{0,n} - E_{0,n}}{2 h_z} - \right. \\ \left. - \frac{E_{0,N_t}^{0.5} - E_{0,N_t}^*}{2} \frac{\hat{E}_{0,n} - E_{0,n}^*}{h_z} + iD \left| \frac{E_{0,N_t}^{0.5} - E_{0,N_t}^*}{\tau} \right|^2 + \frac{D}{2\gamma} \left(\frac{E_{0,N_t}^* - E_{0,N_t-1}^{0.5}}{\tau} E_{0,N_t}^{0.5} - \frac{E_{0,N_t}^{0.5} - E_{0,N_t-1}^*}{\tau} E_{0,N_t}^* \right) \right]. \quad (80)$$

Тогда (79) с учетом (80) преобразуется к виду

$$\frac{iD_r}{2} \sum_{k=1}^{N_r-1} r_k h_r [(E_{k,N_t}^* - E_{k,N_t-1}^{0.5}) \Lambda_{\bar{r}r}^{0.5} E_{k,N_t}^{0.5} + (E_{k,N_t}^{0.5} - E_{k,N_t-1}^*) \Lambda_{\bar{r}r}^{0.5} E_{k,N_t}^*] = \\ = - \frac{iD_r}{2} \sum_{k=0}^{N_r-1} r_{k+0.5} h_r \left[\left(\frac{E_{k+1,N_t}^* - E_{k,N_t}^{0.5}}{h_r} - \frac{E_{k+1,N_t-1}^{0.5} - E_{k,N_t-1}^*}{h_r} \right) \frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^*}{h_r} + \right. \\ \left. + \left(\frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^*}{h_r} - \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^{0.5}}{h_r} \right) \frac{E_{k+1,N_t}^* - E_{k,N_t}^{0.5}}{h_r} \right] - \\ - \frac{E_{0,N_t}^{0.5} - E_{0,N_t}^*}{2} \frac{\hat{E}_{0,n} - E_{0,n}}{h_z} - \frac{E_{0,N_t}^{0.5} - E_{0,N_t}^*}{2} \frac{\hat{E}_{0,n}^* - E_{0,n}^*}{h_z} + \\ + iD \left| \frac{E_{0,N_t}^{0.5} - E_{0,N_t}^*}{\tau} \right|^2 + \frac{D}{2\gamma} \left(\frac{E_{0,N_t}^* - E_{0,N_t-1}^{0.5}}{\tau} E_{0,N_t}^{0.5} - \frac{E_{0,N_t}^{0.5} - E_{0,N_t-1}^*}{\tau} E_{0,N_t}^* \right). \quad (81)$$

Заменим в (78) слагаемые, содержащие оператор Лапласа по координате r правой частью выражения (81), и в результате получим соотношение

$$\sum_{k=0}^{N_r-1} r_k h_r \left\{ \left(\hat{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - \hat{E}_{k,N_t-1}^*}{2h_z} - E_{k,N_t} \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{2h_z} \right) - \frac{E_{k,N_t}^{0.5} - E_{k,N_t}^*}{h_z} \hat{E}_{k,N_t}^* + \right. \\ \left. + \frac{1}{2} \left(\frac{E_{k,N_t-1}^{0.5} - E_{k,N_t}^*}{h_z} \hat{E}_{k,N_t}^* + E_{k,N_t} \frac{\hat{E}_{k,N_t-1}^* - E_{k,N_t-1}^*}{h_z} \right) - iD \left| \frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^*}{\tau} \right|^2 + \right. \\ \left. + \frac{D}{2\gamma} \left(\frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^*}{\tau} E_{k,N_t}^{0.5} - E_{k,N_t} \frac{E_{k,N_t-1}^{0.5} - E_{k,N_t-1}^*}{\tau} \right) - \right. \\ \left. - \frac{iD_r}{2} r_{k+0.5} h_r \left[\left(\frac{E_{k+1,N_t}^* - E_{k,N_t}^{0.5}}{h_r} - \frac{E_{k+1,N_t-1}^{0.5} - E_{k,N_t-1}^*}{h_r} \right) \frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^*}{h_r} + \right. \right. \\ \left. \left. + \left(\frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^*}{h_r} - \frac{E_{k+1,N_t-1}^* - E_{k,N_t-1}^{0.5}}{h_r} \right) \frac{E_{k+1,N_t}^* - E_{k,N_t}^{0.5}}{h_r} \right] \right\} = 0. \quad (82)$$

Сложив выражения (77) и (82), приходим к равенству

$$\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \left[\left(\hat{E} \frac{\hat{E}_{+1}^* - \hat{E}_{-1}^*}{2h_z} - E \frac{E_{+1}^* - E_{-1}^*}{2h_z} \right) + i\alpha |A|^2 \left(\frac{E_{+1}^{0.5} - E_{-1}^{0.5}}{2} E^* + \frac{E_{+1}^* - E_{-1}^*}{2} E^{0.5} \right) \right] + \\ + \frac{1}{2} \left(\hat{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - \hat{E}_{k,N_t-1}^*}{h_z} - E_{k,N_t} \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{h_z} \right) - \frac{E_{k,N_t}^{0.5} - E_{k,N_t}^*}{h_z} \hat{E}_{k,N_t}^* - iD \left| \frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^*}{\tau} \right|^2 + \\ + \frac{D}{2\gamma} \left(\frac{E_{k,N_t}^{0.5} - E_{k,N_t-1}^*}{\tau} E_{k,N_t}^{0.5} - E_{k,N_t} \frac{E_{k,N_t-1}^{0.5} - E_{k,N_t-1}^*}{\tau} \right) - iD_r \frac{r_{k+0.5}}{r_k} \left| \frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^*}{h_r} \right|^2 = 0. \quad (83)$$

Домножив на мнимую единицу (70) и затем сложив его с выражением (83), умноженным на $-\gamma$, получим соотношение

$$\begin{aligned} \sum_{k=0}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-1} \left[i\tau \frac{|\hat{E}|^2 - |E|^2}{h_z} - \gamma \left(\hat{E} \frac{\hat{E}_{+1}^* - \hat{E}_{-1}^*}{2h_z} - E \frac{E_{+1}^* - E_{-1}^*}{2h_z} \right) \right] + i\frac{\tau}{2} \frac{|\hat{E}_{k,N_t}|^2 - |E_{k,N_t}|^2}{h_z} - \right. \\ \left. - \gamma \left(\hat{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - \hat{E}_{k,N_t-1}^*}{2h_z} - E_{k,N_t} \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{2h_z} \right) + \gamma^{0.5} E_{k,N_t} \frac{\hat{E}_{k,N_t}^* - E_{k,N_t}^*}{h_z} + \right. \\ \left. + \frac{iD}{\gamma} |E_{k,N_t}|^2 + \gamma i D_r \frac{r_{k+0.5}}{r_k} \left| \frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^{0.5}}{h_r} \right|^2 \right\} = 0. \end{aligned} \quad (84)$$

Далее, умножив сопряженное уравнение (47) на $r_k h_r \gamma E_{k,N_t}^{0.5}$, просуммировав по k от 1 до $N_r - 1$, применив первую разностную формулу Грина по координате r и учтя краевое условие (46б) для исключения разностной производной по координате r в узлах сетки $k = 0$ из выражения, записанного после применения формулы Грина, получим

$$\sum_{k=0}^{N_r-1} r_k h_r \gamma \left(E_{k,N_t} \frac{\hat{E}_{k,N_t}^* - E_{k,N_t}^*}{h_z} + \frac{iD}{\gamma} |E_{k,N_t}|^2 + iD_r \frac{r_{k+0.5}}{r_k} \left| \frac{E_{k+1,N_t}^{0.5} - E_{k,N_t}^{0.5}}{h_r} \right|^2 \right) = 0. \quad (85)$$

Следовательно, (84) примет вид

$$\begin{aligned} \sum_{k=0}^{N_r-1} r_k h_r \left[\sum_{n=1}^{N_t-1} \left[i\tau \frac{|\hat{E}|^2 - |E|^2}{h_z} - i\gamma \operatorname{Im} \left(\hat{E} \frac{\hat{E}_{+1}^* - \hat{E}_{-1}^*}{2h_z} - E \frac{E_{+1}^* - E_{-1}^*}{2h_z} \right) \right] + \right. \\ \left. + i\frac{\tau}{2} \frac{|\hat{E}_{k,N_t}|^2 - |E_{k,N_t}|^2}{h_z} - i\gamma \operatorname{Im} \left(\hat{E}_{k,N_t} \frac{\hat{E}_{k,N_t}^* - \hat{E}_{k,N_t-1}^*}{2h_z} - E_{k,N_t} \frac{E_{k,N_t}^* - E_{k,N_t-1}^*}{2h_z} \right) \right] = 0, \end{aligned}$$

что означает сохранение разностного аналога инварианта (59) на решении разностной задачи.

Консервативность схемы по инварианту (60). Умножим релаксационное уравнение (42), записанное на нижнем слое по z , на $r_k h_r E^*$, а сопряженное к нему – на $r_k h_r E$. Сложим полученные выражения, а результат умножим на τ и просуммируем по n от 1 до $N_t - 1$ и по k от 0 до $N_r - 1$. В результате можно записать следующую цепочку преобразований:

$$\begin{aligned} 2 \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \operatorname{Re}(AE^*) &= \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \left(\frac{E_{+1} - E_{-1}}{2} E^* + \frac{E_{+1}^* - E_{-1}^*}{2} E \right) = \\ &= \sum_{k=0}^{N_r-1} \frac{r_k h_r}{2} \left[\sum_{n=1}^{N_t-1} \left(\frac{E_{+1} - E}{2} E^* + \frac{E_{+1}^* - E^*}{2} E \right) + \sum_{n=0}^{N_t-2} \left(\frac{E_{+1} - E}{2} E_{+1}^* + \frac{E_{+1}^* - E^*}{2} E_{+1} \right) \right] = \\ &= \sum_{k=0}^{N_r-1} r_k h_r \left\{ \sum_{n=1}^{N_t-2} \left[(E_{+1} - E) \left(\frac{E_{+1}^* + E^*}{2} \right) + (E_{+1}^* - E^*) \left(\frac{E_{+1} + E}{2} \right) \right] + \right. \\ &\quad \left. + \frac{1}{2} [(E_{k,N_t} - E_{k,N_t-1}) E_{k,N_t-1}^* + (E_{k,N_t}^* - E_{k,N_t-1}^*) E_{k,N_t-1}] \right\} = \end{aligned} \quad (86)$$

$$\begin{aligned}
&= \sum_{k=0}^{N_r-1} r_k h_r \left[\sum_{n=1}^{N_t-2} (|E_{+1}|^2 - |E|^2) - |E_{k, N_t-1}|^2 + \frac{1}{2} (E_{k, N_t} E_{k, N_t-1}^* + E_{k, N_t}^* E_{k, N_t-1}) \right] = \\
&= \sum_{k=0}^{N_r-1} \frac{r_k h_r}{2} (E_{k, N_t} E_{k, N_t-1}^* + E_{k, N_t}^* E_{k, N_t-1}).
\end{aligned}$$

Для продолжения преобразования последней суммы в (86) умножим уравнение (46) на $r_k h_r E_{k, N_t}^* \tau/2$, а сопряженное к нему – на $r_k h_r E_{k, N_t} \tau/2$. Сложив их и просуммировав по k от 0 до $N_r - 1$, получим равенство

$$\sum_{k=0}^{N_r-1} r_k h_r \left[-\frac{1}{2} (E_{k, N_t-1} E_{k, N_t}^* + E_{k, N_t-1}^* E_{k, N_t}) + \left(1 - \frac{\tau^2}{\gamma^2}\right) |E_{k, N_t}|^2 \right] = 0. \quad (87)$$

Следовательно, с учетом (87) разностное соотношение (86) преобразуется к виду

$$2 \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \operatorname{Re}(A E^*) = \sum_{k=0}^{N_r-1} r_k h_r |E_{k, N_t}|^2 \left(1 - \frac{\tau^2}{\gamma^2}\right). \quad (88)$$

Таким образом, для доказательства существования инварианта (60) нужно показать, что

$$\sum_{k=0}^{N_r-1} r_k h_r |E_{k, N_t}|^2 = \text{const}, \quad (89)$$

т.е. является инвариантом. Для этого умножим уравнение (46а) на $r_k h_r \hat{E}_{k, N_t}^{0.5}$, а сопряженное к нему – на $r_k h_r E_{k, N_t}^{0.5}$. Сложив их, просуммировав по k от 1 до $N_r - 1$, применив вторую разностную формулу Грина с учетом краевого условия (47б), получим

$$\sum_{k=0}^{N_r-1} r_k h_r \left(\frac{\hat{E}_{k, N_t} - E_{k, N_t}^{0.5} E_{k, N_t}^*}{h_z} + E_{k, N_t}^{0.5} \frac{\hat{E}_{k, N_t}^* - E_{k, N_t}^*}{h_z} \right) = \sum_{k=0}^{N_r-1} r_k h_r \frac{|\hat{E}_{k, N_t}|^2 - |E_{k, N_t}|^2}{h_z} = 0.$$

Таким образом, утверждение (89) доказано, а (60) является разностным аналогом инварианта.

Консервативность схемы по инварианту (61). Умножим уравнение (42), записанное на нижнем слое по z , на $r_k h_r A^*$, а сопряженное к нему – на $r_k h_r A$. Вычтем из первого второе, результат умножим на τ , а затем просуммируем по n от 1 до $N_t - 1$ и по k от 0 до $N_r - 1$. В результате из этого равенства получим

$$\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(\frac{E_{+1} - E_{-1}}{2\tau} A^* - \frac{E_{+1}^* - E_{-1}^*}{2\tau} A \right) = -\frac{i}{\gamma} \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau (E A^* + E^* A) = \text{const}. \quad (90)$$

Постоянство значения последней суммы обусловлено справедливостью разностного аналога инварианта (60).

Преобразуем левую часть равенства (90), учитывая нулевые граничные условия (54) для разностной функции A , следующим образом:

$$\begin{aligned}
&\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left(A^* \frac{E_{+1} - E_{-1}}{2\tau} - A \frac{E_{+1}^* - E_{-1}^*}{2\tau} \right) = \sum_{k=0}^{N_r-1} r_k h_r \left[\sum_{n=1}^{N_t-1} \tau \left(A^* \frac{E_{+1} - E_{-1}}{2\tau} - A \frac{E_{+1}^* - E_{-1}^*}{2\tau} \right) + \right. \\
&+ \sum_{n=1}^{N_t-1} \tau \left(A^* \frac{E - E_{-1}}{2\tau} - A \frac{E^* - E_{-1}^*}{2\tau} \right) \left. \right] = -\frac{1}{2} \sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left[E^* \left(\frac{A - A_{-1}}{\tau} + \frac{A_{+1} - A}{\tau} \right) - \right. \\
&\left. - E \left(\frac{A^* - A_{-1}^*}{\tau} + \frac{A_{+1}^* - A^*}{\tau} \right) \right] = -\sum_{k=0}^{N_r-1} r_k h_r \sum_{n=1}^{N_t-1} \tau \left[E^* \frac{(A_{+1} - A_{-1})}{2\tau} - E \frac{(A_{+1}^* - A_{-1}^*)}{2\tau} \right]. \quad (91)
\end{aligned}$$

Из (91) и (90) следует, что разностный аналог инварианта (61) сохраняется на разностном решении.

Консервативность схемы по инварианту (62). Умножив на $r_k h_r$ и просуммировав по k от 1 до $N_r - 1$ краевое условие (47), получим

$$\sum_{k=1}^{N_r-1} r_k h_r \left(\frac{\hat{E}_{k, N_t} - E_{k, N_t}}{h_z} - \frac{iD^{0.5}}{\gamma^2} E_{k, N_t} + iD_r \Lambda_{\tilde{r}r}^{0.5} E_{k, N_t} \right) = \sum_{k=1}^{N_r-1} r_k h_r \left(\frac{\hat{E}_{k, N_t} - E_{k, N_t}}{h_z} - \frac{iD^{0.5}}{\gamma^2} E_{k, N_t} \right) + iD_r \left(r_{N_r-0.5} \frac{E_{N_r, N_t} - E_{N_r-1, N_t}}{h_r} - r_{0.5} \frac{E_{1, N_t} - E_{0, N_t}}{h_r} \right) = 0. \quad (92)$$

Учитывая краевые условия (47б) и (9г) в точке $n = N_t$, преобразуем (92) к виду

$$\sum_{k=0}^{N_r-1} r_k h_r \left(\frac{\hat{E}_{k, N_t} - E_{k, N_t}}{h_z} - \frac{iD^{0.5}}{\gamma^2} E_{k, N_t} \right) = 0.$$

Отсюда следует соотношение

$$I_{SP}(z_m) = \sum_{k=0}^{N_r-1} r_k h_r E(z_m, r_k, N_t) = \left(\frac{a_{SP}}{b_{SP}} \right)^m \sum_{k=0}^{N_r-1} r_k h_r E_0(r_k, N_t),$$

$$a_{SP} = 1 - \left(\frac{Dh_z}{2\gamma^2} \right)^2 + \frac{iDh_z}{\gamma^2}, \quad b_{SP} = 1 + \left(\frac{Dh_z}{2\gamma^2} \right)^2,$$

которое представляет собой разностный аналог инварианта (62) на введенной сетке.

5. РЕЗУЛЬТАТЫ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

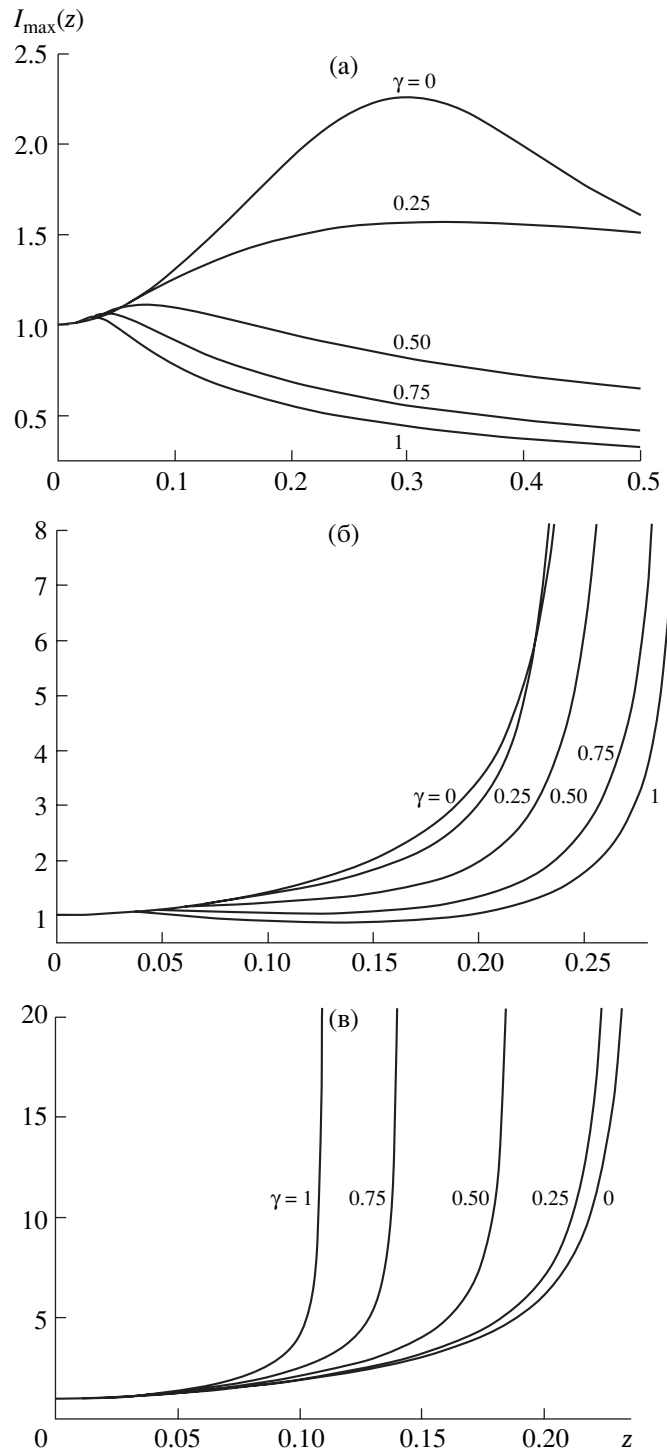
Компьютерное моделирование проводилось для начального распределения комплексной амплитуды в виде гауссовой функции по пространственной координате и времени:

$$A_0(r, t) = e^{-r^2 - \eta^2}, \quad \eta = t - L_t/2.$$

Так как рассматривалось влияние дисперсии нелинейного отклика на самофокусировку аксиально-симметричного пучка, то коэффициенты в (1) выбирались равными $\alpha = 10$, $D = 1$, а параметры D_r и γ были, соответственно, равны $\{10^{-4}; 0.1; 1\}$ и $\{0; 0.25; 0.5; 0.75; 1\}$. Учитывая, что в аксиально-симметричном случае при выбранном значении параметра нелинейности формируется нелинейный фокус, то число узлов по времени и по радиусу мы выбирали равными $N_t = 8192$, $N_r = 1000$ для области $L_t = 30$, $R = 6$. По мере приближения к нелинейному фокусу шаг по продольной координате уменьшался, так как возрастала интенсивность оптического излучения. Результаты расчетов в виде зависимости максимальной интенсивности $I_{\max}(z) = \max_{r,t} |A(z, r, t)|^2$ от z представлены на фигуре. В процессе расчетов контролировалось сохранение разностных аналогов инвариантов, и оно было в пределах точности итерационного процесса.

Прежде всего отметим, что для дифракции, равной 10^{-4} (см. график (а)), эволюция пиковой интенсивности совпадает со случаем задачи в координатах (z, t) , когда $D_r = 0$, и с ростом параметра γ максимальная интенсивность уменьшается. Рост дифракции до 0.1 приводит к качественно новой зависимости: нелинейный фокус формируется всегда (см. график (б)). Однако трасса распространения, на которой он появляется, увеличивается с ростом γ . Причем для $\gamma = 0.75$ и 1 в начале распространения пиковая интенсивность импульса несколько уменьшается.

При дальнейшем росте влияния дифракции пучка до $D_r = 1$ (график (в)) зависимость положения нелинейного фокуса от γ становится иной: при увеличении параметра γ трасса формирования нелинейного фокуса уменьшается, а скорость роста интенсивности вблизи него возрастает. Следовательно, влияние дисперсии нелинейного отклика (параметра γ) на характер распространения оптического излучения существенно зависит от его дифракции.



Фигура.

В заключение отметим, что расчеты проводились на двухпроцессорном компьютере Intel XEON с частотой 3.4 ГГц и двухпроцессорном двухядерном компьютере Intel XEON с частотой 3 ГГц, а также на двухпроцессорном компьютере Intel Itanium 2 с частотой 1.5 ГГц с использованием библиотеки MKL, разработанной для процессоров Intel, и распараллеливанием алгоритмов для 2 и 4 процессоров. Это позволяло уменьшить время расчетов в 2.8–4.5 раза по сравнению с расчетами, проведенными на одном процессоре.

6. ВЫВОДЫ

Таким образом, в настоящей работе записаны инварианты задачи распространения фемтосекундного лазерного импульса в кубично-нелинейной среде в аксиально-симметричном случае с учетом производной по времени от нелинейного отклика. Для этой задачи построена консервативная разностная схема. Доказана ее консервативность по ряду инвариантов. Разностная схема использована для исследования влияния дисперсии нелинейного отклика среды, описываемого производной по времени от него, на формирование нелинейного фокуса и на рост интенсивности в нелинейном фокусе. Показано, что влияние дисперсии нелинейного отклика зависит от величины дифракции оптического излучения.

СПИСОК ЛИТЕРАТУРЫ

1. Ахманов С.А., Выслоух В. А., Чиркин А.С. Оптика фемтосекундных лазерных импульсов. М.: Наука, 1988.
2. Сухоруков А.П. Нелинейные волновые взаимодействия в оптике и радиофизике. М.: Наука, 1988.
3. Боровский А.В., Галкин А.Л. Лазерная физика. М.: Изд-во АТ, 1996.
4. Коротеев Н.И., Шумай И.Л. Физика мощного лазерного излучения. М.: Наука, 1991.
5. Агравал Г. Нелинейная волоконная оптика. М.: Мир, 1996.
6. Варенцова С.А., Волков А.Г., Трофимов В.А. Консервативная разностная схема для задачи распространения фемтосекундного лазерного импульса в кубично-нелинейной среде// Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 11. С. 1709–1721.
7. Трофимов В. А. К вопросу об инвариантах нелинейного распространения фемтосекундных импульсов// Изв. вузов. Радиофизика. 1992. Т. 35. № 6–7. С. 618–621.
8. Трофимов В. А. Инварианты нелинейного распространения фемтосекундных импульсов // Изв. вузов. Радиофизика. 1999. Т. 62. № 4. С. 369–372.
9. Трофимов В. А. О новом подходе к моделированию нелинейного распространения сверхкоротких лазерных импульсов// Ж. вычисл. матем. и матем. физ. 1998. Т. 38. № 5. С. 835–839.
10. Волков А.Г., Трофимов В.А., Терешин Е.Б. Консервативные разностные схемы для некоторых задач фемтосекундной оптики// Дифференц. уравнения. 2005. Т. 41. № 7. С. 908–917.
11. Аскаръян Г.А. Воздействие градиента поля интенсивного электромагнитного луча на электроны и атомы// Ж. эксперим. и теор. физ. 1962. Т. 42. № 6. С. 1567–1570.
12. Ахманов С.А., Сухоруков А.П., Хохлов Р.В. Самофокусировка и дифракция света в нелинейной среде // Успехи физ. наук. 1967. Т. 93. № 1. С. 19–70.
13. Луговой В.Н., Прохоров А.М. Теория распространения мощного лазерного излучения в нелинейной среде // Успехи физ. наук. 1973. Т. 111. № 2. С. 203–260.
14. Berge L. Wave collapse in physics: principles and applications to light and plasma waves // Phys. Repts. 1998. V. 303. № 5–6. P. 259–370.
15. Feit M.D., Fleck J.A. Beam nonparaxiality, filament formation, and beam breakup in the self-focusing of optical beams // J. Optimizat. Soc. Amer. B. 1988. V. 5. № 3. P. 633–640.
16. Fibich G., Malkin V.M., Papanicolaou G.C. Beam self-focusing in the presence of small normal time dispersion// Phys. Rev. A. 1995. V. 52. № 5. P. 4218–4228.
17. Chernev P., Petrov V. Self-focusing of light pulses in the presence of normal group-velocity dispersion // Optimizat. Letts. 1992. V. 17. № 3. P. 172–174.
18. Luther G.G., Moloney J.V., Newell A.C., Wright E. M. Self-focusing threshold is normally dispersive media // Optimizat. Letts. 1994. V. 19. № 12. P. 862–864.
19. Куницын С.Д., Сухоруков А.П., Трофимов В.А. Самовоздействие оптического излучения в условиях генерации встречной волны // Изв. РАН. Сер. Физ. 1992. Т. 56. № 12. С. 201–208.
20. Silbergberg Y. Collapse of optical beams // Optimizat. Letts. 1990. V. 15. № 22. P. 1282–1284.
21. Kivshar Yu.S., Pelinovsky D.E. Self-focusing and transverse instabilities of solitary waves // Phys. Repts. 2000. V. 331. № 4. P. 117–195.
22. Самарский А.А. Теория разностных схем. М.: Наука, 1989.
23. Самарский А.А., Андреев В.Б. Разностные методы для эллиптических уравнений. М.: Наука, 1976.
24. Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. М.: Наука, 1978.

УДК 519.634

ЧИСЛЕННЫЙ АНАЛИЗ АТОМНОЙ СТРУКТУРЫ И ФОРМЫ МЕТАЛЛИЧЕСКИХ НАНОЧАСТИЦ¹⁾

© 2007 г. А. В. Вахрушев, А. М. Липанов

(426067 Ижевск, ул. Т. Барамзиной, 34, ИП Механ. УрО РАН)

e-mail: postmaster@ntm.udm.ru

Поступила в редакцию 01.03.2006 г.
Переработанный вариант 08.12.2006 г.

Методами численного моделирования в рамках молекулярной динамики исследованы закономерности зарождения, формирования и распространения на объем наночастицы при ее увеличении аморфной, кристаллической и поликристаллической фаз. Приведены результаты численных расчетов параметров, характеризующих указанные процессы, для металлических наночастиц, формируемых путем измельчения, дробления или деструкции макротел. Описаны основные закономерности протекания процессов формирования наночастиц в up-down-процессах и свойства их структуры. Библиография: 31. Фиг. 7.

Ключевые слова: наночастицы, структура, метод молекулярной динамики, численное моделирование, металлические наночастицы.

1. ВВЕДЕНИЕ

Одним из ключевых вопросов при создании порошковых нанокомпозитов является построение зависимости структуры и формы наночастиц, создающих основу композита, от их размеров. Это обусловлено тем, что при увеличении или уменьшении характерного размера наночастицы ее физико-механические характеристики – модуль упругости, прочность, деформационные и другие параметры, – могут изменяться на порядок (см. [1]–[3]). Расчетные и экспериментальные исследования показывают, что это объясняется, в первую очередь, существенной перестройкой (причем не обязательно монотонной) атомной структуры и формы наночастицы.

Необходимо отметить, что экспериментальное определение названных выше параметров наноэлементов представляет технически сложную и трудоемкую задачу вследствие малости размеров данных объектов. Кроме того, опытные данные зачастую противоречивы. В частности, это касается расстояния между атомами в наночастице. В некоторых работах указывается на увеличение данного расстояния для атомов, прилегающих к поверхности наночастицы, по сравнению с атомами, расположенными в ее глубине. В других работах наблюдается уменьшение этой величины при удалении от центра наночастицы (см. [8]). Сказанное выше обуславливает актуальность детального систематического исследования данного вопроса теоретическими методами – методами математического моделирования.

Следует также отметить, что атомная структура и форма наночастицы зависят не только от ее размера, но и от методов ее формирования, которые можно разделить на две основные группы.

1. Формирование наночастицы в процессе консолидации атомов путем их “сборки” и остановки (фиксации) данного процесса при достижении необходимого размера наночастицы (в англоязычной литературе эти методы получили название bottom up-процессы). Остановка процесса роста частицы выполняется изменением физических или химических условий образования частицы, прекращением подачи веществ, необходимых для ее образования, или вследствие ограничения пространства, в котором наночастица формируется.

2. Получение наночастиц путем измельчения, дробления, деструкции более массивных (крупных) образований до фрагментов нужного размера (в англоязычной литературе данные методы классифицированы как up-down-процессы).

Моделированию процессов первого типа (bottom up-процессов) посвящено достаточно много работ (см. [14], [15]), а протекание up-down-процессов изучено весьма слабо. Это объясняется

¹⁾ Работа выполнена при финансовой поддержке комплексной программы фундаментальных исследований Президиума РАН “Фундаментальные проблемы физики и химии наноразмерных систем и наноматериалов”, 2005–2006 гг.

тем, что они являются результатом процессов разрушения на разных структурных уровнях. Эта задача вследствие ее объемности и сложности в рамках данной работы не рассматривается. Подробное описание ее постановки и некоторые численные результаты расчетов приведены в [16]–[23]. Целью настоящей работы являлось исследование закономерностей изменения структуры и формы металлических наночастиц, формируемых в деструкционных (up down) процессах на завершающей стадии формирования наночастицы. В работе методами численного моделирования в рамках молекулярной динамики исследованы закономерности формирования аморфной фазы, зарождения и распространения на весь объем наночастицы (при ее увеличении), кристаллической (или поликристаллической) фаз в зависимости от размеров наночастиц. Приведены результаты расчетов параметров, характеризующих указанные процессы, для частиц меди.

2. ПОСТАНОВКА ЗАДАЧИ И МЕТОДИКА МОДЕЛИРОВАНИЯ

Задача расчета внутренней структуры и равновесной конфигурации (формы) отдельной, свободной от взаимодействия наночастицы методами молекулярной динамики включает два основных этапа.

Этап 1. “Инициализация” задачи, т.е. определение условий, при которых начинается процесс формирования наночастицы.

Этап 2. Расчет непосредственно процесса формирования структуры и формы наночастицы.

Рассмотрим последовательно указанные задачи. Инициализация задачи состоит в задании начальных координат и скоростей атомов наночастицы:

$$\mathbf{x}_i = \mathbf{x}_{i0}, \quad \mathbf{V}_i = \mathbf{V}_{i0}, \quad t = 0, \quad \mathbf{x}_i \in \Omega_k, \quad (1)$$

где $\mathbf{x}_{i0}$, $\mathbf{x}_i$ – начальные и текущие координаты i -го атома; $\mathbf{V}_{i0}$, $\mathbf{V}_i$ – начальная и текущая скорости i -го атома соответственно; Ω_k – область, занимаемая наночастицей.

Начальные координаты $\mathbf{x}_{i0}$ и скорости $\mathbf{V}_{i0}$ атомов наночастицы определяются из расчета деструкционных процессов при статическом и динамическом нагружениях, происходящих как в нано-, так и макромасштабах.

Второй этап решения задачи (формирование наночастицы) может осуществляться методами как молекулярной механики, так и молекулярной динамики. При этом методы молекулярной механики, основанные на минимизации потенциальной энергии наночастицы, не учитывают начальное распределение скоростей атомов наночастицы и поэтому не позволяют выявить полное влияние условий инициализации задачи на структуру и форму наночастицы. В связи с этим в данной работе использован метод молекулярной динамики, согласно которому движение атомов, образующих наночастицу, определяется системой дифференциальных уравнений Ланжевена (см. [24])

$$m_i \frac{d\mathbf{V}_i}{dt} = \sum_{j=1}^{N_k} \mathbf{F}_{ij} + \mathbf{F}_i(t) - \alpha_i m_i \mathbf{V}_i, \quad i = 1, 2, \dots, N_k, \quad (2)$$

$$d\mathbf{x}_i/dt = \mathbf{V}_i,$$

где N_k – число атомов, составляющих каждую наночастицу; m_i – масса i -го атома; $\mathbf{F}_{ij}$ – силы межатомного взаимодействия; α_i – коэффициент “трения” в атомарной системе (см. [25]); $\mathbf{F}_i(t)$ – случайный набор сил, задаваемый распределением Гаусса; t – время.

Силы межатомного взаимодействия являются потенциальными и определяются из соотношения

$$\mathbf{F}_{ij} = - \sum_1^q \frac{\partial \Phi(\mathbf{r}_{ij})}{\partial \mathbf{r}_{ij}}, \quad i = 1, 2, \dots, N_k, \quad j = 1, 2, \dots, N_k, \quad (3)$$

где $\mathbf{r}_{ij}$ – радиус-вектор, задающий положение i -го атома относительно j -го атома; $\Phi(\mathbf{r}_{ij})$ – потенциал, зависящий от взаимного расположения всех атомов; q – число типов взаимодействия между атомами.

Потенциал $\Phi(\mathbf{r}_{ij})$ в общем случае задается в виде суммы нескольких составляющих, соответствующих различным типам взаимодействия:

$$\Phi(\mathbf{r}_{ij}) = \Phi_{cb} + \Phi_{va} + \Phi_{ta} + \Phi_{pg} + \Phi_{vv} + \Phi_{es} + \Phi_{hb}. \quad (4)$$

Здесь имеются в виду следующие потенциалы: Φ_{cb} – химических связей; Φ_{va} – валентных углов; Φ_{ta} – торсионных углов; Φ_{pg} – плоских групп; Φ_{vv} – ван-дер-ваальсовых контактов; Φ_{es} – электростатики; Φ_{hb} – водородных связей.

Так как сформулированная задача не имеет аналитического решения, то интегрирование уравнений (2) выполняется численно (см. [26], [27]). В данной работе интегрирование уравнений движения атомов наночастицы проводилось по скоростной численной схеме Верле (см. [28])

$$\mathbf{x}_i^{n+1} = \mathbf{x}_i^n + \Delta t \mathbf{V}_i^n + [(\Delta t)^2 / (2m_i)] \left(\sum_{j=1}^{N_k} \mathbf{F}_{ij} + \mathbf{F}_i - \alpha_i m_i \mathbf{V}_i^n \right), \quad (5)$$

$$\mathbf{V}_i^{n+1} = (1 - \Delta t \alpha_i) \mathbf{V}_i^n + [\Delta t / (2m_i)] \left[\left(\sum_{j=1}^{N_k} \mathbf{F}_{ij} + \mathbf{F}_i \right) + \left(\sum_{j=1}^{N_k} \mathbf{F}_{ij} + \mathbf{F}_i \right)^{n+1} \right], \quad (6)$$

где $\mathbf{x}_i^n$, $\mathbf{V}_i^n$ – координаты и скорость i -го атома на n -м шаге интегрирования по времени; Δt – шаг интегрирования по времени; индекс 0 обозначает параметры в момент времени $t = 0$.

Данная схема интегрирования позволяет осуществлять распараллеливание расчетов. При расчетах использованы пакеты программ для последовательных и параллельных решений уравнений молекулярной динамики, разработанные в Институте прикладной механики УрО РАН и модернизированный пакет программ NAMD, созданный в University of Illinois and Beckman Institute (США) группой Theoretical Biophysics Group (см. [29]). Графическая визуализация результатов расчетов формирования наночастиц осуществлялась с помощью программы VMD (см. [30]).

3. РЕЗУЛЬТАТЫ И АНАЛИЗ РАСЧЕТОВ

Ниже представлены результаты расчетов процессов формирования структуры и формы для металлических частиц меди.

В расчетах использовались потенциалы взаимодействия атомов Морзе

$$\Phi(\mathbf{r}_{ij})_m = D_m (\exp[-2\lambda_m(|\mathbf{r}_{ij}| - \rho_0)] - 2 \exp[-\lambda_m(|\mathbf{r}_{ij}| - \rho_0)]) \quad (7)$$

и Леннарда–Джонса

$$\Phi(\mathbf{r}_{ij})_{LD} = 4\epsilon \left[\left(\frac{\sigma}{|\mathbf{r}_{ij}|} \right)^{12} - \left(\frac{\sigma}{|\mathbf{r}_{ij}|} \right)^6 \right], \quad (8)$$

где D_m , λ_m , ρ_0 , ϵ , σ – константы исследованных материалов, $|\mathbf{r}_{ij}|$ – величина вектора $\mathbf{r}_{ij}$.

В качестве примера рассматривается наночастица, состоящая из атомов меди. Для моделирования процесса формирования структуры такой наночастицы использован потенциал Морзе с параметрами (см. [31]) $D_m = 54.94 \times 10^{-21}$ (Дж); $\lambda_m = 1.3588$ (1/м); $\rho_0 = 2.866 \times 10^{-10}$ (м). Начальные координаты атомов $\mathbf{x}_{i0}$ задавались исходя из атомной структуры кристаллического материала, а $\mathbf{V}_{i0} = 0$. Начальные координаты $\mathbf{x}_{i0}$ определяют силы $\mathbf{F}_{ij}$ в начальный момент времени согласно уравнениям (3). Величина и направление этих сил определяют процесс перестройки кристаллической структуры и форму наночастицы. В данном примере координаты $\mathbf{x}_{i0}$ задавались для первоначально сжатого кристалла:

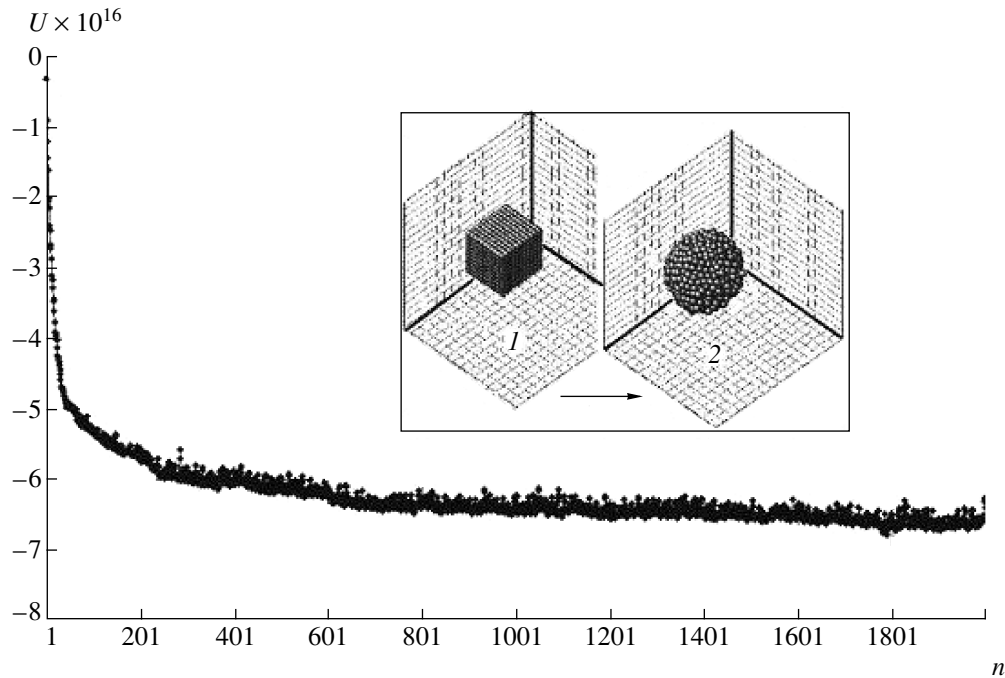
$$\mathbf{x}_{i0} = 0.95 \mathbf{x}_{i0}^k,$$

где $\mathbf{x}_{i0}^k$ – координаты атомов ненагруженного кристалла.

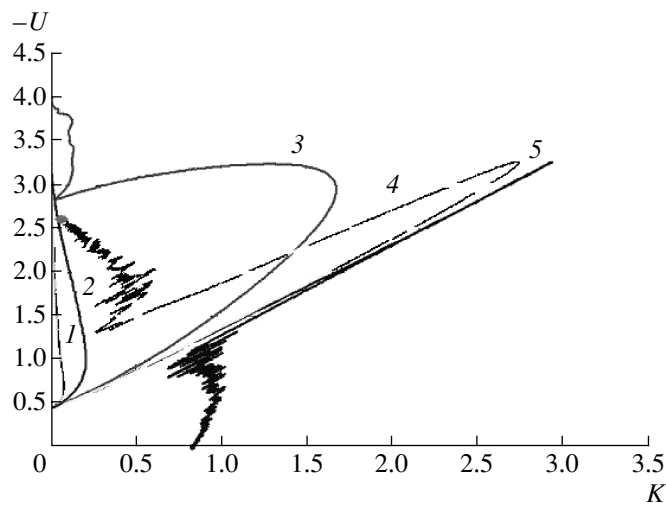
При этом межатомное расстояние на 5% меньше расстояния между атомами в ненагруженном кристалле.

На фиг. 1 приведен график изменения потенциальной энергии наночастицы, состоящей из атомов меди, в процессе перестройки ее структуры от исходной кристаллической (1) к итоговой некристаллической структуре наночастицы (2).

Видно, что в процессе релаксации потенциальная энергия наночастицы уменьшается до некоторого значения, а затем, после 1800 шагов интегрирования по времени, практически остается постоянной. Наблюдаемые в процессе перестройки структуры наночастицы спорадические кратковременные изменения потенциальной энергии системы объясняются ее “нагревом”. Однако



Фиг. 1.

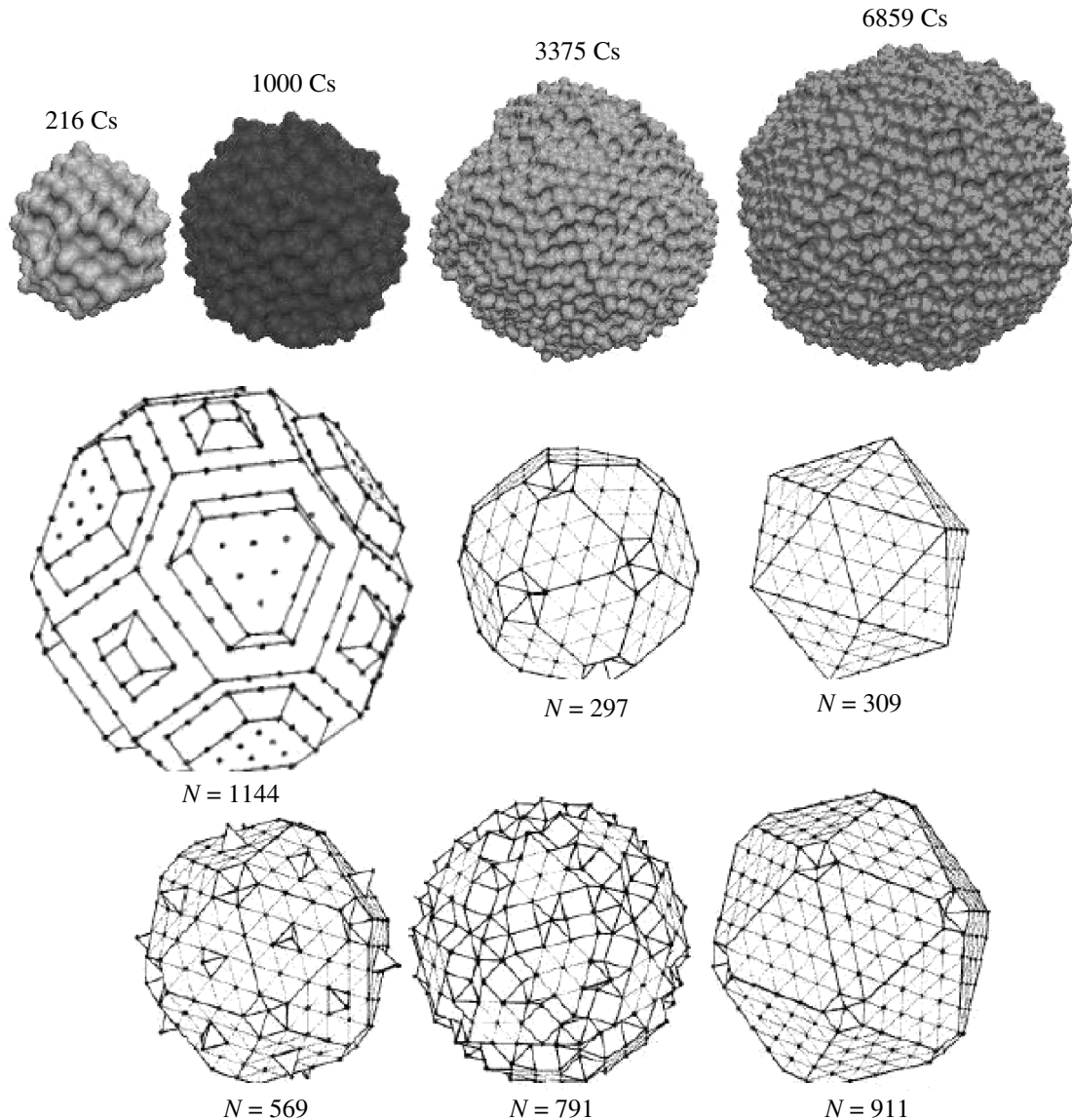


Фиг. 2.

вследствие наличия диссипативной составляющей в уравнении (1) происходит рассеивание тепловой энергии. При этом начальная кристаллическая атомная структура наночастицы полностью перестраивается в некристаллическую, а ее форма приближается к шаровой.

Важную роль при формировании устойчивой наночастицы играет интенсивность диссипации кинетической энергии при перестройке структуры наночастицы. На фиг. 2 приведен график изменения соотношения между потенциальной энергией наночастицы U и кинетической энергией K “теплового” движения ее атомов при перестройке структуры наночастицы для различных коэффициентов уменьшения скорости движения атомов на двух последовательных шагах по времени:

$$k_v = \left(\sum_{i=1}^{i=N_k} |\mathbf{v}_i|_{n+1} \right) \left(\sum_{i=1}^{i=N_k} |\mathbf{v}_i|_n \right)^{-1},$$

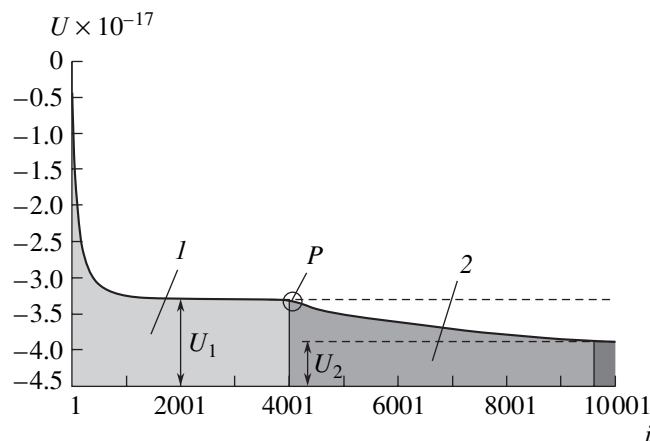


Фиг. 3.

где $|\mathbf{V}_i|_n$ и $|\mathbf{V}_i|_{n+1}$ – величины скоростей i -го атома на двух последовательных n -м и $n + 1$ -м шагах по времени.

Из фиг. 2 следует, что вид зависимости U от K существенно определяется величиной k_v . Для $k_v = 0.999$ (см. кривую 1), соответствующего малой интенсивности диссипации тепловой энергии (энергии колебаний атомов наночастицы), при перестройке структуры наночастицы могут возникать неустойчивости (кривые 4, 5 на фиг. 2). Однако при $k_v < 0.999$ релаксационный процесс всегда устойчив (кривые 1–3 на фиг. 2). Таким образом, даже малая диссипация тепловой энергии обеспечивает устойчивость процесса перестройки структуры наночастицы. Независимо от величины k_v в указанном выше диапазоне его значений атомарная система наночастицы “успокаивается” с течением времени, атомная структура стабилизируется, а ее потенциальная энергия приближается к одинаковому стационарному значению U_c для равновесной наночастицы.

Наночастицы после релаксации могут иметь весьма разнообразные формы, как симметричные (шарообразные, сферические центрированные, нецентрированные и икосаэдрические), так и несимметричные (см. фиг. 3). Количество атомов, входящих в наночастицу, существенно определяет ее форму. Необходимо отметить, что симметричные наночастицы формируются только при определенном числе входящих в нее атомов. В общем случае наночастица имеет отклонения



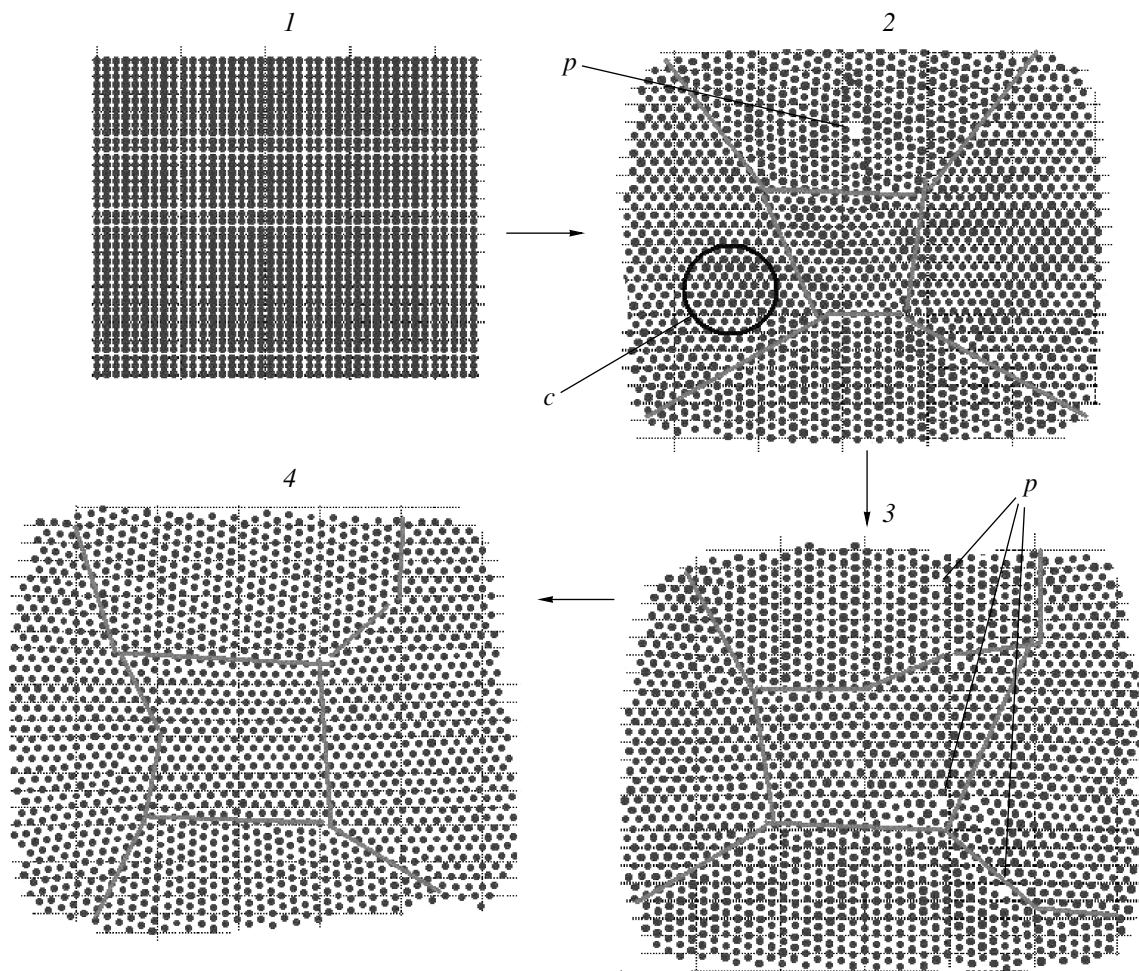
Фиг. 4.

от симметричной формы в виде нерегулярных выступов на поверхности. При увеличении числа атомов, а следовательно, и размера наночастицы число нерегулярностей на поверхности увеличивается. В процессе перестройки структуры они могут появляться и исчезать с поверхности. Основная тенденция при увеличении размеров наночастицы состоит в постепенном переходе от шарообразной формы к кристаллоподобной, конфигурация которой зависит от типа кристаллической решетки, образуемой атомами в макроматериале.

Кроме того, при одном и том же количестве атомов существует несколько различных форм равновесия наночастицы. Это иллюстрирует график изменения потенциальной энергии наночастицы в процессе релаксации, приведенный на фиг. 4. Кривая изменения потенциальной энергии наночастицы, показанная здесь, имеет два характерных участка (1 и 2), на каждом из которых последовательно наблюдаются уменьшение потенциальной энергии и ее стабилизация. Участок 1 соответствует формированию первой равновесной формы наночастицы. На участке 2 наблюдается повторное уменьшение потенциальной энергии наночастицы и формируется участок стабилизации, соответствующий образованию второй равновесной формы наночастицы 2. Безусловно, вторая форма равновесия более устойчива вследствие меньшей потенциальной энергии наночастицы, однако первая форма равновесия также довольно долго “существует” в процессе расчета. Между двумя формами равновесия наночастицы имеется область перехода P от первой равновесной формы наночастицы ко второй. Возникновение этой области является случайным процессом, который не может быть предсказан заранее, исходя из кривой изменения потенциальной энергии наночастицы. Для нахождения наиболее устойчивой формы равновесия наночастицы, соответствующей наименьшему значению потенциальной энергии, необходимо сообщить атомам наночастицы на участке стабильности небольшое количество тепловой энергии (порядка одного-двух % от общей энергии наночастицы). Если возникшая форма равновесия наиболее устойчива, то потенциальная энергия наночастицы будет оставаться постоянной во времени. В противном случае энергия наночастицы уменьшается, а ее конфигурация и структура изменяются. Следует отметить, что это наиболее характерно для частиц “неправильной” формы.

Важным является вопрос о внутренней структуре наночастицы, соответствующей различным ее формам. Расчеты показывают, что атомная структура наночастиц существенно отличается от кристаллического строения массивных материалов: изменяются расстояние между атомами и углы кристаллической решетки, появляются различного типа поверхностные образования.

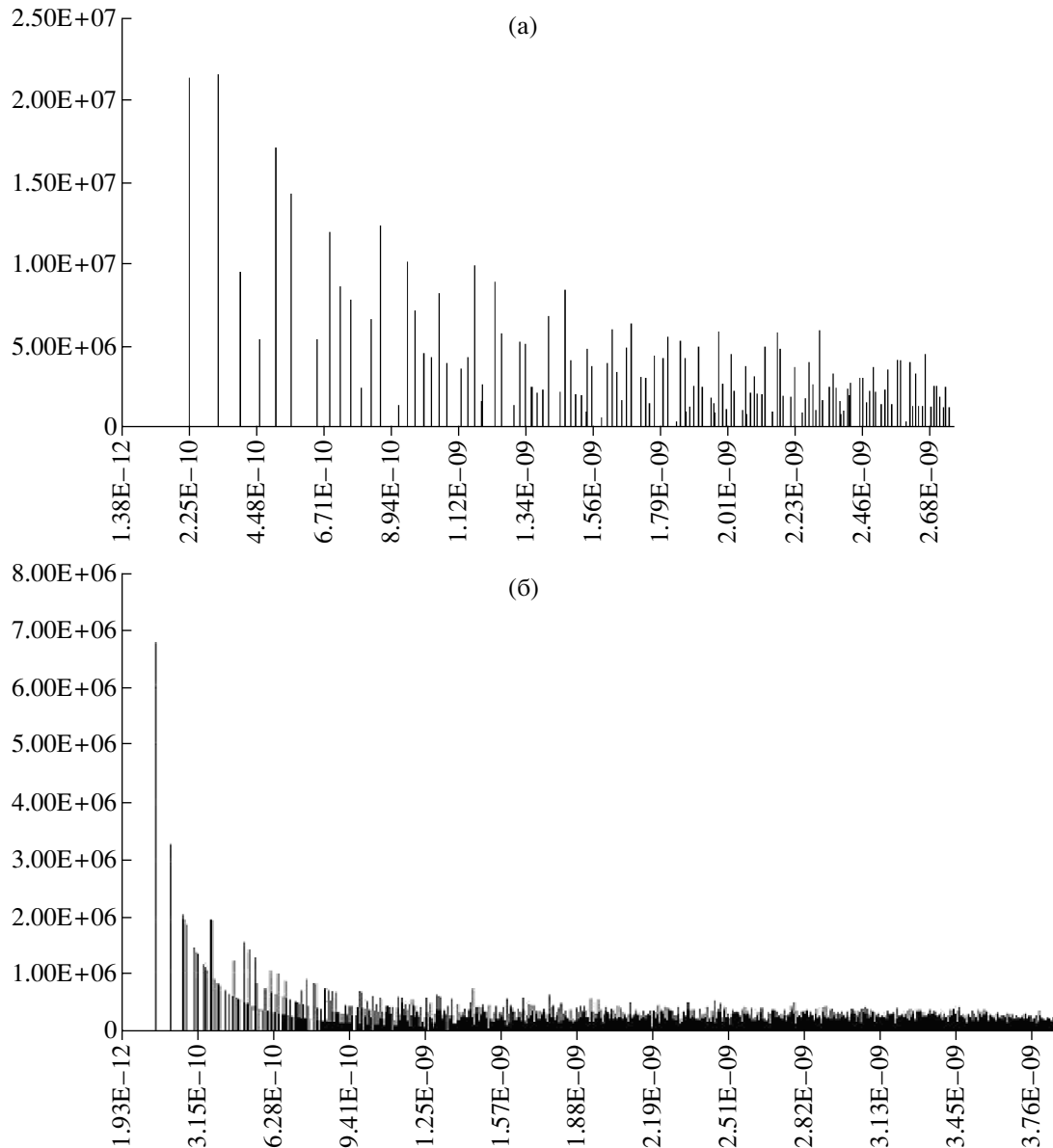
В качестве иллюстрации этого рассмотрим изменение структуры двумерной наночастицы в процессе релаксации, представленное на фиг. 5. Рисунок показывает как начальная кристаллическая структура наночастицы 1 последовательно перестраивается во времени в процессе релаксации (позиции 2, 3, 4). Следует отметить, что итоговая форма наночастицы не круглая, т.е. она “запомнила” начальную атомную структуру. Интересно также отметить, что в процессе релаксации в наночастице образовывались дефекты в виде пор (обозначено через p на рисунке) и области флуктуации плотности (обозначены на рисунке через c), которые в конечной структуре отсутствуют.



Фиг. 5.

Анализ атомной структуры наночастицы выполним, используя функции радиального распределения атомов. На фиг. 6 представлены функции радиального распределения атомов в наночастице (см. [24]): (а) – начальная кристаллическая структура; (б) – структура отрелаксированной наночастицы. Видно, что в процессе релаксации наночастицы ее начальная кристаллическая структура перестраивается в сложную, состоящую из аморфной и кристаллической областей. В центре наночастицы находится кристаллическое ядро (на что указывает линейчатый спектр функции радиального распределения), затем по мере удаления от центра появляется сначала поликристаллическая структура, затем аморфная.

Задача определения диапазона размеров частицы, или количества атомов, входящих в нее, при которых происходит переход от физико-механических свойств наночастицы к свойствам материала в кристаллическом состоянии, до сих пор не имеет окончательного решения, так как образование полной (100%-ной) кристаллической структуры материала происходит при большом количестве атомов, входящих в наночастицу. С этой целью выполнены расчеты для наночастиц, состоящих из большого количества атомов: 10^6 , 8×10^6 , 64×10^6 , 216×10^6 и 512×10^6 . При этом принималось условие сферической симметрии частицы. Исследована зависимость расстояния между “внутренними” атомами в наночастицах (расстояние между атомами определялось по лучу, исходящему из геометрического центра наночастицы), состоящих из различного числа атомов. Расчеты показали, что по мере удаления от центра наночастицы расстояние между атомами увеличивается. В центральной части больших наночастиц (размером порядка 100 Нм и более) расстояние между атомами одинаковое и равно расстоянию между атомами в кристалле. Из этого следует, что физические свойства материала в центральной области частицы будут близки к свойствам материала, находящегося в кристаллическом состоянии, а в объеме, прилегающем к



Фиг. 6.

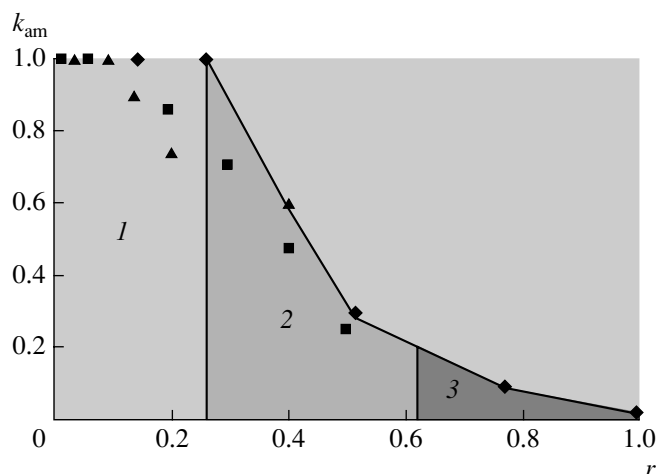
поверхности наночастицы, свойства будут изменяться по мере приближения к этой поверхности. Для малых наночастиц (размером менее 10 Нм) “кристаллическая зона” в структуре частицы не наблюдается.

Зависимость относительного объема “аморфной” части наночастицы $k_{\text{ам}}$ от ее безразмерного радиуса r для различных материалов приведена на фиг. 7. Здесь

$$k_{\text{ам}} = \frac{V_{\text{ам}}}{V_{\text{нп}}}, \quad r = \frac{r_{\text{нп}}}{r_{\text{max}}},$$

где $V_{\text{ам}}$, $V_{\text{нп}}$ – объем аморфной (некристаллической) части наночастицы и объем наночастицы соответственно; $r_{\text{нп}}$, r_{max} – радиус наночастицы и радиус наночастицы, в которой некристаллическая область равна 0 соответственно (метки: ♦ – алюминий, ▲ – медь, ■ – цинк).

Расчеты для всех рассмотренных материалов проводились по изложенной выше методике с использованием потенциала Морзе для алюминия ($D_m = 43.31 \times 10^{-21}$ (Дж), $\lambda_m = 1.1646$ (1/м), $\rho_0 = 3.253 \times 10^{-10}$) и потенциала Леннарда–Джонса для цинка ($\epsilon = 1.7387 \times 10^{-21}$ (Дж), $\sigma = 1.942 \times 10^{-10}$ (м)).



Фиг. 7.

Линия на фиг. 7 обозначает предельный одномерный расчет. Как следует из графика, для всех рассчитанных частиц в диапазоне $r \leq 0.2$ (область 1 на фиг. 7) наночастицы некристаллические (аморфные), на участке $0.2 \leq r \leq 0.8$ (область 2) наночастицы содержат как кристаллическую, так и некристаллическую зоны (эти частицы можно отнести к “рентгеноаморфным”), частицы большего размера (область 3) могут быть классифицированы как кристаллические.

Полученные результаты позволяют утверждать, что обнаруженные закономерности формирования наночастиц в деструкционных (up down) процессах справедливы не только для частиц из меди, но и для наночастиц из других металлов, например алюминия и цинка.

4. ЗАКЛЮЧЕНИЕ

Отметим следующие основные закономерности протекания процессов формирования наночастиц в up down-процессах и свойства их структуры, выявленные при численном моделировании.

1. Возможно существование нескольких типов форм и структур наночастиц в зависимости от термодинамических условий.

2. Отсутствие кристаллического ядра у малых наночастиц.

3. Формирование единичного (идеального) кристаллического ядра с дефектами на поверхности соединения данного ядра с аморфной оболочкой.

4. Формирование поликристаллического ядра с дефектами, распределенными между кристаллическими зернами, имеющими пониженную атомную плотность и измененные межатомные расстояния. Границы зерен, кроме того, неравновесные и содержат большое количество зернограничных дефектов.

5. Структура наночастиц при увеличении размеров изменяется от аморфной к рентгеноаморфной и далее – к кристаллической.

6. Образование дефектных структур различного типа на границах и поверхности наночастицы.

7. Переход от сфероидальной к кристаллоподобной форме наночастицы по мере увеличения ее размеров. Формирование “правильной” и “неправильной” форм наночастиц в зависимости от количества входящих в наночастицу атомов.

Авторы выражают благодарность А.А. Вахрушеву и А.А. Шушкову за помощь в проведении расчетов на ЭВМ.

СПИСОК ЛИТЕРАТУРЫ

1. Qing-Qing Ni, Yaqin Fu, Masaharu Iwamoto. Evaluation of elastic modulus of nano particles in PMMA/Silica Nanocomposites // J. Soc. Materials Sci. Japan. 2004. V. 53. № 9. P. 956–961.
2. Ruoff R.S., Pugno N.M. Strength of nanostructures // Mech. 21st Century. Proc. 21-th Internat. Congress Theor. and Appl. Mech. Warsaw: Springer, 2004. P. 303–311.

3. *Diao J., Gall K., Dunn M.L.* Atomistic simulation of the structure and elastic properties of gold nanowires // *J. Mech. and Phys. Solids*. 2004. V. 52. № 9. P. 1935–1962.
4. *Dingreville R., Qu J., Cherkaoui M.* Surface free energy and its effect on the elastic behavior of nano-sized particles, wires and films // *J. Mech. and Phys. Solids*. 2004. V. 53. № 8. P. 1827–1854.
5. *Duan H.L., Wang J., Huang Z.P., Karimhaloo B.L.* Size-dependent effective elastic constants of solids containing nano-inhomogeneities with interface stress // *J. Mech. and Phys. Solids*. 2005. V. 53. № 7. P. 1574–1596.
6. *Валиев Р.З., Александров И.В.* Наноструктурные материалы, полученные интенсивной пластической деформацией. М.: Логос, 2000.
7. *Галашев А.Е.* Исследование устойчивости метастабильных фаз и кластеров методом молекулярной динамики: Дис. ... докт. физ.-матем. наук, 1997.
8. *Гусев А.И., Ремпель А.А.* Нанокристаллические материалы. М.: Физматлит, 2001.
9. *Заводинский В.Г.* Атомная структура и электронное строение нанометровых систем на основе кремния: Дис. ... докт. физ.-матем. наук, 1997.
10. *Лагунов В.А., Синани А.Б.* Компьютерное моделирование деформирования и разрушения кристаллов // *Физ. твердого тела*. 2001. Т. 43. № 4. С. 644–650.
11. *Петров Ю.И.* Кластеры и малые частицы. М.: Наука, 1986.
12. *Петрунин В.Ф., Зеленюк Ф.М. и др.* Особенности атомной структуры ультрадисперсных систем // *Физикохимия ультрадисперсных систем*. М., 1987. С. 60–67.
13. *Морохов И.Д., Трусов Л.В., Лаповок В.И.* Физические явления в ультрадисперсных системах. М.: Наука, 1984.
14. *Hoare M.R.* Structure and dynamics of simple microclusters // *Ach. Chem. Phys.* 1987. № 40. P. 49–135.
15. *Brooks B.R., Brucoleri R.E., Olafson B.D. et al.* CHARMM: A program for macromolecular energy minimization, and dynamics calculations // *J. Comput. Chemistry*. 1983. V. 4. № 2. P. 187–217.
16. *Вахрушев А.В.* Моделирование статистики и динамики кластеров на макро-, мезо- и микроструктурных уровнях // *Вестн. ИжГТУ*. 2001. № 1. С. 25–29.
17. *Вахрушев А.В., Липанов А.М., Жичкин Е.А.* Компьютерное моделирование статистики и динамики кластерных структур // *Применение матем. моделирования для решения задач в науке и техн.* Ижевск: ИПМ УрО РАН, 1996. С. 158–163.
18. *Вахрушев А.В., Липанов А.М.* Модели процессов формирования кластеров и кластерных макроструктур // *Кластерные системы и материалы*. Ижевск: ИПМехан. УрО РАН, 1997. С. 104–121.
19. *Vakhrouchev A.V., Lipanov A.M.* A numerical analysis of the rupture of powder materials under the power impact influence // *Comput. & Structures*. 1992. V. 44. № 1/2. P. 481–486.
20. *Липанов А.М., Вахрушев А.В.* Макро- и микроструктурное моделирование процессов кластерообразования при импульсном диспергировании порошков // *Кластерные материалы*. Ижевск: ИПМехан. УрО РАН, 1991. С. 114–128.
21. *Липанов А.М., Вахрушев А.В.* Задача о диспергировании порошковых материалов взрывом // *Прикл. механ.* 1991. Т. 27. № 2. С. 47–53.
22. *Липанов А.М., Вахрушев А.В.* Динамика процесса дробления порошков при нестационарном воздействии давления // *Моделирование в механ.* Новосибирск: СО АН СССР, 1990. Т. 4(21). № 6. С. 110–122.
23. *Вахрушев А.В.* Физические основы и математическое моделирование механики порошковых металлокомпозиций: Дис. ... докт. физ.-матем. наук. Ижевск: ИПМехан. УрО РАН, 2002.
24. *Хеерман Д.В.* Методы компьютерного эксперимента в статистической физике. М.: Наука, 1990.
25. *Буркерт У., Эллинджнер Н.* Молекулярная механика. М.: Мир, 1986.
26. *Чудинов В.Г.* Атомные механизмы первичных процессов при радиационном воздействии на твердое тело: Дис. ... докт. физ.-матем. наук, 1992.
27. *Фомин В.М., Гулидов А.И., Сапожников Г.А. и др.* Высокоскоростное взаимодействие тел. Новосибирск: СО РАН, 1999.
28. *Verlet L.* Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules // *Phys. Rev.* 1967. V. 159. № 1. P. 98–103.
29. <http://www.ks.uiuc.edu/Research/namd/>
30. <http://www.ks.uiuc.edu/Research/vmd/>
31. *Рут М.* Наноконструирование в науке и технике. Введение в мир нанорасчета. М.–Ижевск: НИЦ “Регулярная и хаотич. динамика”, 2005.

УДК 519.634

ЗАДАЧА ПОСТРОЕНИЯ КРЫЛОВОГО ПРОФИЛЯ С ВЫДУВОМ РЕАКТИВНОЙ СТРУИ НАВСТРЕЧУ ДОЗВУКОВОМУ ПОТОКУ¹⁾

© 2007 г. Н. Б. Ильинский, Р. Ф. Марданов

(420008 Казань, Университетская, 17, НИИ матем. и механ. КазГУ)

e-mail: Nikolay.Ilinskiy@ksu.ru

Поступила в редакцию 22.11.2006 г.

Решается задача о нахождении формы симметричного крылового профиля с устройством выдува из головной части профиля реактивной струи навстречу дозвуковому стационарному безвихревому набегающему потоку идеальной несжимаемой жидкости. В решении реализована идея Седова об образовании застойной зоны в окрестности критической точки. Разработан итерационный процесс построения решения, проведена серия проектировочных расчетов. Результаты расчетов проанализированы, и сделаны выводы о влиянии параметров реактивной струи на форму профиля и результирующую силу, действующую на профиль. Библ. 10. Фиг. 4.

Ключевые слова: задача о расширении крылового профиля реактивной струи, образование застойной зоны в окрестности критической точки, итерационный метод рассеяния.

ВВЕДЕНИЕ

Задача о выдуве реактивной струи (т.е. струи с плотностью и полным давлением, отличным от тех же характеристик набегающего потока) навстречу набегающему потоку давно привлекает внимание исследователей. Известно, что это один из возможных способов управления потоком, в частности способ уменьшения сопротивления. Известные в этом направлении исследования носят по большей части экспериментальный характер и касаются режимов со сверхзвуковыми скоростями обтекания.

Среди работ, посвященных дозвуковому обтеканию, пожалуй, первой явилась работа [1] (см. также [2]), где рассмотрено кавитационное обтекание тела с поглощением возвратной струи (по схеме Эфроса) и замечено, что если обратить направления на линиях тока, то можно получить новое течение идеальной несжимаемой жидкости (ИНЖ), удовлетворяющее всем уравнениям движения. Это новое течение будет ни чем иным, как обтеканием тела с выбросом струи навстречу движению жидкости. Из уравнений движения следует, что если тело, обтекаемое с поглощением возвратной струи, испытывает силу сопротивления, то тело с выбрасыванием струи вперед по движению будет испытывать, как это ни парадоксально, силу тяги (!), направленную в ту же сторону, куда выбрасывается струя. Объясняется это наличием зоны разрежения в окрестности передней кромки и областью восстановления давления вблизи задней кромки тела. В [3] показано, что увеличение скорости движения тела в жидкости невозможно традиционными способами и одним из путей решения этой задачи является рассмотрение схем с выбросом струи вперед по потоку. Однако отмечается, что в силу вязких и других потерь реализация такого обращенного движения затруднительна.

Попытка преодолеть влияние вязких эффектов, приводящих к возникновению отрывной зоны за обтекаемым телом, сделана в [4], где решена задача проектирования симметричного крылового профиля с каналами отбора и выдува потока, причем из второго жидкость выдувается навстречу набегающему потоку. Благодаря наличию на верхней и нижней поверхностях круговых каналов отбора жидкости удастся построить крыловой профиль с отсутствием участков падения скорости, т.е. с гарантированным безотрывным обтеканием. Существенным упрощением в этой работе является то, что жидкость, выдуваемая вперед по потоку, имеет те же параметры, что и набегающий поток. В этом случае в критической точке, т.е. в точке разветвления потока, скорость обращается в нуль как во внешнем потоке, так и в струе. Исследованию особенностей подобного рода посвящен ряд работ, связанных с задачами о соударении струй (см., например, [5],

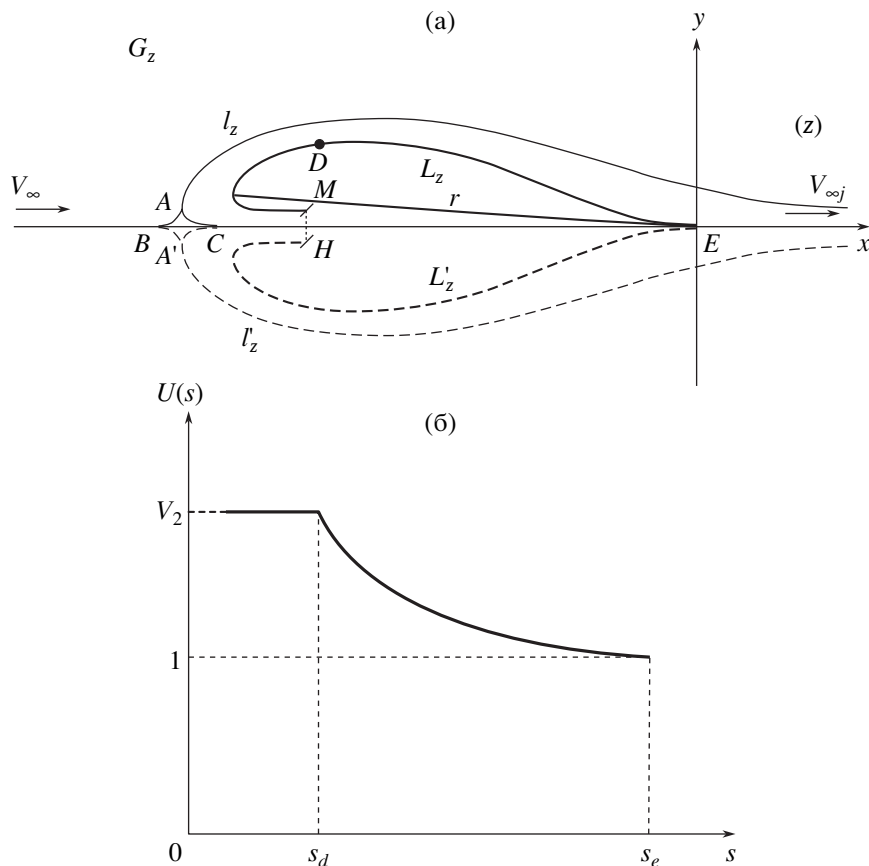
¹⁾ Работа выполнена при финансовой поддержке РФФИ (код проекта 05-08-01153а).

[6]). Случаи обтекания плоских тел со струями ИНЖ, когда константы Бернулли в струях, вытекающих из тела, отличаются от констант Бернулли набегающего потока, изучались в [7].

В данной работе исследуется обратная краевая задача аэрогидродинамики, а именно задача построения симметричного крылового профиля с выдувом реактивной струи навстречу набегающему дозвуковому потоку в случае разных констант Бернулли, т.е. когда в струе жидкость имеет полное давление и плотность, отличные от полного давления и плотности внешнего потока. Для обеспечения безотрывности обтекания на искомом контуре профиля задается гидродинамически целесообразное распределение скорости (ГЦРС) (см., например, [8]). Существенным моментом в предложенной модели явилось введение в окрестности критической точки в потоке, осложняющей математические исследования застойной зоны. Такая модель была предложена в [3].

1. ПОСТАНОВКА ЗАДАЧИ

В физической плоскости $z = x + iy$ искомый симметричный крыловой профиль обтекается дозвуковым потоком ИНЖ с заданной плотностью ρ , скоростью V_∞ и давлением p_∞ на бесконечности (на фиг. 1а MDE – верхняя половина этого профиля). Начало координат выбрано в задней кромке E профиля, а ось абсцисс направлена вдоль скорости V_∞ набегающего потока. В окрестности передней кромки профиля имеется прямолинейный канал M с постоянной скоростью на стенках, уходящих в расчетной схеме на второй лист римановой поверхности. Из канала навстречу потоку выдувается жидкость с другой плотностью ρ_j и скоростью $V_{j\infty}$ на бесконечности. Индексом j обозначаются все параметры, относящиеся к струе. В точке E схода потока, внутренний к области течения, угол равен 2π . Заданы хорда r искомого крылового профиля и относительная ширина $h = H/r$ канала M выдува или безразмерный расход $q = Q/(rV_\infty)$.



Фиг. 1.

Обозначим через l_z и l'_z линии тока, разделяющие реактивную струю и внешний поток. При переходе через эти линии давление меняется непрерывно, а скорость меняется скачком согласно соотношению

$$\rho_j V_j^2 = \rho V^2 + \mu \rho V_\infty^2, \quad (1.1)$$

следующему из интеграла Бернулли, где безразмерный параметр μ определяется формулой $\mu \equiv 2(p_{j0} - p_0)/(\rho V_\infty^2) = (\rho_j V_{j\infty}^2)/(\rho V_\infty^2) - 1$ и характеризует энергию выдуваемой жидкости ($0 \leq \mu < \infty$). Здесь $p_0 = p_\infty + \rho V_\infty^2/2$ и $p_{j0} = p_\infty + \rho_j V_{j\infty}^2/2$ – полные давления во внешнем потоке и в струе соответственно.

В модель обтекания в окрестности точки разветвления потока вводится застойная зона (“целик”) $BACA'$ с постоянными скоростями на границе. На участке $A'BA$ задается скорость V_0 , а на участке $A'CA$ – скорость V_{0j} , которая связана с V_0 формулой (1.1). На контуре L_z (и симметрично на L'_z) крылового профиля задается распределение скорости $V(s) = V_1 U(s)$ как функция дуговой абсциссы s . Функция $U(s)$ выбирается из класса ГЦРС (фиг. 1б). На участке MD контура профиля $U(s) \equiv V_2$, а на диффузорном участке DE распределение скорости определяется формулой

$$U(s) = V_2 [1 + K(s - s_d)]^{-k},$$

где $k > 0$, s_d – абсцисса точки D , а K подбирается так, что скорость $U(s_e) = 1$ (s_e – абсцисса задней кромки E). Положение точки D фиксируется во введенной ниже канонической плоскости и определяется заданным параметром $0 < d < 1$.

Требуется определить форму контура $L_z(L'_z)$ симметричного крылового профиля, форму линии $l_z(l'_z)$, отделяющую реактивную струю от внешнего потока, форму и положение застойной зоны $BACA'$ и силу R_x сопротивления (или тяги).

2. МЕТОД РЕШЕНИЯ ЗАДАЧИ

В силу симметрии задачи достаточно рассмотреть лишь верхнюю половину течения. Введем в рассмотрение каноническую плоскость $t = \tau + i\eta$. Области G_z (верхняя половина области течения в физической плоскости z) поставим в соответствие область G_t – верхнюю полуплоскость в плоскости t . Соответствующие точки в плоскостях z и t обозначим одинаковыми буквами. Для взаимно однозначного отображения областей G_z и G_t , следуя теореме Римана, предположим соответствие бесконечно удаленных точек этих областей, переход точки $z = 0$ в точку $t = 1$, а также переход бесконечно удаленной точки канала M в точку $t = 0$. Координаты точек B , A и C в плоскости t обозначим, соответственно, через b , a и c , причем $b < a < c < 0$. Параметр d – координата точки D в плоскости t . Образ линии l_z во вспомогательной плоскости обозначим через l_t .

При сделанных предположениях во внешнем потоке и в струе существуют комплексные потенциалы течений. Будем рассматривать их как единую кусочно-аналитическую функцию $w(z) = \phi(x, y) + i\psi(x, y)$, терпящую разрыв на линии раздела сред. Выделив особенности (источник интенсивности Q в точке $t = 0$ и точку A разветвления потока) запишем в плоскости t комплексно-сопряженную скорость

$$dw/dt = u_\infty \exp[-\Omega(t)](t - a)/t, \quad (2.1)$$

где u_∞ – скорость набегающего потока в плоскости t , а $\Omega(t) = T(\tau, \eta) + i\Lambda(\tau, \eta)$ – кусочно-аналитическая функция, терпящая скачок на линии l_t .

Пусть $\zeta(\sigma)$ – точка на линии l_t с дуговой абсциссой σ , отсчитываемой от точки $A(\sigma \geq 0)$, а $\vartheta(\sigma)$ – угол наклона касательной к l_t в этой точке, тогда имеем

$$d\zeta = \exp[i\vartheta(\sigma)]d\sigma. \quad (2.2)$$

Рассматривая (2.1) на линии l_t с учетом того, что эта линия непроницаема, записываем соотношение

$$\vartheta(\sigma) = \Lambda(\sigma) - \text{Im}[\ln(\zeta - a) - \ln \zeta], \quad (2.3)$$

т.е. скачок на этой линии испытывает лишь $T(\sigma) = \operatorname{Re}[\Omega(\zeta)]$. Учтя, что вещественная ось τ также является линией тока, и подставив $t = \tau$ в (2.1), получим

$$\Lambda(\tau) \equiv 0. \quad (2.4)$$

Обозначим через $\chi(t) = \ln(dw/dz) = S - i\theta$ функцию Мичела–Жуковского. Здесь $S = \ln|V|$, V – модуль скорости, а θ – аргумент вектора скорости в физической плоскости z . Введем в рассмотрение функцию

$$\tilde{\chi}(t) = \chi(t) - \chi_0(t) + \Omega(t) = \tilde{S} + i\tilde{\theta}, \quad (2.5)$$

где $\chi_0(t) = S_0 + i\theta_0$. Зададим $\theta_0(\tau)$ на вещественной оси так, чтобы $\tilde{\theta}(\tau) = -\theta(\tau) - \theta_0(\tau) \equiv 0$ при $\tau \in (-\infty, b] \cup [c, 0] \cup [1, \infty)$:

участок	$(-\infty; b]$	$[b, c]$	$[c; 0]$	$[0; 1]$	$[1; \infty)$
$\theta_0(\tau)$	0	$-\pi(\tau - b)/(c - b)$	$-\pi$	$\pi(\tau - 1)$	0

Решив задачу Шварца для верхней полуплоскости и определив произвольную постоянную из условия $\chi_0(\infty) = 0$, найдем

$$\chi_0(t) = \frac{t-b}{c-b} \ln(t-b) - \frac{t-c}{c-b} \ln(t-c) - t \ln t + (t-1) \ln(t-1). \quad (2.6)$$

Тогда из (2.5) имеем

$$dw/dz = \exp[\tilde{\chi} + \chi_0 - \Omega] \quad (2.7)$$

и с учетом (2.1) получим

$$\frac{dz}{dt} = \frac{dw/dt}{dw/dz} = u_\infty \frac{t-a}{t} \exp[-\tilde{\chi} - \chi_0]. \quad (2.8)$$

Так как функция в левой части этого равенства аналитическая, то и $\tilde{\chi}(t)$, стоящая справа, также является аналитической. Из (2.5) следует, что скачки функций $\chi(t)$ и $\Omega(t)$ компенсируют друг друга. Обозначим функции скачков с учетом (2.7) через

$$\lambda(\sigma) = T(\sigma) - T_j(\sigma) = \ln[V_j(\sigma)/V(\sigma)], \quad (2.9)$$

где скорости V и V_j связаны соотношением (1.1). Если функция $\lambda(\sigma)$ известна, то с учетом (2.4) найдем функцию Ω в виде

$$\Omega(t) = \Phi(t) + \overline{\Phi(\bar{t})}, \quad \Phi(t) = \frac{1}{2\pi i} \int_{l_t} \frac{\lambda(\zeta) d\zeta}{\zeta - t}. \quad (2.10)$$

Рассмотрев модуль выражения (2.8) при $t = \tau$, запишем

$$\frac{ds}{d\tau} = u_\infty \left| \frac{\tau - a}{\tau} \right| \exp[-\tilde{S}(\tau) - S_0(\tau)],$$

и, проинтегрировав полученную формулу, определим связь между дуговой абсциссой s профиля и параметром τ в виде

$$s(\tau) = \int_1^\tau \frac{ds}{d\tau} d\tau + s_e, \quad (2.11)$$

которая необходима для определения $V(\tau) = V(s(\tau))$ на участке $0 < \tau < 1$.

Выделив в (2.5) при $t = \tau$ действительную часть, получим

$$\tilde{S}(\tau) = \ln|V(\tau)| - S_0(\tau) + T(\tau), \quad (2.12)$$

где из (2.6) следует

$$S_0(\tau) = \frac{\tau-b}{c-b} \ln|\tau-b| - \frac{\tau-c}{c-b} \ln|\tau-c| - \tau \ln|\tau| + (\tau-1) \ln|\tau-1|.$$

Так как на участках $[b; c] \cup [0; 1]$ задана $V(\tau)$ (скорость на границе целика и на контуре профиля) и, следовательно, известна $\tilde{S}(\tau)$, а на остальных участках вещественной оси $\tilde{\theta}(\tau) \equiv 0$, то функцию $\tilde{\chi}(t)$ определим как решение смешанной краевой задачи

$$\tilde{\chi}(t) = \frac{R(t)}{\pi} \int_{l_\tau} \frac{\tilde{S}(\tau)}{|R(\tau)|} \frac{d\tau}{(\tau-t)}, \quad (2.13)$$

где за $R(t) = \sqrt{(t-b)(t-c)t(t-1)}$ принята ветвь, положительная на участке $\tau > 1$, а участок интегрирования l_τ таков, что

$$\int_{l_\tau} f(\tau) d\tau = \int_b^c f(\tau) d\tau - \int_0^1 f(\tau) d\tau.$$

Смешанная задача имеет два условия разрешимости: $a_0 = 0$ и

$$\tilde{S}_\infty = -a_1, \quad (2.14)$$

где

$$a_k = \frac{1}{\pi} \int_{l_\tau} \frac{\tilde{S}(\tau) \tau^k}{|R(\tau)|} d\tau, \quad k = \overline{0, \infty}, \quad (2.15)$$

суть коэффициенты разложения в ряд Тейлора функции $-\tilde{\chi}(t)/R(t)$ в окрестности бесконечности. Второе из этих условий служит для определения $\tilde{S}(\tau)$ на бесконечности. Первое условие необходимо для того, чтобы $\tilde{\chi}(t)$ была ограниченной функцией в области G_t . Для выполнения первого условия удобно использовать параметр V_1 такой, что

$$V_1 = \exp \left[\int_{l_\tau} \frac{\tilde{S}_1(\tau) d\tau}{|R(\tau)|} \left(\int_0^1 \frac{d\tau}{|R(\tau)|} \right)^{-1} \right],$$

где

$$\tilde{S}_1(\tau) = \begin{cases} \tilde{S}(\tau), & b < \tau < c, \\ \ln|U(\tau)| - S_0(\tau) + T(\tau), & 0 < \tau < 1. \end{cases}$$

Форму границы ВАС целика, контура профиля L_z и линии l_z получим, проинтегрировав уравнение (2.8)

$$z(t) = \int_1^t \frac{dz}{dt} dt.$$

3. СХЕМА ИТЕРАЦИОННОГО ПРОЦЕССА

Для нахождения неизвестных функций $\lambda(\sigma)$ и $\vartheta(\sigma)$ необходимо организовать итерационный процесс. С учетом (1.1) функции $\lambda(\sigma)$ и $\vartheta(\sigma)$ должны удовлетворять следующим условиям:

$$\lambda(0) = \lambda_0 = \frac{1}{2} \ln \left[\frac{\rho}{\rho_j} \left(1 + \mu \frac{V_\infty^2}{V_0^2} \right) \right], \quad \lambda(\infty) = \lambda_\infty = \frac{1}{2} \ln \left[\frac{\rho}{\rho_j} (1 + \mu) \right],$$

$$\vartheta(0) = \vartheta_0 = \pi/2, \quad \vartheta(\infty) = \vartheta_\infty = 0.$$

В качестве начального приближения целесообразно выбрать $\lambda^{(0)}(\sigma) = \lambda_0 + \gamma(\sigma)(\lambda_\infty - \lambda_0)$ и $\vartheta^{(0)}(\sigma) = \vartheta_0 + \gamma(\sigma)(\vartheta_\infty - \vartheta_0)$, где $\gamma(\sigma) = 2 \operatorname{arctg}(C\sigma)/\pi$, $C > 0$. Также для первой итерации необходимо задать функцию $s(\tau)$, выбрав в качестве начального приближения линейную функцию $s^{(0)}(\tau) = s_e + C(\tau - 1)$, $C > 0$.

Итерационная процедура содержит следующие шаги:

- интегрируя (2.2), находим линию l_i ;
- по формуле (2.10), определяем $T(\tau) = \operatorname{Re}[\Omega(\tau)]$ на вещественной оси и $\Omega(\zeta)$ на линии l_i ;
- затем, используя (2.12) для $\tilde{S}(\tau)$ и решая смешанную задачу, находим $\tilde{\chi}(\zeta)$ на линии l_i по формуле (2.13);
- из (2.7) определяем $V(\sigma) = |dw/dz|_{t=\zeta(\sigma)}$ и по формуле (1.2) получаем $V_j(\sigma)$;
- новое приближение для функции $\lambda(\sigma)$ находим по формуле (2.9), новое приближение для $\vartheta(\sigma)$ – из (2.3), а новое приближение для $s(\tau)$ получаем, интегрируя (2.11).

Итерационный процесс необходимо продолжать до тех пор, пока не выполняются условия

$$\max_{\sigma} |\vartheta^{(n)}(\sigma) - \vartheta^{(n-1)}(\sigma)| < \varepsilon, \quad \max_{\sigma} |\lambda^{(n)}(\sigma) - \lambda^{(n-1)}(\sigma)| < \varepsilon,$$

где ε – некоторое малое положительное число.

4. УСЛОВИЯ РАЗРЕШИМОСТИ

Как и в основной обратной краевой задаче аэрогидродинамики (ОКЗА), здесь необходимо потребовать выполнения условий разрешимости задачи (см., например, [8]). Условие совпадения скорости на бесконечности, определяемой в процессе решения с заданной скоростью, получим, рассмотрев (2.7) при $t \rightarrow \infty$: $\ln V_\infty = \tilde{S}_\infty$, где $\tilde{S}_\infty$ определяется по формуле (2.14). Условие замкнутости (в силу симметрии только по ординате y) выписывается из уравнения $\operatorname{res}(dz/dt)|_{t \rightarrow \infty} = 0$ и имеет вид

$$2(a_2 - a) + (b + c)(1 - a_1) - 1 - a_1 = 0,$$

где a_k определяются по формуле (2.15). Помимо этого, так как область течения в физической плоскости является двусвязной, возникает еще одно условие замкнутости, которое записывается как условие замкнутости целика: $(y_c - y_b)/(y_a - y_b) = 0$, где y_a, y_b, y_c – координаты y в плоскости z точек A, B, C соответственно. Выражение для расхода Q получим из поведения $dw/dz|_{z \rightarrow 0} \sim Q/(2\pi z)$ с использованием (2.7) и (2.8) в виде $Q = -2\pi u_\infty a \exp[-T(0)]$.

К указанным трем условиям разрешимости необходимо добавить условие заданности хорды r профиля и условие заданности относительной ширины h канала выдува или безразмерного расхода q . Для удовлетворения этих условий целесообразно подбирать скорости u_∞ и V_2 и параметры a, b, c в канонической плоскости. Таким образом, помимо внутреннего итерационного процесса, для определения функций $\lambda(\sigma)$ и $\vartheta(\sigma)$ организуется внешний итерационный процесс, в качестве которого удобно выбрать итерационную процедуру метода Ньютона решения систем нелинейных уравнений. Добавим, что связь между хордой r и скоростью u_∞ является линейной, и нетрудно показать, что остальные четыре уравнения не зависят от u_∞ и внешний итерационный процесс удается упростить.

5. ОПРЕДЕЛЕНИЕ АЭРОДИНАМИЧЕСКИХ СИЛ

Очевидно, что, в силу симметрии искомого крылового профиля, подъемная сила имеет значение $R_y = 0$. В данной задаче представляет интерес вычисление силы R_x сопротивления, или, как утверждается в [1], силы тяги, т.е. $R_x < 0$. Строгий аналитический вывод формул для аэродинамических сил, действующих на проницаемый профиль, содержится в [9], откуда следует

$$R_x = -\rho_j V_{j\infty} Q. \quad (5.1)$$

Для проверки этой формулы, кажущейся на первый взгляд парадоксальной, вычислим R_x прямым интегрированием по контуру профиля. Для этого используем формулу, следующую из

уравнения движения ИНЖ, записанного в интегральной форме:

$$\mathbf{R} = \oint_{L_1 \cup L_2} [p_j \mathbf{n} + \rho_j \mathbf{V}_j(\mathbf{V}_j, \mathbf{n})] ds = \oint_{L_1 \cup L_2} p_j \mathbf{n} ds + \int_{L_2} \rho_j \mathbf{V}_j(\mathbf{V}_j, \mathbf{n}) ds = \mathbf{R}_p + \mathbf{R}_j.$$

Здесь $\mathbf{R}$ – результирующая сила, действующая на профиль, L_1 – непроницаемая, L_2 – проницаемая часть поверхности профиля (некоторая линия поперек канала выдува), $\mathbf{n}$ – нормаль, внутренняя к поверхности профиля. Первое слагаемое $\mathbf{R}_p$ – это интеграл сил давления по контуру профиля, а второе слагаемое $\mathbf{R}_j$ представляет собой реактивную силу. Давление p_j на поверхности профиля определяется по формуле $p_j = p_{j0} - \rho_j V_j^2/2$. Учтя, что $\mathbf{R} = R_x + iR_y$ и $\mathbf{n} ds = dy - i dx$, а также что границы канала M довольно быстро выходят на горизонтальные асимптоты и течение в канале становится плоскопараллельным со скоростью $V_m = V_1 V_2$, после несложных преобразований получим

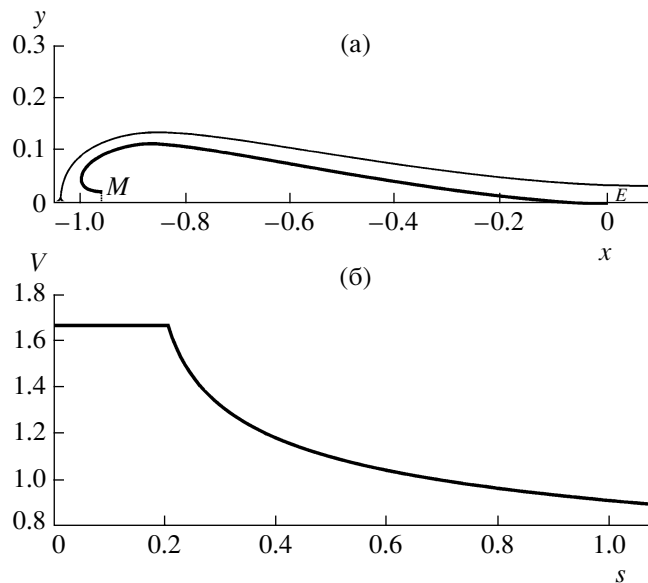
$$R_x = \oint_{L_1 \cup L_2} p_j dy + R_j, \quad R_j = \rho_j V_m Q. \quad (5.2)$$

Отметим также, что силы давления, действующие на целик $BACA'$, в силу (1.1) уравновешиваются.

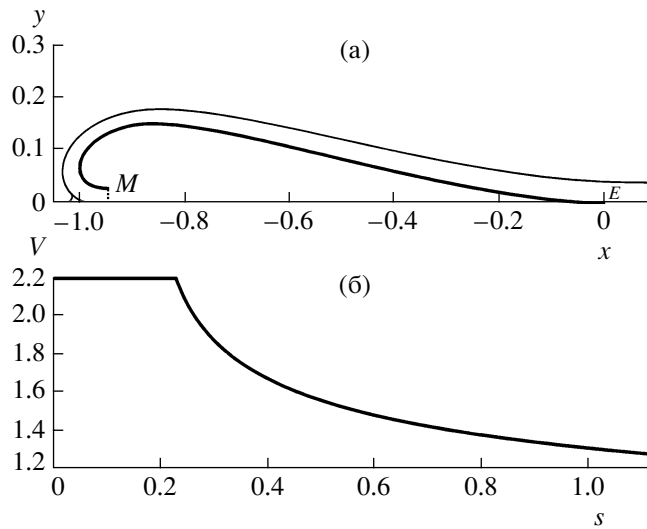
6. РЕЗУЛЬТАТЫ РАСЧЕТОВ

При проведении числовых расчетов, результаты которых представлены ниже, полагалось $V_\infty = \rho = \rho_j = r = 1$. Сходимость итерационного процесса подтверждена числовым экспериментом, причем она тем лучше, чем меньше число μ .

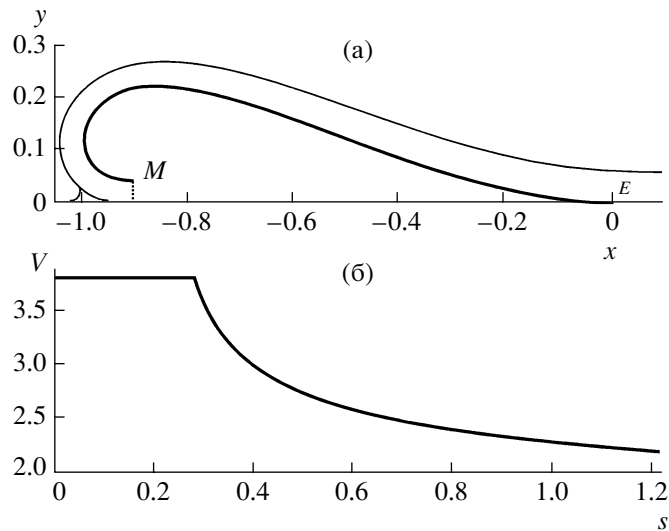
Первый тестовый расчет был проведен для не реактивной струи при следующих параметрах: $\mu = 0$, $h = 0.04$, $V_0 = 0.1$; результат приведен на фиг. 2. Фиг. 2а – верхняя половина построенного симметричного крылового профиля, верхняя линия l_z раздела сред и целик, размеры которого в данном случае составляют ~2% хорды профиля. Из графика видно, что целик в данном случае моделирует обычную критическую точку (линии тока составляют прямой угол). На фиг. 2б изображено задаваемое в виде ГЦРС распределение скорости $V(s)$. Сила R_x , определенная интегрированием сил давления по формуле (5.2), равна $R_x = -0.0645$, т.е. эта сила действительно является силой тяги, так как направлена против набегающего потока. Для сравнения, эта же сила, вычисленная по аналитической формуле (5.1), равна $R_x = -0.0661$, следовательно, погрешность в расчетах



Фиг. 2.



Фиг. 3.



Фиг. 4.

составляет $\sim 2.5\%$. При этом следует отметить, что реактивная составляющая силы $R_j = 0.1105$ по модулю почти в 2 раза превышает результирующую силу тяги.

Во втором расчете полагалось $\mu = 1$, $h = 0.048$, $V_0 = 0.01$, т.е. из канала M выдувается уже реактивная струя. Вид исходного распределения скорости показан на фиг. 3б, а верхняя половина течения – на фиг. 3а. Из графиков видно, что целик в данном случае уже моделирует точку возврата. Несмотря на то, что значение V_0 было взято в 10 раз меньшим, чем в предыдущем, размеры целика увеличились и составили 3% хорды. Силы тяги, вычисленные интегрированием давления по контуру профиля и по аналитической формуле (5.1), равны, соответственно, $R_x = -0.1457$ и $R_x = -0.1512$, а реактивная сила $R_j = 0.2353$.

Последний пример расчета – случай сильного выдува. Здесь $\mu = 5$, $h = 0.075$, $V_0 = 0.01$. Результаты расчета представлены на фиг. 4. Скорость выдува составляет $V_m = 3.73$. Сила тяги $R_x = -0.7034$, а реактивная сила $R_j = 1.093$.

В [10] детально исследовалась модельная задача обтекания точечного источника, из которого выдувается реактивная струя, и сделан вывод о том, что решение практически не зависит от ско-

рости V_0 , которая является параметром задачи, если его значения брать небольшими ($<0.1V_\infty$). Аналогичный результат имеет место и в рассматриваемой задаче.

ЗАКЛЮЧЕНИЕ

Решена задача построения симметричного крылового профиля с выдувом реактивной струи навстречу дозвуковому потоку по заданному на искомом контуре профиля распределению скорости как функции дуговой абсциссы из класса ГЦРС. В решении реализована идея Седова об образовании застойной зоны в окрестности критической точки. Относя такие задачи к сложным в математическом плане (см. [3, с. 100]), удалось преодолеть трудности, используя методы постановок и решения ОКЗА (см., например, [8]). Проведена серия проектировочных расчетов, которые подтвердили утверждение Седова, что результирующая сила является силой тяги. Последнее объясняется наличием области разрежения в окрестности передней кромки и восстановлением давления вблизи задней кромки крылового профиля. Анализ полученных числовых результатов приводит к следующим выводам: в случае задания распределения скорости из класса ГЦРС с ростом числа μ увеличиваются толщина профиля, размеры целика и сила тяги R_x ; реактивная сила R_j при выдуве струи навстречу набегающему потоку по модулю больше результирующей силы тяги R_x , модуль отношения R_j/R_x с увеличением μ уменьшается; погрешность в определении силы R_x является погрешностью итерационного метода решения задачи и составляет порядка 3%.

СПИСОК ЛИТЕРАТУРЫ

1. Седов Л.И. Об обтекании идеальной жидкостью тела со встречной струей // Докл. АН СССР. 1972. Т. 206. № 1. С. 41–42.
2. Седов Л.И. Механика сплошной среды. Т. 2. М.: Наука, 1976.
3. Седов Л.И. Новые методы и новые направления механики сплошной среды // Успехи механ. 2005. Т. 3. № 1. С. 94–106.
4. Степанов Г.Ю. Построение безотрывно обтекаемых тел в комплексе с двигателем // Пробл. современной механ. К юбилею Л.И. Седова. М.: Изд-во МГУ, 1998. С. 109–117.
5. Биркгоф Г., Сарантонелло Э. Струи, следы и каверны. М.: Мир, 1964.
6. Гуревич М.И. Теория струй идеальной жидкости. М.: Наука, 1979.
7. Шурыгин В.М. Аэродинамика тел со струями. М.: Машиностр., 1977.
8. Елизаров А.М., Ильинский Н.Б., Поташев А.В., Степанов Г.Ю. Основные методы, результаты, приложения и нерешенные проблемы теории обратных краевых задач аэрогидродинамики // Тр. Матем. центра им. Н.И. Лобачевского. Т. 10. Казань: Изд-во “ДАС”, 2001.
9. Абзалилов Д.Ф., Ильинский Н.Б. Об аэродинамических силах, действующих на крыловой профиль с проницаемым участком // Инж.-физ. журнал. 2006. Т. 79. № 2. С. 126–130.
10. Ильинский Н.Б., Марданов Р.Ф. Задачи обтекания “реактивного” источника // Докл. РАН. 2005. Т. 405. № 1. С. 46–50.

УДК 519.6:519.712

ЛОГИЧЕСКИЕ ЗАКОНОМЕРНОСТИ В ЗАДАЧАХ РАСПОЗНАВАНИЯ (ПАРАМЕТРИЧЕСКИЙ ПОДХОД)¹⁾

© 2007 г. В. В. Рязанов

(119991 Москва, ул. Вавилова, 40, ВЦ РАН)

e-mail: rvy@ccas.ru

Поступила в редакцию 04.05.2007 г.

Рассматривается задача поиска логических закономерностей классов в задаче распознавания по прецедентам и использование логических закономерностей при решении задач распознавания и прогнозирования. Логические закономерности классов определяются как конъюнкции одноместных предикатов, определяющих принадлежность значений признаков отрезкам вещественной оси. Данные конъюнкции принимают значение 1 на подмножествах эталонов некоторого класса и обладают свойствами оптимальности. Рассмотрены различные критерии оптимальности, и сформулированы задачи поиска логических закономерностей как задачи целочисленного программирования. Проведен качественный анализ данных задач. Рассмотрены модели вычисления оценок по системам логических закономерностей. Предложены модификации линейных решающих правил, основанные на поиске максимального зазора для оценок эталонных объектов за классы и аппроксимации логических закономерностей классов гладкими функциями. Предложено понятие динамической логической закономерности классов, алгоритм их поиска и метод прогнозирования. Библ. 17. Фиг. 1.

Ключевые слова: логические закономерности классов, распознавание по прецедентам, прогноз, алгоритмы вычисления оценок, целочисленное программирование, решающее правило, динамическая закономерность.

ВВЕДЕНИЕ

В настоящее время существуют два основных класса логических алгоритмов распознавания: алгоритмы частичной прецедентности (вычисления оценок) и решающие деревья. Настоящая работа посвящена первому из них.

Начало активных исследований в данной области связано с работами Ю.И. Журавлёва (тестовый алгоритм [1], алгоритмы вычисления оценок [2], [3]). Широкую известность приобрел практический алгоритм “Кора” (см. [4]). В [5] описана модель распознавания, основанная на голосовании по системам представительных наборов. Она является разновидностью алгоритмов вычисления оценок и обобщением алгоритма “Кора”. Одним из базовых принципов данных подходов является принцип частичной прецедентности: по обучающей информации предлагается находить множества несократимых фрагментов признаков описаний объектов (или их аналогов различного типа), свойственные различным классам, и использовать данные множества при распознавании новых объектов. Первоначально данные модели использовались для работы с бинарными или k -значными признаками, что делало возможным прямое использование аппарата дискретного анализа. Для работы с вещественнозначными данными Журавлёвым было предложено рассматривать окрестности значений признаков (при этом вводятся соответствующие параметры окрестностей) или осуществлять предварительно дискретизацию данных с сохранением делимости классов на обучающей выборке. В обоих случаях применяются базовые подходы, разработанные и исследованные для дискретного случая. В данной области имеется обширная библиография (см., например, [3], [6]–[9]).

Практически ценной, имеющей самостоятельный интерес особенностью настоящих моделей является нахождение информативных фрагментов объектов или их окрестностей, свойственных различным классам. Далее мы будем использовать термин “логическая закономерность” класса (ЛЗ). Под ЛЗ классов понимаются предикаты вида $P(S) = A_1(S) \& A_2(S) \& \dots \& A_k(S)$, где $A_1, A_2, \dots, A_k$ – одноместные предикаты, зависящие от одного из признаков и определяющие принадлежность

¹⁾ Работа выполнена при финансовой поддержке РФФИ (коды проектов 05-01-00332, 06-01-08045 офи, 05-07-90333, 06-01-00492), Целевой программы № 14 Президиума РАН, Целевой программы № 2 Отделения математических наук РАН).

значения признака некоторому сегменту (интервалу, полуоси и т.д.) вещественной оси. Предполагается, что предикаты (ЛЗ) данного вида выполняются если не на всех объектах обучающей выборки $\tilde{S}$ из некоторого класса K_i , то по крайней мере на многих эталонах данного класса, при этом $P(S) = 0 \forall S \in CK_i \cap \tilde{S}$.

В работах [10] были предложены алгоритмы распознавания, инвариантные относительно преобразований признаков. Получены теоремы об общем виде таких алгоритмов, которые обладают рядом достоинств, связанных с выбором способа параметризации признаков, дискретизации значений непрерывных признаков, задания метрик на множестве значений количественных признаков и согласования покоординатных метрик. Там же описана реализация алгоритмов распознавания данного типа со свойствами неизбыточности и монотонности. Важным их элементом является поиск “максимальных частичных интервалов” классов. По обучающей выборке находятся такие нерасширяемые подмножества эталонов каждого фиксированного класса, для которых “натянутые” на них прямые координатные гиперпараллелепипеды не содержат эталонов других классов. Данные максимальные частичные интервалы непосредственно определяют ЛЗ классов. Следует отметить, что задачи их поиска весьма трудоемки.

В настоящей статье рассматривается параметрический подход, когда каждая ЛЗ определяется конечным набором вещественных параметров, а задача ее поиска состоит в решении специальной оптимизационной задачи. Это позволяет рассматривать различные критерии и их приближения, решать задачи большой размерности. Будут рассмотрены задача поиска логических закономерностей классов в задаче распознавания по прецедентам при различных критериях качества ЛЗ, решение задачи распознавания на базе найденных множеств ЛЗ, поиск динамических ЛЗ по многомерному временному ряду. Некоторые части статьи являются продолжением работ [11]–[13]. Далее используются стандартные обозначения, принятые в моделях распознавания, основанных на вычислении оценок (см. [3], [6]).

1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Рассматривается стандартная задача распознавания по прецедентам с n признаками $x_1, x_2, \dots, x_n$, l непересекающимися классами $K_1, K_2, \dots, K_l$ и m эталонными объектами $\tilde{S} = \{S_1, S_2, \dots, S_m\}$ (обучающей выборкой). Будем использовать обозначения $\tilde{K}_i = \tilde{S} \cap K_i$, $i = 1, 2, \dots, l$, и считать, что $\tilde{K}_i \neq \emptyset$, $i = 1, 2, \dots, l$. Произвольный объект $S \in \bigcup_{i=1}^l K_i$ отождествляется со своим признаковым описанием в виде числового вектора $S = (x_1(S), x_2(S), \dots, x_n(S))$, $S_t = (a_{t1}, a_{t2}, \dots, a_{tn})$, $a_{ij} = x_j(S_t)$, $x_i \in \mathbb{R}$ (бинарные и k -значные признаки рассматриваются как частный случай вещественнозначных).

Рассмотрим следующее параметрическое множество элементарных предикатов:

$$P_j^{c_j^1, c_j^2}(x) = \begin{cases} 1 & \text{при } c_j^1 \leq x \leq c_j^2, \\ 0 & \text{иначе,} \end{cases}$$

где $c_j^1 \in \mathbb{R}$, $c_j^2 \in \mathbb{R}$, $j = 1, 2, \dots, n$.

Пусть $\Omega \subseteq \{1, 2, \dots, n\}$.

Определение 1. Предикат

$$P^{\Omega, c^1, c^2}(\mathbf{x}) = \bigwedge_{j \in \Omega} P_j^{c_j^1, c_j^2}(x_j) \quad (1)$$

называется *логической закономерностью класса* K_λ , $\lambda = 1, 2, \dots, l$, если верно следующее:

- 1) $\exists S_t \in \tilde{K}_\lambda : P^{\Omega, c^1, c^2}(S_t) = 1$;
- 2) $\forall S_t \notin \tilde{K}_\lambda : P^{\Omega, c^1, c^2}(S_t) = 0$;
- 3) $\Phi(P^{\Omega, c^1, c^2}(\mathbf{x})) = \text{extr}_{\{P^{\Omega^*, c^{1*}, c^{2*}}(\mathbf{x})\}} \Phi(P^{\Omega^*, c^{1*}, c^{2*}}(\mathbf{x}))$, где Φ – критерий качества предиката.

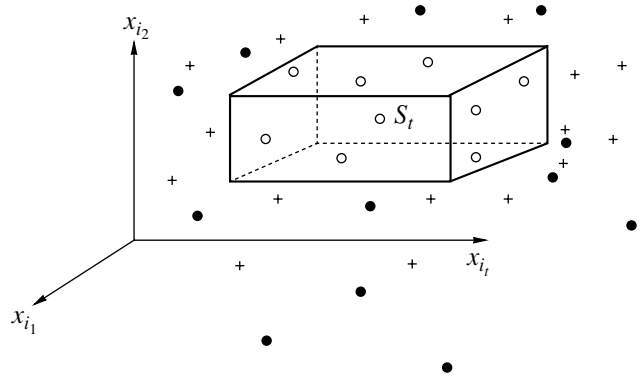
Предикат (1), удовлетворяющий только первым двум ограничениям, называется допустимым предикатом рассматриваемого класса.

Предикат (1), удовлетворяющий только первому и третьему ограничениям, называется частичной ЛЗ класса K_λ .

Стандартным критерием качества ЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$ класса K_λ будем называть следующий критерий: $F(P^{\Omega, c^1, c^2}(\mathbf{x})) = |\{S_i : S_i \in \tilde{K}_\lambda, P^{\Omega, c^1, c^2}(S_i) = 1\}|$.

Множество $N(P^{\Omega, c^1, c^2}) = \{\mathbf{x} \in \mathbb{R}^n : c_j^1 \leq x_j \leq c_j^2, j \in \Omega\}$ будем называть интервалом пред-

диката $P^{\Omega, c^1, c^2}(\mathbf{x})$ (аналог интервалов элементарных конъюнкций в алгебре логики). Два предиката $P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x})$, $P_2^{\Omega_2, c_2^1, c_2^2}(\mathbf{x})$ будем называть эквивалентными, если $P_1^{\Omega_1, c_1^1, c_1^2}(S_t) = P_2^{\Omega_2, c_2^1, c_2^2}(S_t)$, $t = 1, 2, \dots, m$. Два интервала $N(P_1^{\Omega_1, c_1^1, c_1^2})$, $N(P_2^{\Omega_2, c_2^1, c_2^2})$ называются эквивалентными, если $N(P_1^{\Omega_1, c_1^1, c_1^2}) \cap \tilde{S} = N(P_2^{\Omega_2, c_2^1, c_2^2}) \cap \tilde{S}$. Логические закономерности со стандартным критерием качества имеют простую геометрическую интерпретацию: по данным обучающей выборки требуется найти прямоугольный координатный гиперпараллелепипед, лежащий в некотором признаковом подпространстве, содержащий максимальное число эталонов из класса K_λ и только класса K_λ (см. фигуру).



Фигура.

Под экстремумами критерия $\Phi(P^{\Omega, c^1, c^2}(\mathbf{x}))$ понимаются локальные экстремумы. Допустимый предикат $P^{\Omega, c^1, c^2}(\mathbf{x})$ является локально-оптимальным по критерию $\Phi(P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x}))$, если не существует допустимого $P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x})$: $N(P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x})) \supseteq N(P^{\Omega, c^1, c^2}(\mathbf{x}))$, для которого $\Phi(P^{\Omega, c^1, c^2}(\mathbf{x})) > \Phi(P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x}))$ (для задачи минимизации), или $\Phi(P^{\Omega, c^1, c^2}(\mathbf{x})) < \Phi(P_1^{\Omega_1, c_1^1, c_1^2}(\mathbf{x}))$ (для задачи максимизации).

Далее рассмотрим задачу нахождения ЛЗ классов при различных критериях для $P^{\Omega, c^1, c^2}(\mathbf{x})$.

2. ЛОГИЧЕСКИЕ ЗАКОНОМЕРНОСТИ КЛАССОВ С ОПОРНЫМИ ЭТАЛОНАМИ И ЧАСТОТНЫМ КРИТЕРИЕМ КАЧЕСТВА

В настоящем разделе рассмотрим упрощенный вариант основной задачи, имеющий самостоятельный интерес. Данный вариант основан на следующих двух ограничениях:

логические закономерности определяются как характеристические функции определенных ε -окрестностей эталонов классов в признаковых подпространствах (аналоги представительных наборов k -значного случая);

вместо стандартного функционала качества предикатов рассматривается некоторый линейный по признакам эвристический функционал.

Пусть фиксирован произвольный эталон $S_t \in K_\lambda$. Будем искать ЛЗ вида

$$P_t^{\Omega, \varepsilon}(\mathbf{x}) = \bigwedge_{j \in \Omega} P_t^{\varepsilon_j}(x_j), \quad (2)$$

где

$$\Omega \subseteq \{1, 2, \dots, n\}, \quad P_t^{\varepsilon_j}(x_j) = \begin{cases} 1 & \text{при } a_{tj} - \varepsilon_j \leq x_j \leq a_{tj} + \varepsilon_j, \\ 0 & \text{иначе.} \end{cases}$$

Отметим, что $P_t^{\Omega, \epsilon}(S_i) = 1$, т.е. первое условие определения (1) выполнено. Объект S_i будем называть опорным для предикатов (2).

В качестве критерия оптимальности предиката возьмем функционал

$$f(P_t^{\Omega, \epsilon}) = \sum_{j \in \Omega} \varphi_{ij}(\epsilon_j), \quad (3)$$

$$\varphi_{ij}(\epsilon_j) = \frac{1}{|\tilde{K}_\lambda|} \sum_{S_i \in \tilde{K}_\lambda} \rho_j(\epsilon_j, S_i, S_i) + \frac{1}{|C\tilde{K}_\lambda|} \sum_{S_i \in C\tilde{K}_\lambda} [1 - \rho_j(\epsilon_j, S_i, S_i)],$$

где антиблизость $\rho_j(\epsilon_j, S_i, S_i) = \begin{cases} 0 & \text{при } |a_{ij} - a_{ij}| \leq \epsilon_j, \\ 1 & \text{иначе.} \end{cases}$

При выборе предикатов и критерия качества согласно (2), (3) задача поиска ЛЗ класса K_λ будет состоять в минимизации (3) на множестве $\{\Omega, \epsilon : \Omega \subseteq \{1, 2, \dots, n\}, \epsilon \geq 0\}$.

Упорядочим все различные значения $|a_{ij} - a_{ij}|, i = 1, 2, \dots, m$, по возрастанию в виде таких последовательностей:

$$r^j : r_{j1}, r_{j2}, \dots, r_{ju}, \quad r_{jv} < r_{jw} \quad \text{при} \quad v < w. \quad (4)$$

Очевидно, $r_{j1} = 0$.

1. Подпоследовательность $r_{ju}, u = v, v+1, \dots, v+w$, назовем *подпоследовательностью первого типа* последовательности (4), если верно следующее:

- а) для любого ее элемента $r_{ju}, v \leq u \leq v+w, w \geq 0$, не существует строк $S_i \in C\tilde{K}_\lambda : |a_{ij} - a_{ij}| = r_{ju}$;
- б) любое возможное расширение подпоследовательности дополнением элементов из (4) слева или справа нарушает свойства а).

2. Подпоследовательность $r_{ju}, u = v, v+1, \dots, v+w, w \geq 0$, назовем *подпоследовательностью второго типа* последовательности (4), если верно следующее:

- а) для любого ее элемента $r_{ju}, v \leq u \leq v+w, w \geq 0$, не существует строк $S_i \in K_\lambda : |a_{ij} - a_{ij}| = r_{ju}$;
- б) любое возможное расширение подпоследовательности дополнением элементов из (4) слева или справа нарушает свойство а).

Область $D = \{\epsilon_j \geq 0, j = 1, 2, \dots, n\}$ допустимых значений параметров ϵ можно ограничить некоторым конечным множеством $D^* \subseteq D$ таким, что

$$\min_{\epsilon \in D} f(P_t^{\Omega, \epsilon}) = \min_{\epsilon \in D^*} f(P_t^{\Omega, \epsilon}). \quad (5)$$

Действительно. Пусть

$$P_t^{\Omega, \epsilon'} = \arg \min_{\epsilon \in D} f(P_t^{\Omega, \epsilon})$$

и $\epsilon_i^* = \max\{r_{iv} : r_{iv} \leq \epsilon_i'\}$. Тогда P_t^{Ω, ϵ^*} удовлетворяет условиям 1), 2) определения 1, а $f(P_t^{\Omega, \epsilon^*}) = f(P_t^{\Omega, \epsilon'})$. Условие (5) будет выполнено, если в качестве D^* взять множество $D^* = \{\epsilon^* : \epsilon_i^* \in r^i, i = 1, 2, \dots, n\}$.

Покажем, что множество D^* может быть далее сокращено с помощью трех алгоритмов сокращения последовательностей (4) (или их подпоследовательностей, которые будем обозначать так же).

2.1. Первый алгоритм сокращения

В последовательности r^j выделяется подпоследовательность первого типа $r_{ju}, u = v, v+1, \dots, v+w$. Из r^j вычеркиваются все элементы данной подпоследовательности, кроме $r_{j, v+w}$. Данная процедура повторяется для всех подпоследовательностей первого типа последовательностей r^j .

2.2. Второй алгоритм сокращения

В последовательности r^j выделяется подпоследовательность второго типа r_{ju} , $u = v, v + 1, \dots, v + w$. Из r^j вычеркиваются все элементы данной подпоследовательности. Данная процедура повторяется для всех подпоследовательностей второго типа последовательностей r^j .

3. Пусть дана произвольная числовая последовательность $C: c_1, c_2, \dots, c_k$ и функция действительного аргумента $\psi(x) \geq 0$. Числовую последовательность $\psi(c_1), \psi(c_2), \dots, \psi(c_k)$ будем называть *функциональной*.

Третий алгоритм сокращения произвольной конечной числовой последовательности C состоит в выделении такой ее подпоследовательности

$$c_{i_1}, c_{i_2}, \dots, c_{i_h}, \quad (6)$$

которой соответствует специальная монотонно убывающая подпоследовательность функциональной последовательности $\psi(c_1), \psi(c_2), \dots, \psi(c_k)$.

2.3. Третий алгоритм сокращения

Шаг 1. Полагаем $v = w = 1, i_1 = 1$.

Шаг 2. $w \geq 1$. Пусть из $c_1, c_2, \dots, c_w$ выделена подпоследовательность $c_{i_1}, c_{i_2}, \dots, c_{i_v}$, $v \leq w$. Если $w = k$, подпоследовательность (6) считается построенной и происходит переход на шаг 3.

В противном случае w увеличивается на единицу.

Если $\phi(c_{i_v}) > \phi(c_w)$, то v увеличивается на единицу, $i_v = w$. Осуществляется переход на шаг 2.

Шаг 3. Конец 3-го алгоритма сокращения.

Пусть $P_t^{\Omega, \epsilon}(\mathbf{x})$ – произвольный предикат такой, что $\epsilon_i \in r^i, i = 1, 2, \dots, n, \epsilon_j \in \tilde{r}^j = \{r_{ju}, u = v, v + 1, \dots, v + w\} \subseteq r^j$, где r^j – последовательность (4) или ее подпоследовательность, а $\tilde{r}^j$ – подпоследовательность первого типа последовательности r^j .

Рассмотрим предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, для которого $e_i = \begin{cases} \epsilon_i, & i \neq j, \\ r_{j, v+w}, & i = j. \end{cases}$ Если предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$ – допустимый, то будет допустимым также и предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, а также $f(P_t^{\Omega, \epsilon}(\mathbf{x})) \leq f(P_t^{\Omega, \epsilon}(\mathbf{x}))$.

Пусть $\tilde{r}^j$ – подпоследовательность второго типа последовательности r^j . Рассмотрим предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, для которого $e_i = \begin{cases} \epsilon_i, & i \neq j, \\ r_{j, v-1}, & i = j. \end{cases}$ Тогда если предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$ допустимый, то будет допустимым также и предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$ в силу неравенства $e_j < \epsilon_j$. Кроме того, выполнено неравенство $f(P_t^{\Omega, \epsilon}(\mathbf{x})) < f(P_t^{\Omega, \epsilon}(\mathbf{x}))$.

Таким образом, после последовательного применения первого и второго алгоритмов сокращения к последовательностям $r^j, j = 1, 2, \dots, n$, вычисляется такое подмножество D^{**} множества D^* , минимум на котором функционала $f(P_t^{\Omega, \epsilon})$ совпадает с минимумом на D^* . Обозначим сокращенные последовательности снова в виде $r^j, D^{**} = r^1 \times r^2 \times \dots \times r^n, |r^j| = u_j$. Применим к последовательностям $r^j, j = 1, 2, \dots, n$, третий алгоритм сокращения, рассматривая в качестве соответствующих функциональных последовательностей последовательности $\phi_{ij}(r_{ji}), i = 1, 2, \dots, u_j$. В итоге получим множество $D^{***} = \epsilon^1 \times \epsilon^2 \times \dots \times \epsilon^n$, где $\epsilon^j = \{\epsilon_{ji}, i = 1, 2, \dots, k_j\}, \epsilon_{ju} < \epsilon_{jv}$ при $u < v$.

В силу свойств третьего алгоритма сокращения, для любого допустимого предиката $P_t^{\Omega, \epsilon}(\mathbf{x})$, $e_j \in r^j$, существует $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n), \epsilon_j = e_{jv(j)} \leq e_j$, для которого $\phi_{ij}(\epsilon_j) \leq \phi_{ij}(e_j)$, т.е. существует допустимый предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, для которого $f(P_t^{\Omega, \epsilon}(\mathbf{x})) \leq f(P_t^{\Omega, \epsilon}(\mathbf{x}))$. Таким образом, доказана

Теорема 1. Пусть D^{***} получено из D^* с помощью трех алгоритмов сокращения, а логические закономерности $P_t^{\Omega, \epsilon}(\mathbf{x})$ определены согласно выражению (2), тогда

$$\min_{\Omega, \epsilon \in D^{***}} f(P_t^{\Omega, \epsilon}(\mathbf{x})) = \min_{\Omega, \epsilon \in D^*} f(P_t^{\Omega, \epsilon}(\mathbf{x})).$$

После явного описания процедуры нахождения множества D^{***} покажем, что задача поиска логических закономерностей может быть сформулирована в виде специальной задачи целочисленного линейного программирования.

Введем следующие обозначения:

$$c_{ij} = \varphi_{ti}(\epsilon_{ij}),$$

$$y = (y_{11}, y_{12}, \dots, y_{1k_1}, y_{21}, y_{22}, \dots, y_{2k_2}, \dots, y_{n1}, y_{n2}, \dots, y_{nk_n}),$$

$$b_{ij}^u = \rho_i(\epsilon_{ij}, S_t, S_u).$$

Рассмотрим задачу

$$\sum_{i=1}^n \sum_{j=1}^{k_i} c_{ij} y_{ij} \longrightarrow \min, \quad (7)$$

$$\sum_{i=1}^n \sum_{j=1}^{k_i} b_{ij}^u y_{ij} \geq 1, \quad u = 1, 2, \dots, m, \quad S_u \notin K_\lambda, \quad (8)$$

$$y_{ij} \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, k_i.$$

По определению коэффициентов и в силу свойств третьего алгоритма сокращения для задачи (7), (8) имеют место следующие свойства коэффициентов:

$$c_{iu} > c_{iv} \geq 0 \quad \text{при} \quad u < v, \quad b_{ij}^u \geq b_{ij}^v \geq 0, \quad b_{ij}^u \in \{0, 1\}. \quad (9)$$

В силу свойств коэффициентов (9), задачу (7), (8) будем называть задачей целочисленного линейного программирования с блочно-монотонными столбцами (задача ЦЛП-БМС).

Теорема 2. Решению задачи (7), (8) однозначно соответствует решение задачи

$$f(P_t^{\Omega, \epsilon}(\mathbf{x})) \xrightarrow{\Omega \subseteq \{1, 2, \dots, n\}, \epsilon \in D^{***}} \min.$$

В силу свойств коэффициентов функции (7), оптимальное решение $y^* = (y_{11}^*, y_{12}^*, \dots, y_{1k_1}^*, y_{21}^*, y_{22}^*, \dots, y_{2k_2}^*, \dots, y_{n1}^*, y_{n2}^*, \dots, y_{nk_n}^*)$ содержит не более одной единичной компоненты для каждой из n групп параметров $y_{i1}^*, y_{i2}^*, \dots, y_{ik_i}^*$. Действительно, в случае $c_{ik_i} > 0$ это свойство проверяется от противного. Случай $c_{ik_i} = 0$ возможен лишь тогда, когда $k_i = 1$, а значит, данное свойство также выполняется. Множество единичных компонент решения задачи (7), (8) однозначно определяет предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, где $\Omega = \{i : y_{ij}^* = 1\}$, а $\epsilon_i = \epsilon_{ij_i}$, $i \in \Omega$. Предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$ является ЛЗ класса K_λ с опорным объектом S_t , при этом выполнение условий 2) определения ЛЗ обеспечивается выполнением ограничений (8) найденного решения задачи (7), (8).

Замечания. 1. В случаях, когда $\mathbf{I} = \{i : c_{ik_i} = 0, i \in \Omega\} \neq \emptyset$, имеет место отделимость K_λ от остальных классов по одному из признаков. Предикат $P_t^{\Omega, \epsilon}(\mathbf{x})$, ($\Omega = \{v\}$, $v \in \mathbf{I}$, $\epsilon_v = \max_{S_i \in K_\lambda} |a_{iv} - a_{iv}|$, $f(P_t^{\Omega, \epsilon}(\mathbf{x})) = 0$) будет являться ЛЗ класса K_λ .

2. Интервалы $N(P_t^{\Omega, \epsilon})$ являются прямыми аналогами представительных наборов (см. [6]). Находятся несократимые фрагменты эталонных описаний S_t и их ϵ -окрестности.

3. Найденным ЛЗ $P_t^{\Omega, \epsilon}(\mathbf{x})$ соответствует множество им эквивалентных ЛЗ $P_t^{\Omega, \epsilon}(\mathbf{x})$, для которых $N(P_t^{\Omega, \epsilon}(\mathbf{x})) \supseteq N(P_t^{\Omega, \epsilon}(\mathbf{x}))$.

Далее рассмотрим более простой вариант для (3), отбросив нормирующие множители:

$$\varphi_{tj}(\varepsilon_j) = \sum_{S_i \in K_\lambda} \rho_j(\varepsilon_j, S_i) + \sum_{S_i \in CK_\lambda} [1 - \rho_j(\varepsilon_j, S_i)]. \quad (3)'$$

Область допустимых значений параметров ε_j можно ограничить подмножеством множества $A_j = \{[a_{ij} - a_{ij}], i = 1, 2, \dots, m\}$, которое вычисляется с помощью трех алгоритмов сокращения, $\varepsilon_j \in \{\varepsilon_{j1} < \varepsilon_{j2} < \dots < \varepsilon_{jk_j}\}$. Множество $\{\varepsilon_{j1}, \varepsilon_{j2}, \dots, \varepsilon_{jk_j}\}$ монотонно возрастающих вариантов выбора ε_j , как и ранее, будем обозначать как последовательность чисел r^j .

Какова может быть размерность пространства оптимизации в задаче (7), (8) и как она связана с реальными данными и их качеством? Рассмотрим один общий модельный пример, наложив дополнительные ограничения на исходные данные обучающей выборки. Будем считать, что выполнено условие

$$|a_{tj} - a_{ij}| \neq |a_{tj} - a_{\mu j}| \quad \forall \mu \neq i, \quad i, \mu = 1, 2, \dots, m.$$

Тогда последовательность r^j может быть представлена в виде $r^j = r_1^1 r_1^2 r_2^1 r_2^2 r_3^1 r_3^2 \dots r_{k_j}^1$ либо $r^j = r_1^1 r_1^2 r_2^1 r_2^2 r_3^1 r_3^2 \dots r_{k_j}^1 r_{k_j}^2$, где r_s^1 – подпоследовательности первого типа, а r_s^2 – подпоследовательности второго типа.

Пусть $|r_i^1| = b_i > 0$ и $|r_i^2| = d_i > 0$ – число элементов в соответствующих последовательностях.

Пусть $E = e_1, e_2, \dots, e_\tau$ – последовательность, полученная из r^j применением 1-го, 2-го и 3-го алгоритмов сокращения (поскольку здесь и далее рассматривается фиксированный признак, мы опустим индекс j). Введем обозначение $h = |K_\lambda|$.

Теорема 3. Справедливо неравенство $k_j - 1 \leq \min_{\varepsilon_j} \varphi_{tj}(\varepsilon_j) \leq \min\{h - \tau, m - h\}$.

Доказательство. Элементы e_i являются правыми крайними элементами некоторых подпоследовательностей $r_{v_i}^1$, $v_i \geq i$. Ясно, что

$$\min_{\varepsilon_j} \varphi_{tj}(\varepsilon_j) = \varphi_{tj}(e_\tau).$$

Пусть

$$\varphi_i^* = \varphi_{tj}(e_i) = h - \sum_{\mu=1}^{v_i} b_\mu + \sum_{\mu=1}^{v_i-1} d_\mu.$$

По построению e_i имеет место

$$\varphi_1^* > \varphi_2^* > \dots > \varphi_\tau^*. \quad (10)$$

В силу оптимальности e_τ имеет место неравенство

$$\varphi_\tau^* \leq h - \sum_{\mu=1}^{v_\tau+i} b_\mu + \sum_{\mu=1}^{v_\tau+i-1} d_\mu, \quad i = 0, 1, 2, \dots, \quad \text{при } v_\tau + i \leq k. \quad (11)$$

По условию имеем

$$\sum_{\mu=1}^k b_\mu = h, \quad \sum_{\mu=1}^{k-1} d_\mu \leq m - h.$$

Из (10) следует, что $\varphi_\tau^* \leq \varphi_{\tau-1}^* - 1 \leq \varphi_{\tau-2}^* - 2 \leq \dots \leq \varphi_1^* - (\tau - 1) \leq h - \tau$ (здесь $\varphi_1^* \leq h - 1$). Из (11) следует

$$\varphi_\tau^* \leq h - \sum_{\mu=1}^k b_\mu + \sum_{\mu=1}^{k-1} d_\mu \leq m - h.$$

Окончательно имеем $\varphi_\tau^* \leq \min\{h - \tau, m - h\}$.

С другой стороны,

$$\varphi_{\tau}^* = h - \sum_{\mu=1}^{\nu_{\tau}} b_{\mu} + \sum_{\mu=1}^{\nu_{\tau}-1} d_{\mu} = \sum_{\mu=\nu_{\tau}+1}^k b_{\mu} + \sum_{\mu=1}^{\nu_{\tau}-1} d_{\mu} \geq k-1.$$

Теорема доказана.

Следствие 1. Имеет место неравенство $\tau - 1 \leq \varphi_{\tau}^* \leq \min\{h - \tau, m - h\}$.

Теорема и следствие показывают, что чем больше величина $\tau = \tau_j$, тем меньше область допустимых значений для функции $\varphi_i(\varepsilon_{ij})$. А это означает, что при больших τ_j (а значит, “плохих признаках” x_j) вариация функционала по параметру ε_j невелика и значения ε_j являются “примерно равнозначными” по функционалу.

Для оценки сверху τ в рассмотренной выше модельной задаче ограничимся случаем $|K_{\lambda}| = |CK_{\lambda}|$. Пусть $N = |K_{\lambda}|$, $t = [N/2]$.

Теорема 4. Справедливо соотношение $\max_{r^j} \tau = [N/2]$.

Доказательство. Ясно, что $\tau \leq \beta$, где β – число выполненных неравенств $|r_j^1| > |r_j^2|$, $j \geq 1$, либо $\tau \leq \beta + 1$, где β – число выполненных неравенств $|r_j^1| > |r_j^2|$, $j > 1$ (допустим случай $|r_1^1| = |r_1^2|$). В первом случае из $|r_j^2| \geq 1$ и $|r_j^1| \geq 2$ следует $\beta \leq t$, во втором случае имеем $\beta \leq t - 1$. В обоих случаях $\tau \leq t$, причем равенство $\tau = t$ достигается на таких последовательностях:

$$r^j = r_1^1 r_1^2 r_2^1 r_2^2 r_3^1 r_3^2 \dots r_t^1 r_t^2,$$

где

$$\begin{aligned} |r_i^1| &= 2, \quad i = 1, 2, \dots, t-1, \\ |r_t^1| &= \begin{cases} 2, & N \text{ четное,} \\ 3, & N \text{ нечетное,} \end{cases} \\ |r_i^2| &= 1, \quad i = 1, 2, \dots, t-1, \\ |r_t^2| &= N - t + 1. \end{aligned}$$

Теорема доказана.

Следствие 2. Справедливы соотношения $\varphi_1^* = N - 2$, $\varphi_2^* = N - 3$, ..., $\varphi_{\tau}^* = [N/2] - 1$.

Естественно использовать при оптимизации стандартного функционала качества $F(P^{\Omega, c^1, c^2}(\mathbf{x}))$ в качестве начальных приближений ЛЗ, полученные в результате оптимизации эвристического критерия $f(P_S^{\Omega, \varepsilon}(\mathbf{x}))$ (здесь $S \in \mathbb{R}^n$). Связь оптимальных решений при критериях $F(P^{\Omega, c^1, c^2}(\mathbf{x}))$ и $f(P_S^{\Omega, \varepsilon}(\mathbf{x}))$ качества ЛЗ показывает следующее утверждение (как и ранее, будем считать $|K_{\lambda}| = |CK_{\lambda}|$, $N = |K_{\lambda}|$).

Теорема 5. Справедливо неравенство $\min_P f(P_S^{\Omega, \varepsilon}(\mathbf{x})) \leq n[2N - \max_P F(P^{\Omega, c^1, c^2}(\mathbf{x}))] - N$.

Доказательство. Пусть $L = \max_P F(P^{\Omega, c^1, c^2}(\mathbf{x}))$. Рассмотрим предикат $P_S^{\Omega, \varepsilon}(\mathbf{x})$, $\varepsilon_i = (c_i^2 - c_i^1)/2$, $i = 1, 2, \dots, n$, относительно опорной точки $S = (s_1, s_2, \dots, s_n)$, $s_i = (c_i^1 + c_i^2)/2$, $i = 1, 2, \dots, n$. Тогда $\varphi_j(\varepsilon_j) \leq N - L + N$ (в худшем случае имеется $N - L$ “далеких” по данному признаку объектов из K_{λ} и близких из CK_{λ}), и

$$f(P_S^{\Omega, \varepsilon}(\mathbf{x})) = \sum_{j \in \omega} \varphi_j(\varepsilon_j) \leq n(2N - L) - N.$$

Оценку $f(P_S^{\Omega, \varepsilon}(\mathbf{x}))$ можно уменьшить на N в силу условия 2) определения ЛЗ. Теорема доказана.

3. ЛОГИЧЕСКИЕ ЗАКОНОМЕРНОСТИ СО СТАНДАРТНЫМ КРИТЕРИЕМ КАЧЕСТВА

Здесь будет рассмотрена задача поиска ЛЗ со стандартным критерием качества и показано, что она может быть сведена к некоторой задаче целочисленного программирования. Рассмотрим данную задачу при ограничении

$$P^{\Omega, c^1, c^2}(S_i) = \bigwedge_{j \in \Omega} P_j^{c_j^1, c_j^2}(x_j(S_i)) = 1,$$

где $S_i \in \tilde{K}_\lambda$ – произвольный фиксированный (опорный) объект. Поскольку для любого решения Ω, c^1, c^2 ему эквивалентным будет ЛЗ со значениями параметров $\Omega^* = \{1, 2, \dots, n\}$, c^{*1}, c^{*2} , где

$$c_j^{*1} = \begin{cases} c_j^1, & j \in \Omega, \\ \min_i a_{ij}, & j \notin \Omega, \end{cases} \quad c_j^{*2} = \begin{cases} c_j^2, & j \in \Omega, \\ \max_i a_{ij}, & j \notin \Omega, \end{cases}$$

мы можем ограничиться случаем $\Omega \equiv \{1, 2, \dots, n\}$ с последующим удалением фиктивных признаков.

Пусть $D_i^1 = \{d_{i1}^1, d_{i2}^1, \dots, d_{iu_i}^1\}$ и $D_i^2 = \{d_{i1}^2, d_{i2}^2, \dots, d_{iv_i}^2\}$ – множества всех упорядоченных по убыванию значений a_j : $S_i \in \tilde{K}_\lambda$, $a_{ij} \geq a_{ij}$, и по возрастанию a_j : $S_i \in \tilde{K}_\lambda$, $a_{ij} \leq a_{ij}$, соответственно. Нетрудно убедиться (аналогично (5)), что

$$\min_{c^1, c^2 \in \mathbb{R}^n} F(P^{\Omega, c^1, c^2}(\mathbf{x})) = \min_{c_i^1 \in D_i^1, c_i^2 \in D_i^2, i=1, 2, \dots, n} F(P^{\Omega, c^1, c^2}(\mathbf{x})),$$

поэтому ограничимся поиском ЛЗ на $D_1^1 \times D_2^1 \times \dots \times D_n^1 \times D_1^2 \times D_2^2 \times \dots \times D_n^2$.

Для простоты записи будем считать, что первые по порядку объекты обучающей выборки принадлежат рассматриваемому классу $\tilde{K}_\lambda = \{S_1, S_2, \dots, S_h\}$.

Построим числовую матрицу

$$M = \begin{pmatrix} B_1^1 B_1^2 B_2^1 B_2^2 \dots B_n^1 B_n^2 \\ C_1^1 C_1^2 C_2^1 C_2^2 \dots C_n^1 C_n^2 \end{pmatrix}_{m \times N},$$

$$N = n \times \sum_{i=1}^n (u_i + v_i),$$

$$B_i^1 = (b_{ij}^{1q})_{h \times u_i}, \quad q = 1, 2, \dots, h, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, u_i,$$

$$B_i^2 = (b_{ij}^{2q})_{h \times v_i}, \quad q = 1, 2, \dots, h, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, v_i,$$

$$C_i^1 = (c_{ij}^{1q})_{(m-h) \times u_i}, \quad q = 1, 2, \dots, m-h, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, u_i,$$

$$C_i^2 = (c_{ij}^{2q})_{(m-h) \times v_i}, \quad q = 1, 2, \dots, m-h, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, v_i,$$

где

$$b_{ij}^{1q} = \begin{cases} 1 & \text{при } x_i(S_q) \geq d_{ij}^1, \\ 0 & \text{иначе,} \end{cases} \quad b_{ij}^{2q} = \begin{cases} 1 & \text{при } x_i(S_q) \leq d_{ij}^2, \\ 0 & \text{иначе,} \end{cases} \quad (12)$$

$$c_{ij}^{1q} = \begin{cases} 1 & \text{при } x_i(S_{m_\lambda+q}) < d_{ij}^1, \\ 0 & \text{иначе,} \end{cases} \quad c_{ij}^{2q} = \begin{cases} 1 & \text{при } x_i(S_{m_\lambda+q}) > d_{ij}^2, \\ 0 & \text{иначе.} \end{cases} \quad (13)$$

Как следует из (12), (13), выполняются следующие ограничения для элементов М:

$$b_{ij}^{1q} \leq b_{i,j+1}^{1q}, \quad b_{ij}^{2q} \leq b_{i,j+1}^{2q},$$

$$c_{ij}^{1q} \geq c_{i,j+1}^{1q}, \quad c_{ij}^{2q} \geq c_{i,j+1}^{2q}.$$

Рассмотрим множество векторов $\{\langle x_{ij}^1, x_{ij}^2 \rangle\}$, $\langle x_{ij}^1, x_{ij}^2 \rangle = \langle x_{11}^1, x_{12}^1, \dots, x_{1u_1}^1, x_{11}^2, x_{12}^2, \dots, x_{1v_1}^2, x_{21}^1, x_{22}^1, \dots, x_{2u_2}^1, x_{21}^2, x_{22}^2, \dots, x_{2v_2}^2, \dots, x_{n1}^1, x_{n2}^1, \dots, x_{nu_n}^1, x_{n1}^2, x_{n2}^2, \dots, x_{nv_n}^2 \rangle$ при ограничениях

$$x_{ij}^1, x_{ij}^2 \in \{0, 1\}, \quad \sum_{j=1}^{u_j} x_{ij}^1 = 1, \quad \sum_{j=1}^{v_j} x_{ij}^2 = 1, \quad i = 1, 2, \dots, n.$$

Единицам в $\{\langle x_{ij}^1, x_{ij}^2 \rangle\}$ поставим в соответствие выбор значений параметров $c_i^1 \in D_i^1$, $c_i^2 \in D_i^2$, $i = 1, 2, \dots, n$. Множество всех предикатов (1) с возможными границами D_i^1 , D_i^2 находится во взаимно однозначном соответствии с множеством данных бинарных векторов, поэтому мы будем использовать также запись $F(\langle x_{ij}^1, x_{ij}^2 \rangle)$ для стандартного критерия оптимальности.

Условие 1) определения 1 выполняется в силу $S_t \in \tilde{K}_\lambda$. Для выполнения условия 2) определения 1 должны выполняться неравенства

$$f_q^c(\langle x_{ij}^1, x_{ij}^2 \rangle) = \sum_{i=1}^n \left(\sum_{j=1}^{u_j} c_{ij}^{1q} x_{ij}^1 + \sum_{j=1}^{v_j} c_{ij}^{2q} x_{ij}^2 \right) \geq 1, \quad q = 1, 2, \dots, m-h.$$

Наконец, стандартный критерий оптимальности предиката $F(\langle x_{ij}^1, x_{ij}^2 \rangle)$ равен числу выполненных равенств в системе

$$f_q^b(\langle x_{ij}^1, x_{ij}^2 \rangle) = \sum_{i=1}^n \left(\sum_{j=1}^{u_j} (b_{ij}^{1q} - 1) x_{ij}^1 + \sum_{j=1}^{v_j} (b_{ij}^{2q} - 1) x_{ij}^2 \right) = 0, \quad q = 1, 2, \dots, h. \quad (14)$$

Таким образом, проблема поиска ЛЗ (1) может быть сформулирована как следующая дискретная оптимизационная

Задача Z. $F(\langle x_{ij}^1, x_{ij}^2 \rangle) = \langle \text{число выполненных уравнений в (14)} \rangle \rightarrow \max$ при ограничениях

$$\sum_{i=1}^n \left(\sum_{j=1}^{u_j} c_{ij}^{1q} x_{ij}^1 + \sum_{j=1}^{v_j} c_{ij}^{2q} x_{ij}^2 \right) \geq 1, \quad q = 1, 2, \dots, m-h, \quad (15)$$

$$x_{ij}^1, x_{ij}^2 \in \{0, 1\}, \quad \sum_{j=1}^{u_j} x_{ij}^1 = 1, \quad \sum_{j=1}^{v_j} x_{ij}^2 = 1, \quad i = 1, 2, \dots, n. \quad (16)$$

Решая задачи Z, для каждого опорного эталона находим множество логических закономерностей Λ , при этом произвольная логическая закономерность класса $\tilde{K}_\lambda$ имеет в Λ эквивалентную ей ЛЗ. Стохастический вариант алгоритма поиска ЛЗ классов описан в [15].

4. АЛГОРИТМЫ РАСПОЗНАВАНИЯ (ВЫЧИСЛЕНИЯ ОЦЕНОК), ОСНОВАННЫЕ НА ГОЛОСОВАНИИ ПО СИСТЕМАМ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ КЛАССОВ

Пусть для каждого класса K_λ найдено множество ЛЗ $\mathcal{P}_\lambda = \{P^{\Omega, c, d}(\mathbf{x})\}$, при этом множество интервалов $\{N(P^{\Omega, c, d}) : P^{\Omega, c, d} \in \mathcal{P}_\lambda\}$ покрывает $\tilde{K}_\lambda$. Функции

$$D_\lambda(\mathbf{x}) = \bigvee_{P^{\Omega, c, d} \in \mathcal{P}_\lambda} P^{\Omega, c, d}(\mathbf{x})$$

принимают значения 1 на объектах $\tilde{K}_\lambda$ и 0 на $\tilde{S} \setminus \tilde{K}_\lambda$. Они дают некоторое наглядное представление о конфигурации классов. Сокращенные и кратчайшие ДНФ, соответствующие $D_\lambda(\mathbf{x})$, дают

более компактное представление о конфигурации классов и их сложности (чем больше интервалов в кратчайшем или минимальном покрытии, тем класс “сложнее”, см. [14]). Функции $D_\lambda(\mathbf{x})$ могут рассматриваться как приближения характеристических функций классов K_λ . Использование же их в качестве решающих правил нецелесообразно, поскольку при распознавании новых объектов они часто оказываются покрытыми интервалами $N(P^{\Omega, c, d})$ различных классов и зона отказов от распознавания может быть велика. Схемы голосования, используемые в алгоритмах вычисления оценок из [2], [3], [6], позволяют преодолеть подобные ситуации неопределенности. Пусть для распознаваемого объекта S на базе множеств $\mathcal{P}_\lambda$ вычислена его оценка $\Gamma_\lambda(S)$ за класс K_λ , $\lambda = 1, 2, \dots, n$ (“степень близости” S к K_λ). Тогда применяется, например, следующее общее решающее правило:

$$\alpha_j^A(S) = \begin{cases} 1 & \text{при } \sum_{i=1}^l \delta_i^j \Gamma_i(S) \geq \delta_i^0, \\ 0 & \text{иначе.} \end{cases}$$

Здесь $\alpha_j^A(S) = 1$ ($\alpha_j^A(S) = 0$) означает решение $S \in K_j$ ($S \notin K_j$) алгоритма распознавания A . Параметры δ_i^j находятся в процессе оптимизации модели распознавания с использованием контрольной выборки.

В настоящем разделе предлагается модификация общей распространенной схемы алгоритмов вычисления оценок для этапа решения задачи распознавания с использованием логических закономерностей классов.

Определим следующие способ вычисления оценок и решающее правило.

Вычисление оценок:

$$\Gamma_j(S) = \sum_{P^{\Omega, c, d} \in \mathcal{P}_j} \gamma(P^{\Omega, c, d}) \phi^{\Omega, c, d}(S), \quad (17)$$

где $\gamma(P^{\Omega, c, d})$ – неотрицательные параметры, характеризующие “вес” соответствующей ЛЗ, $\phi^{\Omega, c, d}(S)$ – потенциальная функция, соответствующая $P^{\Omega, c, d}(S)$. В частности, функция $\phi^{\Omega, c, d}(S)$ может совпадать с $P^{\Omega, c, d}(S)$.

Решающее правило:

$$\alpha_j^A(S) = \begin{cases} 1 & \text{при } \delta_j \Gamma_j(S) > \delta_i \Gamma_i(S) \quad \forall i \neq j, \\ 0 & \text{иначе,} \end{cases} \quad (18)$$

где $\delta_j, j = 1, 2, \dots, l$, – неотрицательные нормировочные параметры классов.

Рассмотрим вопросы выбора неизвестных параметров и потенциальных функций в (17), (18).

1. Обычный путь к решению задачи поиска неизвестных значений параметров – решение задачи оптимизации стандартного критерия качества распознавания с использованием контрольной выборки, что часто является затруднительным как в части решения самой оптимизационной задачи, так и возможности располагать дополнительной информацией (контрольной выборкой). Вопрос наилучшего разбиения исходной информации на обучающую и контрольную (особенно при ее дефиците) является самостоятельной задачей в каждом практическом случае. Прямое использование здесь обучающей выборки и в качестве контрольной не имеет смысла, так как алгоритмы (17), (18) безошибочно классифицируют непротиворечивую обучающую выборку. Использование для оптимизации схем скользящего контроля имеет свои проблемы, и не только вычислительные. Что, например, делать, когда скользящий контроль неприменим: классы имеют “выбросы”, “нетипичные объекты” (но которые нельзя оправдывать наличием грубых ошибок) или случаи малых обучающих выборок, когда каждый эталон по-своему уникален?

Здесь перспективным представляется развитие подходов, являющихся аналогами “максимального зазора” между классами в методе опорных векторов (см. [16]).

Пусть множество найденных ЛЗ классов перенумеровано:

$$\{P_1(\mathbf{x}), P_2(\mathbf{x}), \dots, P_k(\mathbf{x})\} = \bigcup_{\lambda=1, 2, \dots, l} \mathcal{P}_\lambda.$$

Рассмотрим следующую задачу линейного программирования:

$$\sigma \longrightarrow \max, \quad (19)$$

$$\sum_{P_i \in \mathcal{P}_\lambda} \gamma_i \phi_i^{\Omega, c, d}(S_j) \geq \sigma, \quad j = 1, 2, \dots, m, \quad (20)$$

$$\sum_{i=1}^k \gamma_i = k, \quad \gamma_i \geq 0, \quad i = 1, 2, \dots, k, \quad (21)$$

$$\gamma_i = \gamma(P_i).$$

В задаче (19)–(21) находятся такие весовые коэффициенты для ЛЗ классов, при которых оценки эталонов за классы будут максимальны. Аналогично, поиск параметров решающего правила (18) сведем к решению задачи

$$\delta \longrightarrow \max, \quad (22)$$

$$\delta_j \Gamma_j(S_i) \geq \delta, \quad i = 1, 2, \dots, m, \quad S_i \in \tilde{K}_j, \quad (23)$$

$$\sum_{j=1}^l \delta_j = l, \quad \delta_j \geq 0, \quad j = 1, 2, \dots, l \quad (24)$$

(предполагается, что $\Gamma_j(S_i) > 0$ при $S_i \in \tilde{K}_j$).

Пусть $a_i = \min_{S_t \in \tilde{K}_i} \Gamma_i(S_t)$, тогда система (23) эквивалентна системе

$$\delta_j a_j \geq \delta, \quad j = 1, 2, \dots, l. \quad (25)$$

Нетрудно убедиться, что задача (22), (24), (25) (а значит, и задача (22)–(24)) имеет аналитическое решение

$$\delta = \left(\sum_{i=1}^l \frac{1}{a_i} \right)^{-1}, \quad \delta_j = \left(a_j \sum_{i=1}^l \frac{1}{a_i} \right)^{-1}.$$

Отметим, что поиск параметров $\delta_j, j = 1, 2, \dots, l$, является излишним (им можно присвоить, например, единичные значения), если оценки вычисляются согласно (17) для которых параметры $\gamma_i = \gamma(P_i)$ найдены согласно (19)–(21).

Представляют практический интерес и другие способы наложения ограничений на оценки классов. Так, использование

$$a_j = \frac{1}{|\tilde{K}_j|} \sum_{S_i \in \tilde{K}_j} \Gamma(S_i)$$

в (25) может быть более целесообразным в случаях плохой отделимости классов, при стохастической реализации тестового алгоритма (см. [15]), использовании частичных логических закономерностей.

2. Алгоритмы, основанные на принципе частичной прецедентности (в частности, голосовании по системам ЛЗ) эффективно реализуют такие принципы теории распознавания, как взаимосвязь признаков, локальный учет информации, согласование метрик различных признаков. Использование вместо предикатов $P^{\Omega, c, d}(S)$ аппроксимирующих их “классических” потенциальных функций (см. [17], [16]) позволяет распространить “зону влияния” ЛЗ за пределы $N(P^{\Omega, c, d})$ при решении задач распознавания и более объективно численно оценивать факт $P^{\Omega, c, d}(S) = 1$ для различных объектов. Данный момент может быть особенно существенным при малом числе “голосующих” предикатов $P^{\Omega, c, d}(S)$.

Для аппроксимации $P^{\Omega, c, d}(S)$ будем использовать функции следующего вида:

$$1) f(\mathbf{x}) = \exp \left[-\frac{1}{2} \sum_{i \in \Omega} \left(\frac{x_i - b_i}{\sigma_i} \right)^2 \right],$$

2) $f(\mathbf{x}) = \delta \exp \left[-\frac{1}{2} \sum_{i \in \Omega} \left(\frac{x_i - b_i}{\sigma_i} \right)^2 \right]$, где δ, σ_i, b_i – параметры аппроксимирующей функции.

Пусть $\varepsilon_i = (d_i - c_i)/2$ и положим $b_i = (d_i + c_i)/2, i \in \Omega$. Рассмотрим примеры возможных аппроксимаций предикатов $P^{\Omega, c, d}(S)$, ограничиваясь обычно дополнительными ограничениями $\sigma_i = 2\varepsilon_i\sigma$.

а. Выбор из условия

$$\int f(\mathbf{x}) d\mathbf{x} = \int P^{\Omega, c, d}(\mathbf{x}) d\mathbf{x}$$

для функции 2), где интегрирование проводится по пространству R^k переменных, задаваемых опорным множеством $\Omega, k = |\Omega|$, приводит к равенству

$$\prod_{i \in \Omega} (d_i - c_i) = \delta (2\pi)^{k/2} \prod_{i \in \Omega} \sigma_i.$$

“Естественным” допустимым решением здесь будет $\delta = (2\pi)^{-k/2}, \sigma_i = 2\varepsilon_i$.

б. Выбор из условия минимума критерия

$$\int [f(\mathbf{x}) - P^{\Omega, c, d}(\mathbf{x})]^2 d\mathbf{x}$$

для функции 1) приводит к задаче минимизации выражения

$$\pi^{k/2} \prod_{i \in \Omega} \sigma_i - 2^{k/2+1} \pi^2 \prod_{i \in \Omega} \sigma_i \Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) + 2^k \prod_{i \in \Omega} \varepsilon_i,$$

где $\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$ – функция Лапласа. Ограничив выбор однопараметрическим случаем $\sigma_i = 2\varepsilon_i\sigma$, неизвестное σ найдем из условия $\pi^{k/2}\sigma^k - 2^{k/2+1}\pi^{k/2}\sigma^k\Phi(1/(2\sigma)) + 1 \xrightarrow{\sigma} \min$. В одномерном случае (или при независимой аппроксимации каждой функции $P_j^{c_j, d_j}(x), j \in \Omega$, в отдельности) приближенным решением задачи $\sqrt{\pi}\sigma - 2\sqrt{2\pi}\sigma\Phi(1/(2\sigma)) + 1 \xrightarrow{\sigma} \min$ будет $\sigma \cong 0.4, \sigma_i = 0.8\varepsilon_i$.

в. Выбор из условия минимума критерия

$$\int [f(\mathbf{x}) - P^{\Omega, c, d}(\mathbf{x})]^2 d\mathbf{x}$$

для функции 2) приводит к решению

$$\delta = 2^{k/2} \prod_{i \in \Omega} \left[\sigma_i \Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) \right] \left(\prod_{i \in \Omega} \sigma_i \right)^{-1},$$

где σ_i находятся из решения задачи

$$\pi^{-k/2} \prod_{i \in \Omega} \varepsilon_i - \prod_{i \in \Omega} \left[\sigma_i \Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) \right]^2 \left(\prod_{i \in \Omega} \sigma_i \right)^{-1} \xrightarrow{\sigma_i} \min.$$

При выборе $\sigma_i = 2\varepsilon_i\sigma$ имеем $\delta = 2^{k/2}\Phi^k(1/(2\sigma))$, где σ является решением задачи $\sigma^k\Phi^{2k}(1/(2\sigma)) \xrightarrow{\sigma} \max$. При $k = 1$ имеем, соответственно, $\sigma \cong 0.4, \delta \cong 1.11, \sigma_i = 0.8\varepsilon_i$.

г. Используя критерий $\max_{\mathbf{x}} |f(\mathbf{x}) - P^{\Omega, c, d}(\mathbf{x})| \rightarrow \min$, основную задачу поиска параметров $\sigma_i, i \in \Omega$, сводим к решению уравнения

$$1 - \exp \left(-\frac{1}{2} \sum_{i \in \Omega} (\varepsilon_i/\sigma_i)^2 \right) = \max_{i \in \Omega} \exp \left(-\frac{1}{2} (\varepsilon_i/\sigma_i)^2 \right).$$

При $\sigma_i = 2\varepsilon_i\sigma$ уравнение имеет вид $1 - \exp[-n/(8\sigma^2)] = \exp[-1/(8\sigma^2)]$. Для $k = 1$ имеем $\sigma \cong 0.5, \sigma_i \cong \varepsilon_i$.

Отметим, что варианты выбора значений σ для $k = 1, 2, \dots, n$ в б–г могут быть предварительно вычислены и использоваться для вычисления σ_i в соответствующих выражениях.

д. Параметры σ_i ищутся в виде $\sigma_i = 2\varepsilon_i\sigma$, а параметр σ находится из решения задач (19)–(21) или (22)–(24) при использовании функций вида 1) наряду с соответствующими параметрами данных задач.

е. Потенциальные функции $\phi^{\Omega, c, d}(S)$ находятся с использованием информации о распределении объектов из $N(P^{\Omega, c, d}) \cap \tilde{K}_\lambda$. Предикату $P^{\Omega, c, d} \in \mathcal{P}_\lambda$ поставим в соответствие функцию

$$f(\mathbf{x}) = \exp \left[-\frac{1}{2} \sum_{i \in \Omega} \left(\frac{x_i - \tilde{x}_i}{\tilde{\sigma}_i} \right)^2 \right],$$

где $\tilde{x}_i$, $\tilde{\sigma}_i$ – выборочные средние и выборочные средние квадратические отклонения, оцениваемые по объектам $S_i \in N(P^{\Omega, c, d}) \cap \tilde{K}_\lambda$.

5. НАХОЖДЕНИЕ ДИНАМИЧЕСКИХ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ ПО ДАННЫМ МНОГОМЕРНОГО ВРЕМЕННОГО РЯДА

Рассматривается многомерный временной ряд вида $y(t)$, $\mathbf{x}(t)$, $t = 1, 2, \dots, T, \dots$, где $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_n(t)) \in \mathbb{R}^n$ – признаковое описание рассматриваемого объекта S в момент t , а $y(t) \in \{1, 2, \dots, l\}$ – номер его состояния в данный момент времени. Факт $y(t) = j$ будем интерпретировать также как принадлежность объекта классу K_j в данный момент времени. Пусть в моменты времени $t \in [\theta, T]$, $1 < \theta \leq T$, объект принадлежит классу K_β , в то время как в моменты времени $\theta_0 \leq t < \theta$ объект принадлежал классу K_α , $\alpha \neq \beta$, $y(\theta_0 - 1) \neq \alpha$. Нас интересует, какие комбинации значений признаков свойственны классу K_α на протяжении некоторого интервала времени, но не свойственны K_β . Зная подобные комбинации признаков, можно их использовать для решения задач прогнозирования (вычисления значений $y(T+1)$, $y(T+2)$, $y(T+3)$, ...). Случай k -значных и бинарных признаков будем рассматривать как частный случай числовых.

Пусть

$$P_j^{c_j^1, c_j^2}(\mathbf{x}) = \begin{cases} 1 & \text{при } c_j^1 \leq x \leq c_j^2, \\ 0 & \text{иначе,} \end{cases} \quad P^{\Omega, c^1, c^2}(\mathbf{x}) = \bigwedge_{j \in \Omega} P_j^{c_j^1, c_j^2}(\mathbf{x}).$$

Определение 2. Предикат (1) назовем *динамической логической закономерностью* (ДЛЗ) класса K_α относительно K_β , если выполнено следующее:

$$1) \exists [\tau_1, \tau_2]: \forall \tau \in [\tau_1, \tau_2], \quad \theta_0 \leq \tau_1 \leq \tau_2 < \theta : P^{\Omega, c^1, c^2}(\mathbf{x}(\tau)) = 1,$$

$$2) \forall \tau \in [\theta, T] : P^{\Omega, c^1, c^2}(\mathbf{x}(\tau)) = 0.$$

Мощностью ДЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$ назовем длину $\tau_2 - \tau_1 + 1$ сегмента $[\tau_1, \tau_2]$. Отрезок $H(P^{\Omega, c^1, c^2}) = [\theta - \tau_2, \theta - \tau_1]$ назовем отрезком ожидания наступления класса K_β после выполнения ДЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$. Множество $N = \{\mathbf{x} : P^{\Omega, c^1, c^2}(\mathbf{x}) = 1\}$ назовем интервалом ДЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$. Таким образом, динамическая закономерность K_α относительно K_β имеет простую геометрическую интерпретацию. Ее интервал есть область пространства признаковых описаний вида $\{\mathbf{x} : c_i^1 \leq x_i \leq c_i^2, i \in \Omega\}$, содержащая некоторое число точек из класса K_α в последовательные моменты времени, но не содержащая ни одной точки из наблюдений последующего класса K_β .

Пусть фиксирован некоторый сегмент $[\tau_1, \tau_2]$.

ДЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$, $P^{\Omega, g^1, g^2}(\mathbf{x})$ назовем эквивалентными по $[\tau_1, \tau_2]$, если $P^{\Omega, c^1, c^2}(\mathbf{x}(\tau)) = P^{\Omega, g^1, g^2}(\mathbf{x}(\tau))$, $\tau = \tau_1, \tau_1 + 1, \dots, \tau_2, \theta + 1, \theta + 2, \dots, T$. Пусть $P^{\Omega^*, g^1, g^2}(\mathbf{x})$ – динамическая логическая закономерность K_α относительно K_β ($\Omega^* = \{1, 2, \dots, n\}$), тогда $P^{\Omega^*, c^1, c^2}(\mathbf{x})$,

$$c_i^1 = \min\{x_i(t), t = \tau_1, \tau_1 + 1, \dots, \tau_2\}, \quad c_i^2 = \max\{x_i(t), t = \tau_1, \tau_1 + 1, \dots, \tau_2\}, \quad i = 1, 2, \dots, n, \quad (26)$$

является также динамической логической закономерностью класса K_α относительно K_β , эквивалентной $P^{\Omega, g^1, g^2}(\mathbf{x})$ по $[\tau_1, \tau_2]$. Тогда для поиска ДЛЗ по сегменту $[\tau_1, \tau_2]$ можно ограничиться случаем (26) выбора коэффициентов $c_i^1, c_i^2, i = 1, 2, \dots, n$. Предикат $P^{\Omega^*, c^1, c^2}(\mathbf{x})$ удовлетворяет первому условию основного определения, поэтому достаточно проверить условия (2) определения 2: $\forall \tau \in [\theta, T] : P^{\Omega^*, c^1, c^2}(\mathbf{x}(\tau)) = 0$. Если условия 2) выполняются, то $P^{\Omega^*, c^1, c^2}(\mathbf{x})$ является ДЛЗ и на ее базе можно найти эквивалентную ей ДЛЗ, содержащую минимальное число признаков.

Рассмотрим следующую задачу:

$$\sum_{i=1}^n y_i \longrightarrow \min,$$

$$\sum_{i=1}^n [1 - P_i^{c_i^1, c_i^2}(x_i(t))] y_i \geq 1, \quad t = \theta + 1, \theta + 2, \dots, T, \quad (27)$$

$$y_i \in \{0, 1\}, \quad i = 1, 2, \dots, n.$$

Пусть $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_n^*)$ – решение задачи (27), а $\Omega \subseteq \{1, 2, \dots, n\}$ – множество номеров всех единичных компонент решения. Тогда $P^{\Omega, c^1, c^2}(\mathbf{x})$ является также ДЛЗ, эквивалентной ДЛЗ $P^{\Omega^*, c^1, c^2}(\mathbf{x})$.

Задача поиска множества всех ДЛЗ K_α относительно K_β может быть решена с помощью следующего алгоритма.

Рассматриваются всевозможные пары чисел $\tau_1, \tau_2, \theta_0 \leq \tau_1 < \theta, \theta_0 \leq \tau_2 < \theta, \tau_1 < \tau_2$ и решаются задачи (25) поиска ДНЗ по соответствующим им отрезкам $[\tau_1, \tau_2]$. Заметим, что если для отрезка $[\tau_1, \tau_2]$ не существует ДЛЗ, то ДНЗ не существуют и для всех отрезков $[\tau_1, t], \tau_2 < t \leq \theta$. Перебор отрезков осуществляем в следующем порядке: $[\theta_0, \theta_0 + 1], [\theta_0, \theta_0 + 2], \dots, [\theta_0, \theta - 1], [\theta_0 + 1, \theta_0 + 2], [\theta_0 + 1, \theta_0 + 3], \dots$. Общее число частных задач (26) будет не более $\sigma(\sigma - 1)/2, \sigma = \theta - \theta_0$. Для найденных ДЛЗ с нерасширяемой мощностью решаются задачи (26). Таким образом, в итоге формируется множество $\mathcal{P}_{\alpha\beta}$ ДЛЗ класса K_α относительно K_β . По заданному временному ряду рассматриваются слева направо все возможные соседние классы $\langle K_\alpha, K_\beta \rangle, \alpha, \beta \in \{1, 2, \dots, l\}, \alpha \neq \beta$, и вычисляются соответствующие множества $\mathcal{P}_{\alpha\beta}$. Если пара $\langle K_\alpha, K_\beta \rangle$ встречалась ранее и повторяется снова, то множество $\mathcal{P}_{\alpha\beta}$ пополняется новыми найденными ДЛЗ.

Пусть вычислены все $\mathcal{P}_{\alpha\beta}, \alpha, \beta = 1, 2, \dots, l$ (некоторые из них могут быть пустыми). Задача прогноза в момент T последующего класса объекта S , находящегося в момент T в классе K_α , а также времени его перехода в другой класс может быть решена следующим образом.

Вычисляются величины

$$\Gamma_\beta(S) = \frac{1}{|\mathcal{P}_{\alpha\beta}|} \sum_{P^{\Omega, c^1, c^2} \in \mathcal{P}_{\alpha\beta}} P^{\Omega, c^1, c^2}(\mathbf{x}(T))$$

(полагаем $\Gamma_\beta(S) = 0$, если $\mathcal{P}_{\alpha\beta} = \emptyset$). Прогнозируемым классом будет K_β , если $\Gamma_\beta(S) = \max_\alpha \Gamma_\alpha(S)$.

Для оценки прогнозируемого момента времени τ^* наступления K_β можно использовать информацию об отрезках ожидания наступления класса K_β после выполнения соответствующих ДЛЗ $P^{\Omega, c^1, c^2}(\mathbf{x})$ из $\mathcal{P}_{\alpha\beta}$. Например,

$$\tau^* = \frac{1}{|\hat{\mathcal{P}}_{\alpha\beta}|} \sum_{P^{\Omega, c^1, c^2} \in \hat{\mathcal{P}}_{\alpha\beta}} (2\theta - \tau_2 - \tau_1)/2, \quad \text{где} \quad \hat{\mathcal{P}}_{\alpha\beta} = \{P^{\Omega, c^1, c^2}(\mathbf{x}) : P^{\Omega, c^1, c^2} \in \mathcal{P}_{\alpha\beta}, P^{\Omega, c^1, c^2}(\mathbf{x}(T)) = 1\}.$$

СПИСОК ЛИТЕРАТУРЫ

1. Дмитриев А.Н., Журавлёв Ю.И., Кренделев Ф.П. О математических принципах классификации предметов и явлений // Дискретный анализ. Новосибирск, ИМ СО АН СССР, 1966. Вып. 7. С. 3–11.
2. Журавлёв Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. № 3. С. 1–11.
3. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. М.: Наука, 1978. Вып. 33. С. 5–68.
4. Вайнцивайг М.Н. Алгоритм обучения распознаванию образов “Кора” // Алгоритмы обучения распознаванию образов. М.: Сов. радио, 1973. С. 8–12.
5. Баскакова Л.В., Журавлёв Ю.И. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. 1981. Т. 21. № 5. С. 1264–1275.
6. Журавлёв Ю.И. Избранные научные труды. М.: Изд-во Магистр., 1998.
7. Журавлёв Ю.И. Об алгоритмах распознавания с представительными наборами (о логических алгоритмах) // Ж. вычисл. матем. и матем. физ. 2002. Т. 42. № 9. С. 1425–1435.
8. Дюкова Е.В., Журавлёв Ю.И. Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. 2000. Т. 40. № 8. С. 1264–1278.
9. Дюкова Е.В. Алгоритмы распознавания типа “Кора”: сложность реализации и метрические свойства // Распознавание, классификация, прогноз: Матем. методы и их применение. М.: Наука, 1989. Вып. 2. С. 99–125.
10. Кочетков Д.В. Распознающие алгоритмы, инвариантные относительно преобразований пространства признаков. 1 // Распознавание, классификация, прогноз: Матем. методы и их применение. М.: Наука, 1989. Вып. 1. С. 82–113; 2 // Вып. 2. С. 178–206; 3 // 1992. Вып. 3. С. 178–206, 64–88.
11. Рязанов В.В., Сенько О.В. О некоторых моделях голосования и методах их оптимизации // Распознавание, классификация, прогноз: Матем. методы и их применение. М.: Наука, 1992. Вып. 3. С. 106–145.
12. Ryazanov V.V. Recognition algorithms based on local optimality criteria // Pattern Recognition and Image Analys. 1994. V. 4. № 2. P. 98–109.
13. Lairn S.B., Ryazanov V.V. The search of precedent-based logical regularities for recognition and data analysis problems // Pattern Recognition and Image Analys. 1997. V. 7. № 3. P. 322–333.
14. Журавлёв Ю.И., Рязанов В.В. Об извлечении знаний из выборок прецедентов в моделях классификации, основанных на принципе частичной прецедентности // Докл. IX Междунар. конф. “Знание–диалог–решение”. Санкт-Петербург, 18–23 июня 2001. С. 232–237.
15. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006.
16. Burges Ch.J.C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. 1998. Vol. 2. P. 121–167.
17. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

Сдано в набор 09.06.2007 г.

Подписано к печати 17.08.2007 г.

Формат бумаги 60 × 88¹/₈

Цифровая печать

Усл. печ. л. 20.0

Усл. кр.-отт. 7.4 тыс.

Уч.-изд. л. 20.0

Бум. л. 10.0

Тираж 362 экз.

Зак. 577

Учредители: Российская академия наук, Вычислительный центр им. А.А. Дородницына РАН

Издатель: Академиздатцентр “Наука”, 117997 Москва, Профсоюзная ул., 90

Оригинал-макет подготовлен МАИК “Наука/Интерпериодика”

Отпечатано в ППП “Типография “Наука”, 121099 Москва, Шубинский пер., 6