ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1085–1092. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © F.K. Akhmadishina, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1139–1147.

Sufficient Conditions for the Asymptotic Optimality of Projection Methods as Applied to Operator Equations

F. K. Akhmadishina

Institute of Mathematics and Mechanics, Kazan State University, ul. Universitetskaya 17, Kazan, 420008 Russia
e-mail: povch@yandex.ru

Received June 3, 2008; in final form, November 18, 2008

Abstract—Sufficient conditions are found for the asymptotic optimality of projection methods as applied to linear operator equations in Hilbert spaces. The conditions are applicable to a wide class of equations when asymptotically optimal projection methods are sought for their solution. Applications illustrating the result are presented.

DOI: 10.1134/S096554250907001X

Key words: approximate methods for solving operator equations, asymptotically optimal projection method, projection width, extreme subspace.

1. INTRODUCTION AND BASIC DEFINITIONS

Numerous results available on approximate solutions for various classes of operator equations have recently motivated the creation and development of the optimization theory of computational methods, which is related to the construction and study of the most accurate solution techniques. In this paper, we follow the approach proposed in [1, 2], which has been extensively used to substantiate (basically, the order of) direct methods for operator equations. In Section 3, we establish sufficient conditions for the asymptotic optimality of projection methods. The application of Theorem 3 to the determination of asymptotically optimal projection methods is illustrated in Section 4.

Let X and Y be normed spaces over $\mathbb{C}$ or $\mathbb{R}$, I_X be the identity operator in X, and $K: X \longrightarrow Y$ be a linear bijection. To solve the equation

$$Kx = y, \quad x \in X, \quad y \in Y, \tag{1.1}$$

by a projection method, we choose two sequences of subspaces X_n and Y_n:

$$X_n \subset X, \quad Y_n \subset Y, \quad n = 1, 2, \ldots,$$

and projectors P_n projecting Y onto Y_n. Equation (1.1) is approximated by

$$P_nKx_n = P_ny, \quad x_n \in X_n. \tag{1.2}$$

Naturally, we consider only projection methods such that Eqs. (1.2) are uniquely solvable with any $n \geq n_0$ and $y \in Y$. The solutions to Eqs. (1.1) and (1.2) are denoted by x^* $(= x^*(y))$ and x_n^* $(= x_n^*(y))$, respectively. Let $E_n(x)$ $(= E_{X_n}(x))$ denote the best approximation of an element $x \in X$ by the subspace X_n; i.e.,

$$E_n(x) = \inf_{x_n \in X_n} \|x - x_n\|.$$

The problem arises of finding, among all the uniquely solvable projection methods, those for which the exact solution x^* deviates least from the approximate solution x_n^* on a given solution set $F(\subset X)$. To characterize this deviation, the nth *optimal error estimate* for F among all the projection methods for solving Eq. (1.1) was introduced in [1, 2]:

$$V_n(F) = \inf_{X_m,\, Y_m,\, P_m} \sup_{x^* \in F} \|x^* - x_m^*\|,$$

where inf is taken over all possible subspaces X_m and Y_m of dimension $m \leq n$ and over projectors P_m from Y onto Y_m such that Eq. (1.2) is uniquely solvable with any $y \in Y$.

Definition 1. Projection method (1.2) for solving Eq. (1.1) is called asymptotically optimal for a solution set F if

$$\sup_{x^* \in F} \|x^* - x_n^*\| \sim \sup_{x^* \in F} E_n(x^*), \quad n \longrightarrow \infty.$$

The following definition was introduced by Gabdulkhaev (see, e.g., [1]).

Definition 2. Projection method (1.2) for solving Eq. (1.1) is called asymptotically optimal among all the projection methods for a solution set F if

$$\sup_{x^* \in F} \|x^* - x_n^*\| \sim V_n(F), \quad n \longrightarrow \infty.$$

The following standard notation is used below. Given a linear operator A, the symbols $\operatorname{Im} A$, $\operatorname{Ker} A$, and D_A stands for the image, kernel, and domain of A, respectively.

Recall the definition of a projection width (see [3]).

Definition 3. For a centrally symmetric set F in a normed space X, the nth projection width in X is defined as

$$\pi_n(F) = \inf_{X_m,\, R_m} \sup_{x \in F} \|x - R_m x\|,$$

where inf is taken over all m-dimensional subspaces X_m of X such that $m \le n$ and over all projectors R_m in X such that $\operatorname{Im} R_m = X_m$.

A subspace X_k^0 $(k \le n)$ is called an extreme subspace for the projection width $\pi_n(F)$ if the above infimum is reached on it; i.e.,

$$\pi_n(F) = \inf_{R_k,\, \operatorname{Im} R_k = X_k^0} \sup_{x \in F} \|x - R_k x\|,$$

where inf is taken over all projectors R_k in X such that $\operatorname{Im} R_k = X_k^0$.

For any $F \subset X$ and a subspace $Z_n \subset X$ $(n \in \mathbb{N})$, define

$$\Delta(F, Z_n) = \sup_{x \in F} E_{Z_n}(x).$$

Remark 1. Obviously, if X is a Hilbert space, then

$$\pi_n(F) = \inf_{X_n} \sup_{x \in F} E_n(x),$$

where inf is taken over all the n-dimensional subspaces X_n of X. Moreover, if X_n^0 is an extreme subspace for the projection width $\pi_n(F)$, where F is a centrally symmetric set, then

$$\pi_n(F) = \sup_{x \in F} E_{X_n^0}(x).$$

2. AUXILIARY RESULTS

Lemma 1. *Let $n \in \mathbb{N}$ and*

$$KX_n \cap \operatorname{Ker} P_n = \{0\}. \tag{2.1}$$

Then there exists a projector R_n^K in X such that, for any $y \in Y$, the solution to Eq. (1.2) *can be represented as*

$$x_n^* = R_n^K x^*.$$

Moreover, $\operatorname{Im} R_n^k = X_n$ *and* $\operatorname{Ker} R_n^K = K^{-1}(\operatorname{Ker} P_n)$.

Proof. Denote by P_n' the restriction of a projector P_n to the subspace KX_n. By virtue of (2.1), P_n' is a bijection of KX_n to Y_n. Therefore, there exists an inverse operator $P_n'^{-1} : Y_n \longrightarrow KX_n$. Then the solution to Eq. (1.2) can be represented as

$$x_n^* = K^{-1} P_n'^{-1} P_n y = K^{-1} P_n'^{-1} P_n K x^*.$$

Let us show that the operator $P_n'^{-1} P_n$ is a projector in Y. Indeed, let $y \in Y$. Then there exists an element $x_n \in X_n$ such that $P_n y = P_n K x_n$. From this and the equality $P_n'^{-1} P_n K x_n = K x_n$, we have

$$(P_n'^{-1}P_n)^2 y = P_n'^{-1}P_n(P_n'^{-1}P_n K x_n) = P_n'^{-1}P_n K x_n = P_n'^{-1}P_n y.$$

Since $\operatorname{Im} P_n'^{-1}P_n = KX_n$ and $\operatorname{Ker} P_n'^{-1}P_n = \operatorname{Ker} P_n$, we conclude that $K^{-1}P_n'^{-1}P_n K$ is a projector in X with the range X_n and the kernel $K^{-1}(\operatorname{Ker} P_n)$. Setting $R_n^K = K^{-1}P_n'^{-1}P_n K$ yields the assertion of the lemma.

Corollary 1. Under the conditions of Lemma 1, we have the error estimate

$$\|x^* - x_n^*\| \le \|I_X - R_n^K\| E_n(x^*). \tag{2.2}$$

Remark 2. Condition (2.1) is necessary and sufficient for Eq. (1.2) to be uniquely solvable with any $y \in Y$. (Necessity is obvious, while sufficiency is proved in Lemma 1.)

Lemma 2. *Let X be a unitary space, and let P be a projector in X such that $P \ne 0$ and $P \ne I_X$. Then*

$$\|P\| = (1 - \alpha^2)^{-1/2},$$

where $\alpha = \sup\{|(x, y)| : x \in \operatorname{Ker} P, y \in \operatorname{Im} P, \|x\| = \|y\| = 1\}$. (*If P is an unbounded projector, we set* $\|P\| = +\infty$.)

Proof. First, we show that, if $x, y \in X$ and $\|x\| = \|y\| = 1$, then any scalar λ satisfies the inequality

$$\|\lambda x + y\|^2 \ge 1 - |(x, y)|^2.$$

Indeed,

$$\|\lambda x + y\|^2 + |(x, y)|^2 - 1 = |\lambda|^2 + 2\operatorname{Re}(\lambda x, y) + 1 + |(x, y)|^2 - 1$$

$$= |\lambda|^2 - [-2\operatorname{Re}(\lambda x, y)] + |(x, y)|^2 \ge |\lambda|^2 - 2|(\lambda x, y)| + |(x, y)|^2 = [|\lambda| - |(x, y)|]^2 \ge 0.$$

Now, we prove the lemma. Let $z \in X$ and $\|z\| = 1$. If $z \in \operatorname{Ker} P$, then

$$\|Pz\| = 0 \le (1 - \alpha^2)^{-1/2}.$$

If $z \in \operatorname{Im} P$, then

$$\|Pz\| = \|z\| = 1 \le (1 - \alpha^2)^{-1/2}.$$

Let $z \notin \operatorname{Im} P$ and $z \notin \operatorname{Ker} P$. Define

$$x = \frac{z - Pz}{\|z - Pz\|}, \quad y = \frac{Pz}{\|Pz\|}, \quad \lambda = \frac{\|z - Pz\|}{\|Pz\|}.$$

Then $x \in \operatorname{Ker} P$, $y \in \operatorname{Im} P$, and $\|x\| = \|y\| = 1$. Furthermore,

$$\|Pz\|^2[1 - |(x, y)|^2] \le \|Pz\|^2\|\lambda x + y\|^2 = \|z\|^2 = 1,$$

which yields

$$\|Pz\| \le [1 - |(x, y)|^2]^{-1/2} \le (1 - \alpha^2)^{-1/2}.$$

Thus,

$$\|P\| = (1 - \alpha^2)^{-1/2}.$$

Let us prove the reverse inequality. Suppose that $x \in \operatorname{Ker} P$, $y \in \operatorname{Im} P$, and $\|x\| = \|y\| = 1$. Let $z = y - (y, x)x$. We have

$$\|z\|^2 = 1 - |(x, y)|^2 - |(x, y)|^2 + |(x, y)|^2 = 1 - |(x, y)|^2.$$

Thus,

$$\|P\| \ge \left\|P\frac{z}{\|z\|}\right\| = \frac{\|Pz\|}{\|z\|} = \frac{1}{\sqrt{1 - |(x, y)|^2}}.$$

Since x and y are arbitrary,

$$\|P\| \ge (1 - \alpha^2)^{-1/2}.$$

Corollary 2. Under the conditions of Lemma 2,

$$\|P\| = \|I_X - P\|.$$

Proof. This assertion follows from Lemma 2 and the equalities $\operatorname{Im} P = \operatorname{Ker}(I_X - P)$ and $\operatorname{Ker} P = \operatorname{Im}(I_X - P)$.

Lemma 3. *Let X be a unitary space, $n \in \mathbb{N}$, condition* (2.1) *be satisfied. Then the solutions x^* and x_n^* to Eqs.* (1.1) *and* (1.2) *satisfy the error estimate*

$$\|x^* - x_n^*\| \le (1 - \alpha_n^2)^{-\frac{1}{2}} E_n(x^*),$$

where $\alpha_n = \sup\{|(x, K^{-1}y)| : x \in X_n, y \in \operatorname{Ker} P_n, \|x\| = \|K^{-1}y\| = 1\}$.

Proof. Lemma 1 implies that $(I_X - R_n^K)$ is a projector with the range $K^{-1}(\operatorname{Ker} P_n)$ and the kernel X_n. The assertion of the lemma follows from this, Lemma 2, and (2.2).

Lemma 4. *Let X be a unitary space; $n \in \mathbb{N}$; X_n and Y_n* be *n-dimensional subspaces of X; and X_n^1 and Y_n^1 be the unit spheres in X_n and Y_n, respectively. Then*

$$\Delta(X_n^1, Y_n) = \Delta(Y_n^1, X_n).$$

Proof. Let $e_1, \ldots, e_n$ be an orthonormal basis in X_n; $g_1, \ldots, g_n$ be an orthonormal basis in Y_n; Q_n and R_n be the orthoprojectors onto X_n and Y_n, respectively; $x = x^1 e_1 + \ldots + xT^n e_n$ $(\in X_n^1)$; and $y = y^1 g_1 + \ldots + yT^n g_n$ $(\in Y_n^1)$. Then

$$\Delta(X_n^1, Y_n) = \sup_{x \in X_n^1} (1 - \|R_n x\|) = 1 - \inf_{x \in X_n^1} \|R_n x\|.$$

Similarly,

$$\Delta(Y_n^1, X_n) = 1 - \inf_{y \in Y_n^1} \|Q_n y\|.$$

Let us prove that

$$\inf_{x \in X_n^1} \|R_n x\| = \inf_{y \in Y_n^1} \|Q_n y\|,$$

which implies the lemma.

In the Euclidean space V^n ($V = \mathbb{C}$ or $\mathbb{R}$), consider a linear operator A with the matrix

$$T = \begin{pmatrix} (e_1, g_1) & \ldots & (e_n, g_1) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ (e_1, g_n) & \ldots & (e_n, g_n) \end{pmatrix}, \quad \text{and let} \quad u = \begin{pmatrix} x^1 \\ \ldots \\ x^n \end{pmatrix}, \quad v = \begin{pmatrix} y^1 \\ \ldots \\ y^n \end{pmatrix}.$$

We have

$$\|R_n x\|^2 = \sum_{j=1}^n |(x, g_j)|^2 = \sum_{j=1}^n \left| \sum_{i=1}^n x^i (e_i, g_j) \right|^2 = \sum_{j=1}^n |(Tu)^j|^2 = \|Au\|^2,$$

where $(Tu)^j$ is the jth coordinate of the vector Tu. Therefore,

$$\inf_{x \in X_n^1} \|R_n x\| = \inf_{u \in V^n, \|u\| = 1} \|Au\|.$$

Moreover,

$$\|Q_n y\|^2 = \sum_{j=1}^n |(y, e_j)|^2 = \sum_{j=1}^n \left| \sum_{i=1}^n y^i \overline{(e_j, g_i)} \right|^2 = \sum_{j=1}^n |(T^* v)^j|^2 = \|A^* v\|^2.$$

It follows that

$$\inf_{y \in Y_n^1} \|Q_n y\| = \inf_{v \in V_n, \|v\| = 1} \|A^* v\|.$$

It remains to prove that

$$\inf_{v \in V^n, \|v\| = 1} \|Av\| = \inf_{v \in V^n, \|v\| = 1} \|A^* v\|.$$

For this purpose, we consider the polar decomposition of A:

$$A = |A| U,$$

where $|A| = \sqrt{A^*A}$ and U is a unitary operator. We have

$$\inf_{v \in V^n, \|v\| = 1} \|Av\| = \inf_{v \in V^n, \|v\| = 1} \||A|Uv\| = \inf_{v \in V^n, \|v\| = 1} \||A|v\| = \inf_{v \in V^n, \|v\| = 1} \|U^*|A|v\| = \inf_{v \in V^n, \|v\| = 1} \|A^*v\|.$$

3. THE MAIN RESULTS

Theorem 1. *For any centrally symmetric solution set F ($\subset X$) of Eq.* (1.1), *we have*

$$V_n(F) = \pi_n(F).$$

Proof. Let us show that $\pi_n(F) \le V_n(F)$. Suppose that $m \le n$; X_m and Y_m are arbitrary m-dimensional subspaces of X and Y, respectively; and P_m is an arbitrary linear surjection from Y to Y_m such that Eq. (1.2) is uniquely solvable with any $y \in Y$. By Lemma 1, there exists a projector R_m^K in X with the range X_m such that $x_m^* = R_m^K x^*$ for any $x^* \in F$. It follows that $\pi_n(F) \le V_n(F)$.

Let us prove the reverse inequality. Suppose that $m \le n$, X_m is an arbitrary m-dimensional subspace of X, and R_m is a projector in X with the range X_m. Define $P_m = KR_mK^{-1}$. Obviously, P_m is a projector in Y such that $\operatorname{Im} P_m = KX_m$ and $\operatorname{Ker} P_m = K(\operatorname{Ker} R_m)$. Consider the projection method specified by the subspace X_m and the operator P_m. Since $KX_m \cap \operatorname{Ker} P_m = \{0\}$, Lemma 1 implies that $x_m^* = R_m^K x^*$ for any solution x^*, where R_m^K is a projector in X such that $\operatorname{Im} R_m^K = X_m$ and $\operatorname{Ker} R_m^K = K^{-1}(\operatorname{Ker} P_m)$. Since $\operatorname{Ker} R_m = K^{-1}(\operatorname{Ker} P_m)$, we conclude that $R_m^K = R_m$ and $x_m^* = R_m x^*$ for any $x^* \in X$. Therefore, $V_n(F) \le \pi_n(F)$.

In the following two theorems, H is a Hilbert space and the sequence of subspaces X_n is assumed to be asymptotically dense in H (i.e., $E_n(x) \longrightarrow 0$ for any $x \in H$) and increasing ($X_1 \subset X_2 \subset \ldots$).

Theorem 2. *Let K be a dense* (in H) *linear injective operator with the range H, $X_n \subset D_K$, $Y_n \subset D_{K^*}$, P_n be the orthoprojector onto the subspace Y_n, and S_n be the unit sphere in $K^*(Y_n)$. If $\Delta(S_n, X_n) < 1$, then, for any $y \in Y$, Eq.* (1.2) *is uniquely solvable and the solution satisfies the estimate*

$$\|x^* - x_n^*\| \le \frac{1}{\sqrt{1 - \Delta^2(S_n, X_n)}} E_n(x^*). \tag{3.1}$$

Specifically, if $\Delta(S_n, X_n) \longrightarrow 0$, $n \longrightarrow \infty$, then, for any centrally symmetric solution set F, projection method (1.2) is asymptotically optimal. If, additionally, X_n is an extreme subspace for the projection width $\pi_n(F)$, $n \ge n_0$, then projection method (1.2) is asymptotically optimal for F among all the projection methods for solving Eq. (1.1) and

$$\sup_{x^* \in F} \|x^* - x_n^*\| \sim \pi_n(F), \quad n \longrightarrow \infty. \tag{3.2}$$

Proof. Let $X = D_K$ and $Y = H$. Recall that X_n^1 denotes the unit sphere in X_n. In Lemma 3, we introduced

$$\alpha_n = \sup\{|(x, K^{-1}y)| : x \in X_n, y \in \operatorname{Ker} P_n, \|x\| = \|K^{-1}y\| = 1\}.$$

Let us prove that

$$\alpha_n \le \Delta(S_n, X_n). \tag{3.3}$$

Let R_n be the orthoprojector onto $K^*(Y_n)$. Then $I_H - R_n$ is the orthoprojector onto $(K^*(Y_n))^\perp$. Since $K^{-1}(\operatorname{Ker} P_n) \subset (K^*(Y_n))^\perp$, we conclude that, for any $x \in X_n$, $y \in \operatorname{Ker} P_n$,

$$(x, K^{-1}y) = (x, (I_H - R_n)K^{-1}y) = ((I_H - R_n)x, K^{-1}y),$$

which yields

$$\sup_{y \in \operatorname{Ker} P_n, \|K^{-1}y\| = 1} |(x, K^{-1}y)| \le \|(I_H - R_n)x\|.$$

Then we have

$$\alpha_n \le \sup_{x \in X_n, \|x\| = 1} \|(I_H - R_n)x\| = \Delta(X_n^1, K^*(Y_n)) = \Delta(S_n, X_n),$$

where the last equality holds by Lemma 4.

Let $\Delta(S_n, X_n) < 1$. We show that $K(X_n) \cap \operatorname{Ker} P_n = \{0\}$, which implies the unique solvability of Eq. (1.2). Assume the opposite. Let there exist a nonzero element $z \in K(X_n) \cap \operatorname{Ker} P_n$. Setting $x = K^{-1}z/\|K^{-1}z\|$ and $y = z/\|K^{-1}z\|$, we obtain $\alpha_n = 1$, from which, by virtue of (3.3), we have $\Delta(S_n, X_n) \geq 1$, a contradiction.

Inequality (3.3) and Lemma 3 imply (3.1). Combining (3.1) with $E_n(x^*) \leq \|x^* - x_n^*\|$, $x^* \in F$, $x_n^* \in X_n$ gives

$$\sup_{x^* \in F} E_n(x^*) \leq \sup_{x^* \in F} \|x^* - x_n^*\| \leq \frac{1}{\sqrt{1 - \Delta^2(S_n, X_n)}} \sup_{x^* \in F} E_n(x^*).$$

If $\Delta(S_n, X_n) \longrightarrow 0$, $n \longrightarrow \infty$, this relation implies that

$$\sup_{x^* \in F} \|x^* - x_n^*\| \sim \sup_{x^* \in F} E_n(x^*). \tag{3.4}$$

Thus, we have proved the asymptotic optimality of projection method (1.2) for a solution set F. If, additionally, X_n is an extreme subspace for $\pi_n(F)$, $n \geq n_0$, then (3.4) and Remark 1 imply (3.2), while (3.2) and Theorem 1 imply the asymptotic optimality of projection method (1.2) for F among all the projection methods for solving Eq. (1.1).

Theorem 3. *Consider the equation*

$$Kx \equiv Ax + Bx = y, \quad y \in H, \tag{3.5}$$

where A and B are dense ($D_A \subset D_B$) linear operators in H such that A is an invertible operator, BA^{-1} is a compact operator, and $A^{-1}B$ can be extended to a compact operator T in H. Let $X_n \subset D_A$, $Y_n \subset D_{A^*}$, $n \geq n_0$, P_n be the orthoprojector onto Y_n, and $A^*(Y_n) = X_n$. If the homogeneous equation

$$Ax + Bx = 0$$

has only the trivial solution, then projection method (1.2) *for solving Eq.* (3.5) *is asymptotically optimal for any centrally symmetric solution set F. If, additionally, X_n is an extreme subspace for the projection width $\pi_n(F)$, $n \geq n_0$, then projection method* (1.2) *is asymptotically optimal for F among all the projection methods for solving Eq.* (3.5); *moreover*,

$$\sup_{x^* \in F} \|x^* - x_n^*\| \sim \pi_n(F), \quad n \longrightarrow \infty.$$

Proof. Since $A + B$ is injective, $I_H + BA^{-1}$ is injective as well. Therefore, since BA^{-1} is compact, $I_H + BA^{-1}$ is invertible. Therefore, K $(= (I_H + BA^{-1})A))$ is invertible as well. Specifically, $\operatorname{Im} K = H$ implies the injectivity of K^*. Let us prove the injectivity of $I_H + T^*$. Since $A^{-1}B \subset T$ and A^{-1} is bounded, $T^* = B^*(A^*)^{-1}$. The equality $(A^{-1})^* = (A^*)^{-1}$ implies $\operatorname{Im} A^* = H$. If $I_H + T^*$ is not injective, then there exists $x \neq 0$ such that $x + B^*(A^*)^{-1}x = 0$. Since $x = A^*y$ for some $y \neq 0$, we have $A^*y + B^*y = 0$, which contradicts the injectivity of K^*. Thus, there exists a continuous inverse $(I_H + T^*)^{-1}$. Define $M = \|(I_H + T^*)^{-1}\|$. Let S_n be the unit sphere in $K^*(Y_n)$. We prove that $\Delta(S_n, X_n) \longrightarrow 0$, $n \longrightarrow \infty$. Then the assertion of the theorem follows from Theorem 1. Let $y \in Y_n$, Q_n be the orthoprojector onto X_n, and S be the unit sphere in H. We have

$$\Delta(S_n, X_n) = \sup_{y \in Y_n,\, \|K^*y\| = 1} \|(I_H - Q_n)K^*y\|$$

$$= \sup_{y \in Y_n,\, \|K^*y\| = 1} \|(I_H - Q_n)(I_H + T^*)A^*y\| = \sup_{y \in Y_n,\, \|K^*y\| = 1} \|(I_H - Q_n)T^*A^*y\|$$

$$\leq M \sup_{y \in Y_n,\, \|A^*y\| = 1} \|(I_H - Q_n)T^*A^*y\| \leq M \sup_{z \in \overline{T^*(S)}} \|(I_H - Q_n)z\|.$$

Consider the sequence of continuous functions $f_n(z) = \|(I - Q_n)z\|$ defined on a compact set $\overline{T^*(S)}$. Since this sequence is nonincreasing and pointwise converging to zero, Dini's theorem implies that

$$\sup_{z \in \overline{T^*(S)}} \|(I_H - Q_n)z\| \longrightarrow 0, \quad n \longrightarrow \infty.$$

Hence, $\Delta(S_n, X_n) \longrightarrow 0$, $n \longrightarrow \infty$. The theorem is proved.

4. APPLICATIONS

Below, we follow the notation used in [3, 4]. For $r = 1, 2, \ldots$, let $L_2^r[a, b]$ denote the set of real-valued functions $x(t)$ defined on $[a, b]$ such that $x^{(r-1)}$ is absolutely continuous and $x^{(r)} \in L_2[a, b]$; $W_2^r[a, b] = \{x(t) : x \in L_2^r[a, b], \|x^{(r)}\|_2 \le 1\}$; $W_2^r = \{x(t) : x \in W_2^r[0, 2\pi], x^{(j)}(0) = x^{(j)}(2\pi), j = 0, 1, \ldots, r-1\}$; $H^\omega[a, b] = \{f(t) : f \in C[a, b], \omega(f, \delta)_{C[a, b]} \le \omega(\delta), 0 \le \delta \le b - a\}$, where $\omega(f, \delta)_{C[a, b]}$ is the modulus of continuity of f in $C[a, b]$ and $\omega(\delta)$ is a given modulus of continuity; $S_{n,0}[a, b]$ is the subspace of piecewise constant functions on the uniform partition $t_k = a + kh_n$, where $h_n = (b - a)/n$ and $k = 0, 1, \ldots, n$, $n \in \mathbb{N}$; and

$$A_n(\omega) = \frac{1}{2}\left[n\int_0^{(b-a)/n}\omega^2(t)dt\right]^{1/2}, \quad n \in \mathbb{N}.$$

Define

$$W_{2,0}^r = \{x(t) : x \in W_2^r, x(0) = 0\}.$$

Application 1. Consider the Fredholm equation of the second kind

$$Kx \equiv x + Bx = y, \tag{4.1}$$

where the solution x is sought in $L_2[a, b]$, $y \in L_2[a, b]$, and B is a completely continuous operator in $L_2[a, b]$.

Let $X_n = Y_n = S_{n,0}[a, b]$ and P_n be the orthoprojector onto Y_n. If the homogeneous equation corresponding to Eq. (4.1) has only the trivial solution, then, for the solution set $F = H^\omega[a, b]$, projection method (1.2) is asymptotically optimal among all the projection methods for solving Eq. (4.1) and we have the estimate

$$\sup_{x^* \in H^\omega[a, b]} \|x^* - x_n^*\| \sim A_n(\omega), \quad n \longrightarrow \infty.$$

Indeed, let $H = L_2[a, b]$ and $A = I_H$. It is well known (see [4, p. 371]) that $\pi_n(H^\omega[a, b]) = A_n(\omega)$ and $S_{n,0}[a, b]$ is an extreme subspace for $\pi_n(H^\omega[a, b])$. Now, the required assertion follows from Theorem 3.

Application 2. Consider a strongly singular integral equation of the first kind

$$Kx \equiv \frac{1}{\pi}\int_{-1}^{1}\frac{\sqrt{1-s^2}}{s-t}x(s)ds + Bx(t) = y(t), \tag{4.2}$$

where the integral is understood in the sense of the Hadamard finite part, x is sought in $L_{2,p}[-1, 1]$, $p(t) = \sqrt{1-t^2}$, $y(t) \in L_{2,p}[-1, 1]$, and $B : L_{2,p}[-1, 1] \longrightarrow L_{2,p}[-1, 1]$ is a continuous linear operator such that the homogeneous equation corresponding to Eq. (4.2) has only the trivial solution.

Let X_n be the subspace of algebraic polynomials of degrees no higher than $n - 1$ given on the interval $[-1, 1]$; $Y_n = X_n$; and P_n be the orthoprojector onto Y_n. Then, for any centrally symmetric solution set F, projection method (1.2) for solving Eq. (4.2) is asymptotically optimal.

Indeed, define

$$Ax = \frac{1}{\pi}\int_{-1}^{1}\frac{\sqrt{1-s^2}}{s-t}x(s)ds, \quad x \in L_{2,p}[-1, 1].$$

Let $\{T_k(t)_{k=0}^\infty\}$ be the system of Chebyshev polynomials of the first kind and $c_k(x) = (x, T_k)$. It is well known (see [2, p. 106]) that A^{-1} is a compact operator and

$$Ax(t) = \sum_{k=1}^{\infty} kc_{k-1}(x)T_{k-1}(t).$$

This representation implies that $A^*(X_n) = X_n$. Now, the required assertion follows from Theorem 3.

Application 3. Consider the problem

$$Kx \equiv x'(t) + (Bx)(t) = y(t) \tag{4.3}$$

with the boundary condition

$$x(0) = 0,$$

where $D_K = \{x \in L_2^1[0, 2\pi] : x(0) = 0\}$, $y \in L_2[0, 2\pi]$, and B is a linear continuous operator in $L_2[0, 2\pi]$ such that the homogeneous equation corresponding to Eq. (4.3) has only the trivial solution.

Let X_{2n} be the linear span of the system of functions $\sin t$, $\cos t - 1$, ..., $\sin(nt)$, $\cos(nt) - 1$ defined on $[0, 2\pi]$; Y_{2n} be the linear span of the system of functions $\cos t$, $t - \sin t$, ..., $\cos(nt)$, $t - [\sin(nt)]/n$ defined on $[0, 2\pi]$; and P_{2n} be the orthoprojector onto Y_{2n}. Then, for the solution set $F = W_{2,0}^r$, projection method (1.2) is asymptotically optimal among all the projection methods for solving problem (4.3) and we have the estimate

$$V_{2n} \sim n^{-r}.$$

Indeed, let $H = L_2[0, 2\pi]$. Define the operator $A : D_K \longrightarrow L_2[0, 2\pi]$ by setting $A(x(t)) = x'(t)$. Obviously, A^{-1} is a compact operator in H and $A^*(x(t)) = -x'(t)$, whence $A^*(Y_{2n}) = X_{2n}$. Furthermore, $\pi_{2n}(W_{2,0}^r) = n^{-r}$ and the extreme subspace is the subspace X_{2n} introduced above (the proof is similar to that of $\pi_{2n}(W_2^r) = n^{-r}$[4, pp. 342–343]). Now, the required assertion follows from Theorem 3.

Application 4. Consider the problem

$$Kx \equiv x'(t) + p(t)x(t) + (Bx)(t) = y(t) \tag{4.4}$$

with the periodic boundary condition

$$x(0) = x(2\pi),$$

where $D_K = \{x \in L_2^1[0, 2\pi] : x(0) = x(2\pi)\}$, $y \in L_2[0, 2\pi]$, $p \in C[0, 2\pi]$, $\int_0^{2\pi} p(s)ds \neq 0$, and B is a linear continuous operator in $L_2[0, 2\pi]$.

Define the operator $A : D_K \longrightarrow L_2[0, 2\pi]$ by setting $A(x(t)) = x'(t) + p(t)x(t)$. Let X_{2n-1} be the subspace of trigonometric polynomials of degrees no higher than $n - 1$ defined on the interval $[0, 2\pi]$, $Y_{2n-1} = (A^*)^{-1}X_{2n-1}$ (here, $(A^*)^{-1}$ can be explicitly found, since $A^*(x(t)) = -x'(t) + p(t)x(t)$), and P_{2n-1} be the orthoprojector onto Y_{2n-1}. If the homogeneous equation corresponding to Eq. (4.4) has only the trivial solution, then, for the solution set $F = W_2^r$, projection method (1.2) is asymptotically optimal among all the projection methods for solving problem (4.4) and we have the estimate

$$V_{2n-1} \sim n^{-r}.$$

Indeed, let $H = L_2[0, 2\pi]$. It is easy to see that A^{-1} is a compact operator in H. It is well known (see [4, p. 343]) that $\pi_{2n-1}(W_2^r) = n^{-r}$ and the extreme subspace for $\pi_{2n-1}(W_2^r)$ is X_{2n-1}. Now, the required assertion follows from Theorem 3.

Lemma 2 and Theorem 1 were announced in [5].

REFERENCES

1. B. G. Gabdulkhaev, *Optimal Approximations of Solutions to Linear Problems* (Kazansk. Univ., Kazan, 1980) [in Russian].
2. B. G. Gabdulkhaev, *Numerical Analysis of Singular Integral Equations* (Kazansk. Univ., Kazan, 1995) [in Russian].
3. V. M. Tikhomirov, *Certain Problems in Approximation Theory* (Mosk. Gos. Univ., Moscow, 1976) [in Russian].
4. N. P. Korneichuk, *Accurate Constants in Approximation Theory* (Nauka, Moscow, 1987) [in Russian].
5. F. K. Akhmadishina, "On Asymptotic Optimality of Projection Methods," Izv. Vyssh. Uchebn. Zaved., Mat., No. 3, 77–79 (2000).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1093–1102. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © T.E. Bulgakova, A.V. Voytishek, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1148–1157.

Constrained Optimization of the Randomized Iterative Method

T. E. Bulgakova and A. V. Voytishek

Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch, Russian Academy of Sciences, pr. Akademika Lavrent'eva 6, Novosibirsk, 630090 Russia

e-mail: vav@osmf.sscc.ru

Received November 14, 2008

Abstract—Selection of conditionally optimal parameters of the randomized iterative method for solving large-scale linear systems of equations is considered. The error of this method is analyzed by analogy with the functional Monte Carlo algorithms. For the simple iteration method, the "column" randomization of the matrix is thoroughly analyzed.

DOI: 10.1134/S0965542509070021

Key words: system of linear equations, iterative method, column randomization, Monte Carlo method.

INTRODUCTION

When the problems of mathematical physics are solved numerically, grids or finite elements are often used to reduce the corresponding differential or integral equation to a system of linear algebraic equations for the vector of values at the grid points. The matrices of these systems usually have a large dimension, and the exact methods for solving such systems, which are computationally costly and unstable, are inefficient or inapplicable. Hence, iterative algorithms are often used (see ([1–3]). However, when the iterative methods are used to solve linear systems with large dense matrices, the requirements for the computer memory can be very high. If the coefficients of the system are computed using simple formulas, the Monte Carlo method can be used (for example, see [1, 4, 5]).

Historically, the Ulam–Neumann method (the randomized simple iteration method) was first used to solve such problems. However, it turned out to be suitable only for a narrow class of problems and inefficient for problems with large matrices. The randomization method described in [6–9], which is considered in this paper, is efficient in the case of large systems. Using the randomization technique, the original system can be replaced by a system of a lower dimension. This method can be considered as a modification of the Ulam–Neumann scheme.

In Section 1, we outline the randomized algorithm for solving systems with large dense matrices. In Section 2, we discuss the error estimation for this algorithm; for this purpose, we use the decomposition of the error into the deterministic and stochastic components (e.g., see [5, 10]). In Section 3, a specific randomization technique for the system's matrix is described (as in [6–9], the randomization is performed by columns), and the constrained optimization of the resulting stochastic discrete numerical scheme is discussed (e.g., see [5, 10]). Numerical results are presented in Section 4.

1. RANDOMIZATION OF THE ITERATIVE METHOD

Consider the system of linear algebraic equations

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{1}$$

where $\mathbf{x} = (x_1, ..., x_T)^{(\mathrm{T})}$, $\mathbf{b} = (b_1, ..., b_T)^{(\mathrm{T})} \in R^T$ (here, (T) denotes the transposition operation), and $\mathbf{A} = \{a_{tt'}; t, t' = 1, 2, ..., T\}$ for the case $T \gg 1$ (that is, we consider large algebraic systems with dense real matrices). Let the spectral radius $\rho(\mathbf{A})$ of $\mathbf{A}$ be less than unity. It is known (e.g., see [11]) that the solution $\mathbf{x}$ to the linear system can be written as the Neumann series $\mathbf{x} = \sum_{i=0}^{\infty} \mathbf{A}^i \mathbf{b}$ corresponding to the iterative process

$$\mathbf{x}^{(m+1)} = \mathbf{A}\mathbf{x}^{(m)} + \mathbf{b}, \quad \mathbf{x}^{(0)} = \mathbf{b}, \quad m = 0, 1, \ldots. \tag{2}$$

In practical applications, process (2) is truncated; that is, a certain number of iteration steps M is chosen and the approximate solution

$$\mathbf{x} \approx \mathbf{x}^{(M)} = \sum_{i=0}^{M} \mathbf{A}^i \mathbf{b} \tag{3}$$

is used.

In many applications, the dimension T of problem (1) is large, and the calculation of sum (3) is difficult (in particular, the matrix can be too large to fit in the computer memory). In this case, the following technique was proposed in [6–9].

Consider independent identically distributed random matrices $\mathbf{G}^{(0)}, \ldots, \mathbf{G}^{(M-1)}$ with the expectation $\mathbf{A}$. Define the random vectors $\boldsymbol{\xi}^{(m)}$ as

$$\boldsymbol{\xi}^{(m+1)} = \mathbf{G}^{(m)}\boldsymbol{\xi}^{(m)} + \mathbf{b}, \quad m = 0, 1, \ldots, M-1; \tag{4}$$

the vector $\boldsymbol{\xi}^{(0)}$ is chosen so that $\mathbf{E}\boldsymbol{\xi}^{(0)} = \mathbf{b}$. Relation (4) is a randomization of process (2). Since $\mathbf{G}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$ are independent random variables by construction, we have $\mathbf{E}\boldsymbol{\xi}^{(m)} = \mathbf{x}^{(m)}$.

Algorithm 1. Numerically generate N realizations $\{\mathbf{G}_n^{(0)}, \ldots, \mathbf{G}_n^{(M-1)}; n = 1, 2, \ldots, N\}$ of independent identically distributed (as $\mathbf{G}$) random matrices $\mathbf{G}^{(0)}, \ldots, \mathbf{G}^{(M-1)}$ and calculate the corresponding vectors $\boldsymbol{\xi}_n^{(M)}$ using iterative formula (4). Then, approximately determine the vector $\mathbf{x}^{(M)}$ in (3) using the Monte Carlo method by the formula

$$\mathbf{x}^{(M)} = \mathbf{E}\boldsymbol{\xi}^{(M)} \approx \bar{\boldsymbol{\xi}}^{(M,N)} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\xi}_n^{(M)}. \tag{5}$$

2. ERROR OF THE RANDOMIZATION ALGORITHM

2.1. Decomposition of the Error

We will consider the error of Algorithm 1 in the norm $\|\mathbf{x}\| = \sqrt{x_1^2 + \ldots x_T^2}$. Note that $(\mathbf{x} - \bar{\boldsymbol{\xi}}^{(M,N)})$ and $\delta^{(M,N)} = \|\mathbf{x} - \bar{\boldsymbol{\xi}}^{(M,N)}\|$ are random variables. By analogy with the L_2 approach to the estimation of errors in functional algorithms (see [5, 10]), we examine how $\mathbf{E}(\delta^{(M,N)})^2$ tends to zero as the parameters M and N increase in concord. We show that it holds that

$$\mathbf{E}\left(\delta^{(M,N)}\right)^2 \le \left\|\mathbf{x} - \mathbf{x}^{(M)}\right\|^2 + \mathbf{E}\left\|\bar{\boldsymbol{\xi}}^{(M,N)} - \mathbf{x}^{(M)}\right\|^2 = \Delta_1(M) + \Delta_2(M, N). \tag{6}$$

Indeed, we have

$$\left(\mathbf{E}\delta^{(M,N)}\right)^2 = \left(\mathbf{E}\left(\sum_{t=1}^{T}\left(x_t - \bar{\xi}_t^{(M,N)}\right)^2\right)^{1/2}\right)^2.$$

Here, $\bar{\boldsymbol{\xi}}^{(M,N)} = (\bar{\xi}_1^{(M,N)}, \ldots, \bar{\xi}_T^{(M,N)})^{(\mathrm{T})}$. Make use of the Cauchy–Schwarz inequality for expectations (see [12])

$$\mathbf{E}\left|\xi_1\xi_2\right| \le \left(\mathbf{E}\xi_1^2 \times \mathbf{E}\xi_2^2\right)^{1/2},$$

where ξ_1 and ξ_2 are arbitrary random variables. Set $\xi_1 = 1$ and

$$\xi_2 = \left(\sum_{t=1}^{T}\left(x_t - \bar{\xi}_t^{(M,N)}\right)^2\right)^{1/2}.$$

Then, we have

$$(\mathbf{E}\delta^{(M,N)})^2 \le \mathbf{E}1^2 \times \mathbf{E}\left(\sum_{t=1}^{T}(x_t - \bar{\xi}_t^{(M,N)})^2\right) = \mathbf{E}\left(\sum_{t=1}^{T}(x_t - \bar{\xi}_t^{(M,N)})^2\right) = \sum_{t=1}^{T}\mathbf{E}(x_t - \bar{\xi}_t^{(M,N)})^2$$

$$= \sum_{t=1}^{T}\mathbf{E}(x_t^2 - 2x_t\bar{\xi}_t^{(M,N)} + (\bar{\xi}_t^{(M,N)})^2) = \sum_{t=1}^{T}(x_t^2 - 2x_t x_t^{(M)} + \mathbf{E}(\bar{\xi}_t^{(M,N)})^2)$$

$$= \sum_{t=1}^{T}(x_t^2 - 2x_t x_t^{(M)} + (x_t^{(M)})^2 + \mathbf{E}(\bar{\xi}_t^{(M,N)})^2 - (x_t^{(M)})^2)$$

$$= \sum_{t=1}^{T}((x_t - x_t^{(M)})^2 + \mathbf{E}(\bar{\xi}_t^{(M,N)})^2 - (\mathbf{E}\bar{\xi}_t^{(M,N)})^2) = \sum_{t=1}^{T}(x_t - x_t^{(M)})^2 + \sum_{t=1}^{T}\mathbf{D}\bar{\xi}_t^{(M,N)}.$$

The first term

$$\Delta_1(M) = \sum_{t=1}^{T}(x_t - x_t^{(M)})^2 = \|\mathbf{x} - \mathbf{x}^{(M)}\|^2$$

can be considered as the *deterministic component of the error* and the second term is the *stochastic component of the error* (6).

$$\Delta_2(M, N) = \sum_{t=1}^{T}\mathbf{D}\bar{\xi}_t^{(M,N)} = \mathbf{E}\|\bar{\xi}^{(M,N)} - \mathbf{x}^{(M)}\|^2$$

2.2. Estimation of the Deterministic Error Component

We have the following proposition

Proposition 1 (see [11]). *If*

$$\|\mathbf{A}\| \le q < 1, \tag{7}$$

then process (2) *converges to the unique solution* $\mathbf{x}$ *to Eq.* (1) *and*

$$\|\mathbf{x} - \mathbf{x}^{(M)}\| \le \frac{q^M}{1-q}\|\mathbf{A}\mathbf{b} - \mathbf{b}\|.$$

It follows from Proposition 1 that, under condition (7), it holds that

$$\Delta_1(M) \le \frac{q^{2M}}{(1-q)^2}\|\mathbf{A}\mathbf{b} - \mathbf{b}\|^2.$$

Therefore, when the level of the deterministic error δ_{det} is prescribed, the number of iteration steps can be found from the inequality

$$M \gtrsim \frac{\log_q\delta_{\text{det}} + 2\log_q(1-q) - 2\log_q\|\mathbf{A}\mathbf{b} - \mathbf{b}\|}{2}. \tag{8}$$

In the case of large matrices $\mathbf{A}$, the direct calculation of q and $\|\mathbf{A}\mathbf{b} - \mathbf{b}\|$ on the right-hand side of inequality (8) is difficult. To find them approximately, one can use an analog of randomized Algorithm 1 with a relatively low number of tests $N_1 \ll N$.

2.3. Estimation of the Stochastic Error Component: Covariance Matrix

To analyze the behavior of the second term $\Delta_2(M, N)$, consider the covariance matrix

$$\mathbf{B}^{(M)} = \mathbf{E}(\xi^{(M)} - \mathbf{x}^{(M)})(\xi^{(M)} - \mathbf{x}^{(M)})^{(\mathrm{T})} = \mathbf{E}(\xi^{(M)}(\xi^{(M)})^{(\mathrm{T})}) - \mathbf{x}^{(M)}(\mathbf{x}^{(M)})^{(\mathrm{T})}.$$

Taking into account the fact that, for $m = 0, 1, \ldots, M-1$, we have relations (4) and

$$\mathbf{B}^{(m)} = \mathbf{E}(\xi^{(m)}(\xi^{(m)})^{(\mathrm{T})}) - \mathbf{x}^{(m)}(\mathbf{x}^{(m)})^{(\mathrm{T})},$$

we obtain

$$\begin{aligned}\mathbf{B}^{(m+1)} &= \mathbf{E}(\xi^{(m+1)} - \mathbf{x}^{(m+1)})(\xi^{(m+1)} - \mathbf{x}^{(m+1)})^{(\mathrm{T})} = \mathbf{E}(\mathbf{G}^{(m)}\xi^{(m)} - \mathbf{A}\mathbf{x}^{(m)})(\mathbf{G}^{(m)}\xi^{(m)} - \mathbf{A}\mathbf{x}^{(m)})^{(\mathrm{T})}\\ &= \mathbf{E}(\mathbf{G}^{(m)}\xi^{(m)} - \mathbf{G}^{(m)}\mathbf{x}^{(m)} + \mathbf{G}^{(m)}\mathbf{x}^{(m)} - \mathbf{A}\mathbf{x}^{(m)})(\mathbf{G}^{(m)}\xi^{(m)} - \mathbf{G}^{(m)}\mathbf{x}^{(m)} + \mathbf{G}^{(m)}\mathbf{x}^{(m)} - \mathbf{A}\mathbf{x}^{(m)})^{(\mathrm{T})}\\ &= \mathbf{E}(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)}) + (\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)})(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)}) + (\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)})^{(\mathrm{T})}\\ &= \mathbf{E}(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)}) + (\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)})((\xi^{(m)} - \mathbf{x}^{(m)})^{(\mathrm{tr})}(\mathbf{G}^{(m)})^{(\mathrm{T})} + (\mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)} - \mathbf{A})^{(\mathrm{T})})\\ &= \mathbf{E}(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})(\xi^{(m)} - \mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)})^{(\mathrm{T})}) + \mathbf{E}((\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)})^{(\mathrm{T})})\\ &\quad + \mathbf{E}(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})(\mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)} - \mathbf{A})^{(\mathrm{T})}) + \mathbf{E}((\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}(\mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)} - \mathbf{A})^{(\mathrm{T})}).\end{aligned} \tag{9}$$

Because $\mathbf{G}^{(m)}$ and $\xi^{(m)}$ are independent random variables and $\mathbf{E}\xi^{(m)} = \mathbf{x}^{(m)}$, we have

$$\mathbf{E}((\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)})^{(\mathrm{T})}) = \mathbf{E}(\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})(\mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(m)} - \mathbf{A})^{(\mathrm{T})}) = 0.$$

Writing the first term in (9) component by component, we obtain

$$\begin{aligned}&\mathbf{E}(g_{it}^{(m)}g_{jt'}^{(m)}(\xi_t^{(m)} - x_t^{(m)})(\xi_{t'}^{(m)} - x_{t'}^{(m)}))\\ &= \mathbf{E}(g_{it}^{(m)}g_{jt'}^{(m)})\mathbf{E}((\xi_t^{(m)} - x_t^{(m)})(\xi_{t'}^{(m)} - x_{t'}^{(m)})) = \mathbf{E}(g_{it}^{(m)}g_{jt'}^{(m)})b_{tt'}^{(m)} = \mathbf{E}(g_{it}^{(m)}b_{tt'}^{(m)}g_{jt'}^{(m)}),\\ &i, j, t, t' = 1, 2, \ldots, T.\end{aligned} \tag{10}$$

This reasoning implies that the covariance matrices $\mathbf{B}^{(m)} = \{b_{tt'}^{(m)}\}$ satisfy the recurrence

$$\mathbf{B}^{(m+1)} = \mathbf{E}(\mathbf{G}^{(\boldsymbol{m})}\mathbf{B}^{(m)}(\mathbf{G}^{(\boldsymbol{m})})^{(\mathrm{T})}) + \mathbf{C}^{(m)}, \tag{11}$$

where

$$\mathbf{C}^{(m)} = \mathbf{E}((\mathbf{G}^{(\boldsymbol{m})} - \mathbf{A})\mathbf{x}^{(m)}(\mathbf{x}^{(m)})^{(\mathrm{T})}(\mathbf{G}^{(\boldsymbol{m})} - \mathbf{A})^{(\mathrm{T})}). \tag{12}$$

We stress that $\mathbf{B}^{(m)}$ and $\mathbf{C}^{(m)}$ are symmetric positive definite matrices. The boundedness condition for the sequence of the covariance matrices $\mathbf{B}^{(m)}$ follows from the following fact.

Proposition 2 (see [7]). *Consider the operator* $S\mathbf{K} = \mathbf{E}(\mathbf{K} \otimes \mathbf{K})$ *and the limit* $\mathbf{C} = \lim_{m\to\infty}\mathbf{C}^{(m)}$; *assume that* $\mathbf{C} \neq \mathrm{O}$. *The limit* $\mathbf{B} = \lim_{m\to\infty}\mathbf{B}^{(m)}$ *exists and satisfies the equation* $\mathbf{B} = S\mathbf{B} + \mathbf{C}$ *if and only if* $\rho(S) < 1$.

In the formulation of Proposition 2, we used the *Kronecker* or *tensor multiplication* of matrices (see [13])

$$\mathbf{K} \otimes \mathbf{K} = \begin{pmatrix} k_{11}\mathbf{K} & k_{12}\mathbf{K} & \ldots & k_{1T}\mathbf{K} \\ \ldots & \ldots & \ldots & \ldots \\ k_{T1}\mathbf{K} & k_{T2}\mathbf{K} & \ldots & k_{TT}\mathbf{K} \end{pmatrix}, \quad \mathbf{K} = \{k_{tt'};\ t, t' = 1, 2, \ldots, T\}.$$

It was proved in [13] that, under certain additional conditions, the inequalities $\rho(S) < 1$ and $\rho(\mathbf{E}(\mathbf{G}^{(\mathrm{T})}\mathbf{G})) < 1$ are equivalent.

In [7], the following reasoning is used. Taking into account relations (2) and (4), we have

$$\xi^{(m+1)} - \mathbf{x}^{(m+1)} = \mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)}) + (\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}.$$

Then, we obtain

$$\begin{aligned}&\mathbf{E}\left\|\xi^{(m+1)} - \mathbf{x}^{(m+1)}\right\|^2 = \mathbf{E}((\mathbf{G}^{(m)})^{(\mathrm{T})}\mathbf{G}^{(m)}(\xi^{(m)} - \mathbf{x}^{(m)})^{(\mathrm{T})}(\xi^{(m)} - \mathbf{x}^{(m)}))\\ &+ \mathbf{E}\left\|(\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}\right\|^2 \le \rho(\mathbf{E}((\mathbf{G}^{(m)})^{(\mathrm{T})}\mathbf{G}^{(m)}))\mathbf{E}\left\|\xi^{(m)} - \mathbf{x}^{(m)}\right\|^2 + \mathbf{E}\left\|(\mathbf{G}^{(m)} - \mathbf{A})\mathbf{x}^{(m)}\right\|^2.\end{aligned}$$

This implies that the inequality $\rho(\mathbf{E}(\mathbf{G}^{(\mathrm{T})}\mathbf{G})) < 1$ is sufficient for the convergence of the sequence $\mathbf{E}\|\xi^{(m)} - \mathbf{x}^{(m)}\|^2$ (this allows us to estimate the stochastic component of the error $\Delta_2(M, N)$). Furthermore, using the equation

$$\mathbf{E}\left\|\xi^{(m)} - \mathbf{x}^{(m)}\right\|^2 = \sum_{t=1}^{T}\mathbf{D}\xi_t^{(m)} = \mathrm{Tr}(\mathbf{B}^{(m)})$$

(the trace of $\mathbf{B}^{(m)}$, see [13]), the same inequality is sufficient for the convergence of the sequence $\mathbf{B}^{(m)}$. Using Proposition 2, we conclude that, if $\rho(\mathbf{E}(\mathbf{G}^{(\mathrm{T})}\mathbf{G})) < 1$, then the sequence $\mathbf{E}\|\boldsymbol{\xi}^{(m)} - \mathbf{x}^{(m)}\|^2$ has the limit $\mathrm{Tr}(\mathbf{B})$ and the following inequality holds:

$$\mathrm{Tr}(\mathbf{B}) \le \frac{\mathrm{Tr}(\mathbf{C})}{1 - \rho(\mathbf{E}(\mathbf{G}^{(\mathrm{T})}\mathbf{G}))}.$$

3. COLUMN RANDOMIZATION OF THE MATRIX

3.1. Special Randomization Technique

In [6–9], the following algorithm for constructing the random matrix $\mathbf{G} = \{g_{tt'}\}$ ($t, t' = 1, 2, \ldots, T$) was considered.

Algorithm 2. Step 1. At random, choose L indexes $\mathbf{j}^{(L)} = \{j_1, \ldots, j_L\}$ of the columns of $\mathbf{A}$ (that is, $\mathbf{j}^{(L)}$ is a random vector consisting of equiprobable positive integers j_l not exceeding T).

Step 2. For all $t, t' = 1, 2, \ldots, T$, set

$$g_{tt'} = \{ Ta_{tt'}/L \text{ for } t' \in \mathbf{j}^{(L)}; 0 \text{ otherwise} \}.$$

The expectation of the matrix $\mathbf{G}$ is $\mathbf{A}$. Indeed,

$$\mathbf{E}g_{tt'} = (Ta_{tt'}/L) \times \mathbf{P}\{t' \in \mathbf{j}^{(L)}\} = a_{tt'}. \tag{13}$$

A similar randomization technique can be used for constructing the "initial" (for process (4)) vector $\boldsymbol{\xi}^{(0)}$.

Note that the Ulam–Neumann scheme (e.g., see [4, 5]) uses Algorithm 1 with the randomization by Algorithm 2 with $L = 1$.

3.2. Selection of Random Indexes

For large T, we need an efficient algorithm for producing L equiprobable indexes at the first step of Algorithm 2. Such algorithms are studied in [14]. To obtain the results presented in Section 4, we used the following algorithm.

Consider the array S filled with distinct integers in the range from 1 to T. In particular, we can set

$$S(t) := t, \quad t = 1, 2, \ldots, T, \tag{14}$$

where $S(t)$ is the tth element in S and := is the re-assignment operation.

Algorithm 3 (Selection with mixing [14])**.** Select a component j_1 of $\mathbf{j}^{(L)}$. To this end, we use the formula for simulating the uniform discrete distribution (e.g., see [5])

$$k_1 = [\alpha_1 T + 1]. \tag{15}$$

Here and in what follows, α with a subscript denotes a realization of the standard random variable (the random variable that is uniformly distributed on the interval (0, 1)) obtained using a computer random number generator (e.g., see [5]). The symbol $[A]$ denotes the integral part of A.

Set $j_1 = S(k_1)$; then, exchange the values in $S(k_1)$ and $S(T)$:

$$q := S(k_1), \quad S(k_1) := S(T), \quad S(T) := q.$$

Furthermore, for $l = 2, 3, \ldots, L$, choose the components j_l of $\mathbf{j}^{(L)}$ as follows. Select a uniformly distributed random number k_l among the $T - l + 1$ numbers by the formula (similarly to (15))

$$k_l = [\alpha_l(T - l + 1) + 1]. \tag{16}$$

Set $j_l = S(k_l)$; then, exchange the values in $S(k_l)$ and $S(T - l + 1)$:

$$q := S(k_l), \quad S(k_l) := S(T - l + 1), \quad S(T - l + 1) := q.$$

After Algorithm 3 has finished executing, the array S contains distinct integers from 1 to T; therefore, it can be again used by Algorithm 3. Note that the components $j_1, \ldots, j_L$ of the vector $\mathbf{j}^{(L)}$ are stored in the elements $S(T), S(T-1), \ldots, S(T-L+1)$, respectively.

It is clear that the complexity of Algorithm 3 is of order L. However, note that it can be implemented only for $T \le T_0$, where T_0 is the effective computer memory capacity that can be used to store the arrays. This restriction can be removed as follows.

Assume that the numbers $\{1, ..., T\}$ are arranged in ascending order; that is, assume that relation (14) holds. Create the array B of dimension L to store the winning numbers. After step i of the algorithm (i.e., after the ith winning number has been selected), the first i elements of B are occupied by the first i winning numbers arranged in ascending order.

Algorithm 4 (for looking through the preceding selected numbers, [14])**.** First, using formula (15), select a random number k_1 among T possible numbers and make it the first winning number $j_1 = k_1$. Set $B[1] := k_1$.

For $l = 2, 3, ..., L$, select the components j_l of the vector $\mathbf{j}^{(L)}$ as follows. Select a random number k_l among $(T - l + 1)$ numbers using (16). Sequentially, check the inequalities

$$k_l + 1 > B[1], \quad k_l + 2 > B[2], \quad ..., \quad k_l + l - 1 > B[l-1]. \tag{17}$$

If all of them are satisfied, set $j_l = k_l + l - 1$ and $B[l] := k_l + l - 1$.

Otherwise, for the first i such that $k_l + i \le B[i]$ do the following operations.

For $t = l, l-1, ..., i+1$, sequentially perform the assignments $B[t] := B[t-1]$, and then place $k_l + i - 1$ into $B[i]$ and make it the lth winning number: $j_l = k_l + i - 1$.

Since the number of operations required to select the lth winning number has the order l (see (17)), the complexity of Algorithm 4 is of order L^2 (indeed, $1 + 2 + ... + L = (L+1)L/2$).

3.3. The Computational Cost of Algorithm 1 with the Randomization Performed by Algorithm 2

Denote by $\mathbf{j}^{(L, m)}$ the array of numbers obtained at step m of process (4) ($m = 0, 1, ..., M-1$). The matrix $\mathbf{G}$ has only L nonzero columns, which are equal to the corresponding columns of $\mathbf{A}$; therefore, to calculate the vector $\boldsymbol{\xi}^{(m+1)}$, only L components of $\boldsymbol{\xi}^{(m)}$ with the indexes taken from the array $\mathbf{j}^{(L, m)}$ are needed; in turn, these indexes are obtained at the preceding step from the components of $\boldsymbol{\xi}^{(m-1)}$ with the indexes taken from the array $\mathbf{j}^{(L, m-1)}$. Therefore, the amount of computations needed to determine k components of $\boldsymbol{\xi}^{(m+1)}$ is proportional to $kL + mL^2$. If N realizations are needed to achieve the desired accuracy, the total computational cost of obtaining the estimate $\bar{\boldsymbol{\xi}}^{(M, N)}$ in (5) (if k components of the vector $\boldsymbol{\xi}^{(M)}$ are approximated) is $N(kL + ML^2)$.

If the entire vector $\boldsymbol{\xi}^{(M)}$ must be estimated, the computational cost of Algorithm 1 with the randomization performed by Algorithm 2 can be reduced using the following technique. Perform an additional $(M+1)$th step of process (4), and make up the array $\mathbf{j}^{(L, M+1)}$ using Algorithm 3; in the process, we compute only L components of $\boldsymbol{\xi}^{(M)}$ with the indexes taken from this array. In order to make this estimate unbiased, the resulting components are multiplied by T/L. In this case, the computational cost is $t_1^{(M)} = N((M+1)L^2)$ whereas the cost of computing $\mathbf{x}^{(M)}$ using directly the simple iteration method is $t_2^{(M)} = (M+1)T^2$. For large T, it can happen that $t_1^{(M)} \ll t_2^{(M)}$ (indeed, $T \gg \sqrt{N}L$ in this case).

A similar technique can be used to approximate the functionals $\Sigma_{t=1}^{T} \xi_t^{(M)} h_t$, and a special (not necessarily uniform) discrete distribution with the probabilities $\{p_t\}$ can be used to randomize the vector $\boldsymbol{\xi}^{(M)}$:

$$\sum_{t=1}^{T} \xi_t^{(M)} h_t = \sum_{t=1}^{T} p_t \left(\frac{1}{p_t} \xi_t^{(M)} \right) h_t.$$

3.4. Analysis of the Dependence of the Covariance Matrix on the Parameter L

Much as in [6–8], write Eq. (11) in the componentwise form (see (10)) taking into account relation (13):

$$b_{ij}^{(m+1)} = \mathbf{E}\left(\sum_{t,t'=1}^{T} g_{it} b_{tt'}^{(m)} g_{jt'} \right) + c_{ij}^{(m)} = \frac{T^2}{L^2} \sum_{t,t'=1}^{T} a_{it} b_{tt'}^{(m)} a_{jt'} \mathbf{P}\{t, t' \in \mathbf{j}^{(L)}\} + c_{ij}^{(m)};$$

here,

$$\mathbf{P}\{t, t' \in \mathbf{j}^{(L)}\} = \begin{cases} (L(L-1))/(T(T-1)) & \text{for} \quad t \ne t' \\ L/T \text{ otherwise}. \end{cases}$$

Denote by $\mathbf{D}^{(m)}$ the diagonal of the matrix $\mathbf{B}^{(m)}$. Then, Eq. (11) can be written as

$$\mathbf{B}^{(m+1)} = \frac{T}{L}\mathbf{A}\left(\frac{L-1}{T-1}\mathbf{B}^{(m)} + \frac{T-L}{T-1}\mathbf{D}^{(m)}\right)\mathbf{A}^{(\mathrm{T})} + \mathbf{C}^{(m)} = \frac{TL-1}{L\,T-1}\mathbf{A}\mathbf{B}^{(m)}\mathbf{A}^{(\mathrm{T})} + \frac{T}{L}\frac{T-L}{T-1}\mathbf{A}\mathbf{D}^{(m)}\mathbf{A}^{(\mathrm{T})} + \mathbf{C}^{(m)}.$$

Similarly, we obtain the representation for the matrices $\mathbf{C}^{(m)}$ in (12) in the form

$$c_{ij}^{(m)} = \mathbf{E}\left(\sum_{t,t'=1}^{T} g_{it}x_t^{(m)}x_{t'}^{(m)}g_{jt'}\right) - \sum_{t,t'=1}^{T} a_{it}x_t^{(m)}x_{t'}^{(m)}a_{jt'} = \sum_{t,t'=1}^{T} a_{it}x_t^{(m)}x_{t'}^{(m)}a_{jt'}\left(\frac{T^2}{L^2}\mathbf{P}\{t,t' \in \mathbf{j}^{(L)}\} - 1\right)$$

$$= \frac{1}{L}\frac{T-L}{T-1}\left(T\sum_{t,t'=1}^{T}(a_{it}a_{jt'}(x_t^{(m)})^2 - a_{it}x_t^{(m)}x_t^{(m)}a_{jt'})\right).$$

Consider the operator H acting in the space of matrices $\mathbf{K}$ of order $T \times T$ by the rule

$$H\mathbf{K} = \frac{T}{L}\mathbf{A}\left(\frac{L-1}{T-1}\mathbf{K} + \frac{T-L}{T-1}\mathbf{D}(\mathbf{K})\right)\mathbf{A}^{(\mathrm{T})},$$

where $\mathbf{D}(\mathbf{K})$ is the diagonal of the matrix $\mathbf{K}$. Iterative process (11) can be written in the form

$$\mathbf{B}^{(m+1)} = H\mathbf{B}^{(m)} + \mathbf{C}^{(m)} \quad \text{or} \quad \mathbf{B}^{(m+1)} = H^m\mathbf{C}^{(0)} + H^{m-1}\mathbf{C}^{(1)} + \ldots + H\mathbf{C}^{(m-1)} + \mathbf{C}^{(m)}.$$

If $\mathbf{B}$ is a symmetric matrix (i.e., if $\mathbf{B} = \mathbf{B}^{(\mathrm{T})}$, we have

$$\rho(H\mathbf{B}) = \|H\mathbf{B}\| = \max_{\mathbf{y}}\frac{(H\mathbf{B}\mathbf{y}, \mathbf{y})}{(\mathbf{y}, \mathbf{y})}$$

$$= \max_{\mathbf{y}}\frac{T}{L}\frac{\frac{L-1}{N-1}(\mathbf{B}\mathbf{A}^{(\mathrm{T})}\mathbf{y}, \mathbf{A}^{(\mathrm{T})}\mathbf{y}) + \frac{T-L}{T-1}(\mathbf{D}\mathbf{A}^{(\mathrm{T})}\mathbf{y}, \mathbf{A}^{(\mathrm{T})}\mathbf{y})}{(\mathbf{y}, \mathbf{y})} \le \frac{T}{L}\rho(\mathbf{B})\max_{\mathbf{y}}\frac{(\mathbf{A}^{(\mathrm{T})}\mathbf{y}, \mathbf{A}^{(\mathrm{T})}\mathbf{y})}{(\mathbf{y}, \mathbf{y})} = \frac{T}{L}\rho(\mathbf{B})\|\mathbf{A}^{(\mathrm{T})}\|^2.$$

Therefore, if $\|\mathbf{A}^{(\mathrm{T})}\| < \sqrt{L/T}$, the sequence $H^m\mathbf{B}$ is convergent and $\rho(H) < 1$.

Note that, when L changes from 1 to T, the cost of computing the components of $\boldsymbol{\xi}^{(M)}$ increases as L^2 while the variance of these components decreases. Therefore, we may expect that there is an optimal L for which the average cost of the computation of $\bar{\boldsymbol{\xi}}^{(M,N)}$ is minimal.

4. NUMERICAL RESULTS

4.1. Test Problem

The conjecture of the existence of an optimal value of the parameter L of the numerical scheme based on Algorithm 1 with the randomization performed by Algorithm 2 was checked, in particular, using the following test problem. The entries of $\mathbf{A}$ were specified as

$$a_{tt'} = (1/T)\exp(-|t-t'|). \tag{18}$$

To calculate the spectral radius of $\mathbf{A}$, we used the power method (e.g., see [3]). We assumed that the solution to system (1) is known to be $\mathbf{x} = (1, \ldots, 1)^{(\mathrm{T})}$. In this case, we have $\mathbf{b} = \mathbf{x} - \mathbf{A}\mathbf{x}$.

4.2. On the Choice of the Parameters in Algorithms 1 and 2

By analogy with [9], we investigated the dependence of the computational cost

$$S(L, M, \bar{N}) = L^2(M-1)\|\mathbf{D}\bar{\boldsymbol{\xi}}^{(M,\bar{N})}\| \tag{19}$$

on L for fixed levels of the deterministic and stochastic errors δ_{det} and δ_{stoch}; $\bar{N}$ will be determined later in this paper (see (21)), and

$$\|\mathbf{D}\bar{\boldsymbol{\xi}}^{(M,\bar{N})}\| = \max_{t=1,T}\left(\frac{1}{\bar{N}-1}\sum_{n=1}^{\bar{N}}(\xi_{t,n}^{(M)})^2 - \frac{1}{\bar{N}(\bar{N}-1)}\sum_{n=1}^{\bar{N}}(\xi_{t,n}^{(M)})^2\right). \tag{20}$$

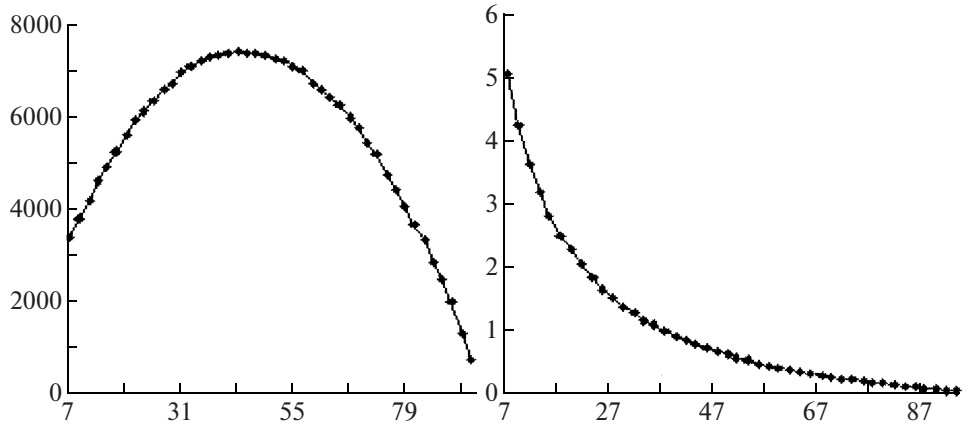


Fig. 1.

Anticipating things, we note that the nature of the dependence of (19) on L in the test calculation was in good agreement with the dependence of the actual computational effort $\tilde{S}(L)$ needed to attain the prescribed accuracy of the approximate solution $\bar{\xi}^{(M,\bar{N})}$ on L (taking into consideration the exact solution $\mathbf{x}^{(M)}$ of the test problem, see Subsection 4.3).

In the test calculations, the number of iteration steps found by formula (8) was $M = 5$ (for $T = 1000$ and matrix (18), this corresponds to $\delta_{\text{det}} = 0.032$).

For fixed L and M, the number of tests N was chosen from the condition that the relative error $\varepsilon_{\text{rel}} = \|\mathbf{x} - \xi^{(M,N)}\|/\|\mathbf{x}\|$ is less than a certain predefined level δ_{stoch}. Using the equivalence of the norms and the expressions for the error and approximate variance (e.g., see [5]), we conclude that, for sufficiently large N,

$$\varepsilon_{\text{rel}} = \frac{\max\limits_{t=1,T}\left|x_t - \xi_t^{(M,N)}\right|}{\max\limits_{t=1,T}\dfrac{1}{N}\sum\limits_{n=1}^{N}\xi_{t,n}^{(M)}} < \frac{\max\limits_{t=1,T}\gamma\sqrt{\dfrac{\mathbf{D}\xi_t^{(M,N)}}{N}}}{\max\limits_{t=1,T}\dfrac{1}{N}\sum\limits_{n=1}^{N}\xi_{t,n}^{(M)}}$$

$$\approx \frac{\max\limits_{t=1,T}\gamma\sqrt{\dfrac{1}{N}\left(\dfrac{1}{N-1}\sum\limits_{n=1}^{N}(\xi_{t,n}^{(M)})^2 - \dfrac{1}{N(N-1)}\sum\limits_{n=1}^{N}(\xi_{t,n}^{(M)})^2\right)}}{\max\limits_{t=1,T}\dfrac{1}{N}\sum\limits_{n=1}^{N}\xi_{t,n}^{(M)}} = \delta_N,$$

with a probability close to unity. We assumed that γ is 2, which corresponds to the level $\delta_{\text{stoch}} = 0.025$. The quantity N_0 was calculated K times, and

$$\bar{N} = \frac{N_0^{(1)} + \ldots + N_0^{(K)}}{K} \tag{21}$$

was used in (19) and (20).

4.3. Description of the results

The test calculations were performed for different values of T. In Fig. 1, for a relatively small $T = 100$, the behavior of computational cost (19) and the norm of variance (20) depending on L varying in the range from 1 to 100 is shown. The dependence of the actual computational effort $\tilde{S}(L)$ needed to attain the prescribed accuracy on L is in complete agreement with the plot of the computational cost $S(L, M, \bar{N})$.

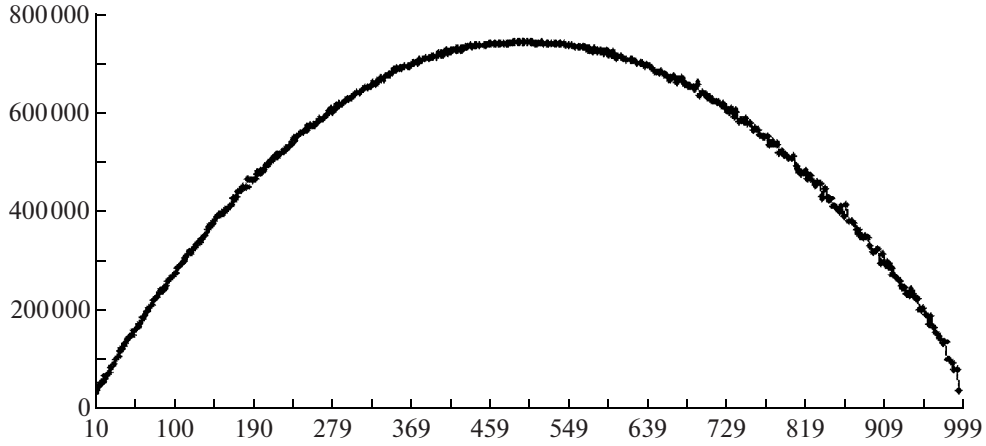


Fig. 2.

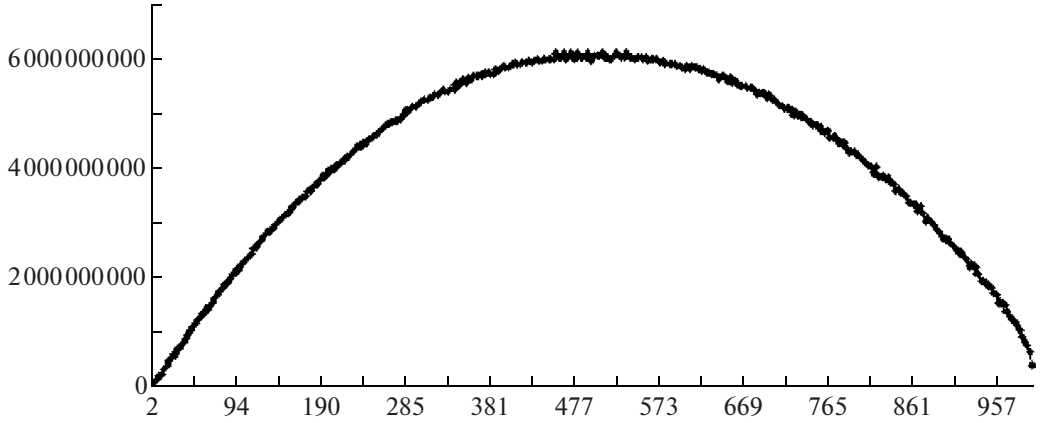


Fig. 3.

A similar picture was observed for the practically important case of large T. In particular, in Fig. 2 and 3, the plots of computational cost (19) and of the actual computational effort $\tilde{S}(L)$ are shown for $T = 1000$.

4.4. Conclusions

The numerical results suggest the following conclusions.

1. Formula (19) for the computational cost is in good agreement with the dependence of the actual computational effort of Algorithm 1 with the randomization performed using Algorithm 2 on the parameters L, M, and $\overline{N}$ for the given level of the acceptable error.

2. The conjecture made in [6–9] that the computational cost of Algorithm 1 with the randomization performed using Algorithm 2 can be small when L is relatively small was confirmed.

3. The conjecture made in [6–9] that computational cost (19) considered as a function of L has a minimum when the error level is fixed was not confirmed. Moreover, it has a maximum, and the same is true not only for matrix (18) but also for other matrices (in particular, see [15]).

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 07-01-00024a and 09-01-00035a.

REFERENCES

1. N. S. Bakhvalov, *Numerical Methods* (Nauka, Moscow, 1975) [in Russian].
2. G. I. Marchuk, *Methods of Numerical Mathematics*, 2nd ed. (Nauka, Moscow, 1980; Springer, New York, 1982).
3. A. N. Konovalov, *Introduction into Computational Methods of Linear Algebra (*Nauka, Novosibirsk, 1993) [in Russian].
4. I. M. Sobol', *Numerical Monte Carlo Methods* (Nauka, Moscow, 1973) [in Russian].
5. G. A. Mikhailov and A. V. Voytishek, *Numerical Statistical Simulation: Monte Carlo Methods* (Izd. tsentr "Akademiya", Moscow, 2006) [in Russian].
6. G. A. Mikhailov, *Weighted Monte Carlo Methods* (IVMiMG SO RAN, Novosibirsk, 2000) [in Russian].
7. Yu. V. Bulavski, "Randomized Method of Successive Approximations for Linear Systems of Algebraic Equations," Rus. J. Numer. Analys. Math. Modelling **10** (6), 481–493 (1995).
8. Yu. V. Bulavsky and S. A. Temnikov, "Randomized Method of Successive Approximations," in *Mathematical Methods in Stochastic Simulation and Experimental Design* (Univ. Publ. House, St. Petersburg, 1996), pp. 64–68.
9. S. A. Temnikov, "A Solution of the Light Scattering Problem from an Ensemble of Fractal Clusters Using the Monte Carlo Method," in *Trudy konferentsii molodykh uchenykh* (VTs SO RAN, Novosibirsk, 1995), pp. 160–172 [in Russian].
10. A. V. Voitishek, *Functional Estimates in the Monte Carlo Method* (Novosibirsk Gos. Univ., Novosibirsk, 2007) [in Russian].
11. L. V. Kantorovich and G. P. Akilov, *Functional Analysis* (Nauka, Moscow, 1977; Pergamon, Oxford, 1982).
12. A. A. Borovkov, *Probability Theory* (Nauka, Moscow, 1986) [in Russian].
13. V. V. Voevodin and Yu. A. Kuznetsov, *Matrices and Computations* (Nauka, Moscow, 1984) [in Russian].
14. P. S. Rouzankin and A. V. Voytishek, "On the Cost of Algorithms for Random Selection," Monte Carlo Meth. Appl. **5** (1), 39–54 (1999).
15. N. S. Motsartova and T. E. Bulgakova, "The Use of a Randomization Algorithms for Large Matrices in Solving the Dirichlet Problem," in *Materialy XLVI Mezhdunarodnoi studencheskoi konferentsii Student i nauchno-tekhnicheskii progress, Matematika (*Novosibirsk Gos. Univ., Novosibirsk, 2008), pp. 238–239 [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1103–1110. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © V.A. Abilov, F.V. Abilova, M.K. Kerimov, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1158–1166.

On Estimates for the Fourier–Bessel Integral Transform in the Space $L_2(\mathbb{R}_+)$

V. A. Abilov[a], F. V. Abilova[b], and M. K. Kerimov[c]

[a] *Dagestan State University, ul. Gadzhieva 43a, Makhachkala, 367025 Russia*

[b] *Dagestan State Technical University, pr. Kalinina 70, Makhachkala, 367015 Russia*

[c] *Dorodnicyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119991 Russia*

e-mail: compmath@mail.ru

Received November 10, 2008; in final form, February 20, 2009

Abstract—Two estimates useful in applications are proved for the Fourier–Bessel integral transform in $L_2(\mathbb{R}_+)$ as applied to some classes of functions characterized by a generalized modulus of continuity.

DOI: 10.1134/S0965542509070033

Key words: Fourier–Bessel integral transform, Bessel operator, generalized shift operator, generalized modulus of continuity, generalized derivatives, estimate.

INTRODUCTION

Integral transforms and their inverses (e.g., the Fourier–Bessel transform) are widely used to solve various problems in calculus, mechanics, mathematical physics, and computational mathematics (see, e.g., [1–6]).

We describe an interesting application (see [7]) of the Fourier–Bessel integral transform in the theory of second-order ordinary differential equations with singularities.

Consider the Bessel second-order differential equation in the form of Liouville,

$$-y''(x) + \left(\nu^2 - \frac{1}{4}\right)x^{-2}y(x) = \lambda y(x), \quad x \in (0, \infty), \quad \lambda \in \mathbb{C}, \quad \nu \in [0, 1], \tag{A}$$

in the Hilbert space $L_2(0, \infty)$, where zero is a singular point. For $\nu \in (0, 1)$, this equation has the solutions

$$y_1(x) = x^{1/2}J_\nu(x\sqrt{\lambda}), \quad y_2(x) = x^{1/2}J_{-\nu}(x\sqrt{\lambda}), \quad x \in (0, \infty),$$

where $J_\nu(\cdot)$ is a Bessel function of the first kind. Here, the analytic function $\sqrt{\lambda} : \mathbb{C} \longrightarrow \mathbb{C}$ is defined in the following sense: if $\lambda = \rho\exp(i\eta)$, then $\sqrt{\lambda} = \rho^{1/2}\exp\left(\frac{1}{2}i\eta\right)$ for $\rho \in [0, \infty]$ and $\eta \in [0, 2\pi)$. Applying the initial value theorem to the point 0, we define the Titchmarsh–Weyl m-coefficient for the entire interval $(0, \infty)$. As a result, it can be proved that the inverse Fourier–Bessel transform is an eigenfunction of problem (A).

In this paper, we prove two estimates for the Fourier–Bessel integral transform, which can be used in particular problems. In a sense, these estimates resemble those for the Fourier integral transform in $L_2(\mathbb{R})$ (see [8]).

In Section 1, we give some definitions and preliminaries concerning the Fourier–Bessel transform. The estimates are proved in Section 2.

1. THE FOURIER–BESSEL INTEGRAL TRANSFORM AND ITS BASIC PROPERTIES

Given a function $f \in L_2(\mathbb{R}_+)$, the Fourier–Bessel integral transform of order p is defined as

$$g(t) = F_p[f](t) = \int_0^\infty \sqrt{xt}J_p(xt)f(x)dx, \quad t > 0, \quad p > -1/2, \tag{1}$$

where $J_p(x)$ is the Bessel function of the first kind of order p. The inverse Fourier–Bessel transform is defined as

$$f(x) = F_p^{-1}[g](x) = \int_0^\infty g(t)\sqrt{xt}J_p(xt)dt. \quad (2)$$

In the theory of special functions and integral transforms, (1) is also known as the Hankel integral transform of order ν. It can be written as

$$H_\nu[f](y) = F_\nu(y) = \int_0^\infty f(x)\sqrt{xy}J_\nu(y,y)dx, \quad y>0, \quad \nu>-1/2,$$

or as

$$G_\nu(y) = \int_0^\infty xg(x)J_\nu(xy)dx$$

(see [6, pp. 666–668]). However, these definitions are equivalent if $f(x)$ is replaced by $\sqrt{x}g(x)$ and $F_\nu(y)$ is replaced by $\sqrt{y}G_\nu(y)$. For $f \in L_1((0,\infty), x^{2\mu+1})$, the following notation is sometimes used:

$$h_\nu(f)y = \int_0^\infty (xy)^{-\mu}J_\mu(xy)f(x)x^{2\mu+1}dx, \quad y \in (0,\infty).$$

Different notation used for inverse transform (2) is

$$f(x) = \int_0^\infty (xt)^{1/2}J_p(xt)\int_0^\infty (t\xi)^{1/2}J_p(t\xi)f(\xi)d\xi.$$

Hankel theorem (see [1, p. 311]). *If $f(x)$ belongs to $L_1(0,\infty)$ and has a bounded variation in the neighborhood of x, then, for $\nu \geq -1/2$,*

$$\frac{1}{2}[f(x+0)+f(x-0)] = \int_0^\infty J_\nu(xu)\sqrt{xu}du\int_0^\infty J_\nu(xy)\sqrt{xy}f(y)dy, \quad x \in (0,\infty).$$

If $f(x)$ is continuous, then the left-hand side of this formula is $f(x)$.

The Hankel transforms of orders 1/2 and $-1/2$ are transformed into the sine and cosine Fourier transforms

$$F_s(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(t)\sin(wt)dt, \quad F_c(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(t)\cos(wt)dt,$$

since

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi}}\sin x, \quad J_{-1/2}(x) = \sqrt{\frac{2}{\pi}}\cos x.$$

Consider the normed space $L_2(\mathbb{R}_+)$ of square integrable functions $f: \mathbb{R}_+ \longrightarrow \mathbb{R}$ equipped with the norm

$$\|f\| = \|f\|_{L_2(\mathbb{R}_+)} = \left(\int_{\mathbb{R}_+}|f(x)|^2dx\right)^{1/2}.$$

Theorem [9, p. 20]. *If $f(x)x^{p+1/2} \in L_2(\mathbb{R}_+)$, then the Hankel transform $g(t) = F_p[f](t)$ also belongs to $L_2(\mathbb{R}_+)$ with the weight $s^{p+1/2}$; i.e., $F_p[f](t)s^{p+1/2} \in L_2(\mathbb{R}_+)$. Moreover, we have Parseval's identity*

$$\int_0^\infty |g(s)|^2 s^{2p+1}ds = 2^{2p}\Gamma^2(p+1)\int_0^\infty |f(x)|^2 x^{2p+1}dx.$$

Specifically, if $f(x) \in L_2(\mathbb{R})$, *then Parseval's identity becomes*

$$\int_0^\infty |g(t)|^2 dt = \int_0^\infty |f(x)|^2 dx. \tag{3}$$

For our goal, it is convenient to replace the Bessel function $J_p(x)$ with the normalized Bessel function of the first kind $j_p(\sqrt{\lambda}x)$, which is related to $J_p(x)$ by the formula

$$j_p(\sqrt{\lambda}x) = \frac{2^p\Gamma(p+1)}{(\sqrt{\lambda}x)^p} J_p(\sqrt{\lambda}x).$$

It is easy to show that $j_p(\sqrt{\lambda}x)$ solves the Bessel equation

$$\frac{d^2y}{dx^2} + \frac{2p+1}{x}\frac{dy}{dx} + \lambda y = 0,$$

with the initial conditions

$$y(0) = 1, \quad y'(0) = 0.$$

Now, we derive formulas for the expansion of the Fourier–Bessel transform in terms of $j_p(\sqrt{\lambda}x)$.

For $x^{p+1/2}f \in L_2(\mathbb{R}_+)$, the Fourier–Bessel (Hankel) transform and its inverse in terms of $j_p(\sqrt{\lambda}x)$ become

$$g(t) = \frac{t^{p+1/2}}{2^p\Gamma(p+1)} \int_0^\infty j_p(xt) f(x) x^{2p+1} dx$$

and

$$x^{p+1/2}f(x) = \frac{x^{p+1/2}}{2^{2p}\Gamma^2(p+1)} \int_0^\infty j_p(xt) g(t) t^{2p+1} dt.$$

Thus, the expansions of $f(x)$ and $g(t)$ in terms of $j_p(xt)$ take the form

$$g(t) = \int_0^\infty j_p(xt) f(x) x^{2p+1} dx \tag{4}$$

and

$$f(x) = \frac{1}{2^{2p}\Gamma^2(p+1)} \int_0^\infty j_p(xt) g(t) t^{2p+1} dt. \tag{5}$$

Accordingly, Parseval's identity becomes

$$\int_0^{+\infty} |g(t)|^2 t^{2p+1} dt = 2^{2p}\Gamma^2(p+1) \int_0^{+\infty} |f(x)|^2 x^{2p+1} dx. \tag{6}$$

In $L_2(\mathbb{R}_+)$, consider the generalized shift operator T_h (see [10, p. 121])

$$T_h f(x) = \frac{\Gamma(p+1)}{\Gamma(1/2)\Gamma(p+1/2)} \int_0^\pi f\left(\sqrt{x^2 + h^2 - 2xh\cos t}\right) \sin^{2p} t\, dt, \quad p \geq -\frac{1}{2}, \quad 0 \geq h \geq 1,$$

which corresponds to the Bessel operator

$$\mathcal{D} = \frac{d^2}{dx^2} + \frac{2p+1}{x}\frac{d}{dx}.$$

It is easy to see that

$$T_0 f(x) = f(x).$$

If $f(x)$ has a continuous first derivative, then

$$\frac{\partial}{\partial h} T_h f(x)\Big|_{h=0} = 0,$$

If it has a continuous second derivative, then $u(x, h) = T_h f(x)$ solves the Cauchy problem

$$\frac{\partial^2 u}{\partial x^2} + \frac{2p+1}{x}\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial h^2} + \frac{2p+1}{h}\frac{\partial u}{\partial h},$$

$$u|_{h=0} = f(x), \quad \frac{\partial u}{\partial h}\Big|_{h=0} = 0.$$

The operator T_h is linear, homogeneous, and continuous. Below are some properties of this operator (see [10, pp. 124–125]):

(i) $T_h j_p(\sqrt{\lambda}x) = j_p(\lambda h) j_p(\lambda x)$.

(ii) T_h is self-adjoint. If $f(x)$ is a continuous function such that $\int_0^\infty x^{2p+1}|f(x)|dx < \infty$ and $g(x)$ is continuous and bounded for all $x \geq 0$, then

$$\int_0^\infty T_h f(x) g(x) x^{2p+1} dx = \int_0^\infty f(x) T_h g(x) x^{2p+1} dx;$$

(iii) $T_h f(x) = T_h f(h)$.

(iv) $\|T_h f - f\| \longrightarrow 0$ as $h \longrightarrow 0$.

The first- and higher order finite differences of $f(x)$ are defined as follows:

$$\Delta_h f(x) = T_h f(x) - f(x) = (T_h - E) f(x),$$

where E is the identity operator in $L_2(\mathbb{R}_+)$, and

$$\Delta_h^k f(x) = \Delta(\Delta_h^{k-1} f(x)) = (T_h - E)^k f(x) = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} T_h^i f(x),$$

where $T_h^0 f(x) = f(x)$, $T_h^i f(x)$, $T_h(T_h^{i-1} f(x))$ for $i = 1, 2, ..., k$; and $k = 1, 2, ...$

The kth order generalized modulus of continuity of a function $f \in L_2(\mathbb{R}_+)$ is defined as

$$\Omega_k(f, \delta) = \sup_{0 < h \leq \delta} \left\| \Delta_h^k f(x) \right\|$$

Let $W_{2,\Phi}^{r,k}\mathfrak{D}$ denote the class of functions $f \in L_2(\mathbb{R}_+)$ that have generalized derivatives in the sense of Levi (see [12, p. 172]) satisfying the estimate

$$\Omega_k(\mathfrak{D}^r f; \delta) = O(\Phi(\delta^k)),$$

where $\Phi(t)$ is any nonnegative function given on $[0, \infty)$. Moreover, for the Bessel operator $\mathfrak{D}$, we have

$$\mathfrak{D}^0 f = f, \quad \mathfrak{D}^r f = \mathfrak{D}(\mathfrak{D}^{r-1} f), \quad r = 1, 2, \ldots,$$

In view of formulas (4) and (5),

$$T_h f(x) = \frac{1}{2^{2p}\Gamma^2(p+1)} \int_0^\infty j_p(tx) j_p(tx) g(t) t^{2p+1} dt.$$

Therefore, combining the relation

$$T_h f(x) - f(x) = \frac{1}{2^{2p}\Gamma^2(p+1)} \int_0^\infty [j_p(th) - 1] j_p(tx) g(t) t^{2p+1} dx$$

with Parseval's identity (6) gives

$$\left\| T_h f(x) - f(x) \right\|^2 = \int_0^\infty [j_p(th) - 1]^2 |g(t)|^2 t^{2p+1} dt.$$

which, for any function $f \in W_{2,\Phi}^{r,k}(\mathcal{D})$, implies

$$\left\|\Delta_h^k \mathcal{D}^r f(x)\right\|^2 = \int_0^\infty t^{4r+2p+1}[j_p(th)-1]^{2k}|g(t)|^2 dt. \tag{7}$$

Below are the main results of this paper.

2. ESTIMATES FOR THE HANKEL TRANSFORM

Taking into account what was said in Section 1, for some classes of functions characterized by the generalized modulus of continuity, we can prove two estimates for the integral

$$\int_N^\infty |g(t)|^2 t^{2p+1} dt, \tag{8}$$

which are useful in applications.

Theorem 1. *For functions $f(x) \in L_2(\mathbb{R}_+)$ in the class $W_{2,\Phi}^{r,k}(\mathcal{D})$,*

$$\sup_{W_{2,\Phi}^{r,k}}\left(\int_N^\infty |g(t)|^2 t^{2p+1} dt\right) = O\left(N^{-2r}\Phi\left[\left(\frac{c}{N}\right)^k\right]\right), \tag{9}$$

where $r = 0, 1, \ldots$; $k = 1, 2, \ldots$; $c > 0$ is a fixed constant; and $\Phi(t)$ is any nonnegative function defined on the interval $[0, \infty)$.

Proof. We use the following asymptotic formulas for the Bessel function $J_p(x)$ (see [3, pp. 353 and 359]):

$$J_p(x) = \sqrt{\frac{2}{\pi x}}\cos\left(x - \frac{p\pi}{2} - \frac{\pi}{4}\right) + O(x^{-3/2}), \quad x \longrightarrow \infty, \tag{10}$$

$$J_p(x) = \frac{x^p}{2^p\Gamma(\nu+1)}[1 + O(x^2)], \quad x \longrightarrow \infty. \tag{11}$$

In the terms of $j_p(x)$, we have (see [11])

$$1 - j_p(x) = O(1), \quad x \geq 1, \tag{12}$$

$$1 - j_p(x) = O(x^2), \quad 0 \leq x \leq 1, \tag{13}$$

$$\sqrt{hx}J_p(hx) = O(1), \quad hx \geq 0. \tag{14}$$

Let $f \in W_{2,\Phi}^{r,k}\mathcal{D}$. Taking into account the Hölder inequality yields

$$\int_{t\geq N} |g(t)|^2 t^{2p+1} dt - \int_{t\geq N} j_p(th)|g(t)|^2 t^{2p+1} dt = \int_{t\geqslant N} [1 - j_p(th)]|g(t)|^2 t^{2p+1} dt$$

$$= \int_{t\geqslant N} [1 - j_p(th)](|g(t)|t^{p+1/2})^2 = \int_{t\geqslant N} [1 - j_p(th)](|g(t)|t^{p+1/2})^{2-1/k}(|g(t)|t^{p+1/2})^{1/k} dt$$

$$\leqslant \left(\int_{t\geqslant N} |g(t)|^2 t^{2p+1} dt\right)^{2k-1/(2k)} \left(\int_{t\geqslant N} [1 - j_p(th)]^{2k}|g(t)|^2 t^{2p+1} dt\right)^{1/(2k)}$$

$$= \left(\int_{t\geqslant N} |g(t)|^2 t^{2p+1} dt\right)^{2k-1/(2k)} \left(\int_{t\geqslant N} t^{4r+2p+1}[1 - j_p(th)]^{2k}|g(t)|^2 t^{-4r} dt\right)^{1/(2k)}$$

$$\leqslant N^{-2r/k}\left(\int_{t\geqslant N} |g(t)|^2 t^{2p+1} dt\right)^{(2k-1)/2k} \left(\int_{t\geqslant N} t^{4r+2p+1}[1 - j_p(th)]^{2k}|g(t)|^2 dt\right)^{1/(2k)}.$$

In view of (7), we conclude that

$$\int\limits_{t \geqslant N} t^{4r+2p+1}[1-j_p(th)]^{2k}|g(t)|^2 t^{-4r}dt \leqslant \left\|\Delta_h^k D^r f(x)\right\|^2.$$

Therefore,

$$\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt \leqslant \int\limits_{t \geqslant N} j_p(th)|g(t)|^2 t^{2p+1}dt + N^{-2r/k}\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{(2k-1)/2k} \left\|\Delta_h^k D^r f(x)\right\|^2.$$

By the definition of $j_p(x)$, it follows that

$$j_p(th) = 2^p\Gamma(p+1)\frac{j_p(th)}{(th)^p} = 2^p\Gamma(p+1)\frac{\sqrt{th}j_p(th)}{(th)^{p+1/2}}.$$

Combining this with estimate (14) gives

$$j_p(th) = O((th)^{-1/2}).$$

Therefore,

$$\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt = O\left(\int\limits_{t \geqslant N} (th)^{-p-1/2}|g(t)|^2 t^{2p+1}dt + N^{-2r/k}\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{(2k-1)/(2k)} \left\|\Delta_h^k D^r f(x)\right\|^{1/k}\right)$$

$$= O\left((Nh)^{-p-1/2}\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right) + N^{-2r/k}\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{(2k-1)/(2k)} \left\|\Delta_h^k D^r f(x)\right\|^{1/k},$$

or

$$(1-O(Nh)^{-p-1/2})\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt = O(N^{-2r/k})\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{(2k-1)/(2k)} \left\|\Delta_h^k D^r f(x)\right\|^{1/k}.$$

Setting $h = c/N$ in the last inequality and choosing $c > 0$ such that $1 - O(c^{-p-1/2}) \geqslant 1/2$, we obtain

$$\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt = O(N^{-2r/k})\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{(2k-1)/(2k)} \Phi^{1/k}\left[\left(\frac{c}{N}\right)\right].$$

Raising both sides to the $2k$th power yields

$$\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt = O\left(N^{-4r}\Phi^2\left[\left(\frac{c}{N}\right)\right]\right),$$

which proves Theorem 1.

Theorem 2. *Let* $\Phi(t) = t^\nu$, *where* $\nu > 0$. *Then*

$$\left(\int\limits_{t \geqslant N}^{\infty} |g(t)|^2 t^{2p+1}dt\right)^{1/2} = O(N^{-2r-k\nu}) \Leftrightarrow f \in W_{2,\Phi}^{r,k}\mathfrak{D},$$

$$r = 0, 1, \ldots, \quad k = 1, 2, \ldots, \quad 0 < \nu < 2.$$

Proof. *Sufficiency.* Let $f \in W_{2,\Phi}^{r,k}(D)$ and $\Phi(t) = t^\nu$ $(\nu > 0)$. Then, by Theorem 1, we have

$$\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{1/2} = O(N^{-2r-k\nu}).$$

The sufficiency is proved.

Necessity. Let

$$\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1}dt\right)^{1/2} = O(N^{-2r-k\nu}),$$

i.e.,

$$\int\limits_{t \geqslant N} |g(t)|^2 t^{2p+1} dt = O(N^{-4r-2k\nu}).$$

It is easy to show, that there exists a function $f \in L_2(\mathbb{R}_+)$ such that $D^r f \in L_2(\mathbb{R})$ and

$$D^r f(x) = \frac{1}{2^{2p}\Gamma(p+1)} \int\limits_0^{+\infty} j_p(th) j_p(tx) g(t) t^{2p+2r+1} dt.$$

Combining this with Parseval's identity produces

$$\left\| \Delta_h^k D^r f(x) \right\|^2 = \int\limits_0^{+\infty} |g(t)|^2 t^{4r+2p+1} [1 - j_p(th)]^{2k} dt.$$

This integral is divided into two:

$$\int\limits_0^{+\infty} = \int\limits_{0<t<N} + \int\limits_{t>N} = I_1 + I_2,$$

where $N = [h^{-1}]$ Let us estimate them separately.

By virtue of (13), we have

$$I_2 = \int\limits_{t \geqslant N} |g(t)|^2 4^{4r+2p+1} [1 - j_p(th)]^{2k} dt = O\left(\int\limits_{t \geqslant N} |g(t)|^2 t^{4r+2p+1} dt \right) = O\left(\sum_{n=N}^{\infty} \int\limits_n^{n+1} |g(t)|^2 t^{4r+2p+1} dt \right)$$

$$= O\left(\sum_{n=N}^{\infty} n^{4r} \int\limits_n^{n+1} |g(t)|^2 t^{2p+1} dt \right) = O\left(\sum_{n=N}^{\infty} n^{2r} \left[\int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt - \int\limits_{n+1}^{+\infty} |g(t)|^2 t^{2p+1} dt \right] \right)$$

$$= O\left(\sum_{n=N}^{\infty} n^{4r} \int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt - \sum_{n=N}^{\infty} n^{4r} \int\limits_{n+1}^{+\infty} |g(t)|^2 t^{2p+1} dt \right)$$

$$= O\left(N^{4r} \int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt + \sum_{n=N}^{\infty} [(n+1)^{4r} - n^{4r}] \int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt \right)$$

$$= O\left(n^{4R} \int\limits_N^{+\infty} |g(t)|^2 t^{2p+1} dt + \sum_{n=N}^{\infty} n^{4r-1} \int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt \right)$$

$$= O(N^{4r} N^{-4r-2k\nu}) + O(N^{-2k\nu}) = O(N^{-2k\nu}) = O(h^{2k\nu}),$$

i.e.,

$$I_2 = O(h^{2k\nu}).$$

Now we estimate I_1. By virtue of (13),

$$I_1 = \int\limits_{0<t<N} |g(t)|^2 t^{4r+2p+1} [1 - j_p(th)]^{2k} dt = O(h^{4k}) \int\limits_{0<t<N} |g(t)|^2 t^{4r+2p+4k+1} dt$$

$$= O(h^{4k}) \sum_{n=0}^{N} \int\limits_n^{n+1} |g(t)|^2 t^{4r+2p+4k+1} dt = O(h^{4k}) \sum_{n=0}^{N} (n+1)^{4r+4k} \int\limits_n^{n+1} |g(t)|^2 t^{2p+1} dt$$

$$= O(h^{4k}) \sum_{n=0}^{N} (n+1)^{4r+4k} \left[\int\limits_n^{+\infty} |g(t)|^2 t^{2p+1} dt - \int\limits_{n+1}^{+\infty} |g(t)|^2 t^{2p+1} dt \right]$$

$$= O(h^{4k})\left[\sum_{n=0}^{N}(n+1)^{4r+4k}\int_{n}^{+\infty}|g(t)|^2 t^{2p+1}dt - \sum_{n=0}^{\infty}(n+1)^{4r+4k}\int_{n+1}^{\infty}|g(t)|^2 t^{2p+1}dt\right]$$

$$= O(h^{4k})\left[1+\sum_{n=1}^{N}(n+1)^{4r+4k-1}\int_{n}^{+\infty}|g(t)|^2 t^{2p+1}dt\right]$$

$$= O(h^{4k})\left(1+\sum_{n=1}^{N}n^{4r+4k-1}n^{-4r-2k\nu}\right) = O(h^{4k})(1+N^{4k-2k\nu}) = O(h^{2k\nu}),$$

i.e., $I_1 = O(h^{2k\nu})$.

Combining the estimates for I_1 and I_2 gives

$$\left\|\Delta_h^k \mathcal{D}^r f(x)\right\| = O(h^{k\nu}),$$

which means that $f \in W_{2,\Phi}^{r,k}\mathcal{D}$. The necessity is proved.

REFERENCES

1. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals* (Clarendon, Oxford, 1948; Komkniga, Moscow, 2005).
2. A. N. Tikhonov and A. A. Samarskii, *Equations of Mathematical Physics* (Pergamon Press, Oxford, 1964; Gostekhteorizdat, Moscow, 1972).
3. V. S. Vladimirov, *Equations of Mathematical Physics* (Marcel Dekker, New York, 1971; Nauka, Moscow, 1976).
4. A. G. Sveshnikov, A. N. Bogolyubov, and V. V. Kravtsov, *Lectures on Mathematical Physics* (Nauka, Moscow, 2004) [in Russian].
5. G. N. Watson, *Treatise on the Theory of Bessel Functions* (Inostrannaya Literatura, Moscow, 1949; Cambridge Univ. Press, Cambridge, 1952).
6. A. I. Zayed, *Handbook of Function and Generalized Function Transformations* (CRC, Boca Raton, 1996).
7. W. N. Everitt and W. Kalf, "The Bessel Differential Equation and the Hankel Transform," J. Comput. Appl. Math. **208** (1), 3–19 (2007).
8. V. A. Abilov, F. V. Abilova, and M. K. Kerimov, "Some Remarks Concerning the Fourier Transform in the Space L_2(R)," Zh. Vychisl. Mat. Mat. Fiz. **48**, 939–945 (2008) [Comput. Math. Math. Phys. **48**, 885–891 (2008)].
9. I. A. Kipriyanov, *Singular Elliptic Boundary Value Problems* (Nauka, Fizmatlit, Moscow, 1997) [in Russian].
10. B. M. Levitan, "Expansion in Fourier Series and Integrals over Bessel Functions," Usp. Mat. Nauk **6** (2(42)), 102–148 (1951).
11. V. A. Abilov and F. V. Abilova, "Approximation of Functions by Fourier–Bessel Sums," Izv. Vyssh. Uchebn. Zaved., Mat., No. 8, 3–9 (2001).
12. S. M. Nikol'skii, *Approximation of Functions of Several Variables and Embedding Theorems* (Nauka, Moscow, 1977) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1111–1118. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © A.V. Arutyunov, B.D. Gel'man, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1167–1174.

Minimum of a Functional in a Metric Space and Fixed Points

A. V. Arutyunov[a] and B. D. Gel'man[b]

[a] *Peoples Friendship University, ul. Miklukho-Maklaya 6, Moscow, 117198 Russia*
[b] *Voronezh State University, Universitetskaya pl. 1, Voronezh, 394693 Russia*
e-mail: arutun@orc.ru, gelman_boris@mail.ru

Received November 6, 2008

Abstract—The existence of minimizers is examined for a function defined on a metric space. Theorems are proved that assert the existence of minimizers, and examples of the functions for which these theorems are valid are given. Then, these theorems are applied to proving theorems on fixed points of univalent and multivalued mappings of metric spaces. Finally, coincident points of two mappings are examined.

DOI: 10.1134/S0965542509070045

Key words: minimum of function, contraction mapping, fixed point, solution to equation

INTRODUCTION

In this paper, we examine the existence of minimizers for a function f defined on a metric space X. Theorems are proved that assert the existence of minimizers, and examples of the functions for which these theorems are valid are given.

Then, these theorems are applied to proving theorems on fixed points of univalent and multivalued mappings of metric spaces. In particular, we prove a theorem generalizing Banach's principle of contraction mappings and its extension given in [3]. We also show that the Nadler theorem (the principle of multivalued contraction mappings) and certain extensions of this principle are implications of our theorems. Finally, we examine coincident points of two mappings and obtain certain results that extend the ones presented in [1].

1. MINIMUM OF A FUNCTIONAL IN A METRIC SPACE

In what follows, X and Y are metric spaces with the metric ρ and d, respectively. Furthermore, X is assumed to be a complete space. In the Cartesian product $X \times Y$, we define the distance between the points (x_1, y_1) and (x_2, y_2) by the formula $\rho(x_1, x_2) + d(y_1, y_2)$, which makes it a metric space.

Let $f: X \longrightarrow \mathbb{R} \cup \{\infty\}$ be a function satisfying the following conditions:

(f1) $f(x) \geq 0$, $x \in X$;

(f2) if a sequence $\{x_n\} \subset X$ converges to x_* and $\lim\limits_{i \to \infty} f(x_n) = 0$, then $f(x_*) = 0$.

It is obvious that, if f satisfies condition (f1) and is continuous (or lower semicontinuous), then it satisfies condition (f2). We give another example of a function satisfying the above conditions.

Proposition 1. *Let $A \subset X$ be a given subset and $h : X \longrightarrow Y$ be a given continuous mapping. Assume that $g : A \longrightarrow Y$ is a mapping whose graph $\{(x, y) : x \in A, y = g(x)\}$ is closed in $X \times Y$ (this means that, if a sequence $\{x_n\} \subset A$ converges to x_* and $\{g(x_n)\} \longrightarrow y_*$, then $x_* \in A$ and $g(x_*) = y_*$). Then, the function*

$$f(x) = \begin{cases} d(h(x), g(x)), & x \in A \\ \infty, & x \notin A, \end{cases}$$

satisfies conditions (f1) *and* (f2).

Proof. The validity of condition (f1) is obvious. Let us verify the validity of (f2). Indeed, assume that a sequence $\{x_n\} \subset A$ converges to x_* and $\lim\limits_{i \to \infty} f(x_n) = 0$. Since h is a continuous mapping, we have $h(x_n) \longrightarrow h(x_*)$. The assumption $f(x_n) = d(h(x_n), g(x_n)) \longrightarrow 0$ implies that $g(x_n) \longrightarrow h(x_*)$. Since the graph of g is closed, we conclude that $x_* \in A$ and $g(x_*) = h(x_*)$; that is, $f(x_*) = 0$.

Remark 1. Suppose that X is the space of continuous functions on the interval [a, b], $[a, b]$, $h : X \longrightarrow X$ is a continuous mapping, A is the set of continuously differentiable functions on the above interval, and $g : A \longrightarrow X$ is the operator of differentiation. It is obvious that, in this case, the hypotheses of Proposition 1 are fulfilled. Moreover, Proposition 1 holds even if X is not assumed to be a complete space.

Now, we examine the question of whether there exist points at which f attains its minimum. To this end, we prove the following lemma, which seems to be of independent interest.

Lemma 1. *Let f satisfy condition* (f2) *and* $\inf\limits_{x \in X} f(x) = 0$. *Let $x_0 \in X$ and $f(x_0) < \infty$. Assume that there exist scalars $c > 0$ and $k \in [0, 1]$ and a point sequence $\{x_n\}_{n=0}^{\infty} \subset X$ satisfying the conditions*

$$f(x_{n+1}) \le kf(x_n) \quad \forall n = 0, 1, \ldots, \tag{1}$$

$$\rho(x_n, x_{n+1}) \le cf(x_n) \quad \forall n = 0, 1, \ldots. \tag{2}$$

Then, there exists a minimizer x_* for the function f, and it holds that

$$\rho(x_0, x_*) \le \frac{cf(x_0)}{1-k}. \tag{3}$$

Proof. Let $\{x_n\}_{n=0}^{\infty}$ be a point sequence satisfying the conditions of the lemma. If $f(x_n) = 0$ for some point x_n in this sequence, then the lemma is proved. Suppose that $f(x_n) \ne 0$ for all n; then, inequality (1) implies that $0 < f(x_n) \le k^n f(x_0)$. Hence, we have $\lim\limits_{n\to\infty} f(x_n) = 0$.

On the other hand, from inequality (2) we deduce that

$$\rho(x_{n-1}, x_n) \le cf(x_{n-1}) \le k^{n-1} cf(x_0).$$

Thus, $\{x_n\}$ is a fundamental sequence. Since X is a complete space, there exists $x_* = \lim\limits_{n\to\infty} x_n$. In view of (f2), we conclude that

$$\lim_{n\to\infty} f(x_n) = 0 = f(x_*).$$

Now, we estimate the distance $\rho(x_0, x_*)$. We have

$$\rho(x_0, x_n) \le \sum_{i=0}^{n-1} \rho(x_i, x_{i+1}) \le \sum_{i=0}^{n-1} k^i cf(x_0) < \frac{cf(x_0)}{1-k}.$$

Passing to the limit over n in this inequality, we obtain $\rho(x_0, x_*) \le \dfrac{cf(x_0)}{1-k}$. The lemma is proved.

Theorem 1. *Let f satisfy conditions* (f1) *and* (f2). *Assume that there are scalars $c > 0$ and $k \in [0, 1]$ such that, for any point $x \in X$, there exists a point $x' \in X$ satisfying the inequalities*

$$f(x') \le kf(x), \tag{4}$$

$$\rho(x', x) \le cf(x). \tag{5}$$

Then, for any point $x_0 \in X : f(x_0) < \infty$, there exists a point $x_* \in X$ such that $f(x_*) = 0$ and $\rho(x_0, x_*) \le \dfrac{cf(x_0)}{1-k}$.

Proof. In view of (4) and (5), there exists a point sequence $\{x_n\}_{n=0}^{\infty} \subset X$ satisfying the conditions of Lemma 1.

Remark 2. The key condition of Theorem 1 is the one requiring the existence of a point x' that satisfies (4) and (5). This condition was first proposed in [4], where a method was developed for the discrete descent over the values of special metric functionals.

Corollary 1. Let f satisfy conditions (f1) and (f2). Let $x_0 \in X$ and $f(x_0) < \infty$. Assume that there exist a (possibly discontinuous) mapping $\varphi : X \longrightarrow X$ and scalars $c > 0$ and $k \in [0, 1]$ such that

$$f(\varphi(x)) \le kf(x), \tag{6}$$

$$\rho(x, \varphi(x)) \le cf(x). \tag{7}$$

Then, there exists a point $x_* \in X$ such that $f(x_*) = 0$ and $\rho(x_0, x_*) \le \dfrac{cf(x_0)}{1-k}$.

Proof. Let $x_0 \in X$ be a point for which $f(x_0) < \infty$. Consider the iterative sequence $x_n = \varphi(x_{n-1})$, $n = 1, 2, \ldots$. It is easy to verify that this sequence satisfies the conditions of Lemma 1, which yields the required result.

Now, we state a simple sufficient condition for the uniqueness of a minimizer of the function f.

Proposition 2. *Let f satisfy the hypotheses of Theorem* 1. *Assume that, for every $x, y \in X$, there exists a scalar β such that*

$$\rho(x, y) \le \beta(f(x) + f(y)).$$

Then, f has a unique minimizer.

This assertion is obvious.

We give examples of the functions for which the conditions of the above propositions are fulfilled.

It is easy to see that, for a contraction mapping $g : X \longrightarrow X$ having the Lipschitz constant $k < 1$, the function $f(x) = \rho(x, g(x))$ satisfies conditions (6) and (7); that is, $f(g(x)) \le kf(x)$ and $\rho(x, g(x)) \le f(x)$.

Example 1. Denote by $C(X)$ the set of nonempty closed subsets of X. Let $G: X \longrightarrow C(X)$ be a multivalued contraction mapping; that is, there exists a scalar $k_0 \in [0, 1)$ such that

$$h(G(x), G(y)) \le k_0 \rho(x, y),$$

for all $x, y \in X$. Here, h is the Hausdorff metric generated by ρ.

Consider the function $f: X \longrightarrow \mathbb{R}$ defined by $f(x) = \rho(x, G(x))$. This is a continuous function because

$$|f(x) - f(y)| = |\rho(x, G(x)) - \rho(y, G(y))| \le \rho(x, y) + h(G(x), G(y)) \le (1 + k_0)\rho(x, y).$$

We show that this function satisfies inequalities (4) and (5). Choose a scalar $\alpha > 0$ such that $k = k_0(1 + \alpha) < 1$. Let x' be an arbitrary point in $G(x)$ such that $\rho(x, x') < (1 + \alpha)\rho(x, G(x)) = (1 + \alpha)f(x)$; that is, inequality (5) is satisfied with $c = 1 + \alpha$. Such a point always exists.

Now, we verify the validity of (4). We have

$$f(x') = \rho(x', G(x')) \le h(G(x), G(x')) \le k_0 \rho(x, x') < k\rho(x, G(x)) = kf(x),$$

that is, (4) is fulfilled. Thus, f satisfies the conditions of Theorem 1.

Let us give an example of a function that satisfies of Lemma 1 but does not satisfy the conditions of Theorem 1.

Example 2. Let $x_0 \in X$. Denote by $B_R(x_0)$ the closed ball of radius R centered at x_0. Let $G: B_R(x_0) \longrightarrow C(X)$ be a multivalued contraction mapping with the constant k_0 satisfying the condition

$$\rho(x_0, G(x_0)) < (1 - k_0)R.$$

Consider the function $f: X \longrightarrow \mathbb{R} \cup \infty$, defined by

$$f(x) = \begin{cases} \rho(x, G(x)), & x \in B_R(x_0) \\ \infty, & x \notin B_R(x_0). \end{cases}$$

Since this function is continuous on the closed ball $B_R(x_0)$, it is lower semicontinuous. Let us show that we can construct a sequence that satisfies conditions (1) and (2).

Choose a scalar $\alpha > 0$ such that $k = k_0(1 + \alpha) < 1$ and

$$r = (1 + \alpha)\rho(x_0, G(x_0)) < (1 - k)R.$$

These inequalities are fulfilled if

$$0 < \alpha < \min\left\{\frac{(1 - k_0)R - \rho(x_0, G(x_0))}{\rho(x_0, G(x_0)) + k_0 R}; \frac{1 - k_0}{k_0}\right\}.$$

We want to construct a sequence $\{x_n\}_{n=0}^{\infty}$ such that

(a) $x_n \in B_R(x_0)$,

(b) $x_n \in G(x_{n-1})$,

(c) $\rho(x_{n-1}, x_n) \le (1 + \alpha)\rho(x_{n-1}, G(x_{n-1}))$.

We construct this sequence by induction. The point x_0 is the same as in the condition of this example. Let x_1 be an arbitrary point in $G(x_0)$ such that

$$\rho(x_0, x_1) \le r = (1 + \alpha)\rho(x_0, G(x_0)).$$

Thus, this point satisfies conditions (a), (b), and (c). Such a point always exists.

Suppose that we have already constructed the first $n + 1$

members of our sequence, that is, the points $x_0, \ldots, x_n$. Then, $\rho(x_n, G(x_n)) < (1 + \alpha)\rho(x_n, G(x_n))$. Consequently, there exists a point $x_{n+1} \in G(x_n)$ such that $\rho(x_{n+1}, x_n) \le (1 + \alpha)\rho(x_n, G(x_n))$. Let us verify that this point belongs to the ball $B_R(x_0)$.

We have

$$\rho(x_i, G(x_i)) \le h(G(x_{i-1}), G(x_i)) \le k_0\rho(x_i, x_{i-1}) \le k_0(1 + \alpha)\rho(x_{i-1}, G(x_{i-1}))$$
$$\le k_0^i(1 + \alpha)^i\rho(x_0, G(x_0)) \quad \forall i = 1, 2, \ldots, n.$$

Consequently, it holds that

$$\rho(x_i, x_{i+1}) \le (1 + \alpha)\rho(x_i, G(x_i)) \le k_0^i(1 + \alpha)^{i+1}\rho(x_0, G(x_0)).$$

Then, we obtain

$$\rho(x_0, x_{n+1}) \le \sum_{i=0}^{n}\rho(x_i, x_{i+1}) \le \sum_{i=0}^{n}k_0^i(1 + \alpha)^{i+1}\rho(x_0, G(x_0)) \le \frac{(1 + \alpha)\rho(x_0, G(x_0))}{1 - k} < R.$$

Thus, we have constructed the sequence $\{x_n\}_{n=0}^{\infty}$

Now, we verify that this sequence satisfies inequalities (1) and (2). Indeed, in view of condition (c), we have

$$\rho(x_{n-1}, x_n) \le (1 + \alpha)\rho(x_{n-1}, G(x_{n-1})) = (1 + \alpha)f(x_{n-1});$$

that is, condition (2) is fulfilled.

Furthermore, it holds that

$$f(x_n) = \rho(x_n, G(x_n)) \le h(G(x_{n-1}), G(x_n)) \le k_0\rho(x_{n-1}, x_n)$$
$$\le k_0(1 + \alpha)\rho(x_{n-1}, G(x_{n-1})) = kf(x_{n-1}),$$

that is, condition (1) is fulfilled.

2. SOLVABILITY OF EQUATIONS IN METRIC SPACES

We apply Theorem 1 and its corollaries to prove certain theorems on the existence of solutions to operator equations and inclusions in metric spaces.

Let $g : X \longrightarrow X$ be a mapping.

Definition 1. We say that g is a weakly *lower semicontinuous mapping* if the function $f(x) = \rho(x, g(x))$ is lower semicontinuous.

It is obvious that a continuous mapping is always weakly lower semicontinuous and the inverse statement is false.

Theorem 2. *Let g be a weakly lower semicontinuous mapping. Assume that there are scalars $k \in [0, 1)$ and $c > 0$ such that, for any point $x \in X$, there exists a point $x' \in X$ satisfying the inequalities*

$$\rho(x', g(x')) \le k\rho(x, g(x)), \tag{8}$$
$$\rho(x, x') \le c\rho(x, g(x)). \tag{9}$$

Then, g has a fixed point x_* such that

$$\rho(x_0, x_*) \le \frac{c\rho(x_0, g(x_0))}{1 - k}. \tag{10}$$

We can derive this assertion from Theorem 1 by considering the function $f(x) = \rho(x, g(x))$.

Banach's principle of contraction mappings is a natural implication of this theorem; namely, if g is a contraction mapping, then $g(x)$ should be taken as the point x'. Then, we have

$$f(x') = \rho(x', g(x')) \le k\rho(x, g(x)) = kf(x)$$

and

$$\rho(x, x') = f(x),$$

Thus, the hypotheses of Theorem 2 are fulfilled. In a similar way, a theorem on fixed points analogous to the one given in [3] can also be derived from Theorem 2.

Below, we use the following definition.

For points $x_1, x_2 \in X$, the set of all points $x \in X$ such that

$$\rho(x_1, x) + \rho(x, x_2) = \rho(x_1, x_2).$$

is called the segment $[x_1, x_2]$.

Theorem 3. *Let* $g: X \longrightarrow X$ *be a weakly lower semicontinuous mapping. Assume that there are scalars* $\lambda_0 > 0$ *and* $k_0 \in (0, 1)$ *such that, for any point* $x \in X$, *there exists a point* $x' \in [x, g(x)]$ *satisfying the inequalities*

$$\rho(g(x), g(x')) \le k_0 \rho(x, x'), \tag{11}$$

$$\lambda_0 \rho(x, g(x)) \le \rho(x, x'). \tag{12}$$

Then, for any point $x_0 \in X$, *there exists a fixed point* x_* *of the mapping* g *for which*

$$\rho(x_0, x_*) \le \frac{\rho(x_0, g(x_0))}{\lambda_0(1 - k_0)}. \tag{13}$$

Proof. To prove this theorem, we verify that the function $f(x) = \rho(x, g(x))$ satisfies the conditions of Theorem 2.

Let x be an arbitrary point in X, and let x' be a point satisfying the conditions of Theorem 3. Let us verify that inequalities (8) and (9) are fulfilled. This is obvious for inequality (9) because $\rho(x, x') \le \rho(x, g(x)) = f(x)$. As for inequality (8), we have

$$f(x') = \rho(x', g(x')) \le \rho(x', g(x)) + \rho(g(x), g(x')) \le \rho(x, g(x)) - \rho(x, x') + k_0 \rho(x, x')$$

$$\le \rho(x, g(x)) - (1 - k_0)\rho(x, x').$$

Since x' satisfies inequality (12), it holds that $\lambda_0 \rho(x, g(x)) \le \rho(x, x')$. This implies that

$$\rho(x, g(x)) - (1 - k_0)\rho(x, x') \le k\rho(x, g(x)) = kf(x),$$

where $k = 1 - \lambda_0(1 - k_0)$. Thus, inequality (8) is fulfilled.

Now, the assertion of Theorem 3 follows from Theorem 2. Moreover, inequality (13) is obtained from (10) by setting $c = 1$ and $k = 1 - \lambda_0(1 - k_0)$.

Remark 3. In [3], mappings satisfying inequality (11) are said to be contracting along a direction. An example is given in [3] of a mapping g that is contracting along a direction but is not a contraction mapping.

Theorem 2 can also be used for proving the existence of fixed points in the case of discontinuous mappings. Here is a very simple example.

Example3. Let $X = \mathbb{R}$, and let $g: X \longrightarrow X$ be the mapping defined by

$$g(x) = \begin{cases} 3x, & x \ne 1/n \\ x/2, & |x| = 1/n, \end{cases}$$

where n is a positive integer. It is obvious that $x = 0$ is a unique fixed point of this mapping.

Consider the function $f(x) = |x - g(x)|$; that is,

$$f(x) = \begin{cases} 2|x|, & x \ne 1/n \\ |x|/2, & |x| = 1/n. \end{cases}$$

We show that this function satisfies the conditions of Theorem 2. It is obvious that it is lower semicontinuous; that is, g is a weakly lower semicontinuous mapping. If $x = 0$, then we also take 0 as x'. If $x \ne 0$, then we take as x' a scalar of the same sign as that of x. The scalar $|x'|$ must belong to the interval $(0, |x|/2)$, and $|x'| = 1/n$, where n is a positive integer. Then, $f(x') \le \frac{1}{2} f(x)$ and $|x - x'| \le 2f(x)$; that is, f satisfies the conditions of Theorem 2.

Consider another problem. Let $b \in X$ be a fixed point.

Definition 2. We say that g is a weakly *lower semicontinuous mapping with respect to* b if the function $f(x) = \rho(b, g(x))$ is lower semicontinuous.

We are interested in the solvability of the equation

$$g(x) = b. \tag{14}$$

Theorem 4. *Let* g *be a weakly lower semicontinuous mapping with respect to* b. *Assume that there are scalars* $k \in [0, 1)$ *and* $c > 0$ *such that, for any point* $x \in X$, *there exists a positive integer* $n = n(x)$ *for which we have*

(i) $\rho(b, g^n(x)) \le k\rho(b, g(x))$,

(ii) $\rho(x, g^n(x)) \le c\rho(b, g(x))$.

Let x_0 be an arbitrary point in X. Then, Eq. (14) *has a solution x_*, and it holds that*

$$\rho(x_0, x_*) \le \frac{c\rho(b, g(x_0))}{1-k}.$$

Proof. Consider the mapping $\varphi : X \longrightarrow X$ defined by $\varphi(x) = g^{n(x)}(x)$. It is obvious that this mapping satisfies the conditions of Corollary 1, which proves the theorem.

Remark 4. This theorem provides a basis for the development of algorithms for solving Eq. (14).

Now, we turn to the study of fixed points of multivalued mappings. Let $G : X \longrightarrow C(X)$ be a multivalued contraction mapping. It is obvious that Theorem 1 and Example 1 imply the well-known Nadler theorem (e.g., see [5]). It is also obvious that Lemma 1 and Example 2 imply the extension of the Nadler theorem given in [2]. We consider another application of Theorem 1.

Let A be a closed subset of X, and let $G : X \longrightarrow C(X)$ be a multivalued contraction mapping. Assume that $A \cap G(x) \ne \varnothing$ for every $x \in A$. We give an example of the situation where the mapping G satisfies the above assumptions but has no fixed points.

Example 4. Let $X = \mathbb{R}^2$ and $A = \{(x, 0) | x \in \mathbb{R}\}$, and let $G : \mathbb{R}^2 \longrightarrow C(\mathbb{R}^2)$ be the mapping defined by

$$G(x) = \left\{(u, v) \middle| u - \frac{1}{2}v - x - 1 = 0\right\}.$$

Then, $h(G(x), G(y))$ is the least distance between two parallel lines and

$$h(G(x), G(y)) = \frac{2|x-y|}{\sqrt{5}};$$

that is, G is a contraction mapping.

Observe that $G(x) \cap A = \{(x + 1, 0)\}$. Thus, this intersection is nonempty, it is not a contraction mapping, and it has no fixed points.

We use Theorem 1 to study the following question: under what conditions a multivalued mapping $G : A \longrightarrow C(X)$ has fixed points?

We assume that G satisfies the following conditions:

(i) the function $f_0(x) = \rho(x, G(x) \cap A)$ is lower semicontinuous;

(ii) the intersection $A \cap G(x) \ne \varnothing$ for all $x \in A$.

The multivalued contraction mapping in Example 4 obviously satisfies these conditions.

Theorem 5. *Let a multivalued mapping G satisfy conditions* (*i*) *and* (*ii*). *In addition, assume that the following condition is fulfilled*:

(iii) *there are scalars $k \in [0, 1)$ and $c > 0$ such that, for any point $x \in A$, there exists a point $x' \in (A \cap G(x))$ satisfying the inequalities*

$$\rho(x', G(x')) \le k\rho(x, G(x)), \quad \rho(x, x') \le c\rho(x, G(x)).$$

Then, for any point $x_0 \in A$, the multivalued mapping $G(x)$ has a fixed point $x_* \in A$ such that

$$\rho(x_0, x_*) \le \frac{c\rho(x_0, G(x_0))}{1-k}.$$

Proof. Define the function $f : X \longrightarrow \mathbb{R}_+$ by the formula

$$f(x) = \begin{cases} f_0(x), & x \in A \\ \infty, & x \notin A. \end{cases}$$

It is easy to see that, under the assumptions made above, f is lower semicontinuous. Then, $f(x_0) < \infty$ for every point $x_0 \in A$, and the conditions of Theorem 1 are fulfilled. It follows that there exists a point $x_* \in A$ such that $f(x_*) = 0$; that is, $\rho(x_*, G(x_*)) = 0$. Since $G(x_*)$ is a closed set, we have $x_* \in G(x_*)$. The theorem is proved.

Let B and C be subsets of X. Denote by $\rho_*(B, C) = \sup\limits_{b \in B} \inf\limits_{c \in C} \rho(b, c)$ the semideviation of B from C.

Corollary 2. Let a multivalued mapping $G(x)$ satisfy conditions (i) and (ii). Assume that there is a scalar $k_0 \in (0, 1)$ such that, for any point $x \in A$, there exists a point $x' \in (A \cap G(x))$ satisfying the inequality

$$\rho_*(G(x) \cap A, G(x') \cap A) \le k_0 \rho(x, x').$$

Then, for any point $x_0 \in A$ and any scalar $k \in (k_0, 1)$, the multivalued mapping $G(x)$ has a fixed point $x_* \in A$ such that

$$\rho(x_0, x_*) \le \frac{k\rho(x_0, G(x_0))}{k_0(1-k)}.$$

Proof. Consider an arbitrary point $x \in A$ and an arbitrary scalar k in the interval $(k_0, 1)$. Define $\varepsilon = (k - k_0)/k_0$; then, $k = k_0(1 + \varepsilon)$. The definition of the semideviation between two sets implies that there exists a point $x' \in (G(x) \cap A)$ such that

$$\rho(x, x') \le (1 + \varepsilon)\rho(x, G(x) \cap A) = (1 + \varepsilon)f(x).$$

Let us find a bound on $f(x')$. We have

$$f(x') = \rho(x', G(x') \cap A) \le \rho_*(G(x) \cap A, G(x') \cap A) \le k_0 \rho(x, x')$$

$$\le k_0(1 + \varepsilon)\rho(x, G(x) \cap A) = kf(x).$$

Now, the validity of the required assertion follows from Theorem 5, where we set $c = 1 + \varepsilon$ and $k = k_0(1 + \varepsilon)$.

3. COINCIDENT POINTS OF TWO MAPPINGS

Let A be an arbitrary subset of X. Consider two mappings, namely, a continuous mapping $h : X \longrightarrow Y$ and a mapping $g : A \longrightarrow Y$ having a closed graph.

We call $x_* \in A$ a coincident point of these mappings if $h(x_*) = g(x_*)$.

Theorem 6. *Assume that there are scalars $k \in [0, 1)$ and $c > 0$ such that, for any point $x \in A$, there exists a point $x' \in A$ satisfying the inequalities*

$$d(h(x'), g(x')) \le kd(h(x), g(x)),$$

$$\rho(x, x') \le cd(h(x), g(x)).$$

Then, the mappings h and g have a coincident point x_* such that

$$\rho(x_0, x_*) \le \frac{cd(h(x_0), g(x_0))}{1-k}.$$

To prove this theorem, consider the function f defined in Proposition 1. According to this proposition, f satisfies conditions (f1) and (f2). Hence, the required assertion follows from Theorem 1.

Some theorems on the coincidence of an α-covering mapping and a Lipschitz mapping were obtained in [1]. We examine a certain development of these theorems based on Theorem 6. First, we recall some relevant definitions. Let $B_R(x)$ be the closed ball of radius R centered at a point $x \in X$; similarly, $B_R(y)$ is a closed ball in the space Y.

We say that $g : A \subset X \longrightarrow Y$ is an α-covering mapping if there exists a scalar $\alpha > 0$ such that the inclusion

$$g(B_R(x) \cap A) \supset B_{\alpha R}(g(x)).$$

holds for all $x \in A$ and all $R > 0$.

Corollary 3. Let $g : A \subset X \longrightarrow Y$ be an α-covering mapping having a closed graph, and let $h : X \longrightarrow Y$ be a β-Lipschitz mapping, where $\beta < \alpha$. Then, for any point $x_0 \in A$, there exists a coincident point x_* of the mappings h and g such that

$$\rho(x_0, x_*) \le \frac{d(h(x_0), g(x_0))}{\alpha - \beta}.$$

Proof. We verify that the mappings g and h satisfy the conditions of Theorem 6 for $k = \beta/\alpha$ and $c = \alpha^{-1}$ Let x be an arbitrary point in A. Define $y = h(x) \in Y$ and $R = \dfrac{d(g(x), y)}{\alpha}$.

Since g is an α-covering mapping, there exists a point $x' \in B_R(x) \cap A$ such that $g(x') = y$. We show that x' is the desired point. Indeed, we have

$$d(g(x'), h(x')) = d(y, h(x')) = d(h(x), h(x')) \le \beta\rho(x, x') \le \frac{\beta}{\alpha} d(g(x), h(x)) = kd(g(x), h(x)),$$

which proves the first inequality. Let us verify the second inequality:

$$\rho(x, x') \le \frac{d(g(x), y)}{\alpha} = \frac{d(g(x), h(x))}{\alpha}.$$

Thus, the second inequality is also fulfilled. Now, the required assertion follows from Theorem 6.

For $A = X$, Corollary 3 can be derived from Theorem 2 given in [1]. Let us give an example of mappings satisfying the hypotheses of Theorem 6 but not the ones of Corollary 3.

Example 5. Consider the metric spaces $X = A = \mathbb{R}^2$ and $Y = \mathbb{R}$. For $x = (x_1, x_2) \in X$, we define the mapping g by the formula $g(x) = x_1$. It is obvious that g is a 1-covering mapping. The mapping $h : X \longrightarrow Y$ is defined by the formula $h(x) = \frac{1}{2}x_1 + e^{x_2}$. We verify that these mappings satisfy all the conditions of Theorem 6 for $k = 1/2$ and $c = 1$.

Take an arbitrary point $x = (x_1, x_2) \in X$ and set $x' = (x_1', x_2')$, where $x_1' = h(x)$ and $x_2' = x_2$. Then, $g(x') = h(x)$, whence $|g(x') - h(x')| = |h(x) - h(x')| = |x_1 - x_1'|/2$. On the other hand, the choice of x' implies that $|x_1' - x_1| = |g(x) - h(x)|$. Consequently, we have $|g(x') - h(x')| = k|g(x) - h(x)|$. Since $\rho(x, x') = |g(x) - h(x)|$, we also have $\rho(x, x') = |x_1 - x_1'|$. Thus, all the hypotheses of Theorem 6 are fulfilled. At the same time, the hypotheses of Corollary 3 are not fulfilled because the mapping h does not satisfy the Lipschitz condition on $\mathbb{R}^2$.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 08-01-00192, 08-01-90267, and 08-01-90001.

REFERENCES

1. A. V. Arutyunov, "Covering Mappings in Metric Spaces and Fixed Points," Dokl. Akad. Nauk **76** (2), 665–668 (2007) [Dokl. Math. **76**, 665–668 (2007)].
2. A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremum Value Problems* (Nauka, Moscow, 1974) [in Russian].
3. G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications* (Plenum, New York, 1981; Radio i Svyaz', Moscow, 1987).
4. T. N. Fomenko, *On the Approximation of O priblizhenii of Coincidence Points and Common Fixed Points of a Set of Mappings of Metric Spaces*, Mat. Zametki **8** (2), 147–160 (2009).
5. S. B. Nadler, "Multi-Valued Contraction Mappings," Pasif. J. Math. **30**, 475–488 (1969).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1119–1127. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © I.Yu. Vygodchikova, S.I. Dudov, E.V. Sorin, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1175–1183.

External Estimation of a Segment Function by a Polynomial Strip

I. Yu. Vygodchikova, S. I. Dudov, and E. V. Sorin

Saratov State University, ul. Astrakhanskaya 42, Saratov, 410012 Russia

e-mail: VigodchikovaIY@info.sgu.ru, DudovSI@info.sgu.ru

Received October 20, 2008

Abstract—The problem is considered of constructing a least-width strip with a polynomial axis that contains the graph of a given continuous segment function. Convex analysis methods are used to obtain a criterion for solving the problem in a form comparable to the Chebyshev alternance. Sufficient conditions for the uniqueness of a solution are given, including those taking into account the differential properties of the segment function to be estimated.

DOI: 10.1134/S0965542509070057

Key words: estimation of a segment function, polynomial strip, subdifferential, alternance, snake problem.

1. INTRODUCTION

The derivation of estimates and the approximation of set-valued mappings is a direction of studies in nonsmooth analysis (see [1–5]). The simplest example of a set-valued mapping is a segment function. Below, we study the properties of the solution to the problem of external estimation of a given continuous segment function by a simpler segment function whose graph is a least-width polynomial strip containing the graph of the original segment function. Here, by a polynomial strip, we mean a strip of constant width in ordinate whose axis is the graph of a polynomial of given degree.

Let a segment function $F(t) = [f_1(t), f_2(t)]$ be given on an interval $[c, d]$ by two continuous functions $f_1(t)$ and $f_2(t)$ with finite values at the endpoints of the interval, and let $f_1(t) \le f_2(t)$ for all $t \in [c, d]$. In what follows, $P_n(A, t) = a_0 + a_1 t + \ldots + a_n t^n$ is understood as a polynomial of fixed degree n with the coefficient vector $A = (a_0, \ldots, a_n) \in \mathbb{R}^{n+1}$.

Consider the problem

$$\rho(A) \equiv \max_{t \in [c, d]} \max\{P_n(A, t) - f_1(t), f_2(t) - P_n(A, t)\} \longrightarrow \min_{A \in \mathbb{R}^{n+1}}. \tag{1.1}$$

Obviously, the segment $[P_n(A, t) - \rho(A), P_n(A, t) + \rho(A)]$ covers $F(t)$ for any $t \in [c, d]$ and $A \in \mathbb{R}^{n+1}$. Therefore, if we define

$$\rho^* = \min_{A \in \mathbb{R}^{n+1}} \rho(A), \quad \Omega_\rho = \operatorname{Arg\,min}_{A \in \mathbb{R}^{n+1}} \rho(A),$$

then the graph of the segment function $\Pi_n(A^*, t) = [P_n(A^*, t) - \rho(A^*), P_n(A^*, t) + \rho(A^*)]$ for $A^* \in \Omega_\rho$ is a strip of the least (in ordinate) width $2\rho^*$ that contains the graph of $F(\cdot)$. Problem (1.1) is referred to as the *problem of external estimation* of $F(\cdot)$ by a polynomial strip.

Obviously, when $f_1(t) \equiv f_2(t)$ for $t \in [c, d]$, problem (1.1) degenerates into the Chebyshev problem of uniform approximation of a continuous function with a polynomial of given degree. Simple examples show that problem (1.1) is not reduced to the Chebyshev problem of approximating the arithmetic mean of $f_1(t)$ and $f_2(t)$. On the other hand, $\max\{P_n(A, t) - f_1(t), f_2(t) - P_n(A, t)\}$ involved in the objective function $\rho(A)$ is the Hausdorff distance between the segment $F(t)$ and the value of the polynomial $P_n(A, t)$. Therefore, (1.1) can also be treated as the problem of the best uniform approximation of $F(\cdot)$ by a polynomial of given degree in the Hausdorff metric. In this context, an association arises with the approximation of the graph of a segment function by the graph of a polynomial but in the Hausdorff metric of a two-dimensional space. This problem has been considered by many authors (see, e.g., [6]), but, in contrast to (1.1), it is not a convex programming problem.

It is also pertinent to recall the snake problem (see, e.g., [7, p. 34]), in which polynomials of given degree n (lower and upper snakes) are sought whose graphs $n + 1$ times touch those of given continuous functions $g_1(\cdot)$ and $g_2(\cdot)$ on an interval, provided that $g_1(t) < g_2(t)$ on the entire interval and the graphs of the polynomials are contained in that of the segment function $\Phi(t) = [g_1(t), g_2(t)]$. Below, we show that, under certain conditions, the solution to problem (1.1) solves a snake problem, but, for such a snake, there is a "redundant" alternance in the sense that the graphs of some functions $g_1(\cdot)$ and $g_2(\cdot)$ are touched in turn at least $n + 2$ times by the snake.

We can also consider the problem

$$\pi(A) \equiv \max_{t \in [c, d]} \max\{f_1(t) - P_n(A, t), P_n(A, t) - f_2(t))\} \longrightarrow \min_{A \in \mathbb{R}^{n+1}}. \tag{1.2}$$

Define

$$\pi^* = \min_{A \in \mathbb{R}^{n+1}} \pi(A), \quad \Omega_\pi = \operatorname{Argmin}_{A \in \mathbb{R}^{n+1}} \pi(A).$$

If $\pi^* \le 0$, then $2|\pi^*|$ is the maximum width (in ordinate) of a polynomial strip contained in the graph of $F(\cdot)$. For $A^* \in \Omega_\pi$, the graph of the polynomial $P_n(A^*, t)$ specifies the axis of this strip. In this situation, (1.2) can be regarded as the *problem of internal estimation* of $F(\cdot)$ by a polynomial strip. However, a situation is possible when $F(\cdot)$ has no polynomial selector of degree n, which corresponds to the case of $\pi^* > 0$. Then π^* is the maximum deviation of the value of the optimal polynomial $P_n(A^*, t)$ from $F(t)$ on the interval $[c, d]$. However, if $f_1(\cdot)$ and $f_2(\cdot)$ in problem (1.2) are replaced by the respective functions

$$\hat{f}_1(t) = f_1(t) - c, \quad \hat{f}_2(t) = f_2(t) + c,$$

where $c \ge \pi^*$, then the new problem, which is equivalent to the old one, can again be geometrically interpreted as the problem of internal estimation of the segment function $\hat{F}(t) = [\hat{f}_1(t), \hat{f}_2(t)]$.

Moreover, since $f_1(t) \le f_2(t)$, we conclude that, by replacing $f_1(\cdot)$ and $f_2(\cdot)$ in (1.2) with the respective functions

$$\tilde{f}_1(t) = f_2(t) - m, \quad \tilde{f}_2(t) = f_1(t) + m,$$

where $m \ge \max_{t \in [c, d]}(f_2(t) - f_1(t))/2$, problem (1.2) becomes the equivalent problem of external estimation for $\tilde{F}(t) = [\tilde{f}_1(t), \tilde{f}_2(t)]$. Therefore, problem (1.2) is of no importance other than that of problem (1.1).

Finally, we note that problem (1.1) was considered in [8, 9] in a discrete setting when $F(\cdot)$ is specified on a finite grid of argument values.

It is easy to see that $\rho(A)$ is convex and finite on $\mathbb{R}^{n+1}$. The study below is based primarily on convex analysis. In Section 2, we show that the solution to problem (1.1) always exists and obtain a criterion for solving the problem in a form comparable to the Chebyshev alternance. Simple examples show that problem (1.1) can have a nonunique solution. Section 3 gives sufficient conditions for the uniqueness of a solution, including those taking into account the differential properties of $F(\cdot)$.

2. EXISTENCE OF A SOLUTION TO THE PROBLEM AND A SOLUTION CRITERION

In what follows, co$B$ and int$B$ denote the convex hull and the interior of the set B, respectively, and $0_{n+1} = (0, \ldots, 0) \in \mathbb{R}^{n+1}$.

2.1. First, we prove the following result.

Theorem 1. *Problem* (1.1) *has a solution.*

Proof. For arbitrary $A^0 \in \mathbb{R}^{n+1}$, let $G(A^0) = \{A \in \mathbb{R}^{n+1}: \rho(A) \le \rho(A^0)\}$. Obviously, $\rho(A)$ is continuous (see, e.g., [10, p. 233]). Consequently, the set $G(A^0)$ is closed. Since

$$\inf_{A \in \mathbb{R}^{n+1}} \rho(A) = \inf_{A \in G(A^0)} \rho(A),$$

the existence of a solution will be proved by the Weierstrass theorem if we show that $G(A^0)$ is bounded.

Let $\{t_i\}$ $(i = \overline{1, n+1})$ be an arbitrary set of points from the interval $[c, d]$, and let $t_i \ne t_j$ for $i \ne j$. For given $y = (y_1, \ldots, y_{n+1})^{\mathrm{T}}$, consider a linear system of equations for $A \in \mathbb{R}^{n+1}$:

$$P_n(A, t_i) = y_i, \quad i = \overline{1, n+1},$$

It can be written as

$$VA = y, \tag{2.1}$$

where the matrix V has the form

$$\begin{pmatrix} 1 & t_1 & \dots & t_1^n \\ \dots & \dots & \dots & \dots \\ 1 & t_{n+1} & \dots & t_{n+1}^n \end{pmatrix}.$$

This matrix (see, e.g., [10, p. 16]) is nonsingular. Thus, it has an inverse matrix V^{-1} and the solution $A = V^{-1}y$ to system (2.1) is a continuous function of y. Since $f_1(t) \le f_2(t)$ for $t \in [c, d]$, it is easy to see that

$$G(A^0) = \left\{ A \in \mathbb{R}^{n+1} : \max_{t \in [c, d]} \max\{ \left| P_n(A, t) - f_1(A, t) \right|, \left| P_n(A, t) - f_2(A, t) \right| \le \rho(A^0) \} \right\}$$

$$\subset \left\{ A \in \mathbb{R}^{n+1} : \max_{t \in [c, d]} \left| P_n(A, t) \right| \le c \right\},$$

where $c = \rho(A^0) + \max_{t \in [c, d]} \max\{|f_1(t)|, |f_2(t)|\}$. This implies the estimate

$$G(A^0) \subset \left\{ A \in \mathbb{R}^{n+1} : \max_{i = \overline{1, n+1}} \left| P_n(A, t_i) \right| \le c \right\}$$

$$\subset \bigcup_{y \in Y} \{ A \in \mathbb{R}^{n+1} : P_n(A, t_i = y_i, i = \overline{1, n+1}) \} = \bigcup_{y \in Y} V^{-1} y,$$

where $Y = \{A \in \mathbb{R}^{n+1} : |y_i| \le c, i = \overline{1, n+1}\}$ is a cube in $\mathbb{R}^{n+1}$. Thus, the set $G(A^0)$ is estimated from above under inclusion by the range of a continuous function on a bounded set.

2.2. Let us formulate two auxiliary facts.

Let T be a subset of the real axis $\mathbb{R}$ on which a set-valued mapping $\xi(\cdot) : T \Rightarrow 2^{\mathbb{R}}$ is defined whose images are subsets $\xi(t)$ of $\mathbb{R}$. The following result is a generalization of that stated in [10, p. 242] and is proved in [11].

Lemma 1. *The inclusion*

$$0_{n+1} \in \operatorname{co}\{ \xi(t)(1, t, \dots, t^n) : t \in T \},$$

holds if and only if at least one of the following conditions is satisfied:

(i) *There exists a point* $t_0 \in T$ *at which* $0 \in \xi(t_0)$.

(ii) *There exists a selector* $\eta(t) \in \xi(t)$ *and a set of ordered numbers* $t_1 < \dots < t_{n+2}$ *from* T *such that* $\eta(t_i) \ne 0$ *and* $\operatorname{sgn} \eta(t_i) = -\operatorname{sgn} \eta(t_{i+1})$ *for* $i = \overline{1, n+1}$.

The following result was also proved in [11].

Lemma 2. *If a function* $\eta(t)$ *takes alternating values on an ordered set of points* $t_1 < \dots < t_{n+2}$, i.e., $\operatorname{sgn} \eta(t_i) = -\operatorname{sgn} \eta(t_{i+1})$, $i = \overline{1, n+1}$, *then*

$$0_{n+1} \in \operatorname{int} \operatorname{co}\{ \eta(t_i)(1, t_i, \dots, t_i^n) : i = \overline{1, n+2} \}.$$

2.3. Let us obtain a solution criterion for problem (1.1). Define

$$R_1(A) = \{ t \in [c, d] : \rho(A) = P_n(A, t) - f_1(t) > f_2(t) - P_n(A, t) \},$$
$$R_2(A) = \{ t \in [c, d] : \rho(A) = f_2(t) - P_n(A, t) > P_n(A, t) - f_1(t) \},$$
$$R_3(A) = \{ t \in [c, d] : \rho(A) = P_n(A, t) - f_1(t) = f_2(t) - P_n(A, t) \},$$
$$R(A) = R_1(A) \cup R_2(A) \cup R_3(A).$$

Theorem 2. *The function* $\rho(A)$ *takes its minimal value on* $\mathbb{R}^{n+1}$ *at a point* A^* if *and only if at least one of the following conditions is satisfied*:

(i) $R_3(A^*) \ne \emptyset$.

(ii) *There exists a ordered sequence of points* $t_1 < \ldots < t_{n+2}$ of $R_1(A^*) \cup R_2(A^*)$ *such that*, if $t_i \in R_1(A^*)(R_2(A^*))$, *then* $t_{i+1} \in R_2(A^*)(R_1(A^*))$, $i = \overline{1, n+1}$.

Proof. The function $\rho(A)$ is convex and finite on R^{n+1}. Its subdifferential (see, e.g., [1, 2, 12]) can be written as

$$\partial\rho(A) = \operatorname{co}\{\partial_A f(A, t) : t \in R(A)\}, \tag{2.2}$$

where $f(A, t) = \max\{P_n(A, t) - f_1(t), f_2(t) - P_n(A, t)\}$, and $\partial_A f(\cdot)$ is the subdifferential of $f(A, t)$ with respect to A. Moreover, according to subdifferential calculus, we have

$$\partial_A f(A, t) = \begin{cases} (1, t, \ldots, t^n), & P_n(A, t) - f_1(A, t) > f_2(A, t) - P_n(A, t) \\ -(1, t, \ldots, t^n), & f_2(A, t) - P_n(A, t) > P_n(A, t) - f_1(A, t) \\ [-1, 1](1, t, \ldots, t^n), & P_n(A, t) - f_1(A, t) = f_2(A, t) - P_n(A, t). \end{cases} \tag{2.3}$$

Here and below, the set B from $\mathbb{R}$ is defined as

$$B(1, t, \ldots, t^n) = \operatorname{co}\{b(1, t, \ldots, t^n) : b \in B\}.$$

Given a fixed value A, the multivalued function $\xi(A, t)$ on a set $R(A)$ is defined as

$$\xi(A, t) = \begin{cases} 1, & t \in R_1(A) \\ -1, & t \in R_2(A) \\ [-1, 1], & t \in R_3(A). \end{cases} \tag{2.4}$$

It follows from (2.2)–(2.4) that

$$\partial\rho(A) = \operatorname{co}\{\xi(A, t)(1, t, \ldots, t^n) : t \in R(A)\}. \tag{2.5}$$

As is known from convex analysis (see, e.g., [1, p. 142]), a necessary and sufficient condition for a point A^* to be a minimizer of a convex function $\rho(A)$ on $\mathbb{R}^{n+1}$ is the fulfillment of the inclusion $0_{n+1} \in \partial\rho(A^*)$. Now, in view of subdifferential formula (2.5) and expression (2.4) for $\xi(A, t)$, is remains to use Lemma 1.

2.4. There are some comments to be made on the solution criterion for problem (1.1).

1. If $f_1(t) \equiv f_2(t)$ for $t \in [c, d]$, then problem (1.1) degenerates into the Chebyshev problem of uniform approximation of a continuous function by a polynomial of given degree. Obviously, except for the trivial case, when the original function is itself a polynomial of a degree no higher than n, we always have $R_3(A) = \emptyset$; i.e., condition (i) in Theorem 2 is not satisfied, while condition (ii) is another expression for the alternance.

2. Simple examples show that, depending on the situation, the solution to problem (1.1) can have only one of the properties or both properties indicated in Theorem 2.

3. Let us establish the relation of problem (1.1) to the snake problem. Below is a direct consequence of Theorem 2.

Corollary 1. If A^* is a solution to problem (1.1) such that it satisfies condition (ii) of Theorem 2 and $R_3(A^*) = \emptyset$, then, for the functions $g_1(t)$ and $g_2(t)$, the graph of the polynomial $P_n(A^*, t)$ touches in turn the graphs of $g_1(\cdot)$ and $g_2(\cdot)$ at the points $\{t_i\}_{i=1}^{n+2}$. Moreover, $g_1(t) \le P_n(A^*, t) \le g_2(t)$, $t \in [c, d]$.

Thus, in this situation, in the sense of the definition in [7, p. 34], $P_n(A^*, t)$ is a snake (with a redundant alternance) for $g_1(t)$ and $g_2(t)$. Therefore, it is an upper snake for one of the systems of points $\{t_i\}_{i=1}^{n+1}$ and $\{t_i\}_{i=2}^{n+2}$ and a lower snake for the other. Note also that the formulation of the snake problem assumes $g_1(t) < g_2(t)$ for all $t \in [c, d]$. This is ensured by the condition $R_3(A) = \emptyset$.

Feedback of these problems is expressed as follows.

Corollary 2. Let the polynomial $P_n(A^*, t)$ be a snake with a redundant alternance for $g_1(t)$ and $g_2(t)$ satisfying $g_1(t) < g_2(t)$ for $t \in [c, d]$. Then A^* is a solution to problem (1.1) for the segment function $F(t) = [f_1(t), f_2(t)]$, where $f_1(t) = g_2(t) - m$, $f_2(t) = g_1(t) + m$, and m satisfies the inequality $m \ge \max_{t \in [c, d]}(g_2(t) - g_1(t))/2$.

4. If condition (ii) of Theorem 2 is satisfied, then formulas (2.4), (2.5) and Lemma 2 imply the inclusion $0_{n+1} \in \operatorname{int}\partial\rho(A^*)$, which, as is known from convex analysis, guarantees the uniqueness of a solution. The uniqueness of a solution is discussed in more detail in the following section.

3. CONDITIONS FOR THE UNIQUENESS OF A SOLUTION

It is well known that the Chebyshev problem of uniform approximation of a continuous function by polynomials of a given degree always has a unique solution. Below, we show that the uniqueness of a solution to problem (1.1) may depend on the properties of the estimated segment function, including the differential ones.

3.1. First, we give sufficient conditions for the uniqueness of a solution that make no use of the differential properties of $f_1(\cdot)$ and $f_2(\cdot)$.

Theorem 3. *Suppose that A^* satisfies one of the following conditions*:

(i) *The set $R_3(A^*)$ contains no less than $n+1$ points*.

(ii) *There exists a set of points $T = \{t_i\}_{i=1}^{n+2} \subset R(A^*)$ in which there are points $t_{i_1} < \ldots < t_{i_l}$ from $R_1(A^*) \cup R_2(A^*)$, $l \le n+2$, while the other points are contained in $R_3(A^*)$. Moreover, if $t_{i_k} \in R_1(A^*)(R_2(A^*))$, then $t_{i_{k+1}} \in R_1(A^*)(R_2(A^*))$ for even $(i_{k+1} - i_k)$ and $t_{i_{k+1}} \in R_2(A^*)(R_1(A^*))$ for odd $(i_{k+1} - i_k)$.*

Then A^* is a unique solution to problem (1.1).

Proof. First, we show that A^* is a minimizer of $\rho(A)$. Indeed, if condition (i) or (ii) is satisfied and $L < n+2$, then $R(A^*) \neq \varnothing$; i.e., condition (i) of Theorem 2 holds. If condition (ii) is satisfied and $l = n+2$, then condition (ii) of Theorem 2 holds. Thus, in any case, A^* is a minimizer of $\rho(A)$.

Now, we prove uniqueness. Obviously, $R_3(A)$ is invariant on the solution set Ω_ρ of problem (1.1). Moreover, if $R_3(A)$ is not empty, then

$$R_3(A) = \{\hat{t} \in [c, d] : f_2(\hat{t}) - f_1(\hat{t}) = \max_{t \in [c, d]} (f_2(t) - f_1(t))\} \quad \forall A \in \Omega_\rho.$$

Therefore, assuming that condition (i) of Theorem 3 is satisfied and $\{t_i\}_{i=1}^{n+1} \subset R_3(A^*)$, the definition of $R_3(\cdot)$ implies that

$$P_n(A, t_i) = (f_1(t_i) + f_2(t_i))/2, \quad i = \overline{1, n+1}, \tag{3.1}$$

for any $A \in \Omega_\rho$. The determinant of system (3.1), which is linear with respect to $A = (a_0, \ldots, a_n)$, is the Vandermonde determinant. Since it is nonzero, A^* is a unique solution to system (3.1) and, hence, to problem (1.1).

Now, let condition (ii) of the theorem be satisfied. It is easy to see that we can then choose a selector $\eta(t) \in \xi(t, A^*)$ of the multivalued function $\xi(\cdot, A^*)$ defined by (2.4) such that, if $\eta(t_i) = +1(-1)$, then $\eta(t_{i+1}) = -1(+1)$, $i = \overline{1, n+1}$. Then, by Lemma 2, $0_{n+1} \in \operatorname{int}\operatorname{co}\{\eta(t_i)(1, t_i, \ldots, t_i^n), i = \overline{1, n+2}\}$. Therefore, in view of (2.5), we conclude that $0_{n+1} \in \operatorname{int}\partial\rho(A^*)$. As is known from convex analysis (see, e.g., [2, p. 216]), this inclusion implies that A^* is a unique minimizer of the convex function $\rho(\cdot)$.

3.2. Now, we present conditions for the uniqueness of a solution that take into account the differential properties of the estimated segment function. Note that, if $t^* \in R_3(A)$, then the definition of $R_3(\cdot)$ implies that

$$\rho^* = P_n(A, t^*) - f_1(t^*) = f_2(t^*) - P_n(A, t^*), \tag{3.2}$$

$$P_n(A, t^*) - f_1(t^*) = \max_{t \in [c, d]} (P_n(A, t) - f_1(t)), \tag{3.3}$$

$$f_2(t^*) - P_n(A, t^*) = \max_{t \in [c, d]} (f_2(t) - P_n(A, t)) \tag{3.4}$$

and, additionally,

$$f_2(t^*) - f_1(t^*) = \max_{t \in [c, d]} (f_2(t) - f_1(t)). \tag{3.5}$$

Specifically, (3.5) implies that, if $t^* \in (c, d)$ and the functions $f_1(t)$ and $f_2(t)$ are differentiable at this point, then $f_1'(t^*) = f_2'(t^*)$. However, higher order optimality conditions of may be satisfied at this point. It was found that the uniqueness of a solution is affected by this circumstance. In this context, we introduce the following definition.

Definition 1. We say that $t^* \in R_3(A)$ is a point of multiplicity l if at least one of the following conditions is satisfied:

(a) $f_1(t)$ and $f_2(t)$ are differentiable $l-1$ times at this point on the right (or left), and

$$f_1^{(i)}(t^*+0) = f_2^{(i)}(t^*+0), \quad i = \overline{1, l-1},$$
$$\text{or, } f_1^{(i)}(t^*-0) = f_2^{(i)}(t^*-0), \quad i = \overline{1, l-1}, \text{ respectively.} \tag{3.6}$$

(b) If $l = 2k$, it is sufficient that one of the functions be $2k-1$ times differentiable and the other $2k-2$ times differentiable at this point, and

$$f_1^{(i)}(t^*) = f_2^{(i)}(t^*), \quad i = \overline{1, 2k-2}.$$

Remark 1. Definition 1 assumes that, if condition (a) or (b) is not satisfied for $l \geq 2$, then the point t^* has the multiplicity $l = 1$.

Remark 2. Naturally, if t^* coincides with the left or right endpoint of $[c, d]$, then condition (a) assumes differentiability on the right or left, respectively.

Remark 3. If $k = 1$, then condition (b) means the differentiability of one of the functions at t^*.

The sense of Definition 1 lies in the following important auxiliary result.

Lemma 3. *If $t^* \in R_3(A)$ is a point of multiplicity l and the coefficient vector A_1 is also a solution to problem* (1.1), *then t^* is an l-fold root of the equation $P_n(\Delta A, t) = 0$, where $\Delta A = A_1 - A^*$.*

Proof. If $t^* + \Delta t \in [c, d]$, then (3.2)–(3.4) imply

$$P_n(A^*, t^* + \Delta t) - f_1(t^* + \Delta t) \leq \rho^*, \tag{3.7}$$
$$f_2(t^* + \Delta t) - P_n(A^*, t^* + \Delta t) \leq \rho^*. \tag{3.8}$$

For $l = 1$, the assertion is obvious, since

$$P_n(A^*, t^*) = P_n(A_1, t^*) = (f_1(t^*) + f_2(t^*))/2. \tag{3.9}$$

Let $l \geq 2$ and condition (a) of Definition 1 be satisfied at t^*. To be definite, we assume that $f_1(t)$ and $f_2(t)$ are $l-1$ times differentiable at t^* on the right and equalities (3.6) hold.

For $\Delta t \geq 0$, inequalities (3.7) and (3.8), combined with (3.2), imply that

$$[P_n(A^*, t) - f_1(t)]'_{t=t^*+0}\Delta t + \ldots + \frac{[P_n(A^*, t) - f_1(t)]^{(l-1)}_{t=t^*+0}}{(l-1)!}(\Delta t)^{l-1} + o_1((\Delta t)^{l-1}) \leq 0, \tag{3.10}$$

$$[f_2(t) - P_n(A^*, t)]'_{t=t^*+0}\Delta t + \ldots + \frac{[f_2(t) - P_n(A^*, t)]^{(l-1)}_{t=t^*+0}}{(l-1)!}(\Delta t)^{l-1} + o_2((\Delta t)^{l-1}) \leq 0, \tag{3.11}$$

where $o_j((\Delta t)^{l-1})/(\Delta t)^{l-1} \longrightarrow 0$ as $\Delta t \downarrow 0$, $j = 1, 2$. From (3.10) and (3.11), for sufficiently small $\Delta t > 0$, we obtain

$$[P_n(A^*, t) - f_1(t)]'_{t=t^*+0} \leq 0, \quad [f_2(t) - P_n(A^*, t)]'_{t=t^*+0} \leq 0.$$

According to (3.6), $f_1'(t^*+0) = f_2'(t^*+0)$. Therefore, it follows from this inequality that

$$[P_n(A^*, t) - f_1(t)]'_{t=t^*+0} = [f_2(t) - P_n(A^*, t)]'_{t=t^*+0} = 0. \tag{3.12}$$

Substituting (3.12) into (3.10) and (3.11) and proceeding by analogy, we finally obtain

$$[P_n(A^*, t) - f_j(t)]^{(i)}_{t=t^*+0} = 0, \quad i = \overline{1, l-1}, \quad j = 1, 2. \tag{3.13}$$

Naturally, the same relations hold for A_1, since $R_3(A^*) = R_3(A_1)$ and

$$[P_n(A_1, t) - f_j(t)]^{(i)}_{t=t^*+0} = 0, \quad i = \overline{1, l-1}, \quad j = 1, 2. \tag{3.14}$$

Relations (3.13) and (3.14) imply that $P_n^{(i)}(A^*, t^*) = P_n^{(i)}(A_1, t^*)$ or $P_n^{(i)}(\Delta A, t^*) = 0$, $i = \overline{1, l-1}$. In view of (3.9), this means that t^* is an l-fold root of $P_n(\Delta A, t)$.

If t^* is a point of multiplicity l satisfying condition (b) of Definition 1, where, to be definite, $f_1(t)$ is $2k-1$ times differentiable at this point, then, in view of (3.7) and (3.8), we can write

$$[P_n(A^*, t) - f_1(t)]'_{t=t^*}\Delta t + \ldots + \frac{[P_n(A^*, t) - f_1(t)]^{(2k-1)}_{t=t^*}}{(2k-1)!}(\Delta t)^{2k-1} + o_3((\Delta t)^{2k-1}) \leq 0, \tag{3.15}$$

$$[f_2(t) - P_n(A^*, t)]'_{t=t^*}\Delta t + \ldots + \frac{[f_2(t) - P_n(A^*, t)]^{(2k-2)}_{t=t^*}}{(2k-2)!}(\Delta t)^{2k-2} + o_4((\Delta t)^{2k-2}) \leq 0, \tag{3.16}$$

where $o_3((\Delta t)^{2k-1})/(\Delta t)^{2k-1} \longrightarrow 0$ and $o_4((\Delta t)^{2k-2})/(\Delta t)^{2k-2} \longrightarrow 0$ as $t \longrightarrow 0$. From (3.15) and (3.16), proceeding as in the previous case, we obtain $P_n(A^*, t) - f_1(t)]_{t=t^*}^{(i)} = 0$ for $i = \overline{1, 2k-2}$. Then (3.15) implies

$$[P_n(A^*, t) - f_1(t)]_{t=t^*}^{(2k-1)}(\Delta t)^{2k-1} + o((\Delta t)^{2k-1}) \le 0.$$

For sufficiently small Δt of different signs, this yields the equality

$$[P_n(A^*, t) - f_1(t)]_{t=t^*}^{(2k-1)} = 0.$$

Since A_1 satisfies the same relations

$$[P_n(A_1, t) - f_1(t)]_{t=t^*}^{(i)} = 0, \quad i = \overline{1, 2k-1},$$

we conclude, as before, that t^* is an l-fold root of the equation $P_n(\Delta A, t) = 0$.

Theorem 4. *Let $R_3(A^*) = \{t_i\}_{i=1}^m$, where $m < n+1$ and the points t_i are of multiplicity l_i, $i = \overline{1, m}$, and let one of the following conditions be satisfied*:

(i) $l_1 + \ldots + l_m \ge n+1$,

(ii) $l = l_1 + \ldots + l_m < n+1$ *and there exists $k = n+1-l$ pairs of points*

$$t_1^{(1)} < t_1^{(2)} \le t_2^{(1)} < t_2^{(2)} \le \ldots \le t_k^{(1)} < t_k^{(2)}$$

such that $[t_i^{(1)}, t_i^{(2)}] \cap R_3(A^*) = \emptyset$ and, if $t_i^{(1)} \in R_1(A^*)(R_2(A^*))$, then $t_i^{(2)} \in R_2(A^*)(R_1(A^*))$. Then A^* is a unique solution to problem (1.1).

Proof. Let condition (i) be satisfied. Then, assuming that the coefficient vector A_1 is also a solution to problem (1.1) and using Lemma 3, we conclude that the equation $P_n(A_1 - A^*, t) = 0$ has roots with a total multiplicity greater than n. This means that $A_1 = A^*$.

Now assume that condition (ii) be satisfied. According to the definition of $R_1(A)$ and $R_2(A)$, if, for example, $t_i^{(1)} \in R_1(A^*)$ but $t_i^{(2)} \in R_2(A^*)$, then we can write

$$\rho^* = P_n(A^*, t_i^{(1)}) - f_1(t_i^{(1)}), \tag{3.17}$$

$$\rho^* = f_2(t_i^{(2)}) - P_n(A^*, t_i^{(2)}). \tag{3.18}$$

Assuming that A_1 is also a solution to problem (1.1),

$$\rho^* \ge P_n(A_1, t_i^{(1)}) - f_1(t_i^{(1)}), \tag{3.19}$$

$$\rho^* \ge f_2(t_i^{(2)}) - P_n(A_1, t_i^{(2)}). \tag{3.20}$$

Combining (3.17) with (3.19) and (3.18) with (3.20), we obtain

$$P_n(\Delta A, t_i^{(1)}) \le 0, \quad P_n(\Delta A, t_i^{(2)}) \ge 0, \tag{3.21}$$

respectively. Here, $\Delta A = A_1 - A^*$.

If $t_{i+1}^{(1)} > t_i^{(2)}$, then (3.21) implies that the interval $[t_i^{(1)}, t_i^{(2)}]$ contains at least one root of the equation $P_n(\Delta A, t) = 0$.

Now, consider the case of $t_i^{(2)} = t_{i+1}^{(1)}$. According to condition (ii), $t_{i+1}^{(2)} \in R_1(A^*)$, which implies $P_n(\Delta A, t_{i+1}^{(2)}) \le 0$. By taking into account (3.21), a further consideration of the versions suggests that the interval $[t_i^{(1)}, t_{i+1}^{(2)}]$ either contains at least two roots of the equation $P_n(\Delta A, t) = 0$ or at least one multiple root at the point $t_i^{(2)}$.

Thus, condition (ii) implies that, in the union of the intervals $[t_i^{(1)}, t_i^{(2)}]$, $i = \overline{1, k}$, the equation $P_n(\Delta A, t) = 0$ has roots with a total multiplicity of no less than k. Taking also into account the roots of this equation from the set $R_3(A^*)$, together with their multiplicity, and applying Lemma 3, we find that, overall, the equation on $[c, d]$ has roots with a total multiplicity of greater than n. Therefore, $A_1 = A^*$.

3.3. The resulting uniqueness conditions can be commented on as follows.

1. The example below shows the importance of the conditions in Theorems 3 and 4.

Example 1. Let $n = 1$, and let the following functions be given on the interval $[-1, 1]$:

$$f_1(t) = \begin{cases} 0, & t \in [-1, 0] \\ t, & t \in [0, 1], \end{cases} \quad f_2(t) = \begin{cases} 1 + t, & t \in [-1, 0] \\ 1, & t \in [0, 1]. \end{cases}$$

For the coefficient vector A^* (0.5, 0), we have

$$R_1(A^*) = [-1, 0), \quad R_2(A^*) = (0, 1], \quad R_3(A^*) = \{0\}.$$

In this case, the set $R_3(A^*)$, though nonempty, contains less than $n + 1$ points; i.e., condition (i) of Theorem 3 is not satisfied. For any choice of the points $t_1 \in R_1(A^*)$, $t_3 \in R_2(A^*)$, and $t_2 \in R_3(A^*)$, taking into account that $t_2 = 0 \in [t_1, t_2]$, condition (ii) of Theorem 3 is not satisfied either (the evenness or oddness rule for $i_{k+1} - i_k$).

On the other hand, $f_1(t)$ and $f_2(t)$ are not differentiable at the point $t_2 = 0$, which comprises the set $R_3(A^*)$. However, they are differentiable on the right and left, and the values of the right and left derivatives do not coincide. Thus, according to Definition 1, the point t_2 has multiplicity 1. Hence, condition (i) of Theorem 4 is not satisfied. As was noted above, for any choice $t_1 \in R_1(A^*)$ and $t_3 \in R_2(A^*)$, the interval $[t_1, t_3]$ contains a point t_2; i.e., the condition (ii) of Theorem 4 is not satisfied either. The coefficient vector A^* is a (nonunique) solution. It is easy to see that the entire solution set has the form

$$\Omega_\rho = \{A = (a_0, a_1) \in \mathbb{R}^2 : a_1 \in [0, 1], a_0 = 0.5\}.$$

2. Condition (ii) of Theorem 3 can be treated as a generalized alternance. It means that the graph of the segment function in the ordered set of points $T = \{t_i\}_{i=1}^{n+2}$ touches in turn the upper and lower boundaries of the polynomial strip "from within." This is ensured by the evenness and oddness of $i_{k+1} - i_k$.

3. Theorem 3 implies that the solution A^* may be nonunique only if $R_3(A^*) \neq \varnothing$.

4. Simple examples show that each of the above sufficient uniqueness conditions can be satisfied individually, i.e., the remaining ones do not hold. At the same time, the fulfillment of one of them does not rule out this for others. Consider an example in which the uniqueness of a solution depends entirely on the multiplicity of a point from the set $R_3(\cdot)$, i.e., depends on whether or not condition (i) of Theorem 4 holds.

Example 2. Consider the functions $f_1(t) = t^2$ and $f_2(t) = 2 - t^4$ on the interval [-1, 1].

(a) Let $n = 1$ and $A^* = (1, 0)$. Then

$$R(A^*) = R_3(A^*) = \{0\}.$$

The following conditions are satisfied for the point $t^* = 0$, which comprises the set $R_3(A^*)$:

$$f_1'(0) = f_2'(0) = 0, \quad f_1''(0) \neq 0, \quad f_2''(0) = 0.$$

Thus, the point t^* has the multiplicity $l = 2$. Since $n + 1 = 2$, condition (i) of Theorem 3 is satisfied. Therefore, A^* is a unique solution.

(b) Now, let $n = 2$ and $A_1 = (1, 0, 0)$. For this coefficient vector, we have

$$R(A_1) = R_3(A_1) = \{0\}$$

and the point $t^* = 0$ has the multiplicity $l = 2$. However, in this case, $l < n + 1$. The vector A_1 is a nonunique solution. It is easy to see that the entire solution set is representable as

$$\Omega_\rho = \{A = (1, 0, a_2) \in \mathbb{R}^3 : a_2 \in [0, 1]\}.$$

5. We know examples where the solution to problem (1.1) is unique, but none of the conditions in Theorem 3 or 4 are satisfied. However, in the case of a discrete setting, namely, when the interval $[c, d]$ in (1.1) is replaced with a set of points $T = \{t_i\}_{i=1}^{m}$ for $m \geq n + 1$, it was proved in [9] that condition (i) or (ii) in Theorem 3 (written in another form but still equivalent) is not only sufficient but also necessary for the uniqueness of a solution.

6. The sufficient conditions for the uniqueness of a solution to Theorem 4 are based on the differential properties of the segment function $F(t)$ at points of $R_3(\cdot)$. We has failed to answer the question as to whether or not the uniqueness of a solution is affected by these properties on the sets $R_1(\cdot)$ and $R_2(\cdot)$.

ACKNOWLEDGMENTS

This work was supported by a grant from the President of the Russian Federation, project no. NSh-2970.2008.1.

REFERENCES

1. B. N. Pshenichnyi, *Convex Analysis and Extremal Problems* (Nauka, Moscow, 1981) [in Russian].
2. V. F. Dem'yanov and L. V. Vasil'ev, *Nondifferential Optimization* (Nauka, Moscow, 1981) [in Russian].
3. F. L. Chernousko, *State Estimation for Dynamic Systems* (Nauka, Moscow, 1988; CRC, Boca Raton, 1994).
4. V. F. Dem'yanov and A. M. Rubinov, *Elements of Nonsmooth Analysis and Quasi-Differential Calculus* (Nauka, Moscow, 1990) [in Russian].
5. A. B. Kurzhanski and I. Valui, *Ellipsoidal Calculus for Estimation and Control* (Birkhüser, Boston, 1997).
6. B. Sendov, *Hausdorff Approximations* (Bolgarsk. Akad. Nauk, Sofia, 1979; Kluwer, Dordrecht 1990).
7. V. K. Dzyadyk, *Introduction to the Theory of Uniform Polynomial Approximation of Functions* (Nauka, Moscow, 1977) [in Russian].
8. I. Yu. Vygodchikova, "On the Best Approximation of a Discrete Set-Valued Mapping by an Algebraic Polynomial," *Collected Papers: Mathematics and Mechanics* (Saratov. Univ., Saratov, 2001), No. 3, pp. 25–27 [in Russian].
9. I. Yu. Vygodchikova, "On the Uniqueness of the Solution to the Problem of the Best Approximation of a Discrete Set-Valued Mapping by an Algebraic Polynomial," Izv. Saratov. Univ. **6** (1/2), 11–19 (2006).
10. V. F. Dem'yanov and V. N. Malozemov, *Introduction to the Minimax* (Nauka, Moscow, 1972) [in Russian].
11. S. I. Dudov, "On Two Auxiliary Facts for Analysis of Polynomial Approximation Problems," *Collected Papers: Mathematics and Mechanics* (Saratov. Univ., Saratov, 2007), No. 9, pp. 22–26 [in Russian].
12. F. P. Vasil'ev, *Numerical Methods for Optimization Problems* (Nauka, Moscow, 1988) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1128–1140. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © A.F. Izmailov, A.L. Pogosyan, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1184–1196.

Optimality Conditions and Newton-Type Methods for Mathematical Programs with Vanishing Constraints

A. F. Izmailov and A. L. Pogosyan

Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, 119992 Russia
e-mail: izmaf@ccas.ru, pogosyan85@gmail.com

Received November 11, 2008

Abstract—A new class of optimization problems is discussed in which some constraints must hold in certain regions of the corresponding space rather than everywhere. In particular, the optimal design of topologies for mechanical structures can be reduced to problems of this kind. Problems in this class are difficult to analyze and solve numerically because their constraints are usually irregular. Some known first- and second-order necessary conditions for local optimality are refined for problems with vanishing constraints, and special Newton-type methods are developed for solving such problems.

DOI: 10.1134/S0965542509070069

Key words: mathematical program with vanishing constraints, mathematical program with complementarity constraints, constraint qualification, optimality conditions, sequential quadratic programming, active-set method.

1. INTRODUCTION: THE OPTIMIZATION PROBLEM WITH VANISHING CONSTRAINTS

Consider the mathematical program with vanishing constraints (MPVC) stated as follows:

$$f(x) \longrightarrow \min, \quad H_i(x) \geq 0, \quad G_i(x)H_i(x) \leq 0, \quad i = 1, 2, \ldots, m, \tag{1.1}$$

Here, $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is a smooth function, and $G, H: \mathbb{R}^n \longrightarrow \mathbb{R}^m$ are smooth mappings. Problems of this type were first introduced in [1], and their name is explained by the following consideration: if the first constraint in (1.1) holds as an equality at a point $x \in \mathbb{R}^n$ for some index $i \in \{1, 2, ..., m\}$, then the second constraint is automatically fulfilled; thus, this constraint "vanishes." If the first constraint holds as a strict inequality, then the second constraint is equivalent to the relation $G_i(x) \leq 0$.

As shown in [1], problems with vanishing constraints are a natural and convenient means of modeling problems in the optimal design of topologies for mechanical structures. Such a design becomes a standard tool in industrial applications (for instance, in aircraft or automobile design, etc.).

The problem examined in [1] may include conventional equality and inequality constraints. Such a generalization does not involve serious additional difficulties and is not considered here.

Let $\bar{x} \in \mathbb{R}^n$ be a feasible point of problem (1.1). Following [1], we define the index sets

$$I_+ = I_+(\bar{x}) = \{i = 1, 2, \ldots, m \,|\, H_i(\bar{x}) > 0\},$$
$$I_0 = I_0(\bar{x}) = \{i = 1, 2, \ldots, m \,|\, H_i(\bar{x}) = 0\}.$$

We also introduce the partition of I_+ into the sets

$$I_{+0} = I_{+0}(\bar{x}) = \{i \in I_+ \,|\, G_i(\bar{x}) = 0\},$$
$$I_{+-} = I_{+-}(\bar{x}) = \{i \in I_+ \,|\, G_i(\bar{x}) < 0\},$$

and the partition of I_0 into the sets

$$I_{0+} = I_{0+}(\bar{x}) = \{i \in I_0 \,|\, G_i(\bar{x}) > 0\},$$
$$I_{00} = I_{00}(\bar{x}) = \{i \in I_0 \,|\, G_i(\bar{x}) = 0\},$$
$$I_{0-} = I_{0-}(\bar{x}) = \{i \in I_0 \,|\, G_i(\bar{x}) < 0\}.$$

The condition $I_{00} = \varnothing$ is called the *lower-level strict complementarity condition*. It was shown in [1] that, if this (very restructive) condition is violated, then the constraints in problem (1.1) are necessarily irregular

at the point $\bar{x}$; that is, they do not satisfy the Mangasarian–Fromovitz constraint qualification. This makes the MPVC difficult to analyze and solve numerically. Special results that concern the optimality conditions, the sensitivity, and relaxation methods and use the special structure of the MPVCs were obtained in [1–5].

Note that the introduction of the additional variable $u \in \mathbb{R}^m$ makes it possible to reduce MPVC (1.1) to the *mathematical program with complementarity constraints* (MPCC):

$$f(x) \longrightarrow \min, \quad G(x) - u \le 0, \quad H(x) \ge 0, \quad u \ge 0, \quad H_i(x)u_i = 0, \quad i = 1, 2, \ldots, m. \tag{1.2}$$

The MPCCs are a rather well-studied class of problems, which enjoys much attention of the experts (e.g., see [6–8; 9, Section 4.3]). However, the reduction to an MPCC increases the dimension of the problem and has another serious drawback; namely, for a given (local) solution $\bar{x}$ to MPVC (1.1), the corresponding optimal value of the additional variable u is not uniquely defined, and the corresponding local solutions to MPCC (1.2) cannot be strict. In particular, these solutions cannot satisfy the second-order sufficient optimality conditions. Consequently, the theoretical results (concerning the sensitivity and numerical methods) based on these conditions are inapplicable. This fact makes us regard the MPVCs as an independent class of problems that requires special approaches and techniques.

Our aim in this paper is to refine for MPVCs some familiar first- and second-order necessary conditions for local optimality and to construct (special) Newton-type methods that take into account the structure of these problems and have local superlinear convergence despite the fact that the traditional constraint qualifications are violated.

2. OPTIMALITY CONDITIONS

We begin with certain stationarity concepts used for MPVCs. With a feasible point $\bar{x}$ of problem (1.1), we associate two auxiliary conventional mathematical programming problems, namely, the *relaxed nonlinear programming problem* (RNLP):

$$f(x) \longrightarrow \min, \quad H_{I_{0+}}(x) = 0, \quad H_{I_{00} \cup I_{0-}}(x) \ge 0, \quad G_{I_{+0}}(x) \le 0, \tag{2.1}$$

and the *tightened nonlinear programming problem* (TNLP):

$$f(x) \longrightarrow \min, \quad H_{I_{0+} \cup I_{00}}(x) = 0, \quad H_{I_{0-}}(x) \ge 0, \quad G_{I_{+0} \cup I_{00}}(x) \le 0. \tag{2.2}$$

Here, for a finite set I, the symbol y_I stands for the subvector of y with the components y_i, $i \in I$.

Next, we define the MPVC Lagrangian function of problem (1.1) as

$$\mathcal{L}(x, \mu) = f(x) - \langle \mu^H, H(x) \rangle + \langle \mu^G, G(x) \rangle,$$

where $x \in \mathbb{R}^n$ and $\mu = (\mu^H, \mu^G) \in \mathbb{R}^m \times \mathbb{R}^m$. This is obviously the conventional Lagrangian function for TNLP (2.2) supplemented with the additional constraints

$$H_{I_+}(x) \ge 0, \quad G_{I_{+-} \cup I_{0-}}(x) \le 0, \quad G_{I_{0+}}(x) \ge 0, \tag{2.3}$$

which are inactive at $\bar{x}$. We can also regard the MPVC-Lagrangian function as the conventional Lagrangian function for RNLP (2.1) supplemented with constraints (2.3) and the conditions

$$G_{I_{00}}(x) \le 0$$

(the latter constraints are active at $\bar{x}$).

A feasible point $\bar{x}$ of MPVC (1.1) is called a *strongly* (*weakly*) *stationary point* of this problem if it is a stationary point of RNLP (2.1) (respectively, TNLP (2.2)) in the conventional sense. Thus, the weak stationarity implies the existence of $\mu = (\mu^H, \mu^G) \in \mathbb{R}^m \times \mathbb{R}^m$ such that

$$\frac{\partial \mathcal{L}}{\partial x}(\bar{x}, \mu) = 0, \tag{2.4}$$

$$\mu^H_{I_{0-}} \ge 0, \quad \mu^H_{I_+} = 0, \quad \mu^G_{I_{+0} \cup I_{00}} \ge 0, \quad \mu^G_{I_{+-} \cup I_{0+} \cup I_{0-}} = 0. \tag{2.5}$$

Such μ will be called the Lagrange multipliers of TNLP (2.2). The strong stationarity (this concept for MPVCs was introduced in [4, Definition 2.1]) implies that, in addition, the relations

$$\mu^H_{I_{00}} \ge 0, \quad \mu^G_{I_{00}} = 0. \tag{2.6}$$

are fulfilled. In this case, μ is called an MPVC *multiplier* corresponding to the strongly stationary point $\bar{x}$. The set of MPVC multipliers corresponding to $\bar{x}$ is denoted by $\mathcal{M} = \mathcal{M}(\bar{x})$.

Let $\mathcal{I} = \mathcal{I}(\bar{x})$ be the set of all the partitions of I_{00}, that is, the set of all the pairs (I_1, I_2) such that $I_1 \cup I_2 = I_{00}$ and $I_1 \cap I_2 = \varnothing$. It is obvious that $|\mathcal{I}| = 2^{|I_{00}|}$, where $|I|$ is the number of elements in the finite set I. For each $(I_1, I_2) \in \mathcal{I}$, we define the *piecewise problem*

$$f(x) \longrightarrow \min, \quad H_{I_{0+} \cup I_2}(x) = 0, \quad H_{I_1 \cup I_{0-}}(x) \geq 0, \quad G_{I_{+0} \cup I_1}(x) \leq 0. \tag{2.7}$$

Its feasible set is denoted by $D_{(I_1, I_2)} = D_{(I_1, I_2)}(\bar{x})$. It can be regarded as a *branch* (or a *piece*) of the feasible set D of original problem (1.1). It can easily be verified that, locally (in the vicinity of $\bar{x}$), D splits into branches of the above form.

A feasible point $\bar{x} \in \mathbb{R}^n$ of MPVC (1.1) is called a *B-stationary* (or *piecewise stationary*) *point* of this problem if it is stationary for every piecewise problem (2.7). Thus, for every partition $(I_1, I_2) \in \mathcal{I}$, there exists $\mu = (\mu^H, \mu^G) \in \mathbb{R}^m \times \mathbb{R}^m$ satisfying relation (2.4) and the conditions

$$\mu^H_{I_1 \cup I_{0-}} \geq 0, \quad \mu^H_{I_+} = 0, \quad \mu^G_{I_{+0} \cup I_1} \geq 0, \quad \mu^G_{I_{+-} \cup I_{0+} \cup I_2 \cup I_{0-}} = 0, \tag{2.8}$$

Such μ will be called the Lagrange multipliers of piecewise problem (2.7). Obviously, the piecewise stationarity of $\bar{x}$ implies its weak stationarity, and the corresponding Lagrange multiplier is the same.

Now, we turn to the constraint qualifications that are used below to derive necessary optimality conditions for a feasible point $\bar{x}$ of MPVC (1.1), to justify the convergence of Newton-type methods, and to estimate their convergence rate. The *piecewise Mangasarian–Fromovitz condition* at $\bar{x}$ is understood as the conventional Mangasarian–Fromovitz condition (e.g., see [10, p. 41]) fulfilled for each branch; that is, for each partition $(I_1, I_2) \in \mathcal{I}$, we have

$$\operatorname{rank} H'_{I_{0+} \cup I_2}(\bar{x}) = |I_{0+}| + |I_2|,$$

$$\exists \bar{\xi} \in \ker H'_{I_{0+} \cup I_2}(\bar{x}) \text{ such that } H'_{I_1 \cup I_{0-}}(\bar{x})\bar{\xi} > 0, \quad G'_{I_{+0} \cup I_1}(\bar{x})\bar{\xi} < 0.$$

The next condition is identical to the so-called strict Mangasarian–Fromovitz constraint qualification for TNLP (2.2), which in turn is equivalent to the combination of the conventional Mangasarian–Fromovitz constraint qualification and the uniqueness requirement for the corresponding Lagrange multiplier in this problem (see [10, exercise 1.4.3]). Let $\bar{\mu}$ be a Lagrange multiplier of TNLP (2.2) corresponding to a weakly stationary point $\bar{x}$. We say that the MPVC-strict *Mangasarian–Fromovitz constraint qualification* is fulfilled at $\bar{x}$ if

$$\operatorname{rank}\begin{pmatrix} H'_{I_{0+} \cup I_{00} \cup I^+_{0-}}(\bar{x}) \\ G'_{I^+_{+0} \cup I^+_{00}}(\bar{x}) \end{pmatrix} = |I_{0+}| + |I_{00}| + |I^+_{0-}| + |I^+_{+0}| + |I^+_{00}|, \tag{2.9}$$

$$\begin{gathered} \exists \bar{\xi} \in \ker H'_{I_{0+} \cup I_{00} \cup I^+_{0-}}(\bar{x}) \cap \ker G'_{I^+_{+0} \cup I^+_{00}}(\bar{x}) \quad \text{such that} \\ H'_{I_{0-} \setminus I^+_{0-}}(\bar{x})\bar{\xi} > 0, \quad G'_{(I_{+0} \setminus I^+_{+0}) \cup (I_{00} \setminus I^+_{00})}(\bar{x})\bar{\xi} < 0, \end{gathered} \tag{2.10}$$

where

$$I^+_{0-} = \{i \in I_{0-} \mid \bar{\mu}^H_i > 0\}, \quad I^+_{+0} = \{i \in I_{+0} \mid \bar{\mu}^G_i > 0\}, \quad I^+_{00} = \{i \in I_{00} \mid \bar{\mu}^G_i > 0\}.$$

Finally, following [2, Definition 4.1], we say that the MPVC *linear independence constraint qualification* is fulfilled at a feasible point $\bar{x}$ of problem (1.1) if

$$H'_i(\bar{x}), \quad i \in I_0, \quad G'_i(\bar{x}), \quad i \in I_{+0} \cup I_{00}, \quad \text{are linearly independent.} \tag{2.11}$$

Note that this is the conventional linear independence constraint qualification (i.e., the linear independence condition for the gradients of the equality constraints and the active inequality constraints) for TNLP (2.2) at $\bar{x}$.

It is trivial to verify that the MPVC linear independence constraint qualification (2.11) implies the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10). In turn, the latter condition

implies the piecewise Mangasarian–Fromovitz constraint qualification. The converse implications are false, which is demonstrated by the following examples.

Example 1. Consider problem (1.1), where $n = m = 2$, $f(x) = x_1^2 + x_2^2$, $H(\cdot) \equiv (1, 1)$, and $G(x) = (x_1, x_1)$. The feasible set of this problem has the form $D = \{x \in \mathbb{R}^2 | x_1 \le 0\}$, and its unique (global and local) solution is the point $\bar{x} = 0$. The corresponding index sets are $I_+ = I_{+0} = \{1, 2\}$ (the other sets are empty), and the TNLP can be written as

$$x_1^2 + x_2^2 \longrightarrow \min, \quad x_1 \le 0, \quad x_2 \le 0.$$

It is obvious that the constraints in this problem do not satisfy the linear independence condition. On the other hand, they satisfy the Mangasarian–Fromovitz constraint qualification. The Lagrange multipliers of the TNLP associated with $\bar{x} = 0$ are determined by the system

$$\mu_1^G + \mu_2^G = 0, \quad \mu^H = 0, \quad \mu^G \ge 0,$$

which has a unique solution $\bar{\mu} = 0$. Thus, the MPVC linear independence constraint qualification (2.11) is not fulfilled at $\bar{x} = 0$; however, the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10) is fulfilled.

Example 2. Consider the problem that has the same constraints as in Example 1 and the objective function $f(x) = -x_1 + x_2^2$. The point $\bar{x} = 0$ remains a unique (global and local) solution to this problem. Since $I_{00} = \varnothing$, the unique piecewise problem associated with $\bar{x} = 0$ is identical to the TNLP

$$-x_1 + x_2^2 \longrightarrow \min, \quad x_1 \le 0, \quad x_2 \le 0.$$

The constraints in this problem satisfy the Mangasarian–Fromovitz constraint qualification, which implies the fulfillment of the piecewise Mangasarian–Fromovitz constraint qualification at $\bar{x} = 0$. At the same time, the Lagrange multipliers of the TNLP associated with $\bar{x} = 0$ are determined by the system

$$\mu_1^G + \mu_2^G = 1, \quad \mu^H = 0, \quad \mu^G \ge 0,$$

which is not uniquely solvable. Thus, the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10) cannot hold for any multiplier $\bar{\mu}$ in this problem.

The following theorem gives the first-order necessary optimality conditions for MPVC (1.1). It improves on the result presented in [5, Theorem 2.1] in the sense that its assertion 2 only assumes the fulfillment of the MPVC-strict Mangasarian–Fromovitz constraint qualification instead of the stronger MPVC linear independence constraint qualification.

Theorem 1. *Assume that the function f and the mapping G are differentiable at a point $\bar{x} \in \mathbb{R}^n$ and the mapping H is differentiable in a neighborhood of this point; moreover, its derivative is continuous at $\bar{x}$.*

Then, the following assertions are true:

(1) *If $\bar{x}$ is a local solution to problem* (1.1) *satisfying the piecewise Mangasarian–Fromovitz constraint qualification, then $\bar{x}$ is a B-stationary point of problem* (1.1) *and, hence, a weakly stationary point of this problem.*

(2) *Let $\bar{x}$ be a weakly stationary point of problem* (1.1), *and let the MPVC-strict Mangasarian–Fromovitz constraint qualification* (2.9), (2.10) *be fulfilled for a Lagrange multiplier $\bar{\mu}$ of TNLP* (2.2) *corresponding to $\bar{x}$ (that is, for $\bar{\mu} = (\bar{\mu}^H, \bar{\mu}^G)$ satisfying* (2.4) *and* (2.5)). *Then, $\bar{x}$ is a strongly stationary point of problem* (1.1). *Moreover, $\bar{\mu}$ is a unique MPVC multiplier associated with $\bar{x}$ and a unique Lagrange multiplier of TNLP* (2.2) *and piecewise problem* (2.7) *for any partition $(I_1, I_2) = \mathcal{J}$.*

Proof. The standard result on the first-order optimality conditions is that a local solution to a mathematical program satisfying the Mangasarian–Fromovitz constraint qualification is a stationary point of this problem (e.g., see [10, Theorem 1.4.2]). Since the local solution $\bar{x}$ to problem (1.1) is a local solution to each piecewise problem, we obtain the first assertion of Theorem 1 by applying the above result to the piecewise problem.

Now, assume that $\bar{x}$ is a weakly stationary point of problem (1.1) and the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10) is fulfilled at this point. Then, the piecewise Mangasarian–Fromovitz constraint qualification is fulfilled at $\bar{x}$, and $\bar{\mu}$ is a unique Lagrange multiplier of the

TNLP associated with $\bar{x}$. It follows that $\bar{\mu}$ is a unique Lagrange multiplier associated with $\bar{x}$ for every piecewise problem (because the Lagrange multipliers of any piecewise problem are Lagrange multipliers for the TNLP). Hence, the Lagrange multipliers for all the piecewise problems are identical and equal to $\bar{\mu}$. Now, (2.8) implies that the relations $\mu^H_{I_1} \geq 0$ and $\mu^G_{I_2} = 0$ are fulfilled for any partition $(I_1, I_2) \in \mathcal{I}$. This yields (2.6) and, hence, the strong stationarity of $\bar{x}$, which justifies the second assertion of Theorem 1. The theorem is proved.

We introduce the conventional Lagrangian function for problem (1.1):

$$L(x, \lambda) = f(x) - \langle \lambda^H, H(x) \rangle + \sum_{i=1}^{m} \lambda_i^{GH} G_i(x) H_i(x),$$

Here, $x \in \mathbb{R}^n$, and $\lambda = (\lambda^G, \lambda^{GH}) \in \mathbb{R}^m \times \mathbb{R}^m$. As shown in [1, Remark 1], the strong stationarity of a feasible point $\bar{x}$ of MPVC (1.1) is actually equivalent to the conventional stationarity, that is, to the existence of a Lagrange multiplier $\lambda = (\lambda^G, \lambda^{GH}) \in \mathbb{R}^m \times \mathbb{R}^m$ satisfying the relations

$$\frac{\partial L}{\partial x}(\bar{x}, \lambda) = 0, \tag{2.12}$$

$$\lambda^H_{I_0} \geq 0, \quad \lambda^H_{I_+} = 0, \quad \lambda^{GH}_{I_{+0} \cup I_0} \geq 0, \quad \lambda^{GH}_{I_{+-}} = 0. \tag{2.13}$$

To be more exact, define $\Lambda = \Lambda(\bar{x})$ as the set of the Lagrange multipliers associated with $\bar{x}$ (in other words, the set of all $\lambda = (\lambda^G, \lambda^{GH}) \in \mathbb{R}^m \times \mathbb{R}^m$ satisfying relations (2.12) and (2.13). Then, we have the following proposition.

Proposition 1. *Let the function f and the mappings H and G be differentiable at a feasible point $\bar{x} \in \mathbb{R}^n$ of problem* (1.1).

The point $\bar{x}$ is a stationary point of problem (1.1) *if and only if it is a strongly stationary point of this problem. Moreover, for every fixed $\lambda = (\lambda^H, \lambda^{GH}) \in \Lambda$ and for $\mu = (\mu^H, \mu^G) \in \mathbb{R}^m \times \mathbb{R}^m$ determined by the formulas*

$$\mu_i^H = \lambda_i^H = 0, \quad i \in I_+, \quad \mu_i^H = \lambda_i^H - \lambda_i^{GH} G_i(\bar{x}), \quad i \in I_{0+} \cup I_{0-}, \quad \mu_i^H = \lambda_i^H, \quad i \in I_{00}, \tag{2.14}$$

$$\mu_i^G = \lambda_i^{GH} H_i(\bar{x}), \quad i \in I_{+0}, \quad \mu_i^G = \lambda_i^{GH} H_i(\bar{x}) = 0, \quad i \in I_{+-} \cup I_0, \tag{2.15}$$

it holds that $\mu \in \mathcal{M}$. Conversely, for any $\mu = (\mu^H, \mu^G) \in \mathcal{M}$ and any $\lambda = (\lambda^H, \lambda^{GH}) \in \mathbb{R}^m \times \mathbb{R}^m$ satisfying (2.14), (2.15), *and the relations*

$$\lambda_i^{GH} \geq 0, \quad i \in I_{00}, \quad \lambda_i^{GH} \geq \max\left\{0, -\frac{\mu_i^H}{G_i(\bar{x})}\right\}, \quad i \in I_{0+}, \quad 0 \leq \lambda_i^{GH} \leq -\frac{\mu_i^H}{G_i(\bar{x})}, \quad i \in I_{0-},$$

it holds that $\lambda = (\lambda^H, \lambda^{GH}) \in \Lambda$.

Thus, despite the fact that the Mangasarian–Fromovitz constraint qualification is violated for the MPVC if the lower-level strict complementarity condition is not fulfilled, the traditional concept of stationarity is still quite adequate for problems of this class.

Now, we turn to second-order optimality conditions. For each partition $(I_1, I_2) \in \mathcal{I}$, define the critical cone of the corresponding piecewise problem (2.7) at the point $\bar{x}$:

$$C_{(I_1, I_2)} = C_{(I_1, I_2)}(\bar{x}) = \{\xi \in \mathbb{R}^n \mid H'_{I_{0+} \cup I_2}(\bar{x})\xi = 0, H'_{I_{0-} \cup I_1}(\bar{x})\xi \geq 0, G'_{I_{+0} \cup I_1}(\bar{x})\xi \leq 0, \langle f'(\bar{x}), \xi \rangle \leq 0\}.$$

We set

$$C_2 = C_2(\bar{x}) = \bigcup_{(I_1, I_2) \in \mathcal{I}} C_{(I_1, I_2)} = \{\xi \in \mathbb{R}^n \mid H'_{I_{0+}}(\bar{x})\xi = 0, H'_{I_{00} \cup I_{0-}}(\bar{x})\xi \geq 0, G'_{I_{+0}}(\bar{x})\xi \leq 0,$$
$$\langle G'_i(\bar{x}), \xi \rangle \langle H'_i(\bar{x}), \xi \rangle \leq 0, i \in I_{00}, \langle f'(\bar{x}), \xi \rangle \leq 0\}. \tag{2.16}$$

The meaning of the index 2 is that, unlike the standard critical cone of problem (1.1) at the point $\bar{x}$, which is given by

$$C = C(\bar{x}) = \{\xi \in \mathbb{R}^n \mid H'_{I_{0+}}(\bar{x})\xi = 0, H'_{I_{00} \cup I_{0-}}(\bar{x})\xi \geq 0, G'_{I_{+0}}(\bar{x})\xi \leq 0, \langle f'(\bar{x}), \xi \rangle \leq 0\}$$

the cone C_2 takes into account second-order information on the last constraint in (1.1).

The following theorem improves the result of [5, Theorem 3.1] in the sense that it only assumes the fulfillment of the MPVC-strict Mangasarian–Fromovitz constraint qualification instead of the stronger MPVC linear independence constraint qualification.

Theorem 2. *Let the function f and the mappings H and G be twice differentiable at a strongly stationary point $\bar{x} \in \mathbb{R}^n$ of problem* (1.1). *Assume that the MPVC-strict Mangasarian–Fromovitz constraint qualification* (2.9), (2.10) *is fulfilled at this point for the* (*unique*) *MPVC-multiplier $\bar{\mu} = (\bar{\mu}^H, \bar{\mu}^G)$ associated with $\bar{x}$.*

If $\bar{x}$ is a local solution to (1.1), *then it holds that*

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] \geq 0 \quad \forall \xi \in C_2. \tag{2.17}$$

Proof. As shown in the proof of Theorem 1, the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10) implies that the piecewise Mangasarian–Fromovitz constraint qualification is fulfilled and, for every piecewise problem, $\bar{\mu}$ is the unique Lagrange multiplier associated with $\bar{x}$. Since $\bar{x}$ is locally optimal in every piecewise problem, we can apply to these problems the standard result on the second-order necessary optimality condition (e.g., see [9, Theorem 1.3.8]), which yields

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] \geq 0 \quad \forall \xi \in C_{(I_1, I_2)}.$$

The required result follows from this inequality and the second equality in (2.16). The theorem is proved.

It is natural to associate with the second-order necessary condition (2.17) the sufficient condition

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] > 0 \quad \forall \xi \in C_2 \backslash \{0\}, \tag{2.18}$$

which we call the *piecewise second-order sufficient condition*. As shown in [4, Theorem 4.4], this condition is indeed sufficient for a strongly stationary point $\bar{x}$ of problem (1.1) to be strictly locally optimal. A subtler second-order sufficient condition was obtained in [5, Theorem 3.3]. However, it is condition (2.18) that we use in this paper.

In addition to (2.18), we also use the conventional second-order sufficient optimality condition, which states that

$$\frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})[\xi, \xi] > 0 \quad \forall \xi \in C \backslash \{0\}. \tag{2.19}$$

for some $\bar{\lambda} \in \Lambda$. According to [5, Theorem 3.3], if (2.19) is fulfilled for a certain $\bar{\lambda} \in \Lambda$, then, under the hypotheses of Theorem 2, relation (2.18) is also fulfilled. On the other hand, it is conjectured in [5] that the reverse implication is false.

3. PIECEWISE SEQUENTIAL QUADRATIC PROGRAMMING METHOD

The piecewise sequential quadratic programming (SQP) method proposed for the MPVC in this section is based on the same idea as the corresponding method for the MPCC (see [6, 11]). The idea is to identify *any* piecewise problem corresponding to the desired solution and apply SQP to this problem (see [10, Section 4.4]).

The local identification of a branch in the feasible set requires no effort; namely, it suffices to identify a partition (J_1, J_2) of the set $\{1, 2, \ldots, m\}$ such that

$$I_{0+} \subset J_2, \quad I_+ \cup I_{0-} \subset J_1. \tag{3.1}$$

As soon as this has been done, we can consider the problem

$$f(x) \longrightarrow \min, \quad H_{J_2}(x) = 0, \quad H_{J_1}(x) \geq 0, \quad G_{J_1}(x) \leq 0. \tag{3.2}$$

If we set $I_1 = J_1 \cap I_{00}$ and $I_2 = J_2 \cap I_{00}$, then problem (3.2) differs from (2.7) only by the additional constraints

$$H_{J_2 \backslash (I_{0+} \cup I_{00})}(x) = 0, \quad H_{J_1 \backslash (I_{00} \cup I_{0-})}(x) \geq 0, \quad G_{J_1 \backslash (I_{+0} \cup I_{00})}(x) \leq 0. \tag{3.3}$$

Relations (3.1) and the condition $J_1 \cap J_2 = \varnothing$ imply that

$$J_1 \cap I_{0+} = \varnothing, \quad J_2 \cap (I_+ \cup I_{0-}) = \varnothing,$$

Consequently, we have

$$J_2 \backslash (I_{0+} \cup I_{00}) = J_2 \backslash (I_+ \cup I_0) = \varnothing,$$
$$J_1 \backslash (I_{00} \cup I_{0-}) = J_1 \backslash I_0 \subset I_+,$$
$$J_1 \backslash (I_{+0} \cup I_{00}) = J_1 \backslash (I_{+0} \cup I_{0+} \cup I_{00}) \subset I_{+-} \cup I_{0-}.$$

Using the definition of the above index sets, we conclude that all the additional constraints actually appearing in (3.3) are inequality constraints; moreover, they are inactive at $\bar{x}$. Thus, locally (in the vicinity of $\bar{x}$), problems (3.2) and (2.7) are identical.

It is obvious from the definition of the sets I_+, I_{0+}, and I_{0-} that the desired partition (J_1, J_2) satisfying (3.1) can be identified for $x \in \mathbb{R}^n$ that is close to $\bar{x}$ in the following way:

$$J_1 = J_1(x) = \{i = 1, 2, \ldots, m \,|\, G_i(x) < H_i(x)\}, \tag{3.4}$$

$$J_2 = J_2(x) = \{i = 1, 2, \ldots, m \,|\, G_i(x) \geq H_i(x)\}. \tag{3.5}$$

Let us give a formal description of the piecewise SQP method as applied to MPVC (1.1).

Algorithm 1

Preliminary step. Set $k = 0$ and choose $x^0 \in \mathbb{R}^n$ and $\mu^0 = ((\mu^H)^0, (\mu^G)^0) \in \mathbb{R}^m \times \mathbb{R}^m$.

Identification step. Define the index sets $J_1 = J_1(x^k)$ and $J_2 = J_2(x^k)$ in accordance with (3.4) and (3.5). Modify μ^k by setting $((\mu^G_{J_2})^k = 0$.

SQP step. Calculate $x^{k+1} \in \mathbb{R}^n$ as a stationary point of the quadratic program

$$\langle f'(x^k), x - x^k \rangle + \frac{1}{2}\left\langle \frac{\partial^2 \mathcal{L}}{\partial x^2}(x^k, \mu^k)(x - x^k), x - x^k \right\rangle \longrightarrow \min, \tag{3.6}$$

$$H_{J_2}(x^k) + H'_{J_2}(x^k)(x - x^k) = 0, \quad H_{J_1}(x^k) + H'_{J_1}(x^k)(x - x^k) \geq 0, \tag{3.7}$$

$$G_{J_1}(x^k) + G'_{J_1}(x^k)(x - x^k) \leq 0, \tag{3.8}$$

Calculate $((\mu^H_{J_2})^{k+1}, (\mu^H_{J_1})^{k+1}, (\mu^G_{J_1})^{k+1}) \in \mathbb{R}^{|J_2|}_+ \times \mathbb{R}^{|J_1|}_+ \times \mathbb{R}^{|J_1|}_+$ as the Lagrange multiplier associated with x^{k+1}. Set $\mu^{k+1} = ((\mu^H)^{k+1}, (\mu^G)^{k+1})$ assigning zero values to the undefined components; that is, $(\mu^G_{J_2})^{k+1} = 0$. Increase k by one and go to the identification step.

Theorem 3. *Let the function f and the mappings H and G be twice differentiable in a neighborhood of a strongly stationary point $\bar{x} \in \mathbb{R}^n$ of problem (*1.1*), and let their second derivatives be continuous at this point. Assume that the MPVC-strict Mangasarian–Fromovitz constraint qualification* (2.9), (2.10) *is fulfilled at $\bar{x}$ for the* (*unique*) *MPVC-multiplier $\bar{\mu} = (\bar{\mu}^H, \bar{\mu}^G)$ associated with $\bar{x}$. Let the piecewise second-order sufficient condition* (2.18) *be also fulfilled.*

Assume that, in Algorithm 1, $(x^{k+1}, ((\mu^H_{J_2})^{k+1}, (\mu^H_{J_1})^{k+1}, (\mu^G_{J_1})^{k+1}))$ *is calculated as a pair that consists of a stationary point and a Lagrange multiplier of quadratic program* (3.6)–(3.8) *and is a closest such pair to* $(x^k, ((\mu^H_{J_2})^k, (\mu^H_{J_1})^k, (\mu^G_{J_1})^k)$. *Then, for every initial approximation* (x^0, μ^0) *that is sufficiently close to* $(\bar{x}, \bar{\mu})$, *Algorithm* 1 *determines a trajectory* $\{(x^k, \mu^k)\}$ *converging to* $(\bar{x}, \bar{\mu})$ *with a superlinear rate. The rate of convergence is quadratic if the second derivatives of f,* H, and G *satisfy the Lipschitz condition with respect to $\bar{x}$.*

Note the following fact. Under the hypotheses of this theorem, we have $\bar{\mu}^G_{I_{00}} = 0$ in view of the second equality in (2.15). If the lower-level strict complementarity condition is violated, then there exist piecewise problems corresponding to $\bar{x}$ for which the strict complementarity condition (that is, the requirement that all the Lagrange multipliers associated with the inequality constraints that are active at $\bar{x}$ be pos-

itive) is not fulfilled. This condition is also violated for TNLP (2.2); consequently, the convergence results for the SQP method based on the strict complementarity condition (e.g., see [10, Theorem 4.4.1]) cannot be applied to these problems. Fortunately, a subtler result of this kind was obtained in [12] (see also [9, Theorem 4.5.2]). Instead of strict complementarity, it only assumes the strict Mangasarian–Fromovitz condition and the second-order sufficient condition.

Proof. As noted above, for every $x^k \in \mathbb{R}^n$ that is sufficiently close to $\bar{x}$, the index sets J_1 and J_2 defined by (3.4) and (3.5) at $x = x^k$ generate problem (3.2), which is identical to the piecewise problem associated with the desired solution $\bar{x}$ up to some inequality constraints that are inactive at $\bar{x}$.

As shown in the proof of Theorem 1, the strong stationarity of $\bar{x}$ in problem (1.1) in combination with the MPVC-strict Mangasarian–Fromovitz constraint qualification fulfilled at this point for the MPVC-multiplier $\bar{\mu}$ implies that $\bar{x}$ is a stationary point for every piecewise problem (and is associated with the same unique Lagrange multiplier $\bar{\mu}$); moreover, for every piecewise problem, the strict Mangasarian–Fromovitz condition is fulfilled at $\bar{x}$ for this multiplier. Finally, according to the second equality in (2.16), the piecewise second-order sufficient condition (2.18) implies that the conventional second-order sufficient condition is fulfilled at $\bar{x}$ for every piecewise problem (again with the same unique multiplier $\bar{\mu}$). Using the results given in [12], we conclude the following: for every point (x^k, μ^k) that is sufficiently close to $(\bar{x}, \bar{\mu})$, one SQP step as applied to any problem (and, hence, to problem (3.2)) obeys the estimate

$$\left\|(x^{k+1} - \bar{x}, \mu^{k+1} - \bar{\mu})\right\| = o(\left\|(x^k - \bar{x}, \mu^k - \bar{\mu})\right\|).$$

In particular, x^{k+1} does not leave the region of the correct identification of the piecewise problem, which implies the required result. The theorem is proved.

Note that, theoretically, the identification step can be performed only once for $k = 0$ rather than at each iteration of Algorithm 1. This does not affect the local superlinear convergence proved in Theorem 3. However, performing the identification step at each iteration can be useful in practice because, intuitively, this makes the algorithm less local. On the other hand, we explain below that the globalization of convergence for the piecewise SQP method seems very problematic even if identification is performed at each iteration step.

4. ACTIVE-SET METHODS

The active-set methods as applied to the MPCC were developed in [13]. Below, we propose active-set methods for the MPVC. The idea behind them is to identify TNLP (2.2) corresponding to the desired solution and apply the SQP method to this problem. To identify TNLP (2.2) means to identify the index sets I_{+0}, I_{0+}, I_{00}, and I_{0-}. Under certain (very weak) assumptions, such an identification can locally be implemented with no significant computational costs using the procedure proposed in [14] and the bound for the distance to the set $\{\bar{x}\} \times \Lambda$ (see inequality (4.14) below) that follows from [15, Lemma 2] and [16, Theorem 2]. An identification technique based on the combination of these methods was first applied in [17], where no constraint qualification was imposed on the problem.

Namely, consider the Karush–Kuhn–Tucker (KKT) system for the original problem (1.1):

$$\frac{\partial L}{\partial x}(x, \lambda) = 0, \tag{4.1}$$

$$\lambda^H \geq 0, \quad H(x) \geq 0, \quad \langle \lambda^H, H(x) \rangle = 0, \tag{4.2}$$

$$\lambda^{GH} \geq 0, \quad G_i(x)H_i(x) \leq 0, \quad i = 1, 2, \ldots, m, \quad \sum_{i=1}^{m} \lambda_i^{GH} G_i(x)H_i(x) = 0. \tag{4.3}$$

This system can be rewritten as the equation

$$\Phi(x, \lambda) = 0,$$

where the mapping $\Phi\colon \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^m) \longrightarrow \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^m)$ is given by the formula

$$\Phi(x, \lambda) = \begin{pmatrix} \dfrac{\partial L}{\partial x}(x, \lambda) \\ (\min\{\lambda_1^H, H_1(x)\}, \ldots, \min\{\lambda_m^H, H_m(x)\}) \\ (\min\{\lambda_1^{GH}, -G_1(x)H_1(x)\}, \ldots, \min\{\lambda_m^{GH}, -G_m(x)H_m(x)\}) \end{pmatrix}.$$

Let $\theta \in (0, 1)$ be a fixed scalar. For the available primal-dual approximation $(x^0, \lambda^0) \in \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^m)$, we set

$$I_+(x^0, \lambda^0) = \{i = 1, 2, \ldots, m \,|\, H_i(x^0) > \|\Phi(x^0, \lambda^0)\|^\theta\}, \tag{4.4}$$

$$I_0(x^0, \lambda^0) = \{i = 1, 2, \ldots, m\} \setminus I_+(x^0, \lambda^0), \tag{4.5}$$

$$I_{+-}(x^0, \lambda^0) = \{i \in I_+(x^0, \lambda^0) \,|\, G_i(x^0) < -\|\Phi(x^0, \lambda^0)\|^\theta\}, \tag{4.6}$$

$$I_{+0}(x^0, \lambda^0) = I_+(x^0, \lambda^0) \setminus I_{+-}(x^0, \lambda^0), \tag{4.7}$$

$$I_{0+}(x^0, \lambda^0) = \{i \in I_0(x^0, \lambda^0) \,|\, G_i(x^0) > \|\Phi(x^0, \lambda^0)\|^\theta\}, \tag{4.8}$$

$$I_{0-}(x^0, \lambda^0) = \{i \in I_0(x^0, \lambda^0) \,|\, G_i(x^0) < -\|\Phi(x^0, \lambda^0)\|^\theta\}, \tag{4.9}$$

$$I_{00}(x^0, \lambda^0) = I_0(x^0, \lambda^0) \setminus (I_{0+}(x^0, \lambda^0) \cup I_{0-}(x^0, \lambda^0)). \tag{4.10}$$

Proposition 2. *Let the function f and the mappings H and G be twice differentiable in a neighborhood of a strongly stationary point $\bar{x} \in \mathbb{R}^n$ of problem* (1.1). *Assume that the second-order sufficient condition* (2.19), (2.10) *is fulfilled at this point for some Lagrange multiplier $\bar{\lambda} = (\bar{\lambda}^H, \bar{\lambda}^{GH}) \in \Lambda$.*

Then, for any fixed scalar $\theta \in (0, 1)$, there exists a neighborhood U of the point $(\bar{x}, \bar{\lambda})$ such that for any $(x^0, \lambda^0) \in U$, we have

$$I_+(x^0, \lambda^0) = I_+, \quad I_{+0}(x^0, \lambda^0) = I_{+0}, \quad I_{+-}(x^0, \lambda^0) = I_{+-}, \tag{4.11}$$

$$I_0(x^0, \lambda^0) = I_0, \quad I_{0+}(x^0, \lambda^0) = I_{0+}, \quad I_{00}(x^0, \lambda^0) = I_{00}, \quad I_{0-}(x^0, \lambda^0) = I_{0-} \tag{4.12}$$

provided that the index sets are defined by (4.4)–(4.10).

Proof. We prove the first equality in (4.11). (The other equalities in (4.11) and (4.12) can be proved in a similar way.)

Let $i \in I_+(x^0, \lambda^0)$ but $i \notin I_+$. Then, $H_i(\bar{x}) = 0$; consequently, in view of (4.4) and the mean value theorem, we have

$$\|\Phi(x^0, \lambda^0)\|^\theta < H_i(x^0) = H_i(x^0) - H_i(\bar{x}) = O(\|x^0 - \bar{x}\|). \tag{4.13}$$

In particular, this means that $x^0 \neq \bar{x}$.

On the other hand, from [15, Lemma 2; 16, Theorem 2], we conclude the following: if (2.19) is fulfilled, then there exists a constant $c > 0$ such that the bound

$$\operatorname{dist}((x, \lambda), \{\bar{x}\} \times \Lambda) \le c\|\Phi(x, \lambda)\|, \tag{4.14}$$

holds for every pair $(x, \lambda) \in \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^m)$ that is sufficiently close to $(\bar{x}, \bar{\lambda})$. In particular, this means that the bound

$$\|x^0 - \bar{x}\| \le c\|\Phi(x^0, \lambda^0)\|.$$

holds if (x^0, λ^0) is sufficiently close to $(\bar{x}, \bar{\lambda})$. This inequality, combined with (4.13), implies the estimate

$$\|x^0 - \bar{x}\|^\theta = O(\|x^0 - \bar{x}\|),$$

which is impossible for $\theta \in (0, 1)$ and $x^0 \neq \bar{x}$. Thus, the inclusion $I_+(x^0, \lambda^0) \subset I_+$ is proved.

Now, assume that $i \in I_+$. Then, $H_i(\bar{x}) > 0$, and the inequality

$$H_i(x^0) \geq \frac{1}{2} H_i(\bar{x}). \tag{4.15}$$

holds for every x^0 that is sufficiently close to $\bar{x}$. On the other hand, $\Phi(x, \lambda) \longrightarrow 0$ as $(x, \lambda) \longrightarrow (\bar{x}, \bar{\lambda})$. Consequently, we have

$$\left\|\Phi(x^0, \lambda^0)\right\|^\theta < \frac{1}{2} H_i(\bar{x}). \tag{4.16}$$

if (x^0, λ^0) is sufficiently close to $(\bar{x}, \bar{\lambda})$. Combining (4.15) and (4.16), we obtain the inequality $\|\Phi(x^0, \lambda^0)\|^\theta < H_i(x^0)$. Now, (4.4) implies that $i \in I_+(x^0, \lambda^0)$. This proves the inclusion $I_+ \subset I_+(x^0, \lambda^0)$, which completes the proof of the first equality in (4.11).

Now, we give a formal description of the active-set method with the identification performed only at the initial point.

Algorithm 2

Preliminary step. Fix $\theta \in (0, 1)$. Set $k = 0$ and choose $x^0 \in \mathbb{R}^n$ and $\lambda^0 = ((\lambda^H)^0, (\lambda^{GH})^0) \in \mathbb{R}^m \times \mathbb{R}^m$.

Identification step. Define the index sets $I_{+0} = I_{+0}(x^0, \lambda^0)$, $I_{0+} = I_{0+}(x^0, \lambda^0)$, $I_{00} = I_{00}(x^0, \lambda^0)$, and $I_{0-} = I_{0-}(x^0, \lambda^0)$ in accordance with formulas (4.4)–(4.10).

Initialization step. Define $(\mu^H_{I_0})^0$ and $(\mu^G_{I_{+0} \cup I_{00}})^0$ by formulas given in Proposition 1; namely,

$$(\mu^H_i)^0 = (\lambda^H_i)^0 - (\lambda^{GH}_i)^0 G_i(x^0), \quad i \in I_{0+} \cup I_{0-}, \quad (\mu^H_i)^0 = (\lambda^H_i)^0, \quad i \in I_{00}, \tag{4.17}$$

$$(\mu^G_i)^0 = (\lambda^{GH}_i)^0 H_i(x^0), \quad i \in I_{+0}, \quad (\mu^G_i)^0 = 0, \quad i \in I_{00}, \tag{4.18}$$

and set

$$(\mu^H_{I_+})^0 = 0, \quad (\mu^G_{I_{+-} \cup I_{0+} \cup I_{0-}})^0 = 0.$$

SQP step. Calculate $x^{k+1} \in \mathbb{R}^n$ as a stationary point of the quadratic programing problem

$$\langle f'(x^k), x - x^k \rangle + \frac{1}{2} \left\langle \frac{\partial^2 \mathcal{L}}{\partial x^2}(x^k, \mu^k)(x - x^k), x - x^k \right\rangle \longrightarrow \min, \tag{4.19}$$

$$H_{I_{0+} \cup I_{00}}(x^k) + H'_{I_{0+} \cup I_{00}}(x^k)(x - x^k) = 0, \quad H_{I_{0-}}(x^k) + H'_{I_{0-}}(x^k)(x - x^k) \geq 0, \tag{4.20}$$

$$G_{I_{+0} \cup I_{00}}(x^k) + G'_{I_{+0} \cup I_{00}}(x^k)(x - x^k) \leq 0, \tag{4.21}$$

Calculate $((\mu^H_{I_{0+} \cup I_{00}})^{k+1}, (\mu^H_{I_{0-}})^{k+1}, (\mu^G_{I_{+0} \cup I_{00}})^{k+1}) \in \mathbb{R}^{|I_{0+} \cup I_{00}|} \times \mathbb{R}^{|I_{0-}|}_+ \times \mathbb{R}^{|I_{+0} \cup I_{00}|}_+$ as the Lagrange multiplier associated with x^{k+1}. Set $\mu^{k+1} = ((\mu^H)^{k+1}, (\mu^G)^{k+1})$ assigning zero values to the undefined components; that is,

$$(\mu^H_{I_+})^{k+1} = 0, \quad (\mu^G_{I_{+-} \cup I_{0+} \cup I_{0-}})^{k+1} = 0. \tag{4.22}$$

Increase k by one and go to the beginning of the SQP step.

Theorem 4. *Let the function f and the mappings H and G be twice differentiable in a neighborhood of a strongly stationary point $\bar{x} \in \mathbb{R}^n$ of problem* (1.1), *and let their second derivatives be continuous at this point. Assume that the MPVC-strict Mangasarian–Fromovitz constraint qualification* (2.9), (2.10) *is fulfilled at $\bar{x}$ for the* (*unique*) *MPVC-multiplier $\bar{\mu} = (\bar{\mu}^H, \bar{\mu}^G)$ associated with $\bar{x}$. Let the second-order sufficient condition* (2.19) *be also fulfilled for some Lagrange multiplier $\bar{\lambda} = (\bar{\lambda}^H, \bar{\lambda}^{GH}) \in \Lambda$.*

Assume that, in Algorithm 2, $(x^{k+1}, ((\mu^H_{I_{0+} \cup I_{00}})^{k+1}, (\mu^H_{I_{0-}})^{k+1}, (\mu^G_{I_{+0} \cup I_{00}})^{k+1}))$ *is calculated as a pair that consists of a stationary point and a Lagrange multiplier of quadratic program* (4.19)–(4.21) *and is a closest such pair to* $(x^k, ((\mu^H_{I_{0+} \cup I_{00}})^k, (\mu^H_{I_{0-}})^k, (\mu^G_{I_{+0} \cup I_{00}})^k))$. *Then, for every initial approximation* (x^0, λ^0) *that is suf-*

ficiently close to $(\bar{x}, \bar{\lambda})$, *Algorithm* 2 *determines a trajectory* $\{(x^k, \mu^k)\}$ *converging to* $(\bar{x}, \bar{\mu})$ *with a superlinear rate. The rate of convergence is quadratic if the second derivatives of f, H, and G satisfy the Lipschitz condition with respect to* $\bar{x}$.

Proof. By Proposition 2, for every initial approximation (x^0, λ^0) that is sufficiently close to $(\bar{x}, \bar{\lambda})$, the index sets I_{+0}, I_{0+}, I_{00}, and I_{0-} defined at the identification step of Algorithm 2 are identical to $I_{+0}(\bar{x})$, $I_{0+}(\bar{x})$, $I_{00}(\bar{x})$, and $I_{0-}(\bar{x})$, respectively.

By Theorem 1, the strong stationarity of $\bar{x}$ in problem (1.1) in combination with the MPVC-strict Mangasarian–Fromovitz constraint qualification fulfilled at this point for the MPVC-multiplier $\bar{\mu}$ implies that $\bar{x}$ is a stationary point of TNLP (2.2) (and is associated with the same unique Lagrange multiplier $\bar{\mu}$). Also, recall that the MPVC-strict Mangasarian–Fromovitz constraint qualification is the strict Mangasarian-Fromovitz condition for TNLP (2.2).

As indicated at the end of Section 2, the second-order sufficient condition (2.19) implies the piecewise sufficient condition (2.18). In turn, the latter implies that the second-order sufficient condition is fulfilled at $\bar{x}$ for TNLP (2.2) (again, with the same unique multiplier $\bar{\mu}$). Indeed, according to (2.16), the critical cone of TNLP (2.2) at the point $\bar{x}$, given by

$$C_{\mathrm{NMPP}} = C_{\mathrm{NMPP}}(\bar{x}) = \{\xi \in \mathbb{R}^n \mid H'_{I_{0+} \cup I_{00}}(\bar{x})\xi = 0,\ H'_{0-}(\bar{x})\xi \geq 0,\ G'_{I_{+0} \cup I_{00}}(\bar{x})\xi \leq 0,\ \langle f'(\bar{x}), \xi\rangle \leq 0\} \quad (4.23)$$

is contained in C_2.

Finally, if (x^0, λ^0) is sufficiently close to $(\bar{x}, \bar{\lambda})$, then the pair $((\mu^H_{I_0})^0, (\mu^G_{I_{+0} \cup I_{00}})^0)$ defined by the formulas (4.17) and (4.18) is sufficiently close to $(\bar{\mu}^H_{I_0}, \bar{\mu}^G_{I_{+0} \cup I_{00}})$ because, by Proposition 1, the latter pair satisfies (2.14) and (2.15) at $\lambda = \bar{\lambda}$.

Now, the required assertion follows from available results on the superlinear local convergence of SQP methods (see [12]). The theorem is proved.

A natural analog of Algorithm 2 is the one in which the identification step is performed at each iteration rather than only at the initial point. At each iteration of such algorithm, one should use the available approximation μ^k to the MPVC multiplier to find an approximation λ^k to the genuine Lagrange multiplier; the latter is required at the identification step. The necessity of the repeated mutual updating of λ^k and μ^k makes this algorithm and its local analysis significantly more complex. However, compared to Algorithms 1 and 2, the modified algorithm is much more suitable for the globalization of convergence. Indeed, the proximity to the points satisfying the KKT system (4.1)–(4.3) can be controlled using some globally defined performance criterion such as $|\Phi(\cdot)|$. By contrast, finding MPVC multipliers involves certain index sets depending on a specific $\bar{x}$; consequently, it is hardly possible to indicate a global performance criterion specifying MPVC multipliers.

Recall that the piecewise second-order sufficient condition (2.18), which is required for the local superlinear convergence of the piecewise SQP method, is at least not stronger than the existence of a Lagrange multiplier $\bar{\lambda} \in \Lambda$ satisfying (2.19). The latter condition is required for the local superlinear convergence of the active-set method. On the other hand, as noted above, the active-set method is more promising for the globalization of convergence. Moreover, the step of Algorithm 2 can be made considerably cheaper if, instead of the MPVC-strict Mangasarian–Fromovitz constraint qualification (2.9), (2.10), the stronger linear independence constraint qualification (2.11) is fulfilled. In this case, subproblem (4.19)–(4.21) can be replaced by a quadratic program with equality constraints. Then, finding stationary points reduces to solving a system of linear equations.

Namely, let the sets I_{+0}, I_{0+}, I_{00}, and I_{0-} be identified. Then, instead of TNLP (2.2), we solve the associated problem in which all the inequality constraints are replaced by equalities; that is,

$$f(x) \longrightarrow \min, \quad H_{I_0}(x) = 0, \quad G_{I_{+0} \cup I_{00}}(x) = 0. \quad (4.24)$$

Accordingly, the SQP step is replaced in Algorithm 2 by the Newton method step for the Lagrangian system of problem (4.24). Namely, for the current approximation $(x^k, \mu^k) \in \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^m)$, we calculate the point $(x^{k+1}, ((\mu^H_{I_0})^{k+1}, (\mu^G_{I_{+0} \cup I_{00}})^{k+1}))$ as a solution to the linear system

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(x^k, \mu^k)(x - x^k) + (H'_{I_0}(x^k))^{\mathrm{T}}(\mu^H - (\mu^H)^k)_{I_0} + (G'_{I_{+0} \cup I_{00}}(x^k))^{\mathrm{T}}(\mu^G - (\mu^G)^k)_{I_{+0} \cup I_{00}} = -\frac{\partial \mathcal{L}}{\partial x}(x^k, \mu^k),$$

$$H_{I_0}(x^k) + H'_{I_0}(x^k)(x - x^k) = 0, \quad G_{I_{+0} \cup I_{00}}(x^k) + G'_{I_{+0} \cup I_{00}}(x^k)(x - x^k) = 0.$$

As before, the remaining components $\mu^{k+1} = ((\mu^H)^{k+1}, (\mu^G)^{k+1})$ are set equal to zero (see (4.22)).

Observe that, if the sets I_{+0}, I_{0+}, I_{00}, and I_{0-} are correctly identified, then a weakly stationary point $\bar{x}$ of MPVC (1.1) (and, even more so, a strongly stationary point) is a stationary point of problem (4.24). Moreover, the MPVC linear independence condition (2.11) is the conventional constraint qualification for problem (4.24) at the point $\bar{x}$. If this condition is fulfilled, then the MPVC multiplier $\bar{\mu}$ is the unique Lagrange multiplier of problem (4.24) associated with $\bar{x}$.

Furthermore, as repeatedly noted above, the second-order sufficient condition (2.19) fulfilled for some Lagrange multiplier $\bar{\lambda} \in \Lambda$ implies that the piecewise second-order sufficient condition (2.18) is also fulfilled. The stationarity of $\bar{x}$ in problem (4.24) is equivalent to the relation

$$\langle f'(\bar{x}), \xi \rangle = 0 \quad \forall \xi \in \ker H'_{I_0}(\bar{x}) \cap \ker G'_{I_{+0} \cup I_{00}}(\bar{x}).$$

Then, (2.16) entails the inclusion

$$\ker H'_{I_0}(\bar{x}) \cap \ker G'_{I_{+0} \cup I_{00}}(\bar{x}) \subset C_2.$$

In view of (2.18), the standard second-order sufficient optimality condition for problem (4.24) is fulfilled at $\bar{x}$; that is,

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] > 0 \quad \forall \xi \in \ker H'_{I_0}(\bar{x}) \cap \ker G'_{I_{+0} \cup I_{00}}(\bar{x}) \backslash \{0\}.$$

Thus, we showed that all the conditions for the local superlinear convergence of Newton's method as applied to the Lagrangian system of problem (4.24) are fulfilled at the point $(\bar{x}, \bar{\mu})$ (e.g., see [10, Theorem 4.3.1]). It follows that an analog of Theorem 4 holds for the above modification of Algorithm 2. Moreover, there is no need to require any proximity of the next approximation to the current one. In addition, we can claim that the trajectory $\{(x^k, \mu^k)\}$ is uniquely determined.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 07-01-00270, 07-01-00416, and 08-01-90001-Bel) and by the Russian Federation President's Grant no. NSh-693.2008.1 for the support of leading scientific schools.

REFERENCES

1. W. Achtziger and C. Kanzow, "Mathematical Programs with Vanishing Constraints: Optimality Conditions and Constraint Qualifications," Math. Program., **114** (1), 69–99 (2007).
2. T. Hoheisel and C. Kanzow, "On the Abadie and Guignard Constraint Qualifications for Mathematical Programs with Vanishing Constraints," Optimization. DOI 10.1080/02331930701763405.
3. T. Hoheisel and C. Kanzow, "Stationarity Conditions for Mathematical Programs with Vanishing Constraints using Weak Constraint Qualifications," J. Math. Anal. Appl. **337**, 292–310 (2008).
4. T. Hoheisel and C. Kanzow, First- and Second-Order Optimality Conditions for Mathematical Programs with Vanishing Constraints," Appl. of Math. **52**, 495–514 (2007).
5. A. F. Izmailov and M. V. Solodov, "Mathematical Programs with Vanishing Constraints: Optimality Conditions, Sensitivity, and a Relaxation Method," J. Optim. Theory Appl. (2009) DOI 10.1007/810957-009-9517-4
6. Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints* (Cambridge Univ. Press, Cambridge, 1996).
7. J. V. Outrata, M. Kocvara, and J. Zowe, *Nonsmooth Approach to Mathematical Programs with Equilibrium Constraints: Theory, Applications, and Numerical Results* (Kluwer, Boston, 1998).

8. A. F. Izmailov, "Mathematical Programs with Complementarity Constraints: Regularity, Optimality Conditions, and Sensitivity," Zh. Vychisl. Mat. Mat. Fiz. **44** (7), 1209–1228 (2004) [Comput. Math. Math. Phys. **44**, 1145–1164 (2004)].
9. A. F. Izmailov, *Sensitivity in Optimization* (Fizmatlit, Moscow, 2006) [in Russian].
10. A. F. Izmailov and V. M. Solodov, *Numerical Optimization Methods*, 2nd. ed. (Fizmatlit, Moscow, 2008) [in Russian].
11. D. Ralph, "Sequential Quadratic Programming for Mathematical Programs with Linear Complementarity Constraints," *Computational Techniques and Applications CTAC95* (World Sci., Singapore, 1996), pp. 663–668.
12. J. F. Bonnans, "Local Analysis of Newton-Type Methods for Variational Inequalities and Nonlinear Programming," Appl. Math. Optim. **29**, 161–186 (1994).
13. A. F. Izmailov and M. V. Solodov, "An Active-Set Newton Method for Mathematical Programs with Complementarity Constraints," SIAM J. Optim. **19**, 1003–1027 (2008).
14. F. Facchinei, A. Fischer, and C. Kanzow, "On the Accurate Identification of Active Constraints," SIAM. J. Optim. **9**, 14–32 (1999).
15. W. W. Hager and M. S. Gowda, "Stability in the Presence of Degeneracy and Error Estimation," Math. Program. **85**, 181–192 (1999).
16. A. Fischer, "Local Behavior of an Iterative Framework for Generalized Equations with Nonisolated Solutions," Math. Program. **94,** 91–124 (2002).
17. A. F. Izmailov and M. V. Solodov, "Newton-Type Methods for Optimization Problems without Constraint Qualifications," SIAM J. Optim. **15**, 210–228 (2004).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1141–1150. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © E.V. Zakharov, A.V. Kalinin, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1197–1206.

Method of Boundary Integral Equations as Applied to the Numerical Solution of the Three-Dimensional Dirichlet Problem for the Laplace Equation in a Piecewise Homogeneous Medium

E. V. Zakharov and A. V. Kalinin

Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, 119992 Russia
e-mail: zspec@cs.msu.ru, alec.kalinin@gmail.com

Received October 17, 2008; in final form, December 15, 2008

Abstract—A Dirichlet problem is considered in a three-dimensional domain filled with a piecewise homogeneous medium. The uniqueness of its solution is proved. A system of Fredholm boundary integral equations of the second kind is constructed using the method of surface potentials, and a system of boundary integral equations of the first kind is derived directly from Green's identity. A technique for the numerical solution of integral equations is proposed, and results of numerical experiments are presented.

DOI: 10.1134/S0965542509070070

Key words: Dirichlet problem for the Laplace equation, piecewise homogeneous medium, method of boundary integral equations.

1. INTRODUCTION

The conjugation problem for the Laplace equation is a classical model in the theory of direct currents in piecewise homogeneous conducting media (see [1]). It naturally arises in the theory of direct-current electrical exploration (see [2, 3]) and in the simulation of electrical engineering systems (see, e.g., [4]). This class of problems concerns to boundary value problems in unbounded domains.

In the study of bioelectric phenomena, the conjugation problem models tissues inhomogeneities and most frequently arises as an interior problem with Dirichlet or Neumann boundary conditions. Specifically, interest in Dirichlet problems for piecewise homogeneous three-dimensional domains has relatively recently arisen in computational cardiac electrophysiology. Examples of problems in this area are direct and inverse electrocardiology problems (see [5]) and the modeling of cardiac excitation based on the bidomain model equations (see [6]). The conjugation problem also arises in medical diagnostics related to the processing of electroencephalography and impedancemetry data, specifically, in impedance tomography (see [7]). Many of these issues lead to inverse problems and the design of algorithms for their solution involves the development of effective methods for solving direct problems. An example of the latter is the three-dimensional Dirichlet problem for the Laplace equation in piecewise homogeneous media.

In this paper, we give the mathematical formulation of the three-dimensional Dirichlet problem and prove the uniqueness of its solution. Additionally, a system of Fredholm integral equations of the second kind is constructed and the existence of a solution to this system and the problem itself is proved. We construct a system of integral equations of the first kind with a weak singularity in the kernel and develop numerical algorithms for its solution based on interpolation and collocations (see [8–10]). Numerical results are presented.

2. FORMULATION OF THE PROBLEM AND A UNIQUENESS THEOREM

Consider a domain $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots \cup \Omega_N$ in R^3 (see Fig. 1). The boundaries Γ_i of Ω_i ($i = 0, 1, \ldots, N$) are sufficiently smooth (Lyapunov surfaces). The problem is formulated as follows.

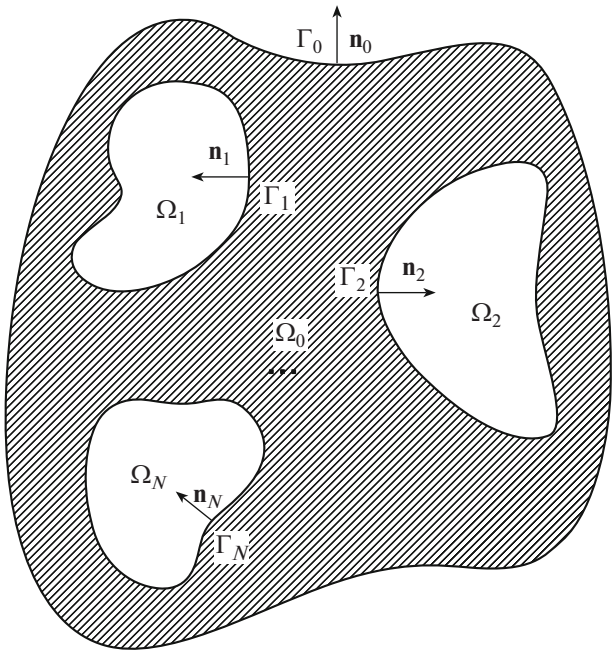


Fig. 1.

Find a function $u(x)$ such that $u \in C(\overline{\Omega})$; $u(x) = u_i(x)$, $x \in \Omega_i$, $i = 0, 1, ..., N$, where $u_i \in C^2(\Omega_i) \cap C^1(\overline{\Omega}_i)$ and

$$\Delta u_i(x) = 0, \quad x \in \Omega_i, \quad i = 0, 1, \ldots, N, \tag{1}$$

$$u_0(x) = U_0(x), \quad x \in \Gamma_0, \quad U_0(x) \in C(\Gamma_0). \tag{2}$$

The transmission conditions

$$u_0(x) = u_i(x), \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N, \tag{3}$$

$$k_0 \frac{\partial u_0(x)}{\partial n} = k_i \frac{\partial u_i(x)}{\partial n}, \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N. \tag{4}$$

are on Γ_i, $i = 1, 2,, ..., N$. Here, k_i ($i = 0, 1, ..., N$) are positive and finite parameters.

Theorem 1. *The solution to problem* (1)–(4) *is unique.*

Proof. Let $\tilde{u}(x)$ and $\tilde{\tilde{u}}(x)$ be solutions to problem (1)–(4). Define the function

$$w(x) = \tilde{u}(x) - \tilde{\tilde{u}}(x), \quad x \in \Omega.$$

Then we have

$$\Delta w_i(x) = 0, \quad x \in \Omega_i, \quad i = 0, 1, \ldots, N,$$

$$w_0(x) = 0, \quad x \in \Gamma_0.$$

On Γ_i, $i = 1, 2, ..., N$, the following transmission conditions hold:

$$w_0(x) = w_i(x), \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N, \tag{5}$$

$$k_0 \frac{\partial w_0(x)}{\partial n} = k_i \frac{\partial w_i(x)}{\partial n}, \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N. \tag{6}$$

Using the first Green's identity in the multiply connected domain Ω_0 yields

$$\int_{\Omega_0} w_0 \Delta w_0 dx = \int_{\Gamma_0} w_0 \frac{\partial w_0}{\partial n} ds - \int_{\Omega_0} |\operatorname{grad} w_0|^2 dx + \sum_{i=1}^{N} \int_{\Gamma_i} w_0 \frac{\partial w_0}{\partial n} ds. \tag{7}$$

Since $\Delta w_0(x) = 0$ for $x \in \Omega_0$ and $w_0(x) = 0$ for $x \in \Gamma_0$, relation (7) becomes

$$\sum_{i=1}^{N} \int_{\Gamma_i} w_0 \frac{\partial w_0}{\partial n} ds - \int_{\Omega_0} |\mathrm{grad}\, w_0|^2 dx = 0. \tag{8}$$

For Ω_i, $i = 1, 2, \ldots, N$, the Green's identity implies

$$\int_{\Omega_i} w_i \Delta w_i dx = - \int_{\Gamma_i} w_i \frac{\partial w_i}{\partial n} ds - \int_{\Omega_i} |\mathrm{grad}\, w_i|^2 dx. \tag{9}$$

Since $\Delta w_i(x) = 0$ for $x \in \Omega_i$, $i = 1, 2, \ldots, N$, Eq. (9) becomes

$$- \int_{\Gamma_i} w_i \frac{\partial w_i}{\partial n} ds - \int_{\Omega_i} |\mathrm{grad}\, w_i|^2 dx = 0.$$

In view of (5), this can be rewritten as

$$- \int_{\Gamma_i} w_0 \frac{\partial w_i}{\partial n} ds - \int_{\Omega_i} |\mathrm{grad}\, w_i|^2 dx = 0. \tag{10}$$

Equation (8) is multiplied by k_0 and each ith expression in (10) is multiplied by k_i and is added to obtain

$$\sum_{i=1}^{N} \int_{\Gamma_i} w_0 \left(k_0 \frac{\partial w_0}{\partial n} - k_i \frac{\partial w_i}{\partial n} \right) ds - \sum_{i=0}^{N} k_i \int_{\Omega_i} |\mathrm{grad}\, w_i|^2 dx = 0. \tag{11}$$

In view of (6), the first sum in (11) vanishes. Thus, (11) becomes

$$\sum_{i=0}^{N} k_i \int_{\Omega_i} |\mathrm{grad}\, w_i|^2 dx = 0. \tag{12}$$

Since $k_i > 0$, identity (12) vanishes if and only if $\mathrm{grad} w_i(x) = 0$ in Ω_i. Therefore, $w_i(x) = \mathrm{const}$ for $i = 0, 1, \ldots, N$, and the condition $w_0(x) = 0$, $x \in \Gamma_0$, implies that $w(x) = 0$. Thus, the solution to problem (1)–(4) is unique.

3. CONSTRUCTION OF A SYSTEM OF FREDHOLM BOUNDARY INTEGRAL EQUATIONS OF THE SECOND KIND

Let $u(x)$ ($x \in \Omega$) be a solution to problem (1)–(4). Suppose that there exist functions $\mu_j(y)$, $y \in \Gamma_j$, $j = 0, 1, \ldots, N$ such that $u(x)$ can be represented as

$$u(x) = \int_{\Gamma_0} \mu_0(y) \frac{\partial}{\partial n_y} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \int_{\Gamma_j} \mu_j(y) \frac{k_0 - k_j}{|x-y|} ds_y, \tag{13}$$

where $|x-y|$ is the distance between the points x and y; μ_0 is the double-layer potential density on the surface Γ_0; and μ_j is the single-layer potential density on Γ_j, $j = 1, 2, \ldots, N$.

Note that (13) automatically satisfies conditions (1) and (3), while the fulfillment of conditions (2) and (4) leads to a system of $(N+1)$ integral equations. The first equation of the system is constructed as follows.

Let a point x be dropped from the domain Ω_0 onto the surface Γ_0 in (13). By the well-known properties of double-layer potentials, we obtain the following equation on Γ_0:

$$2\pi \mu_0(x) + \int_{\Gamma_0} \mu_0(y) \frac{\partial}{\partial n_0} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \int_{\Gamma_j} \mu_j(y) \frac{k_0 - k_j}{|x-y|} ds_y = U_0(x), \quad x \in \Gamma_0. \tag{14}$$

Differentiating (13) along the normal n_i to Γ_i ($i = 1, 2, ..., N$) gives N representations

$$\frac{\partial u(x)}{\partial n_i} = \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{k_0 - k_j}{|x-y|} ds_y, \quad x \in \Omega. \tag{15}$$

Let a point x be dropped from Ω_0 onto each surface Γ_i ($i = 1, 2, ..., N$) in (15). By the properties of normal derivatives of a single-layer potential, we obtain integral equations for μ_i:

$$\frac{\partial u_0(x)}{\partial n} = (k_0 - k_i) 2\pi \mu_i(x) + \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{k_0 - k_j}{|x-y|} ds_y, \tag{16}$$
$$x \in \Gamma_i, \quad i = 1, 2, \ldots, N.$$

Dropping in (15) a point x from Ω_i onto Γ_i, $i = 1, 2, ..., N$, respectively, we obtain the integral equations

$$\frac{\partial u_i(x)}{\partial n} = -(k_0 - k_i) 2\pi \mu_i(x) + \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{k_0 - k_j}{|x-y|} ds_y, \tag{17}$$
$$x \in \Gamma_i, \quad i = 1, 2, \ldots, N.$$

Subtracting (17) times $-k_i$ from (16) times k_0 and taking into account conditions (4) gives

$$2\pi(k_0 + k_i)(k_0 - k_i)\mu_i(x) + (k_0 - k_i) \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y$$
$$+ \sum_{j=1}^{N} \left((k_0 - k_i)(k_0 - k_j) \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{1}{|x-y|} ds_y \right).$$

Thus, we have derived N integral equations

$$2\pi \mu_i(x) + \frac{1}{k_0 + k_i} \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \left(\frac{k_0 - k_j}{k_0 + k_i} \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{1}{|x-y|} ds_y \right) = 0, \quad i = 1, 2, \ldots, N. \tag{18}$$

Combining (14) with (18), we obtain the system of integral equations

$$\begin{aligned} &2\pi \mu_0(x) + \int_{\Gamma_0} \mu_0(y) \frac{\partial}{\partial n_0} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \left((k_0 - k_j) \int_{\Gamma_j} \mu_j(y) \frac{1}{|x-y|} ds_y \right) = U_0(x), \\ &2\pi \mu_i(x) + \frac{1}{k_0 + k_i} \int_{\Gamma_0} \mu_0(y) \frac{\partial^2}{\partial n_0 \partial n_i} \frac{1}{|x-y|} ds_y + \sum_{j=1}^{N} \left(\frac{k_0 - k_j}{k_0 + k_i} \int_{\Gamma_j} \mu_j(y) \frac{\partial}{\partial n_i} \frac{1}{|x-y|} ds_y \right) = 0, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{19}$$

Note that (19) is a system of Fredholm integral equations of the second kind in a space of continuous functions, since the matrix of free terms is diagonal and the kernels of the system are either continuous or have a weak singularity at coinciding arguments. Therefore, the unique solvability of this system follows from the Fredholm first theorem if we show that the system of homogeneous integral equations corresponding to (19) has only the trivial solution.

Since system (19) is equivalent to the original problem, the unique solvability of (19) implies the unique solvability of problem (1)–(4).

Theorem 2. *Problem* (1)–(4) *has a solution.*

Proof. Consider homogeneous system (19) (with $U_0(x) \equiv 0$). Let continuous functions $\mu_0^{\text{hom}}, ..., \mu_N^{\text{hom}}$ be a nontrivial solution to this system.

By using formula (13), homogeneous systems (19) generates a solution $u^{\text{hom}}(x)$ to homogeneous problem (1)–(4). Consider the following interior Dirichlet problems in the domains Ω_i:

$$\Delta u_i^{\text{hom}}(x) = 0, \; x \in \Omega_i,$$
$$u_i^{\text{hom}}(x) = 0, \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N.$$

If each of the functions $u_i^{\text{hom}}(x)$ is represented as a double-layer potential with the density μ_i^{hom}, then, on each boundary Γ_i, we obtain the integral equation

$$\mu_i^{\text{hom}}(x) - \frac{1}{2\pi}\int_{\Gamma_i}\mu_i^{\text{hom}}(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y = 0,$$

which has only the trivial solution. Thus, we have

$$\mu_i^{\text{hom}} = 0, \quad i = 1, 2, \ldots, N.$$

Substituting μ_i^{hom} into the first equation in (19) yields an equation for μ_0^{hom}:

$$2\pi\mu_0^{\text{hom}}(x) + \int_{\Gamma_0}\mu_0^{\text{hom}}(y)\frac{\partial}{\partial n_0}\frac{1}{|x-y|}ds_y = 0,$$

which also has only the trivial solution.

Thus, the vector $\mu(x) = [\mu_0(x), \ldots, \mu_N(x)]^{\mathrm{T}} = 0$ solves homogeneous system (19), and, by the uniqueness theorem, there are no functions $\mu_i(x)$ that are nonzero. Therefore, the inhomogeneous system of integral equations (19) with any given function $U_0(x) \in C(\Gamma_0)$ is uniquely solvable, which proves the unique solvability of problem (1)–(4).

4. CONSTRUCTION OF A SYSTEM OF FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND AND A METHOD FOR ITS NUMERICAL SOLUTION

Along with system (19) of Fredholm integral equations of the second kind, the original differential problem (1)–(4) can be reduced to a system of integral equations of the first kind, which has a simpler structure. Specifically, the latter system can be immediately written for the function and its normal derivative and does not contain any second normal derivatives. Let us construct this system.

For the domain Ω_0 with a multiply connected boundary $\Gamma_0 \cup \ldots \cup \Gamma_N$ and outward normals, using the third Green's identity and condition (3), we can write $N + 1$ equations

$$2\pi u_i(x) = \sum_{j=0}^{N}\left(\int_{\Gamma_j} q_j^+(y)\frac{1}{|x-y|}ds_y - \int_{\Gamma_j} u_j(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y\right), \tag{20}$$

where $i = 0, 1, \ldots, N$; $x \in \Gamma_i$ are collocation points; $y \in \Gamma_j$ are integration points; $|x-y|$ is the distance between the points x and y; and $q_j^+(y) = \partial u_0(y)/\partial n_y$.

In turn, for each domain Ω_i with a simply connected boundary Γ_i and outward normals, we can write N boundary integral equations

$$-2\pi u_i(x) = \int_{\Gamma_i} q_i^-(y)\frac{1}{|x-y|}ds_y - \int_{\Gamma_i} u_i(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y, \tag{21}$$

where $i = 1, 2, \ldots, N$; $x \in \Gamma_i$ are collocation points; $y \in \Gamma_i$ are integration points; $|x-y|$ is the distance between the points x and y; and $q_i^-(y) = \partial u_i(y)/\partial n_y$. In view of conditions (4) we have

$$q_i^-(y) = \frac{k_0}{k_i}q_i^+(y), \quad i = 1, 2, \ldots, N, \quad y \in \Gamma_i,$$

and Eq. (21) can be written as

$$-2\pi u_i(x) = \int_{\Gamma_i}\frac{k_0}{k_i}q_i^+(y)\frac{1}{|x-y|}ds_y - \int_{\Gamma_i} u_i(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y, \quad i = 1, 2, \ldots, N. \tag{22}$$

Combining (20) and (22) yields a system of Fredholm integral equations of the first kind

$$2\pi u_i(x) = \sum_{j=0}^{N}\left(\int_{\Gamma_j} q_j^+(y)\frac{1}{|x-y|}ds_y - \int_{\Gamma_j} u_j(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y\right), \quad x \in \Gamma_i, \quad i = 0, 1, \ldots, N,$$
$$-2\pi u_i(x) = \int_{\Gamma_i}\frac{k_0}{k_i}q_i^+(y)\frac{1}{|x-y|}ds_y - \int_{\Gamma_i} u_i(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y, \quad x \in \Gamma_i, \quad i = 1, 2, \ldots, N. \tag{23}$$

It can be shown that system (23) is equivalent to original problem (1)–(4); moreover, it is uniquely solvable if there exists a solution to problem (1)–(4), and its solution can be obtained by the method of interpolation and collocations (see [8–10]).

Following [11], we pass to a discrete representation of system (23). The surfaces Γ_i ($i = 0, 1, \ldots, N$) are triangulated and each Γ_i is represented as a collection of boundary elements ds_p: $\Gamma_i = ds_1 \cup \ldots \cup ds_m$. Let $\varphi_1, \ldots, \varphi_m$ be a system of m linearly independent basis elements (characteristic functions) defined as

$$\varphi_p(s) = \begin{cases} 1, & s \in ds_p \\ 0, & s \notin ds_p. \end{cases}$$

The function $u(x)$ and its normal derivative are represented as expansions in terms of φ_p (piecewise constant approximation):

$$u(s) = \sum_{p=1}^{m}\alpha_p\varphi_p(s), \quad q(s) = \sum_{p=1}^{m}\beta_p\varphi_p(s), \tag{24}$$

where α_p and β_p are the values of $u(s)$ and $q(s)$, respectively, at the barycenter of the pth boundary element.

Then, after discretizing, system (23) becomes

$$2\pi u_i = \sum_{j=0}^{N}(G_{ij}q_j^+ - \hat{H}_{ij}u_j), \quad i = 0, 1, \ldots, N,$$
$$-2\pi u_i = \frac{k_0}{k_i}G_{ii}q_i^+ - \hat{H}_{ii}u_i, \quad i = 1, 2, \ldots, N, \tag{25}$$

where the matrices G_{ij} are obtained by discretizing integrals of the form

$$\int_{\Gamma_j}\frac{1}{|x-y|}ds_y, \quad x \in \Gamma_i, \tag{26}$$

and the matrices $\hat{H}_{ij}$ are obtained by discretizing integrals of the form

$$\int_{\Gamma_j}\frac{\partial}{\partial n_y}\frac{1}{|x-y|}ds_y, \quad x \in \Gamma_i. \tag{27}$$

The matrices H_{ij}^+ and H_{ij}^- are defined as

$$H_{ij}^+ = \begin{cases} \hat{H}_{ij}, & i \neq j \\ \hat{H}_{ij} + 2\pi E, & i = j, \end{cases}$$
$$H_{ij}^- = \begin{cases} \hat{H}_{ij}, & i \neq j \\ \hat{H}_{ij} - 2\pi E, & i = j, \end{cases} \tag{28}$$

where E is the identity matrix.

Then system (25) can be written as

$$\sum_{j=0}^{N} H_{ij}^{+} u_j = \sum_{j=0}^{N} G_{ij} q_j^{+}, \quad i = 0, 1, \ldots, N,$$
$$H_{ii}^{-} u_i = \frac{k_0}{k_i} G_{ii} q_i^{+}, \quad i = 1, 2, \ldots, N. \tag{29}$$

System (29) contains $(2N + 1)$ equations and $(4N + 2)$ unknowns. A system of $(N + 1)$ equations with $(N + 1)$ unknowns can be derived from it as follows. Expressing q^+ from the second part of system (29) and substituting it into the first part of (29) yields the system

$$\begin{aligned}
R_{01}u_1 + R_{02}u_2 + \ldots + R_{0N}u_N &= G_{00}q_0^{+} - H_{00}^{+}u_0,\\
R_{11}u_1 + R_{12}u_2 + \ldots + R_{1N}u_N &= G_{10}q_0^{+} - H_{10}^{+}u_0,\\
&\cdots\cdots\cdots\\
R_{N1}u_1 + R_{N2}u_2 + \ldots + R_{NN}u_N &= G_{N0}q_0^{+} - H_{N0}^{+}u_0,
\end{aligned} \tag{30}$$

where

$$R_{ij} = H_{ij}^{+} - \frac{k_j}{k_0} G_{ij} G_{jj}^{-1} H_{jj}^{-}, \tag{31}$$

$i = 0, 1, \ldots, N$ and $j = 1, 2, \ldots, N$. Applying Dirichlet conditions (2) to (30), we obtain the system of equations

$$\begin{aligned}
-G_{00}q_0^{+} + R_{01}u_1 + R_{02}u_2 + \ldots + R_{0N}u_N &= -c_0,\\
-G_{10}q_0^{+} + R_{11}u_1 + R_{12}u_2 + \ldots + R_{1N}u_N &= -c_1,\\
&\cdots\cdots\cdots\\
-G_{N0}q_0^{+} + R_{N1}u_1 + R_{N2}u_2 + \ldots + R_{NN}u_N &= -c_N,
\end{aligned} \tag{32}$$

or, in a compact form,

$$\begin{bmatrix} -G_{00} & R_{01} & R_{02} & \ldots & R_{0N} \\ -G_{10} & R_{11} & R_{12} & \ldots & R_{1N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -G_{N0} & R_{N1} & R_{N2} & \ldots & R_{NN} \end{bmatrix} \begin{bmatrix} q_0^{+} \\ u_1 \\ u_2 \\ \ldots \\ u_N \end{bmatrix} = \begin{bmatrix} -c_0 \\ -c_1 \\ -c_2 \\ \ldots \\ -c_N \end{bmatrix}, \tag{33}$$

where $c_i = H_{i0}^{+} U_0$.

The unknowns in the system of block matrix equations (33) can be directly expressed. However, for a large number of inhomogeneous domains, the direct expression becomes laborious. In that case, the system can be solved iteratively, for example, by the Gauss–Seidel method.

The computational algorithm consists of the following steps.

ALGORITHM

Step 1. For each boundary Γ_i, choose normal directions outward with respect to Ω_0 (see Fig. 1).

Step 2. Discretize each boundary and represent it as a triangulation grid.

Step 3. For each pair of boundaries Γ_i and Γ_j, produce the matrices G_{ij} and $\hat{H}_{ij}$ by numerically computing surface integrals (26) and (27) with collocation points $x \in \Gamma_i$ and integration points $y \in \Gamma_j$, where the points are taken at the barycenters of boundary elements.

Step 4. Calculate the matrices R_{ij} by formulas (28) and (31) with the use of Dirichlet boundary conditions. As a result, the system of block matrix equations (33) is formed.

Step 5. Solve the resulting system by directly expressing the unknowns or applying the Gauss–Seidel method.

Table

Ω	$\frac{k_0}{k_1} = 10^{-2}$	$\frac{k_0}{k_1} = 10^{-1}$	$\frac{k_0}{k_1} = 10^{0}$	$\frac{k_0}{k_1} = 10^{1}$	$\frac{k_0}{k_1} = 10^{2}$
$\Omega_0 : a_0 = 3$	3.84×10^{-2}	2.93×10^{-2}	1.74×10^{-4}	1.60×10^{-2}	1.83×10^{-2}
$\Omega_1 : a = 1$	3.41×10^{-2}	2.62×10^{-2}	1.68×10^{-4}	1.41×10^{-2}	1.62×10^{-2}
$\Omega_0 : a_0 = 5$	8.22×10^{-3}	6.31×10^{-3}	6.71×10^{-5}	3.52×10^{-3}	4.03×10^{-3}
$\Omega_1 : a = 1$	4.52×10^{-3}	3.51×10^{-3}	1.20×10^{-4}	1.81×10^{-3}	2.09×10^{-3}
$\Omega_0 : a_0 = 7$	3.04×10^{-3}	2.34×10^{-3}	4.55×10^{-5}	1.26×10^{-3}	1.45×10^{-3}
$\Omega_1 : a = 1$	9.67×10^{-4}	3.64×10^{-4}	1.10×10^{-4}	3.76×10^{-4}	4.17×10^{-4}

5. SOME RESULTS OF NUMERICAL EXPERIMENTS

The numerical experiments were performed according to the following scheme. Arbitrary domains Ω_0, ..., Ω_N with boundaries Γ_0, ..., Γ_N were specified in a system of spline 3D simulation. A software code for automatic grid generation (see [12]) was used to produce a boundary element triangulation grid on Γ_i, $i = 0, 1, ..., N$. The number of boundary elements on each surface ranged from 500 to 700. The potential of electric charges was set as a boundary condition. Next, system (33) was produced and the problem was solved numerically.

To test the computational methods and schemes and estimate their errors, we used the following classical problem with an analytical solution. Let an electric field of potential U_0 with a constant gradient $\operatorname{grad} U_0 = \text{const}$ be given in a ball Ω of radius a_0 bounded by the sphere Γ_0 with the electrical conductivity k_0. A ball Ω_1 of radius a with the electric conductivity k_1 bounded by the sphere Γ_1 is placed in Ω_1 (here, $a < a_0$ and the centers of the balls coincide). The goal is to find the resulting electric field potential u_0 in $\Omega_0 = \Omega \backslash \Omega_1$ and u_1 in Ω_1.

Introducing spherical coordinates with the origin at the center of the ball and solving the Laplace equations in spherical coordinates, we can show (see [13]) that the desired potential is represented in Cartesian coordinates as

$$u(\mathbf{r}) = \operatorname{grad} U_0 \cdot \mathbf{r} + \frac{(\sigma - 1)}{(\sigma + 2)} \frac{a^3}{|\mathbf{r}|^3} \operatorname{grad} U_0 \cdot \mathbf{r}, \quad \mathbf{r} \in \Omega_0, \tag{34}$$

in Ω_0 and as

$$u(\mathbf{r}) = \operatorname{grad} U_0 \cdot \mathbf{r} + \frac{(\sigma - 1)}{(\sigma + 2)} \operatorname{grad} U_0 \cdot \mathbf{r}, \quad \mathbf{r} \in \Omega_1, \tag{35}$$

in Ω_1. Here, $\sigma = k_1/k_0$, $\mathbf{r} = (x, y, z)^{\mathrm{T}}$ is the vector from the ball center to a given point, and a is the radius of Ω_1.

The electric conductivity was specified as $k_0 = 1$ in Ω_0 and as $k_1 = 10^n$ ($n = -2, -1, ..., 2$) in Ω_0. In Cartesian coordinates, the original potential U_0 was defined as

$$U_0(\mathbf{r}) = \alpha x + \beta y + \gamma z$$

with the gradient

$$\operatorname{grad} U_0(\mathbf{r}) = (\alpha, \beta, \gamma)^{\mathrm{T}},$$

where $\mathbf{r} = (x, y, z)^{\mathrm{T}}$ and α, β, and γ are constant coefficients. Next, an analytical solution to the problem was constructed and the potentials in Ω_0 and Ω_1 were calculated by formulas (34) and (35). The results were used as reference ones for comparison with the numerical solution.

For the numerical solution, the Dirichlet condition on Γ_0 was specified by computing the potential by formulas (34) and (35) at nodes of the surface triangulation grid. The numerical solution was found using a software code developed by the authors in MatLab.

To compare the numerical solution with the analytical one, we calculated the relative error

$$\varepsilon = \|\mathbf{u}_a - \mathbf{u}_n\| / \|\mathbf{u}_a\|,$$

where $\mathbf{u}_a$ are the values of the potential computed analytically at interior grid points in Ω and $\mathbf{u}_n$ are the numerical values of the potential at the same points.

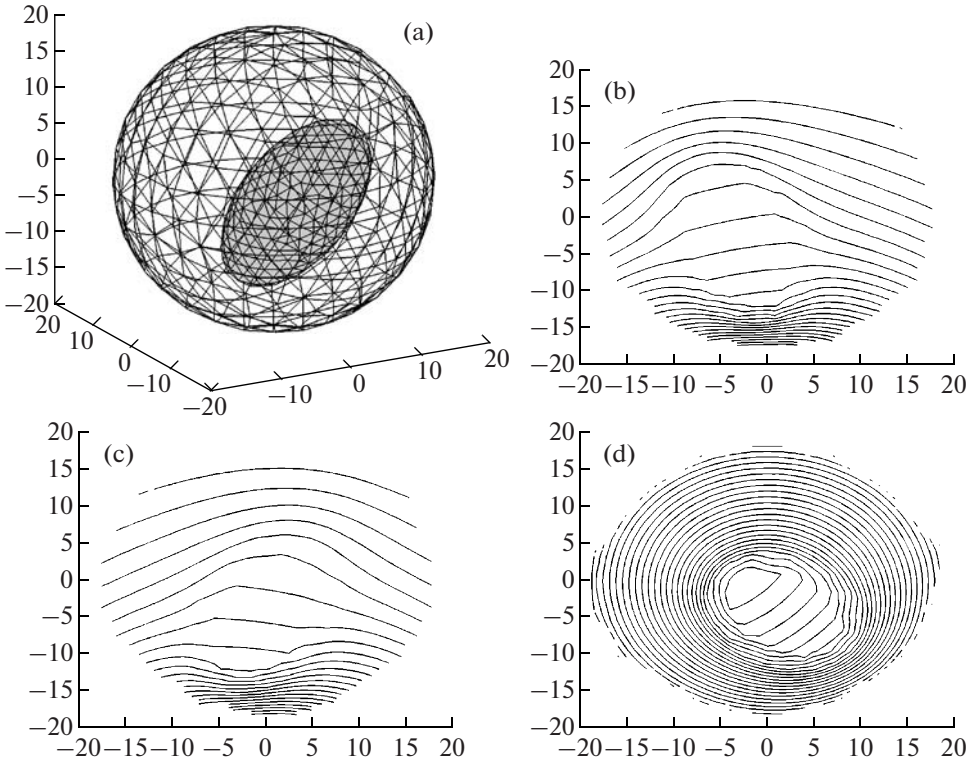


Fig. 2.

The relative errors for various radii of the spheres and various conductivities are presented in the table. These results show that the algorithm designed for the numerical solution of the Dirichlet problem for the Laplace equation in a piecewise homogeneous medium produces a solution with the relative error $\varepsilon \approx 10^{-2}$–$10^{-5}$.

To illustrate the capabilities of the method, we solved the Dirichlet problem for the Laplace equation in a piecewise homogeneous domain of complex geometry. A sphere Γ_0 and an ellipsoid Γ_1 with different centers were defined in a 3D surface simulation editor (see Fig. 2a). The potential of two positive electric charges located on the z axis was specified as a boundary condition. We computed the potential in the domain Ω_0 bounded by Γ_0 and Γ_1 and in the domain Ω_1 bounded by Γ_1. The conductivities were specified as $k_0 = 1$ and $k_1 = 5$. The number of boundary elements on each surface ranged from 1400 to 1600. Figures 2b–2d show the numerical results in the form of contour lines of the potential in the planes $x = 0$, $y = 0$, and $z = 0$, respectively.

Thus, based on the methods proposed in this paper, effective algorithms can be constructed for solving three-dimensional Dirichlet problems in a piecewise homogeneous medium.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 08-01-00314.

REFERENCES

1. G. A. Grinberg, *Certain Aspects of the Mathematical Theory of Electric and Magnetic Phenomena* (Akad. Nauk SSSR, Moscow, 1948) [in Russian].
2. V. R. Bursian, *Theory of Electromagnetic Fields Applied in Electrical Exploration (*Nedra, Leningrad, 1972) [in Russian].
3. V. I. Dmitriev and E. V. Zakharov, "Method for Computing Constant Current Field in Nonuniform Conducting Media," Vychisl. Metody Program., No. 20, 175–185 (1973).
4. O. V. Tozoni and I. D. Maergoiz, *Calculation of Three-Dimensional Electromagnetic Fields* (Tekhnika, Kiev, 1974) [in Russian].

5. R. S. MacLeod and D. H. Brooks, "Recent Progress in Inverse Problems in Electrocardiology," IEEE Eng. Med. Bio. Mag. **17** (1), 73–83 (1998).
6. I. R. Efimov, A. T. Sembelashvili, and V. N. Nikol'skii, "Progress in the Study of Mechanisms of Electrical Cardiac Stimulation," Vestn. Aritmol., No. 26, 91–96 (2002).
7. G. J. Saulnier, R. S. Blue, J. C. Newell, et al., "Electrical Impedance Tomography," IEEE Signal Proc. Mag. **18** (6), 31–43 (2001).
8. V. I. Dmitriev and E. V. Zakharov, "On the Numerical Solution of Certain Fredholm Integral Equations of the First Kind," Vychisl. Metody Program., No. 10, 49–54 (1968).
9. S. I. Smagin, "Numerical Solution of an Integral Equation of the First Kind with a Weak Singularity for Single-Layer Potential Density," Zh. Vychisl. Mat. Mat. Fiz. **28**, 1663–1673 (1988).
10. S. I. Smagin, "Numerical Solution of an Integral Equation of the First Kind with a Weak Singularity on a Closed Surface," Dokl. Akad. Nauk SSSR **303**, 1048–1051 (1988).
11. C. A. Brebbia, J. C. F. Telles, and L. C. Wrobel, *Boundary Element Techniques: Theory and Applications in Engineering* (Springer-Verlag, Berlin, 1984; Mir, Moscow, 1987).
12. E. R. Gol'nik, A. A. Vdovichenko, and A. A. Uspekhov, "Design and Application of a Preprocessor for Generation, Quality Control, and Optimization of Triangulation Grids in Contact Systems," Inform. Tekhnol., No. 4, 2–10 (2004).
13. L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon, Oxford, 1960; Fizmatlit, Moscow, 2003).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1151–1166. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © I.M. Troyanova, V.A. Tupchiev, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1207–1222.

Asymptotics of the Solution to the Point Explosion Problem in the Case of Cylindrical Symmetry

I. M. Troyanova and V. A. Tupchiev

Obninsk State Technical University of Nuclear Power Engineering, Studgorodok 1, Obninsk, Kaluga oblast, 249020 Russia

e-mail: irina.troyanova@gmail.com, tupchiev@obninsk.ru

Received February 27, 2008; in final form, August 26, 2008

Abstract—For the problem of a strong point explosion with cylindrical symmetry, high-order asymptotic expansions of the solution with respect to the viscosity and thermal conductivity are constructed and justified.

DOI: 10.1134/S0965542509070082

Key words: gasdynamic equations, point explosion problem with cylindrical symmetry, asymptotic solution method.

The cylindrical flow of a polytropic gas is governed by the gasdynamic equations (see [1]) in Eulerian coordinates

$$\begin{gathered}\rho_t + (\rho u)_x + \rho u/x = 0,\\ (\rho u)_t + (p + \rho u^2)_x + \rho u^2/x = (\widehat{\mu} u_x)_x,\\ (\rho E)_t + [\rho u(E + p/\rho)]_x + [\rho u(E + p/\rho)]/x = (\widehat{\lambda}\theta_x + \widehat{\mu} u u_x)_x + (\widehat{\lambda}\theta_x + \widehat{\mu} u u_x)/x,\end{gathered} \tag{1}$$

where $0 \le x < \infty$, ρ is the density, u is the velocity, $p = R_g\rho\theta$ is the pressure, θ is the temperature, $E = e + u^2/2$ is the specific energy, $e = c_V\theta$ is the specific internal energy, and R_g is the gas constant.

According to [2], the viscosity $\widehat{\mu}$ and the thermal conductivity $\widehat{\lambda}$ are proportional to the mean free path in the gas. Consequently, they can be represented as $\widehat{\mu} = \varepsilon\hat{\mu}$ and $\widehat{\lambda} = \varepsilon\hat{\lambda}$, where ε is the Knudsen number.

System (1) is supplemented with the initial and boundary conditions

$$(\rho, u, p)|_{t=0} = (\rho_0, u_0, p_0) = \text{const}, \tag{2}$$

$$u(0, t, \varepsilon) = 0, \tag{2a}$$

Here, (2a) is a flow symmetry condition (see [3, 4]) and

$$\rho_0 = \rho_0, \quad u_0 = 0, \quad p_0 = 0, \quad \gamma > 1, \tag{2b}$$

where $\gamma = c_p/c_V$ and $\widehat{p}_0$ is the back pressure. It is assumed that an energy amount of $2\pi E_0$ is instantaneously released on the axis of symmetry at $t = 0$ and remains constant for the entire volume of the moving gas, specifically,

$$\int_0^\infty \rho\left(e + \frac{1}{2}u^2\right)x\,dx = E_0. \tag{2c}$$

A similar problem without viscosity or thermal conductivity was solved in a number of monographs and works (see, e.g., [1, 3–6]) assuming the existence of a shock wave $x = a\sqrt{t}$ arising in the gas with no viscosity or thermal conductivity.

Our goal is to construct a high-order asymptotic expansion of the solution to problem (1), (2) in powers of ε by applying the boundary function method (see [7, 8]). Additionally, we substantiate this expansion and prove the existence and uniqueness of a solution to problem (1), (2), (2a)–(2e).

1. SELF-SIMILAR PROBLEM

1. After switching to the new variables

$$y = \frac{x}{a\sqrt{t}}, \quad \rho' = \frac{\rho}{\rho_0}, \quad u'(y) = \frac{\sqrt{t}u}{q_0 V_0}, \quad \theta'(y) = \frac{R_g t\theta}{q_0^2 V_0^2}, \quad E'(y) = \frac{tE}{q_0^2 V_0^2} \tag{1a}$$

(below, the primes are omitted), system (1) becomes

$$\frac{d}{dy}\rho(u-y) + \rho + \frac{\rho u}{y} = 0,$$

$$\frac{d}{dy}\left(p + \rho u(u-y) - \varepsilon\mu\frac{du}{dy}\right) + \frac{\rho u^2}{y} = 0, \tag{1b}$$

$$\frac{d}{dy}\left[y\left(\rho(u-y)E + up - \varepsilon\mu u\frac{du}{dy}\right)\right] = \varepsilon\frac{d}{dy}\left(\lambda y\frac{d\theta}{dy}\right),$$

where

$$\mu = \frac{\hat{\mu}}{aq_0}, \quad \lambda = \frac{\hat{\lambda}}{aq_0 R_r}, \quad q_0 = \frac{1}{2}\rho_0 a, \quad V_0 = \frac{1}{\rho_0},$$

and a is a parameter chosen so that $y = 1$ at the shock front.

Integrating system (1b) with respect to y and taking into account (2a) and (2b), we obtain the system

$$\rho(u-y) + q = 0, \quad \frac{dq}{dy} = \rho + \frac{\rho u}{y},$$

$$\varepsilon\mu\frac{du}{dy} = p + \rho u(u-y) - r, \quad \frac{dr}{dy} = -\frac{\rho u^2}{y}, \tag{1c}$$

$$\varepsilon\lambda\frac{d\theta}{dy} = \rho(u-y)E - up - \varepsilon\mu u\frac{du}{dy},$$

which is reduced in one in the Tikhonov form

$$\varepsilon\frac{dw}{dy} = F(w, \nu, y, \varepsilon), \quad \frac{d\nu}{dy} = f(w, \nu, y), \tag{3}$$

where $w = (V, \theta)$, $\nu = (q, r)$, $\mu = \hat{\mu}/(aq_0)$, $\lambda = \hat{\lambda}/(aq_0 R_r)$, and $u = y - qV$. In view of (1a), we obtain

$$F = \begin{pmatrix} \frac{1}{\mu}\left(-qV - \frac{\theta}{Vq} + \frac{r}{q} + y\right) + \varepsilon\left(\frac{V}{y} - \frac{1}{q}\right) \\ \frac{1}{\lambda}\left[-q^2y^2V - \frac{q\theta}{\gamma-1} + \frac{q^3V^2 + qy}{2} + r(qV-y)\right] \end{pmatrix}, \quad f = \begin{pmatrix} \frac{2}{V} - \frac{q}{y} \\ -\frac{1}{yV}(qV-y)^2 \end{pmatrix}. \tag{3a}$$

The additional conditions become

$$q|_{y=0} = 0, \quad (V, \theta)|_{y\to\infty} = (1, 0), \quad q/y|_{y\to\infty} = 1, \tag{3b}$$

$$\int_0^\infty \frac{1}{V}\left[\frac{\theta}{\gamma-1} + \frac{1}{2}(qV-y)^2\right]ydy = E_0, \tag{3c}$$

where $E^0 = 4E_0/\rho_0 a^4$.

System (1) is also reduced to the Tikhonov form in terms of $z = (\rho, p)$,

$$\varepsilon\frac{dz}{dy} = \mathrm{F}(z, \nu, y, \varepsilon), \quad \frac{d\nu}{dy} = \mathrm{f}(z, \nu, y), \tag{4}$$

by making the substitutions

$$V = \frac{1}{\rho}, \quad \theta = \frac{p}{\rho}, \tag{4^0}$$

Here,

$$\mathrm{F} = T^{-1}F|_{w\mapsto z}, \quad T^{-1}(z) = \begin{pmatrix} -\rho^2 & 0 \\ -\rho p & \rho \end{pmatrix}, \quad \mathrm{f} = f|_{w\mapsto z}, \quad T(z) = \begin{pmatrix} -\rho^{-2} & 0 \\ -\rho^{-2}p & \rho^{-1} \end{pmatrix}. \tag{4a}$$

According to (2a)–(2c), the additional condition become

$$q|_{y=0} = 0, \quad (\rho, p)|_{y\to\infty} = (1, 0), \quad q/y|_{y\to\infty} = 1, \tag{4b}$$

$$\int_0^\infty \rho E y dy = E^0, \tag{4c}$$

where

$$E = \frac{p}{\gamma-1} + \frac{1}{2\rho}(q - y\rho)^2, \quad E^0 = 4E_0/(\rho_0 a^4).$$

System (1b) can also be reduced to the form

$$\begin{gathered} (u-y)(\ln\rho)' + u' + u/y = 0, \\ (u-y)u' + p'/\rho - u = \varepsilon\mu u''/\rho, \\ (u-y)(\ln p)' + \gamma(u' + u/y) - 2 = \varepsilon g/p, \end{gathered} \tag{1d}$$

where

$$X' = \frac{dX}{dy}, \quad g = (\gamma-1)\{\lambda(\theta'' + \theta'/\xi) + \mu[(u')^2 + uu'/\xi]\}, \tag{4e}$$

Combining the equations of system (1d) yields

$$(\ln p)' - (\gamma-1)(\ln\rho)' + [\ln y(u-y)]' = \varepsilon g/p. \tag{4f}$$

2. FORMAL SCHEME FOR CONSTRUCTING ASYMPTOTICS

Following [7, 8], an asymptotic expansion for problem (3), (3b), (3c) or problem (4), (4b), (4c) is sought in the form of an expansion in powers of ε for $X = (w, \nu)$ or $Z = (z, \nu)$ on the half-axle $0 < y < \infty$ assuming that the degenerate solution is discontinuous (shock wave) at the point $y = 1$, i.e.,

$$X = \bar{X} + \Pi X, \quad \bar{X}(y, \varepsilon) = \sum_{k=0} \bar{X}_k(y)\varepsilon^k, \quad \Pi X(\eta, \varepsilon) = \sum_{k=0} \Pi_k X(\eta)\varepsilon^k, \quad \eta = \frac{y-1}{\varepsilon}, \tag{5}$$

where $\Pi_k X(\eta) = \{\Pi_k^- X(\eta)$ for $\eta < 0$, $\Pi_k^+ X(\eta)$ for $\eta > 0\}$.

Substituting (5) into (3) gives the formal equality

$$\varepsilon\frac{d\bar{w}}{dy} + \frac{d\Pi w}{d\eta} = F(\bar{X}, y, \varepsilon) + \Pi F, \quad \varepsilon\frac{d\bar{\nu}}{dy} + \frac{d\Pi\nu}{d\eta} = \varepsilon f(\bar{X}, y) + \varepsilon\Pi f,$$

where $\Pi F = F(\bar{X}(1 + \varepsilon\eta, \varepsilon) + \Pi X(\eta, \varepsilon), 1 + \varepsilon\eta, \varepsilon) - F(\bar{X}(1 + \varepsilon\eta, \varepsilon), 1 + \varepsilon\eta, \varepsilon)$ and $\Pi f = \Pi F|_{F\mapsto f}$, which is divided into two subsystems: a regular one

$$\varepsilon\frac{d\bar{w}}{dy} = F(\bar{X}, y, \varepsilon), \quad \frac{d\bar{\nu}}{dy} = f(\bar{X}, y) \tag{6}$$

and a singular one

$$\frac{d\Pi w}{d\eta} = \Pi F, \quad \frac{d\Pi\nu}{d\eta} = \varepsilon\Pi f. \tag{7}$$

Moreover, at the point $y = 1$, we have the matching conditions

$$X^l + \Pi^- X|_{\eta=0} = X^r + \Pi^+ X|_{\eta=0}, \tag{8}$$

where

$$X^l = \bar{X}(1-0), \quad X^r = \bar{X}(1+0), \quad X^l = \sum_{k=1} \varepsilon^k X_k^l, \quad X^r = \sum_{k=1} \varepsilon^k X_k^r. \tag{8a}$$

The coefficients of the regular expansion $\bar{X} = (\bar{w}, \bar{\nu})$ are found from the following systems for the zeroth and kth approximations:

$$0 = F(\bar{X}_0, y, 0), \quad \frac{d\bar{\nu}_0}{dy} = f(\bar{X}_0, y), \tag{6a}$$

$$\begin{aligned} \frac{d\bar{w}_{k-1}}{dy} &= \bar{F}_w(y)\bar{w}_k + \bar{F}_\nu(y)\bar{\nu}_k + R_k F(y), \\ \frac{d\bar{\nu}_k}{dy} &= \bar{f}_w(y)\bar{w}_k + \bar{f}_\nu(y)\bar{\nu}_k + R_k f(y), \end{aligned} \tag{6b}$$

where $\bar{F}_w(y) = F_w(\bar{X}_0(y), y, 0)$; $\bar{F}_\nu(y) = F_\nu(\bar{X}_0(y), y, 0)$; and $R_k F(y)$ are polynomials in $\bar{X}_1, \ldots, \bar{X}_{k-1}$, which, according to conditions (3b), are supplemented by the boundary conditions

$$\bar{q}_0\big|_{y=0} = 0, \quad (\bar{V}_0, \bar{\theta}_0)\big|_{y\to\infty} = (1, 0), \quad \bar{q}_0/y\big|_{y\to\infty} = 1, \tag{9a}$$

for $k = 0$ and by the conditions

$$\bar{q}_k\big|_{y=0} = 0, \quad (\bar{V}_k, \bar{\theta}_k)\big|_{y\to\infty} = (0, 0), \quad \bar{q}_k/y\big|_{y\to\infty} = 1, \tag{9b}$$

for $k > 1$. Condition (3c) will be taken into account later.

For $y > 1$, the degenerate solution of problem (6a), (9a) is

$$\bar{V}_0 \equiv 1, \quad \bar{\theta}_0 \equiv 0 (\bar{\rho}_0 \equiv 1, \bar{p} \equiv 0), \quad \bar{q}_0 = y, \quad \bar{r}_0 \equiv 0,$$

For $0 < y < 1$, since $\bar{\nu}_0(y)$ is continuous, the degenerate solution is found from system (6a) with the initial condition $\bar{\nu}_0|_{y=1} = \nu_0^l = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Moreover, the values $w_0^l = \bar{w}_0(1-0)$ are determined from the conditions

$$F(w_0^l, \nu_0^l, 1, 0) = 0, \tag{10}$$

which are known as the *Rankine–Hugoniot relations*.

Thus, the degenerate solution is a discontinuous function of the components $\bar{w}_0(y)$ with the values w_0^l and $w_0^r = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ on the left- and right-hand sides of the discontinuity at $y = 1$.

Similarly, for $k > 1$, problem (6b), (9b) on the interval $1 < y < \infty$ has the solution $\bar{w}_k(y) \equiv 0$, $\bar{\nu}_k(y) \equiv 0$, while, on the interval $0 < y < 1$, the solution is determined from (6b) with the initial data $\bar{\nu}_k|_{y=1} = \nu_k^l$, which are specified below.

To determine $\bar{w}_k(y)$ and $\bar{\nu}_k(y)$ from the indicated initial value problems, it is convenient to use some of their first integrals.

For the regular part of the expansion, Eq. (3c) is the energy integral

$$(\bar{u} - y)\left(\frac{\bar{p}}{\gamma - 1} + \bar{\rho}\frac{\bar{u}^2}{2}\right) + \overline{up} = \varepsilon\bar{\varphi}. \tag{11}$$

Integrating (4e) with respect to y from 1 to some $y \in [0, 1]$, we obtain an adiabatic integral for the regular part of the expansion:

$$\ln\frac{\bar{p}}{p^l} - (\gamma - 1)\ln\frac{\bar{\rho}}{\rho^l} + \ln\frac{y(y - \bar{u})}{(1 - u^l)} = \varepsilon\bar{Q}, \tag{12}$$

where

$$\bar{\varphi} = (\lambda\bar{\theta}' + \mu\overline{uu}'), \quad \bar{Q} = \int_1^y \frac{\bar{g}dy}{\bar{p}(\bar{u} - y)}, \quad \varphi = \frac{8\hat{\varphi}}{\rho_0 a^3}, \quad g = \frac{8\hat{g}}{\rho_0 a^2},$$

and g is defined in (4d).

By simple combinations, system (6) is reduced to the single equation

$$\frac{d\bar{u}}{dy} = B(\bar{u}, y) + \varepsilon\Phi(\bar{w}, y), \tag{13}$$

where

$$B = 2(\gamma-1)\Delta^{-1}\bar{u}\{\bar{u}/2 - y^{-1}(\gamma\bar{u}-y)[\bar{u}/2 - y/(\gamma-1)]\},$$
$$\Delta = \gamma(\gamma+1)u^2 - 2(\gamma+1)uy + 2y^2, \tag{13a}$$
$$\Phi = 4\Delta^{-1}\bar{V}(\bar{u}-y)^{-1}(\gamma\bar{u}-y)^2(\mu u'' - \bar{g}).$$

Making the substitutions

$$u = yU, \quad \rho = R, \quad \theta = y^2\Xi, \quad p = y^2P, \quad E = y^2\Sigma \tag{14}$$

in (11), (12), (13), and (13a), we obtain the energy and adiabatic integrals

$$(\bar{U}-1)\left(\bar{P}/(\gamma-1) + \frac{1}{2}\bar{R}\bar{U}^2\right) + \bar{U}\bar{P} = \varepsilon\breve{\varphi}/y^2, \tag{11a}$$

$$\ln(\bar{P}/P^l) - (\gamma-1)\ln(\bar{R}/R^l) + \ln[(1-\bar{U})/(1-U^l)] + 4\ln y = \varepsilon\bar{Q} \tag{12a}$$

Moreover,

$$\bar{R}\bar{\Sigma} = \bar{P}\bar{U}/(1-\bar{U}) + \varepsilon\breve{\varphi}/y^2(\bar{U}-1), \quad \bar{\Sigma} = \frac{\bar{\Xi}}{\gamma-1} + \frac{1}{2}\bar{U}^2, \tag{15}$$

where

$$\breve{\varphi} = \lambda(y\bar{\Xi}'' + 2\bar{\Xi}') + \mu\bar{U}(y\bar{U}' + \bar{U}), \quad \bar{\varphi} = y\breve{\varphi},$$
$$\breve{g} = (\gamma-1)\{\lambda(y^2\bar{\Xi}'' + 5y\bar{\Xi}' + 4\bar{\Xi}) + \mu[(y\bar{U}' + \bar{U})^2 + \bar{U}(y\bar{U}' + \bar{U})]\}, \tag{15a}$$
$$\breve{Q} = \int_1^y \frac{\breve{g}dy}{y^3\bar{R}\bar{\Xi}(\bar{U}-1)}, \quad (R, U, P)_{y=1} = (R^l, U^l, P^l) = \sum_{k=0}(R^l_k, U^l_k, P^l_k)\varepsilon^k.$$

Specifically, for some $k \geq 1$, we have

$$(\gamma\bar{U}_0 - 1)\bar{P}_k/(\gamma-1) + \frac{1}{2}(\bar{U}_0-1)\bar{U}_0^2\bar{R}_k + \left[\frac{\gamma}{\gamma-1}\bar{P}_0 + \frac{1}{2}\bar{R}_0(3\bar{U}_0-2)\bar{U}_0\right]\bar{U}_k = \breve{\varphi}_{k-1}/y^2,$$

$$\bar{P}_k/\bar{P}_0 - (\gamma-1)\bar{R}_k/\bar{R}_0 + \bar{U}_k/(\bar{U}_0-1) = \bar{Q}_{k-1} + L^l_k,$$

where $L^l_k = P^l_k/P^l_0 - (\gamma-1)R^l_k/R^l_0 + \bar{U}^l_k/(\bar{U}_0-1)$, and Eq. (13) becomes

$$y\frac{d\bar{U}}{dy} = \Psi(\bar{U}) + \varepsilon\Phi(\bar{U}, y), \quad \bar{U}\big|_{y=1} = U^l, \quad 0 < y < 1,$$

$$\Psi(U) = \frac{2U(2-\gamma U)(\gamma U-1)}{\bar{\Delta}_0}, \quad \Phi(U, y) = \frac{4(\gamma U-1)^2}{y\bar{\Delta}_0 R(U-1)[\mu((yU)'' - \breve{g})]}, \tag{16}$$
$$\bar{\Delta}_0 = \gamma(\gamma+1)\bar{U}^2 - 2(\gamma+1)\bar{U} + 2.$$

The coefficients of the singular expansion $\Pi X = (\Pi w, \Pi v)$ are determined from the systems for the zeroth and kth approximations:

$$\frac{d\Pi_0 w}{d\eta} = \Pi_0 F, \quad \frac{d\Pi_0 v}{d\eta} = 0, \tag{7a}$$

$$\frac{d\Pi_k w}{d\eta} = \Pi_k F, \quad \frac{d\Pi_k v}{d\eta} = \Pi_{k-1} f \tag{7b}$$

provided that

$$\Pi_k^- X \longrightarrow 0 \quad \text{as} \quad \eta \longrightarrow -\infty, \quad \Pi_k^+ X \longrightarrow 0 \quad \text{as} \quad \eta \longrightarrow \infty, \quad k = 0, 1, \ldots, \tag{7c}$$

where

$$\Pi_0 F = F(\bar{w}_0(1) + \Pi_0 w, \bar{v}_0(1), 1, 0) - F(\bar{w}_0(1), \bar{v}_0(1), 1, 0),$$
$$\Pi_k F = F_w(\eta)\Pi_k w + F_v(\eta)\Pi_k v + \Pi F_k(\eta).$$

Here, $\Pi F_k(\eta)$ is the power form of the kth degree with respect to $\Pi_i w(\eta)$ and $\Pi_i \nu(\eta)$ for $i = 0, 1, ..., k-1$, and the elements of the matrices $F_w(\eta)$ and $F_\nu(\eta)$ are calculated at the point $(\bar{w}_0(1) + \Pi_0 w(\eta), \bar{\nu}_0(1), 1, 0)$ (a similar expression holds for $\Pi_k f$).

Moreover, the matching conditions imply

$$X_k^l + \Pi_k^- X|_{\eta=0} = X_k^r + \Pi_k^+ X|_{\eta=0}, \quad X_k^l = \bar{X}_k(1-0), \quad X_k^r = \bar{X}_k(1+0), \tag{8b}$$

and

$$\Pi_k^\pm \nu = -\int_\eta^{\pm\infty} \Pi_{k-1}^\pm f d\eta, \quad \nu_k^l = -\int_{-\infty}^{+\infty} \Pi_{k-1} f d\eta. \tag{16a}$$

The values w_k^l are determined from the Rankine–Hugoniot conditions of the zeroth (10) and kth order

$$\bar{F}_w(1) w_k^l = w_{k-1}'^l - \bar{F}_\nu(1)\nu_k^l - R_k F(1). \tag{10a}$$

Thus, (16) gives initial data, while (ρ_k^l, u_k^l, p_k^l) are determined in (15a).

Substituting (5) into (4c) produces the relations

$$\int_0^\infty \sum_{k=0} (\bar{\rho}\bar{E})_k \varepsilon^k y dy + \varepsilon \int_{-1/\varepsilon}^\infty \sum_{k=0} \Pi_k(\rho E)\varepsilon^k (1+\varepsilon\eta) d\eta = E^0, \tag{17}$$

where

$$(\bar{\rho}\bar{E})_k = \sum_{k=0} \bar{\rho}_i \bar{E}_{k-i}, \quad \Pi_k(\rho E) = \sum_{k=0} \Pi_i \rho \Pi_{k-i} E, \quad \bar{E}_0 = 0.$$

Equating the coefficients of like powers of ε in (17), we obtain

$$E^0 = \int_0^1 \bar{\rho}_0 \bar{E}_0 y dy, \tag{17a}$$

$$0 = \int_0^1 (\bar{\rho}\bar{E})_1 y dy + \int_{-\infty}^\infty \Pi_0(\rho E) d\eta, \tag{17b}$$

.....

$$0 = \int_0^1 (\bar{\rho}\bar{E})_k y dy + \int_{-\infty}^\infty \Pi_{k-1}(\rho E) d\eta + \int_{-\infty}^\infty \Pi_{k-2}(\rho E)\eta d\eta, \quad k = 1, 2, \ldots, n, \tag{17c}$$

$$(\rho E)_k = (\bar{u}_0 \bar{p}_k + \bar{u}_k \bar{p}_0)/(y - \bar{u}_0) + \bar{\varphi}_{k-1}/(\bar{u}_0 - y), \tag{18}$$

where the functions $\bar{\varphi}_{k-1}$ are determined in terms of $\bar{w}_j$ for $j = 0, 1, ..., k-1$.

3. SUBSTANTIATION OF THE ALGORITHM FOR CONSTRUCTING THE ASYMPTOTICS

1. First, we construct the zero approximation, specifically, the degenerate solution $\bar{X}_0(y)$ to problem (3), (4).

Obviously, for $y > 1$ (ahead of the shock front), system (6a) with the right end condition (3b) $\bar{w}_0|_{y\to\infty} = w_0 = (1, 0)$, $\bar{q}_0/y|_{y\to\infty} = 1$ yields

$$\bar{X}_0 = (\bar{w}_0, \bar{\nu}_0), \quad \bar{w}_0 = w_0^r = (1, 0), \quad \bar{\nu}_0 = (y, 0). \tag{19}$$

Moreover, from system (7a) with conditions (7c), we have $\Pi_0 \nu \equiv 0$. Therefore, by virtue of (8a) and (18),

$$\nu_0^l = (1, 0). \tag{19a}$$

The Rankine–Hugoniot conditions (10) in expanded form with (19) taken into account become

$$-V_0^l - \frac{\theta_0^l}{V_0^l} + 1 = 0, \quad \frac{1}{2}(V_0^l - 1) - \frac{1}{\gamma - 1}\theta_0^l = 0. \tag{20}$$

These equation give $(V_0^l, \theta_0^l) = (h, 2h/(\gamma + 1))$ and $h = (\gamma - 1)/(\gamma + 1)$, which are the values on the left of the discontinuity at the point $y = 1$. Moreover,

$$\rho_0^l = \frac{1}{h}, \quad u_0^l = p_0^l = \frac{2}{\gamma + 1}. \tag{20a}$$

The values on the right of the discontinuity are $(V_0^r, \theta_0^r) = (1, 0)$.

On the interval $0 < y < 1$, the functions $\bar{X}_0(y)$ are determined from system (6a) with the conditions

$$\bar{\nu}_0\big|_{y=1} = \nu_0^l = (1, 0). \tag{20b}$$

The solution to problem (6a), (20b) is well known (see [1, 3, 4]). In variables (14), it satisfies the energy and adiabatic integrals

$$(\bar{U}_0 - 1)\left(\frac{\bar{\Xi}_0}{\gamma - 1} + \frac{\bar{U}_0^2}{2}\right) + \bar{U}_0\bar{\Xi}_0 = 0, \tag{11b}$$

$$\ln\frac{\bar{P}_0}{P_0^l} - (\gamma - 1)\ln\frac{\bar{R}_0}{R_0^l} + \ln\frac{(1 - \bar{U}_0)}{(1 - U_0^l)} + 4\ln y = 0, \tag{12b}$$

whence

$$\bar{P}_0 = \frac{(\gamma - 1)\bar{R}_0\bar{U}_0^2(1 - \bar{U}_0)}{2(\gamma\bar{U}_0 - 1)}, \quad \bar{R}_0 = \left[\frac{4(\gamma - 1)^{\gamma - 1}(\gamma\bar{U}_0 - 1)}{(\gamma + 1)^{\gamma + 1}(1 - \bar{U}_0)^2\bar{U}^2}\right]^{1/(2 - \gamma)} y^{4/(\gamma - 2)}. \tag{20c}$$

The function $\bar{U}_0 = 1 - \bar{Q}_0/\bar{R}_0$ is determined from Eq. (16) at $\varepsilon = 0$:

$$y\frac{d\bar{U}_0}{dy} = \Psi(\bar{U}_0) \tag{21}$$

with the initial condition

$$\bar{U}\big|_{y=1} = U_0^l = \frac{2}{\gamma + 1}, \tag{21a}$$

The solution to problem (21), (21a) is implicitly determined from

$$y = \chi(\bar{U}_0)/\chi(U_0^l),$$

where $\chi(U) = U^{-1/2}(1/\gamma - U)^{-1/2}(U - 1/2\gamma)^{(\gamma - 1)/2\gamma}$ and $\chi(\bar{U}_0)/\chi'(\bar{U}_0) = \Psi(\bar{U}_0)$.

Clearly, it is necessary, following [3], to introduce the parameter $\sigma = \bar{U}_0/U_0^l$, which varies on the interval $[(\gamma + 1)/2\gamma, 1]$ in view of (20a).

Now, following [3, 4], the degenerate solution $\bar{w}_0(y)$ on the interval $0 \le y \le 1$ can be written in the parametric form

$$y(\sigma) = \frac{1}{\sqrt{\sigma}}\left[\frac{2\gamma}{\gamma - 1}\left(\sigma - \frac{\gamma + 1}{2\gamma}\right)\right]^{(\gamma - 1)/2\gamma}\left[\gamma\left(\frac{\gamma + 1}{\gamma} - \sigma\right)\right]^{-1/2}, \quad \bar{u}_0(\sigma) = y(\sigma)\frac{2\sigma}{\gamma + 1},$$

$$\bar{\rho}_0(\sigma) = \left[\frac{2\gamma}{\gamma - 1}\left(\sigma - \frac{\gamma + 1}{2\gamma}\right)\right]^{1/\gamma}\left[\frac{2}{\gamma - 1}\left(\frac{\gamma + 1}{2} - \sigma\right)\right]^{2/(\gamma - 2)}\left[\gamma\left(\frac{\gamma + 1}{\gamma} - \sigma\right)\right]^{2/(2 - \gamma)}, \tag{22}$$

$$\bar{p}_0(\sigma) = \frac{2\sigma}{\gamma + 1}\left[\frac{2}{\gamma - 1}\left(\frac{\gamma + 1}{2} - \sigma\right)\right]^{\gamma/(\gamma - 2)}\left[\gamma\left(\frac{\gamma + 1}{\gamma} - \sigma\right)\right]^{(4 - \gamma)/(2 - \gamma)}, \quad \bar{\theta}_0(\sigma) = \frac{\bar{p}_0(\sigma)}{\bar{\rho}_0(\sigma)}.$$

Moreover, it follows from (22) that

$$\bar{\rho}_0(y) \longrightarrow 0, \quad \bar{p}_0(y) \longrightarrow \left(\frac{2}{\gamma}\right)^{2/(\gamma - 2)}\frac{(\gamma + 1)^2}{16\gamma} \quad \text{as} \quad y \longrightarrow 0, \tag{22a}$$

$$\bar{q}_0(\sigma) = \frac{2}{\gamma+1}\bar{\rho}_0(\sigma)y(\sigma)\left(\frac{\gamma+1}{2}-\sigma\right),$$
$$\bar{r}_0(\sigma) = \frac{2\sigma(1-\sigma)}{\gamma(\gamma+1)}\bar{\rho}_0(\sigma)y^2(\sigma)\left(\frac{\gamma+1}{2}-\sigma\right)\left(\sigma-\frac{\gamma+1}{2\gamma}\right)^{-1}. \quad (22b)$$

Specifically, this implies that

$$\bar{\Delta}_0 = 2y^2(\sigma)\left(\frac{2\gamma}{\gamma+1}\sigma^2-2\sigma+1\right) > 0 \quad \text{for} \quad \frac{\gamma+1}{2\gamma} < \sigma \le 1,$$

and

$$\bar{V}_0(y) \longrightarrow \infty, \quad \bar{\theta}_0(y) \longrightarrow \infty \quad \text{as } y \longrightarrow 0.$$

To determine the character of the degenerate solution for $0 \le y \le 1$, we find, in view of (3a), the eigenvalues of the matrix

$$\bar{F}_w(y) = \begin{pmatrix} \frac{1}{\mu\bar{q}_0}(-\bar{q}_0^2+\bar{\rho}_0\bar{p}_0) & -\frac{\bar{\rho}_0}{\mu\bar{q}_0} \\ -\frac{\bar{q}_0\bar{p}_0}{\lambda} & -\frac{\bar{q}_0}{\lambda(\gamma-1)} \end{pmatrix}. \quad (23)$$

Consider the characteristic equation $|\bar{F}_w(y) - \Lambda E| = 0$:

$$\mu\lambda\Lambda^2 - L\Lambda - \Delta = 0, \quad (23a)$$

where

$$L = \lambda(\bar{\rho}_0\bar{p}_0/\bar{q}_0 - \bar{q}_0) - \mu\bar{q}_0/(\gamma-1), \quad \Delta = \frac{\bar{q}_0\bar{\rho}_0\bar{\Delta}_0}{2(\gamma-1)(\gamma\bar{u}_0-y)}. \quad (23b)$$

It follows that $\Lambda_{1,2} = \frac{1}{2\mu\lambda}(L \pm \sqrt{L^2+4\mu\lambda\Delta})$. Since $\Delta > 0$ for $0 < y \le 1$, we have conditional stability (see [7]): $\Lambda_1(y) < 0$, $\Lambda_2(y) > 0$, and $\lim_{y\to 0} L = \infty$.

Using (4^0) and (4a), we have

$$\bar{F}_z(y) = T_0^{-1}(y)\bar{F}_w(y)T_0(y), \quad (23c)$$

$$\Phi_0^{-1}(y)F_w(y)\Phi_0(y) = B_0^{-1}(y)F_z(y)B_0(y) = \begin{pmatrix} \Lambda_1(y) & 0 \\ 0 & \Lambda_2(y) \end{pmatrix}, \quad (23d)$$

where

$$\Phi_0(y) = \begin{pmatrix} -\frac{\bar{q}_0\bar{p}_0}{\lambda} & \Lambda_1 + \frac{1}{\mu}\left(\bar{q}_0 - \frac{\bar{\rho}_0\bar{p}_0}{\bar{q}_0}\right) \\ -\frac{\bar{q}_0\bar{p}_0}{\lambda} & \Lambda_2 + \frac{1}{\mu}\left(\bar{q}_0 - \frac{\bar{\rho}_0\bar{p}_0}{\bar{q}_0}\right) \end{pmatrix}, \quad B_0(y) = T_0^{-1}(y)\Phi_0(y), \quad T_0(y) = T(\bar{z}_0(y)).$$

For Tikhonov system (4) in variables (ρ, p),

$$\bar{F}_z(y) = T_0^{-1}(y)\bar{F}_w(y)T_0(y).$$

Therefore, the eigenvalues Λ_1 and Λ_2 of matrix (23) are invariant under transformations (4^0) for the degenerate solution and at singular points.

Combining (17a) with (18) and (20c) yields

$$E^0 = \int_0^1 (\rho E)_0 y\,dy = \int_0^1 \frac{\bar{u}_0\bar{p}_0}{y-\bar{u}_0}y\,dy = \int_0^1 \frac{\bar{U}_0\bar{P}_0}{1-\bar{U}_0}y^3\,dy,$$

where

$$J_0 = \frac{\gamma-1}{8}\frac{\overline{R}_0\overline{U}_0^3}{\gamma\overline{U}_0-1}, \quad J = \int_0^1 J_0 y^3\,dy, \quad E^0 = \frac{4E_0}{\rho_0 a^4}.$$

Thus, a is given by the formula

$$a = \left(\frac{E_0}{\rho_0 J}\right)^{1/4}, \tag{24}$$

where J satisfies, according to [4], the estimate

$$\left(\frac{2}{\gamma}\right)^{2/(\gamma-2)}\frac{(\gamma+1)^2}{32\gamma(\gamma-1)} < J < \frac{16}{(\nu+1)(\gamma^2-1)(\nu+3)^2}.$$

Now, we determine the zeroth approximation of the singular part of the expansion. The associated system corresponding to (7a) has the form

$$\begin{aligned}\mu\frac{dV}{d\eta} &= -\theta/V + 1 - V \equiv M(V,\theta),\\ \lambda\frac{d\theta}{d\mu} &= \frac{1}{2}(V-1)^2 - \theta/(\gamma-1) \equiv N(V,\theta).\end{aligned} \tag{25}$$

The singular points (V_0^l, θ_0^l) and (V_0^r, θ_0^r) of system (25) are a saddle and an unstable node. They are joined by the separatrix $\theta = \varphi(V)$.

System (25) is supplemented by the initial conditions

$$V|_{\eta=0} = V_l^0, \quad \theta|_{\eta=0} = \theta_l^0 = \varphi(V_l^0), \tag{25a}$$

where the parameter V_0^l is to be determined (this will be done in the course of constructing the first approximation). Moreover,

$$\begin{aligned}\Pi_0^+ V &= V - V_0^r, \quad \Pi_0^+\theta = \theta - \theta_0^r, \quad 0 < \eta < \infty,\\ \Pi_0^- V &= V - V_0^l, \quad \Pi_0^-\theta = \theta - \theta_0^l, \quad -\infty < \eta < 0.\end{aligned}$$

The corresponding variational system has the form

$$\begin{aligned}\mu\dot{V}_\eta &= [-1 + \theta V^{-2}]\dot{V} - V^{-1}\dot{\theta},\\ \lambda\dot{\theta}_\eta &= (V-1)\dot{V} - \frac{1}{\gamma-1}\dot{\theta}.\end{aligned} \tag{26}$$

At the singular point (V^r, θ^r), the Jacobian of the right-hand side of (25) has the eigenvalues $\Lambda_1 = -1/\mu$ and $\Lambda_2 = -1/(\gamma-1)\lambda$. Therefore, this point is a stable node.

At the singular point (V^l, θ^l).

$$\Lambda_1 = [(\gamma-3)\lambda + \mu - \sqrt{[(\gamma-3)\lambda+\mu]^2 + 4(\gamma+1)\mu\lambda}][2(\gamma-1)\mu\lambda]^{-1} < 0,$$

$$\Lambda_2 = [(\gamma-3)\lambda + \mu + \sqrt{[(\gamma-3)\lambda+\mu]^2 + 4(\gamma+1)\mu\lambda}][2(\gamma-1)\mu\lambda]^{-1} > 0$$

Therefore, it is a saddle. It is well known (see [1]) that these singular points are joined by an unstable separatrix $\theta = \varphi(V)$ passing through the domain $\Omega = \{(V,\theta) : M > 0, N < 0\}$. Consequently, for $h \le V < 1$, we have on the separatrix

$$\frac{d\varphi}{dV} < 0, \quad \theta' < 0, \quad V' > 0.$$

The corresponding boundary functions satisfy the estimates

$$|\Pi^0 V|, |\Pi^0\theta| \le Ce^{\sigma\eta}, \quad \eta < 0, \quad |\Pi^0 V|, |\Pi^0\theta| \le Ce^{-\sigma\eta}, \quad 0 < \eta < \infty, \tag{27}$$

where $\Pi^0 V = V - h$, and $\Pi^0\theta = \theta - \theta^l$ for $-\infty < \eta < 0$, and $\Pi^0 V = V - 1$ and $\Pi^0\theta = \theta$ for $0 < \eta < \infty$.

The following lemmas are easy to prove.

Lemma 1. *The solution to system* (26) *with the initial data*

$$(\dot{V}, \dot{\theta})|_{\eta=0} = (1, \varphi'(V_l^0)) \tag{27a}$$

satisfies the limiting relation $\lim_{|\eta|\to\infty}(\dot V,\dot\theta)=(0,0)$ *and has the estimates*

$$|\dot V|,|\dot\theta|\le Ce^{-\sigma|\eta|},\quad -\infty<\eta<\infty. \tag{27b}$$

Lemma 2. *The solution* $\hat W=(\hat V,\hat\theta)$ *of the system*

$$\begin{aligned}\mu\hat V_\eta &= [-1+\theta V^{-2}]\hat V - V^{-1}\hat\theta + f_m(\eta),\\ \lambda\hat\theta_\eta &= (V-1)\hat V - \frac{1}{\gamma-1}\hat\theta + g_m(\eta),\end{aligned} \tag{28}$$

where $|f_m(\eta)|,|g_m(\eta)|\le Ce^{\sigma\eta}$, $-\infty<\eta<0$; *and* (V,θ) *solves problem* (25), (25a) *for* $\theta_l^0=\varphi(V_l^0)$, *can be represented as*

$$\hat W = V^m\dot W + W^m, \tag{28a}$$

where $\dot W$ *is a solution to problem* (26), (27a); V^m *is an arbitrary constant*; *and* W^m *is a particular solution to system* (28) *that satisfies the estimate*

$$|V^m|,|\theta^m|\le Ce^{\sigma\eta},\quad -\infty<\eta<0. \tag{19b}$$

Remark 1. (a) $W'=C_0\dot W$, where $W'=W_\eta$ and $C_0=V_\eta|_{\eta=0}=\mu^{-1}M(V_0,\theta_0)$. In (28a) $\dot W$ can be replaced by W', since V^m is arbitrary.

(b) For $y_0=1-a_0\varepsilon$, i.e., for $\eta=-a_0$ and sufficiently large a_0, the point $(V^0,\varphi(V^0))$ on the separatrix belongs to an arbitrarily small neighborhood of the saddle (V_0^l,θ_0^l). Moreover, the corresponding unstable linear initial manifold of system (28) is defined by the equation

$$\theta-\theta_0^l = \varphi'(V^0)(V-V_0^l)+b_m(V^0),$$

where

$$\varphi'(V^0)\longrightarrow\varphi'(h)=\frac{2(\gamma-1)\mu}{(\gamma+1)\lambda[(\gamma-1)\mu\Lambda_1+\gamma-3]},\quad b_m\longrightarrow 0\quad\text{as}\quad V^0\longrightarrow V_l^0.$$

(c) An associated system in the variables ρ and p is derived from (25) by making substitution (4^0). Lemmas 1 and 2 and all the subsequent conclusions remain similar for this system. In the variables (ρ,p), the associated system and the boundary functions look somewhat more complicated.

Thus, the zeroth asymptotic expansion is not completely determined at the first step, since the parameter V_0^l remains undetermined.

2. Now we construct and substantiate the first and kth approximations of the asymptotics. Obviously, for the regular expansion, $\bar X_1=\ldots=\bar X_k\equiv 0$ for $y>1$. On the interval $(0,1)$, we find a regular approximation $\bar W_k$, while $\bar U_k$ is a solution to the initial value problem for the linear equation

$$y\frac{d\bar U_k}{dy} = \Psi'(\bar U_0)\bar U_k+\Phi^k(\bar W_0,\ldots,\bar W_{k-1},y),\quad \bar U_k|_{y=1}=U_k^l=u_k^l,\quad -\infty<\xi<0. \tag{29}$$

The solution to homogeneous equation (29) has the form $\bar U_1=U_1^l\Psi(\bar U_0)/\Psi(U_0^l)$, while the solution to the inhomogeneous equation is

$$\bar U_k = U_k^l\Psi(\bar U_0)/\Psi(U_0^l)+f^k(y),\quad f^k(y)=\Psi(\bar U_0(y))\int_1^y\frac{\Phi^k(\bar W_0(s),\ldots,\bar W_{k-1}(s))ds}{\Psi(\bar U_0(s))s}. \tag{29a}$$

Expressions (11a) and (12a) yield the energy and adiabatic integrals

$$(\gamma\bar U_0-1)\bar P_1/(\gamma-1)+\frac{1}{2}(\bar U_0-1)\bar U_0^2\bar R_k+\left[\frac{\gamma}{\gamma-1}\bar P_0+\frac{1}{2}\bar R_0(3\bar U_0-2)\bar U_0\right]\bar U_k = \breve\varphi_{k-1}/y^2, \tag{11c}$$

$$\overline{P}_k/\overline{P}_0 - (\gamma-1)\overline{R}_k/\overline{R}_0 + \overline{U}_k/(\overline{U}_0 - 1) = \overline{Q}_{k-1} + L_k^l,$$
$$\breve{\varphi}_{k-1} = \lambda(y\overline{\Xi}''_{k-1} + 2\overline{\Xi}'_{k-1}) + \mu\overline{U}_{k-1}(y\overline{U}'_{k-1} + \overline{U}_{k-1}),$$
$$L_k^1 = \overline{P}_k^l/P_0^l - (\gamma-1)\overline{R}_k^l/\overline{R}_0^l + \overline{U}_k^l/(\overline{U}_k^l - 1), \quad \overline{Q}_{k-1} = \int_1^y \frac{\breve{g}_{k-1}dy}{y^3\overline{R}_{k-1}\overline{\Xi}_{k-1}(\overline{U}_{k-1}-1)}, \tag{12c}$$
$$\breve{g}_{k-1} = (\gamma-1)\{\lambda(y^2\overline{\Xi}''_{k-1} + 5y\overline{\Xi}'_{k-1} + 4\overline{\Xi}_{k-1}) + \mu[(y\overline{U}'_{k-1} + \overline{U}_{k-1})^2 + \overline{U}_{k-1}(y\overline{U}'_{k-1} + \overline{U}_{k-1})]\}.$$

Let us estimate the boundary functions and discuss some of their properties.

Combining (7b) with (16a), we obtain, for $0 < \eta < \infty$, the system

$$\frac{d\Pi_k^+ w}{d\eta} = F_w^+(\eta)\Pi_k^+ w + G_k^+(\eta), \tag{30$^+$}$$

where

$$F_w^+(\eta) = \begin{pmatrix} \frac{1}{\mu}\left(-1 + \frac{\Pi_0^+\theta}{(1+\Pi_0^+ V)^2}\right) & -\frac{1}{\mu(1+\Pi_0^+ V)} \\ \frac{\Pi_0^+ V}{\lambda} & -\frac{1}{\lambda(\gamma-1)} \end{pmatrix}, \quad G_k^+(\eta) = -F_V^+(\eta)\int_\eta^\infty \Pi_{k-1}^+ f(s)ds,$$

and, for $-\infty < \eta < 0$, the system

$$\frac{d\Pi_k^- w}{d\eta} = F_w^-(\eta)\Pi_k^- w + G_k^-(\eta), \tag{30$^-$}$$

where

$$F_w^-(\eta) = \begin{pmatrix} \frac{1}{\mu}\left(-1 + \frac{\theta_0^l + \Pi_0^-\theta}{(h+\Pi_0^- V)^2}\right) & -\frac{1}{\mu(h+\Pi_0^- V)} \\ \frac{\Pi_0^- V + h - 1}{\lambda} & -\frac{1}{\lambda(\gamma-1)} \end{pmatrix}, \quad G_k^-(\eta) = -F_v^-(\eta)\int_\eta^\infty \Pi_{k-1}^- f(s)ds + S_k F(\eta).$$

Applying Lemmas 1 and 2 gives

$$\Pi_k^- w = V_0^k \dot{W} + W^k, \tag{31}$$

where, according to (28b), we have the estimates (see [7])

$$|G_k^-(\eta)|, |W^k| \le Ce^{\kappa\eta}, \quad -\infty < \eta < 0, \quad |G_k^+(\eta)| \le Ce^{-\kappa\eta}, \quad 0 < \eta < \infty, \tag{32}$$

while $\Pi_k^+ w$ are determined from system (30$^-$) with the initial condition

$$\Pi_k^+ w\big|_{\eta=0} = V_0^k \dot{W} + W^k\big|_{\eta=0}. \tag{33}$$

Lemma 3. *The following estimates hold*:

$$|\Pi_k^- w||\Pi_k^- z| < Ce^{\kappa\eta}, \quad |\Pi_k^+ w||\Pi_k^+ z| < Ce^{-\kappa\eta}, \quad k = 1, 2, \ldots, n, \tag{34}$$

where

$$\kappa = \frac{1}{2}\min_{i=1,2}[\Lambda_i(w_0^l), \Lambda_i(w_0^r)].$$

Proof. Estimates (34) are easy to derive by applying estimates (27), (27b), and (28b) and Lemmas 1 and 2 to systems (30) and taking into account (31) and (32).

In view of (31), we obtain the following formulas for the derivatives at $k = 0$:

$$\dot{I}_e^0 = \int_{-\infty}^\infty \dot{\Pi}_0(\rho u^2)d\eta = \frac{1}{C_0}\int_{-\infty}^\infty \Pi_0'(RU^2)d\eta = \frac{1}{C_0}(-R_0^l(U_0^l)^2) = -\frac{4}{C_0(\gamma^2-1)}, \tag{35a}$$

$$\dot{I}_\rho^0 = \int_{-\infty}^{\infty} \dot{\Pi}_0 R d\eta = \frac{1}{C_0}\int_{-\infty}^{\infty} \Pi_0' R d\eta = -\frac{2}{C_0(\gamma-1)}, \quad \dot{I}_E^0 = \int_{-\infty}^{\infty} \dot{\Pi}_0 (R\Sigma) d\eta = -\frac{4}{C_0(\gamma^2-1)}, \tag{35b}$$

and, for $k \geq 1$,

$$I_\rho^k = -\frac{2}{\gamma-1} V_l^k + I_{\rho 0}^k, \quad I_e^k = -\frac{4}{\gamma^2-1} V_l^k + I_{e0}^k, \quad I_E^k = -\frac{4}{\gamma^2-1} V_l^k + I_{E0}^k, \tag{36}$$

where $I_{\rho 0}^k$, I_{e0}^k, and I_{E0}^k are determined in terms of indices smaller than k.

Now we determine the parameters $V_l^0, \ldots, V_l^{k-1}$.

From formula (16a), we find that

$$q_k^l = -2I_\rho^{k-1} + q_{k0}^l, \quad r_k^l = I_e^{k-1} + r_{k0}^l, \tag{37}$$

where

$$I_\rho^{k-1} = \int_{-\infty}^{\infty} \Pi_{k-1} \rho d\eta, \quad I_e^{k-1} = \int_{-\infty}^{\infty} \Pi_{k-1} (\rho u^2) d\eta, \quad q_{00}^l = 0, \quad r_{00}^l = 0,$$

and g_{k0}^l and m_{k0}^l are determined in terms of indices smaller than k.

The kth-order Rankine–Hugoniot conditions (10a) yield the system

$$\begin{aligned} \frac{3-\gamma}{\gamma-1} V_k^l - \frac{1}{h}\theta_k^l + \frac{3-\gamma}{\gamma+1} q_k^l + r_k^l &= g_{k0}^l, \\ -\frac{2}{\gamma+1} V_k^l - \frac{1}{\gamma-1}\theta_k^l - \frac{2h}{\gamma+1} q_k^l + \frac{2}{\gamma+1} r_k^l &= m_{k0}^l, \end{aligned} \tag{37a}$$

where $g_{10}^l = \mu \bar{u}_0'(1)$, $m_{10}^l = \lambda \bar{\theta}_0'(1)$, and g_{k0}^l and m_{k0}^l are determined in terms of indices smaller than k.

Now, combining (37) and (37a) produces

$$\begin{aligned} R_k^l &= -h^{-1}(2I_\rho^{k-1} + I_e^{k-1}) + R_{k0}^l, \quad U_k^l = -hI_e^{k-1} + U_{k0}^k, \\ P_k^l &= \frac{2}{\gamma+2}(I_e^{k-1} - 2I_\rho^{k-1}) + P_{k0}^l, \quad L_k^l = 2(\gamma-2)I_\rho^{k-1} + (\gamma+1)I_e^{k-1} + L_{k0}^l, \end{aligned} \tag{38}$$

where R_{k0}^l, U_{k0}^k, P_{k0}^l, and L_{k0}^l are determined in terms of indices smaller than k, $R_{10}^l = h^{-1}\psi_0$, $U_{10}^l = h\psi_0$, $\psi_0 = (\gamma+1)\lambda\bar{\theta}_0'(1) + \mu\bar{u}_0'(1)$,

$$P_{10}^l = 2\mu\gamma(\gamma+1)^{-1}\bar{u}_0'(1) + \lambda(\gamma-1)\bar{\theta}'(1), \quad \text{and } L_{10}^l = -\frac{(\gamma+1)^2}{2}\theta_0'^l.$$

Simple transformations in (11c) and (12c) with (20c) taken into account give

$$\frac{\gamma-2}{\gamma-1}\bar{P}_k = \frac{\bar{P}_0 X_0(\bar{U}_0)}{(\gamma-1)\bar{U}_0(U_0-1)(\gamma\bar{U}_0-1)}\bar{U}_k - \frac{\bar{P}_0}{\gamma-1}L_k^l - H^{k-1}, \tag{38a}$$

where

$$X_0 = (2\gamma-1)\gamma U^2 - (\gamma^2+2\gamma-2)U + 2(\gamma-1), \quad H^{k-1} = \frac{(\gamma-1)\breve{\varphi}_{k-1}}{y^2(\gamma\bar{U}_0-1)} + \frac{\bar{P}_0}{\gamma-1}\bar{Q}_{k-1}. \tag{38b}$$

It follows from (15) that

$$(\bar{R}\bar{\Sigma})_k = \bar{P}_k\bar{U}_0/(1-\bar{U}_0) + \bar{U}_k\bar{P}_0/(1-\bar{U}_0)^2 + \breve{\varphi}_{k-1}/y^2(\bar{U}_0-1) \tag{15b}$$

Therefore,

$$\frac{\gamma-2}{\gamma-1}(\bar{R}\bar{\Sigma})_k = \frac{\bar{U}_0}{1-\bar{U}_0\gamma-1}\frac{\gamma-2}{\gamma-1}\bar{P}_k + \frac{\gamma-2}{\gamma-1}\frac{\bar{P}_0}{(1-\bar{U}_0)^2}\bar{U}_k + \frac{\gamma-2}{\gamma-1}\frac{\breve{\varphi}_{k-1}}{y^2(\bar{U}_0-1)}.$$

Substituting (38a) into this relation, we obtain

$$\frac{\gamma-2}{\gamma-1}(\overline{R}\overline{\Sigma})_k = -\frac{\overline{R}_0\overline{U}_0^2X_1(\overline{U}_0)}{2(\gamma U_0-1)^2(1-\overline{U}_0)}\overline{U}_k - \frac{\overline{R}_0U_0^3}{2(\gamma U_0-1)}L_k^l - H_{k-1},$$

where

$$X_1 = \gamma(2\gamma-1)\overline{U}_0^2 - 2(\gamma^2-1)\overline{U}_0 + 3\gamma - 4, \quad H_{k-1} = \frac{\overline{U}_0}{1-\overline{U}_0}\left[H^{k-1} + \frac{(\gamma-2)\breve{\varphi}_{k-1}}{(\gamma-1)\overline{U}_0y^2}\right],$$

$$\overline{U}_k = U_k^l\Psi(\overline{U}_0)/\Psi(U_0^l) + f^k(y), \quad J_0 = \frac{\gamma-1}{8}\frac{\overline{R}_0\overline{U}_0^3}{\gamma\overline{U}_0-1} = \frac{P_0\overline{U}_0}{1-\overline{U}_0}, \quad \int_0^1 J_0y^3dy = J.$$

The parameter V_l^{k-1} is determined using (17c), which, in variables (14), has the form

$$Q_k(V_l^{k-1}) \equiv \int_0^1(\overline{R}\overline{\Sigma})_ky^3dy + \int_{-\infty}^{\infty}\Pi_{k-1}(R\Sigma)d\eta + \int_{-\infty}^{\infty}\Pi_{k-2}(R\Sigma)d\eta = 0. \tag{39}$$

Combining (38) with (38b) and setting

$$A_0 = \frac{\overline{P}_0X_1(\overline{U}_0)}{(\gamma-1)(\gamma U_0-1)(1-\overline{U}_0)^2}\frac{\Psi(\overline{U}_0)}{\Psi(U_0^l)} = \frac{2(\gamma+1)^2(2-\gamma\overline{U}_0)X_1(\overline{U}_0)}{(\gamma-1)\overline{\Delta}_0(1-\overline{U}_0)^2}J_0,$$

where

$$\overline{\Delta}_0 = \gamma(\gamma+1)\overline{U}^2 - 2(\gamma+1)\overline{U} + 2, \quad \hat{B}_0 = \frac{\overline{R}_0\overline{U}_0^3}{2(\gamma\overline{U}_0-1)} = \frac{4J_0}{\gamma-1},$$

we find that

$$\frac{\gamma-2}{\gamma-1}(\overline{R}\overline{\Sigma})_k = (hA_0 - 4h^{-1}J_0)I_e^{k-1} - 8\frac{\gamma-2}{\gamma-1}J_0I_\rho^{k-1} - H_{k-1}, \tag{39a}$$

Here,

$$hA_0 - 4h^{-1}J_0 = \frac{4}{h}\left[\frac{(\gamma-1)}{2}\frac{\overline{X}_1(\overline{U}_0)(2-\gamma\overline{U}_0)}{(1-\overline{U}_0)\overline{\Delta}_0} - 1\right]J_0 = \frac{2(\gamma-2)\hat{\Gamma}(\overline{U}_0)}{h(1-\overline{U}_0)\overline{\Delta}_0(\overline{U}_0)}J_0,$$

$$\hat{\Gamma}(\overline{U}_0) = -\frac{1}{2}\gamma(2\gamma^2+\gamma+1)\overline{U}_0^3 + (\gamma^3+3\gamma^2+\gamma+1)\overline{U}_0^2 - \frac{1}{2}(7\gamma^2+3\gamma+2)\overline{U}_0 + 3\gamma - 1.$$

For $k = 1$, Eq. (39) becomes

$$\overline{Q}_1(V^0) \equiv \frac{\gamma-2}{\gamma-1}\int_0^1(\overline{R}\overline{\Sigma})_1y^3dy + \frac{\gamma-2}{\gamma-1}\int_{-\infty}^{\infty}\Pi_0(R\Sigma)d\eta = 0.$$

Combining this relation with (39a) yields

$$\overline{Q}_1(V_l^0) \equiv (hA - 4h^{-1}J)I_e^0 - 8\frac{\gamma-2}{\gamma-1}JI_\rho^0 + \frac{\gamma-2}{\gamma-1}I_E^0 - H_0, \tag{39b}$$

where

$$\overline{Q}_1(V_l^0) = \frac{\gamma-2}{\gamma-1}Q_1(V_l^0), \quad A = \int_0^1A_0(\overline{U}_0)y^3dy, \quad H = \int_0^1H_0(U_0)y^3dy.$$

Taking into account (35a) and (35b), we consider the derivative of (39b):

$$\dot{\overline{Q}}_1(V_l^0) = (hA - (\gamma+1)B)\dot{I}_e^0 - 2(\gamma-2)B\dot{I}_\rho^0 + \frac{\gamma-2}{\gamma-1}\dot{I}_E^0 \tag{39c}$$

$$= \left[(\gamma-2)h^{-1}\int_0^1\frac{2\hat{\Gamma}(\overline{U}_0)}{(1-\overline{U}_0)\overline{\Delta}_0}J_0y^3dy\right]\left(-\frac{4}{C_0(\gamma^2-1)}\right) - (\gamma-2)4h^{-1}\int_0^1J_0y^3dy\left(-\frac{4}{C_0(\gamma^2-1)}\right) + \frac{\gamma-2}{\gamma-1}\left(-\frac{4}{C_0(\gamma^2-1)}\right).$$

Thus, applying Lemmas 1–3 and taking into account Remark 1 gives the following expression for the derivative:

$$\dot{Q}_1(V_l^0) = -\frac{4\mu}{M(V_l^0, \theta_l^0)(\gamma^2-1)}\left\{(\gamma+1)\left[\int_0^1 \frac{2\hat{\Gamma}(\overline{U}_0)}{\Delta_0(U)(1-\overline{U}_0)} J_0 y^3 dy - 4J\right] + 1\right\}.$$

Noting that

$$\Gamma(\overline{U}_0) = 2\hat{\Gamma}(\overline{U}_0) - 4(1-\overline{U}_0)\overline{\Delta}_0, \quad \overline{\Delta}_0 = \gamma(\gamma+1)\overline{U}_0^2 - 2(\gamma+1)\overline{U}_0 + 2, \quad \frac{1}{\gamma} \le \overline{U}_0 \le \frac{2}{\gamma+1},$$

we obtain

$$\dot{Q}_1(V_l^0) = -\frac{4\mu}{M(V_l^0, \theta_l^0)(\gamma^2-1)}\left\{\left[(\gamma^2-1)\int_0^1 \frac{\Gamma(\overline{U}_0)}{\overline{\Delta}_0(U)(1-\overline{U}_0)} J_0 y^3 dy\right] + 1\right\}, \tag{40}$$

where $\Gamma(\overline{U}_0) = -\frac{1}{2}\gamma(2\gamma+1)\overline{U}_0^3 + (\gamma^2+3\gamma+1)\overline{U}_0^2 - \frac{1}{2}(7\gamma+6)\overline{U}_0 + 3$.

For $k > 1$, similarly, using (39a) and (36), we find from (39) that

$$Q_k(V_l^{k-1}) = -\frac{4}{\gamma^2-1}\left\{\left[(\gamma^2-1)\int_0^1 \frac{\Gamma(\overline{U}_0)}{\overline{\Delta}_0(U)(1-\overline{U}_0)} J_0 y^3 dy\right] + 1\right\} V_l^{k-1} + Q^k, \tag{40a}$$

where $(\gamma^2-1)\int_0^1 \frac{\Gamma(\overline{U}_0)}{\overline{\Delta}_0(U)(1-\overline{U}_0)} J_0 y^3 dy$ will be denoted by $I(\gamma)$.

Lemma 4. *Let*

$$I(\gamma) + 1 \ne 0. \tag{41}$$

Then Eqs. (17b) *and* (17c) *have a unique solution.*

Proof. Integrating Eq. (40) with respect to V_l^0 with the first equation in (25) taken into account, we obtain the linear equation $Q(V_l^0) = (I(\gamma)+1)\eta + Q^0$. Its unique solvability follows from (41). Thus, we determine the parameter $V_l^0 = V_*^0$. Similarly, under condition (41), Eq. (40a) (and, hence, Eq. (17c)) is uniquely solvable and we determine the parameters $V_l^{k-1} = V_*^{k-1}$ for $k = 1, 2, \ldots, n$.

3. Let us estimate the residual for system (4). For this purpose, consider the partial sums

$$\mathrm{X}_n + \overline{\mathrm{X}}_n + \Pi\mathrm{X}_n, \quad \overline{\mathrm{X}}_n(y,\varepsilon) = \sum_{k=1}^n \overline{X}_k(y)\varepsilon^k, \quad \Pi\mathrm{X}_n(\eta,\varepsilon) = \sum_{k=1}^n \Pi_k X(\eta)\varepsilon^k, \tag{42}$$

where $\mathrm{X}_n = (\mathrm{Z}_n, \mathrm{V}_n)$, $\mathrm{Z}_n = (\rho_n, \mathrm{P}_n)$, and $\mathrm{V}_n = (\mathrm{q}_n, \mathrm{r}_n)$.

Substituting them into problems (4) and (4b) yields a problem with residuals:

$$\varepsilon\frac{d\mathrm{Z}_n}{dy} = \mathrm{F}(\mathrm{Z}_n, \mathrm{V}_n, y, \varepsilon) + \omega_0, \quad \frac{d\mathrm{q}_n}{dy} = -\frac{1}{y}\mathrm{q}_n + 2\rho_n + \omega_1, \quad \frac{d\mathrm{r}_n}{dy} = -\frac{(y\rho_n - \mathrm{q}_n)^2}{y\rho_n} + \omega_2, \tag{43}$$

$$\mathrm{q}_n\big|_{y=0} = 0, \quad \mathrm{r}_n\big|_{y=1} = 0, \quad \mathrm{Z}_n\big|_{y\to\infty} = (1,0), \quad \mathrm{q}_n/y\big|_{y\to\infty} = 1. \tag{43a}$$

$$\int_0^\infty \rho_n \mathrm{E}_n y dy + \varpi = E^0. \tag{43b}$$

Lemma 5. *The following estimates for the residuals hold*:

$$|\omega_0| \le C\varepsilon^{n+1} \quad \textit{for} \quad 0 \le y < \infty, \tag{44}$$

$$|\omega_1||\omega_2||\varpi| \le \begin{cases} C[\varepsilon^{n+1} + \varepsilon^n e^{\kappa(y-1)/\varepsilon}] & \textit{for } 0 \le y \le 1 \\ C[\varepsilon^{n+1} + \varepsilon^n e^{\kappa(1-y)/\varepsilon}] & \textit{for } 1 \le y < \infty. \end{cases} \tag{44a}$$

Proof. Estimates (44) and (44a) are derived in the same manner as in [7, Section 15].

4. SUBSTANTIATION OF THE ASYMPTOTICS

We substantiate the asymptotics for the boundary value problem in the variables (ρ, p), i.e., for problem (4), (4b), (4c), since its degenerate solution is bounded with respect to ρ and p on the half-line $0 \le y < \infty$.

While substantiating the asymptotics, following [7, Section 15; 8], we prove the existence and uniqueness of a solution to problem (4), (4b), (4c).

For the deviations $\zeta = z - Z_n$, $\rho = \rho - \rho_n$, $\vartheta = \nu - V_n$, and $\vartheta = (q, r)$, the system has the form

$$\begin{gathered}\varepsilon\frac{d\zeta}{dy} = \bar{F}_z(y,)\zeta + \bar{F}_\nu(y)\vartheta + H(\zeta, \vartheta, y, \varepsilon),\\ \frac{dq}{dy} = -\frac{1}{y}q + 2\rho - \omega_1, \quad \frac{dr}{dy} = f_0(\rho, q, y) - \omega_2,\end{gathered} \tag{45}$$

where

$$f_0 = -\frac{[y(\rho+\rho_n) - (q+q_n)]^2}{y(\rho+\rho_n)} + \frac{(y\rho_n - q_n)^2}{y\rho_n}, \qquad \int_0^\infty \rho E y dy + \varpi = E^0,$$

and the nonlinear terms satisfy the following properties:

(i) $\|H(0, 0, y, \varepsilon)\| \le c\varepsilon^{n+1}$, $\|\omega_1\|, \|\omega_2\| \le c[\varepsilon^{n+1} + \varepsilon^n e^{-\kappa(1-y)/\varepsilon}]$;

(ii) for any $\delta > 0$, there are constants $\beta(\delta)$ and $\varepsilon_0 = \varepsilon_0(\delta)$ such that, if $\|\xi_1\| \le \beta$, $\|\xi_2\| \le \beta$, $\|\vartheta_1\| \le \beta$, $\|\vartheta_2\| \le \beta$, and $0 < \varepsilon \le \varepsilon_0$, then

$$\|H(\xi_1, \vartheta_1, y, \varepsilon) - H(\xi_2, \vartheta_2, y, \varepsilon)\| \le \delta(\|\xi_1 - \xi_2\| + \|\vartheta_1 - \vartheta_2\|).$$

Considering system (45) on the interval (α, y_0), where y_0 is defined in Remark 1 and α is arbitrarily small, we apply the transformation

$$\zeta = B_0(y)\begin{pmatrix}\xi\\ \chi\end{pmatrix} - \bar{F}_z^{-1}(y)\bar{F}_\nu(y)\vartheta, \tag{46}$$

to obtain the system

$$\begin{gathered}\varepsilon\frac{d\xi}{dy} = \Lambda_1\xi + R(\xi, \chi, q, r, y, \varepsilon), \quad \varepsilon\frac{d\chi}{dy} = \Lambda_2\chi + S(\xi, \chi, q, r, y, \varepsilon),\\ \frac{dq}{dy} = -\frac{1}{y}q + 2\rho - \omega_1, \quad \frac{dr}{dy} = f_0(\rho, q, y) - \omega_2,\end{gathered} \tag{46a}$$

where ρ is determined by (46).

System (46a) is supplemented by the boundary conditions

$$\xi(\alpha) = 0, \quad \chi(y_0) = 0, \quad q(\alpha) = 0, \quad r(\alpha) = 0. \tag{46b}$$

The proof sketch for the existence and uniqueness of a solution to problem (46a), (46b) on the interval (α, y_0) is as follows.

Problem (46a), (46b) is reduced to a system of integral equations, which, after some transformations with the use of the resolvent, is reduced to

$$\begin{gathered}\xi = \frac{1}{\varepsilon}\int_\alpha^y \exp\left[\frac{1}{\varepsilon}\int_s^y \Lambda_1(\tau)d\tau\right]R(\xi, \chi, \vartheta, s, \varepsilon)ds \equiv \hat{R}(\xi, \chi, \vartheta, s, \varepsilon),\\ \chi = -\frac{1}{\varepsilon}\int_y^{y_0} \exp\left[\frac{1}{\varepsilon}\int_s^y \Lambda_2(\tau)d\tau\right]S(\xi, \chi, \vartheta, s, \varepsilon)ds \equiv \hat{S}(\xi, \chi, \vartheta, s, \varepsilon),\\ q = \hat{Q}_1(\xi, \chi, \vartheta, s, \varepsilon), \quad r = \hat{Q}_2(\xi, \chi, \vartheta, s, \varepsilon),\end{gathered} \tag{47}$$

where the operators $\hat{R}$, $\hat{S}$, and $\hat{Q} = (\hat{Q}_1, \hat{Q}_2)$ have the following properties:

(i) $\|\hat{R}(0, 0, 0, y, \varepsilon)\|$, $\|\hat{S}(0, 0, 0, y, \varepsilon)\|$, $\|\hat{Q}(0, 0, 0, y, \varepsilon)\| \le C\varepsilon^{n+1}$;

(ii) for any $\delta > 0$, there are constants $\beta(\delta)$, $\varepsilon_0 = \varepsilon_0(\delta)$, and A such that

$$\|\hat{Z}(\xi, \tilde{\chi}, \tilde{\vartheta}, s, \varepsilon) - Z(\xi, \chi, \vartheta, s, \varepsilon)\| \le \delta\max_{[0,1]}[\|\tilde{\xi} - \xi\| + \|\tilde{\chi} - \chi\| + \|\tilde{\vartheta} - \vartheta\|],$$

if $\|\tilde{\xi}\|, \|\tilde{\chi}\|, \|\tilde{\vartheta}\|, \|\xi\|, \|\chi\|, \|\vartheta\| \le \beta$, $0 < \varepsilon \le \varepsilon_0$; and $\hat{Z} = (\hat{R}, \hat{S}, \hat{Q})$.

Following [7, 8], we apply the method of successive approximations to system (47) to prove the existence and uniqueness of a solution to system (47) and, simultaneously, we prove the estimates

$$\|\xi\|, \|\chi\|, \|\vartheta\| \le C\varepsilon^{n+1}. \tag{48}$$

On the interval (y_0, ∞), the solution is first extended to the interval $(y_0, 1)$, where estimates (48) remain valid, since, after switching to η, the system regularly depends on ε for $-a_0 < \eta < 0$. Then the resulting solution is extended to the half-line $(0, \infty)$, where Theorem 3.1 from [7] holds in view of the solution behavior at infinity (see [10]). Next, the solution extends to the interval $(0, \alpha)$. Finally, the desired solution is the limit of the constructed solution in $C(0, \infty)$ as $\alpha \longrightarrow 0$. Since the desired solution is smooth, it also satisfies the symmetry condition $\lim\limits_{y \to 0} u = 0$.

Theorem. *Let condition* (41) *be satisfied. Then, in the neighborhood of the degenerate solution, problem* (4), (4b) *has a solution satisfying condition* (4c) *with an* $O(\varepsilon^{n+1})$ *residual such that, for sufficiently small* ε,

$$\|X - X_n\| \le C\varepsilon^{n+1}, \quad 0 < \varepsilon \le \varepsilon_0,$$

where C is independent of ε *and n is arbitrary.*

Remark 2. Condition (41) was checked numerically for various values of γ. It was found that (41) holds for $\gamma \le 3.1$ and for $\gamma \ge 3.2$.

REFERENCES

1. B. L. Rozhdestvenskii and N. N. Yanenko, *Systems of Quasilinear Equations and Their Applications to Gas Dynamics* (Nauka, Moscow, 1978; Am. Math. Soc., Providence, 1983).
2. E. M. Lifshitz and L. P. Pitaevsky, *Physical Kinetics* (Nauka, Moscow, 1979; Pergamon, Oxford, 1980).
3. L. I. Sedov, *Similarity and Dimensional Methods in Mechanics* (Academic, New York, 1959; Nauka, Moscow, 1987).
4. V. P. Korobeinikov, *Problems of Point Blast Theory* (Nauka, Moscow, 1985).
5. P. P. Volosevich and E. I. Livanov, *Self-Similar Solutions in Gas Dynamics and Heat Transfer* (Nauka, Moscow, 1997) [in Russian].
6. N. L. Krasheninnikova, "On the Unsteady Motion of a Gas Displaced by a Piston," Izv. Akad. Nauk SSSR, Otd. Tekh. Nauk, No. 8 (1955).
7. A. B. Vasil'eva and V. F. Butuzov, *Asymptotics Expansions of Solutions to Singularly Perturbed Equations* (Nauka, Moscow, 1973) [in Russian].
8. V. A. Tupchiev, "Asymptotics of Solutions to a Boundary Value Problem for a System of First-Order Differential Equations with a Small Parameter Multiplying the Derivative," Dokl. Akad. Nauk SSSR **143**, 1296–1299 (1962).
9. V. A. Tupchiev, "On Corner Solutions of Boundary Value Problems for a System of First-Order Differential Equations with a Small Parameter Multiplying the Derivative," Vestn. Mosk. Gos. Univ. Mat. Mekh., No. 3, 17–24 (1963).
10. F. Hoppenstead, "Singular Perturbations on Infinite Interval," Trans. Am. Math. Soc. **123**, 521–535 (1966).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1167–1174. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © M.Kh. Shkhanukov-Lafishev, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1223–1231.

Locally One-Dimensional Scheme for a Loaded Heat Equation with Robin Boundary Conditions

M. Kh. Shkhanukov-Lafishev

Institute of Computer Science and Problems of Regional Management, Kabardino-Balkar Scientific Center, Russian Academy of Sciences, ul. I. Armand 37a, Nalchik, 360000 Russia

e-mail: math@kbsu.ru

Received June 6, 2008

Abstract—The third boundary value problem for a loaded heat equation in a p-dimensional parallelepiped is considered. An a priori estimate for the solution to a locally one-dimensional scheme is derived, and the convergence of the scheme is proved.

DOI: 10.1134/S0965542509070094

Key words: boundary value problem, loaded heat equation, difference scheme, scheme convergence, total approximation, embedding theorem, a priori estimate.

INTRODUCTION

Boundary value problems for loaded differential equations arise in the study of soil moisture motion, in the heat conduction theory (see [1–4]), and in the quality control of water resources (see [5, p. 26]). Numerical methods for loaded differential equations were addressed, for example, in [5–7].

In this paper, we numerically solve the third boundary value problem for a loaded heat equation in a p-dimensional parallelepiped.

1. STATEMENT OF THE PROBLEM

In the cylinder $Q_T = G \times [0 < t \le T]$, whose bottom is the p-dimensional rectangular parallelepiped $G = \{x = \{x_1, ..., x_p) : 0 < x_\alpha < l_\alpha, \alpha = 1, 2, ..., p\}$ with the boundary Γ, we consider the problem

$$\partial u/\partial t = Lu + f(x, t), \quad (x, t) \in Q_T, \tag{1}$$

$$\begin{aligned} k_\alpha(x, t)\frac{\partial u}{\partial x_\alpha} &= \beta_{-\alpha}(x, t)u(x, t) - \mu_{-\alpha}(x, t), \quad x_\alpha = 0, \\ -k_\alpha(x, t)\frac{\partial u}{\partial x_\alpha} &= \beta_{+\alpha}(x, t)u(x, t) - \mu_{+\alpha}(x, t), \quad x_\alpha = l_\alpha, \end{aligned} \tag{2}$$

$$u(x, 0) = u_0(x), \quad x \in \overline{G}, \tag{3}$$

where

$$Lu = \sum_{\alpha=1}^{p} L_\alpha u, \quad L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x, t)\frac{\partial u}{\partial x_\alpha}\right) - q_\alpha(x, t)u(x_1, \ldots, x_{\alpha-1}, x_\alpha^\circ, x_{\alpha+1}, \ldots, x_p, t),$$

$$k_\alpha(x, t) \ge c_0 > 0, \quad |q_\alpha(x, t)| \le c_1, \quad |\beta_{-\alpha}(x, t)|, |\beta_{+\alpha}(x, t)| \le c_2, \quad \alpha = 1, 2, \ldots, p,$$

x_α° is a fixed point in the interval $(0, l_\alpha)$; and c_0, c_1, and c_2 are positive constants.

The numerical solution of third boundary value problems for the heat equation and a hyperbolic equation was considered in [8]. Below, we numerically solve problem (1)–(3).

Let us introduce a uniform spatial grid in each direction Ox_α with the mesh size h_α, $h_\alpha = l_\alpha/N_\alpha$, $\alpha = 1, 2, ..., p$,

$$\overline{\omega} = \prod_{\alpha=1}^{p} \overline{\omega}_\alpha, \quad \overline{\omega}_\alpha = \{x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha, i_\alpha = 0, 1, \ldots, N_\alpha, \alpha = 1, 2, \ldots, p\},$$

$$\hbar_\alpha = \begin{cases} h_\alpha, & i_\alpha = 1, 2, \ldots, N_\alpha - 1 \\ \dfrac{h_\alpha}{2}, & i_\alpha = 0, N_\alpha, \end{cases}$$

Let $\gamma_{-\alpha}$ denote the left boundary node $x_\alpha = 0$ and $\gamma_{+\alpha}$ denote the right boundary node $x_\alpha = l_\alpha$.

On the interval $[0, T]$, we introduce a uniform grid $\overline{\omega}_\tau = \{t_j = j\tau, j = 0, 1, \ldots, j_0\}$ with the time step τ. Each of the intervals $[t_j, t_{j+1}]$ is divided into p subintervals by the points $t_{j+\alpha/p} = t_j + \dfrac{\alpha\tau}{p}$, $\alpha = 1, 2, \ldots, p-1$. Denote by $\Delta_\alpha = \left(t_{j+\frac{\alpha-1}{p}}, t_{j+\frac{\alpha}{p}}\right]$ a half-open interval, where $\alpha = 1, 2, \ldots, p$.

Equation (1) is rewritten as

$$\mathcal{L}u = \frac{\partial u}{\partial t} - Lu - f = 0,$$

or

$$\sum_{\alpha=1}^{p} \mathcal{L}_\alpha u = 0, \quad \mathcal{L}_\alpha u = \frac{1}{p}\frac{\partial u}{\partial t} - L_\alpha u - f_\alpha, \quad \sum_{\alpha=1}^{p} f_\alpha = f.$$

On each half-open interval Δ_α, $\alpha = 1, 2, \ldots, p$, we sequentially solve the problems

$$\mathcal{L}_\alpha \vartheta_{(\alpha)} = \frac{1}{p}\frac{\partial \vartheta_{(\alpha)}}{\partial t} - L_\alpha \vartheta_{(\alpha)} - f_\alpha = 0, \quad x \in G, \quad t \in \Delta_\alpha, \quad \alpha = 1, 2, \ldots, p, \tag{4}$$

$$\begin{aligned} k_\alpha \frac{\partial \vartheta_{(\alpha)}}{\partial x_\alpha} &= \beta_{-\alpha}(x,t)\vartheta_{(\alpha)} - \mu_{-\alpha}(x,t), \quad x_\alpha = 0, \\ -k_\alpha \frac{\partial \vartheta_{(\alpha)}}{\partial x_\alpha} &= \beta_{+\alpha}(x,t)\vartheta_{(\alpha)} - \mu_{+\alpha}(x,t), \quad x_\alpha = l_\alpha, \end{aligned} \tag{5}$$

with the conditions (see [9, p. 522])

$$\begin{aligned} \vartheta_{(1)}(x,0) &= u_0 x, \quad \vartheta_{(1)}(x,t_j) = \vartheta_{(p)}(x,t_j), \quad j = 1, 2, \ldots, m, \\ \vartheta_{(\alpha)}\left(x, t_{j+\frac{\alpha-1}{p}}\right) &= \vartheta_{(\alpha-1)}\left(x, t_{j+\frac{\alpha-1}{p}}\right), \quad \alpha = 2, 3, \ldots, p, \quad j = 0, 1, \ldots, m. \end{aligned} \tag{6}$$

Each equation in (4) indexed by α is approximated on the half-open interval $t_{j+\frac{\alpha-1}{p}} < t \le t_{j+\frac{\alpha}{p}}$ by a two-level implicit scheme. Applying the well-known technique for improving the order of accuracy of Robin boundary conditions up to the second with respect to h_α (see [9, p. 180]), we obtain the chain of one-dimensional schemes

$$y_{\bar{t}}^{(\alpha)} = \overline{\Lambda}_\alpha y^{(\alpha)} + \Phi_\alpha^{j+\frac{\alpha}{p}}, \quad \alpha = \overline{1, p}, \quad x \in \overline{\omega}_h, \tag{7}$$

$$y(x,0) = u_0(x), \tag{8}$$

where

$$\overline{\Lambda}_\alpha y^{(\alpha)} = \begin{cases} \Lambda_\alpha y^{(\alpha)} = (a_\alpha y_{\bar{x}_\alpha}^{(\alpha)})_{x_\alpha} - d_\alpha \xi(y_{i_{\alpha_0}}, y_{i_{\alpha_0}+1}), & x_\alpha \in \omega_{h_\alpha} \\ \Lambda_\alpha^- y^{(\alpha)} = \dfrac{a_\alpha^{(+1_\alpha)} y_{x_\alpha,0}^{(\alpha)} - \beta_{-\alpha} y^{(\alpha)}}{0.5h_\alpha} - d_\alpha \xi(y_{i_{\alpha_0}}, y_{i_{\alpha_0}+1}), & x_\alpha \in \gamma_{-\alpha} \\ \Lambda_\alpha^+ y^{(\alpha)} = \dfrac{-a_\alpha^{(N_\alpha)} y_{\bar{x}_\alpha, N_\alpha}^{(\alpha)} - \beta_{+\alpha} y^{(\alpha)}}{0.5h_\alpha} - d_\alpha \xi(y_{i_{\alpha_0}}, y_{i_{\alpha_0}+1}), & x_\alpha \in \gamma_{+\alpha}, \end{cases}$$

$$\Phi_\alpha^{j+\frac{\alpha}{p}} = \begin{cases} \varphi_\alpha^{j+\frac{\alpha}{p}}, & x_\alpha \in \omega_{h_\alpha} \\ \dfrac{\mu_{-\alpha}}{0.5h_\alpha} + f_{\alpha,0}, & x_\alpha \in \gamma_{+\alpha} \\ \dfrac{\mu_{+\alpha}}{0.5h_\alpha} + f_{\alpha,N_\alpha}, & x_\alpha \in \gamma_{+\alpha}, \end{cases}$$

and

$$\xi(y_{i_{\alpha_0}}, y_{i_{\alpha_0}+1}) = y^{(\alpha)}_{i_{\alpha_0}} \frac{x_\alpha^{(i_{\alpha_0}+1)} - x_\alpha^\circ}{h_\alpha} + y^{(\alpha)}_{i_{\alpha_0}+1} \frac{x_\alpha^\circ - x_\alpha^{(i_{\alpha_0})}}{h_\alpha}, \quad x_\alpha^{(i_{\alpha_0})} \le x_\alpha^\circ \le x_\alpha^{(i_{\alpha_0}+1)}.$$

Thus, we have a chain of one-dimensional schemes with nonlocal boundary conditions for each $\alpha = 1, 2, \ldots, p$.

Since no maximum principle has been established for loaded equations, we obtain a priori estimates by using the method of energy inequalities. Take the scalar product of Eq. (7) and $y^{(\alpha)} = y^{j+\frac{\alpha}{p}}$:

$$[y_{\bar t}^{(\alpha)}, y^{(\alpha)}] - [\overline{\Lambda}_\alpha y^{(\alpha)}, y^{(\alpha)}] = [\Phi^{(\alpha)}, y^{(\alpha)}], \tag{9}$$

where

$$[u, v] = \sum_{x \in \overline{\omega}} u v H, \quad H = \prod_{\alpha=1}^{p} \hbar_\alpha, \quad [u, v]_\alpha = \sum_{i_\alpha = 0}^{N_\alpha} u_{i_\alpha} v_{i_\alpha} \hbar_\alpha.$$

Each sum in (9) is rearranged as follows:

$$[y_{\bar t}^{(\alpha)}, y^{(\alpha)}] = \frac{1}{2}\left(\left\|y^{(\alpha)}\right\|^2_{L_2(\overline{\omega}_h)}\right)_{\bar t} + \frac{\tau}{2}\left\|y_{\bar t}^{(\alpha)}\right\|^2_{L_2(\overline{\omega}_h)}, \tag{10}$$

$$\begin{aligned}[\overline{\Lambda}_\alpha y^{(\alpha)}, y^{(\alpha)}]_\alpha &= (\Lambda_\alpha y^{(\alpha)}, y^{(\alpha)})_\alpha + \Lambda_\alpha^- y_0^{(\alpha)} y_0^{(\alpha)} \hbar_\alpha + \Lambda_\alpha^+ y_{N_\alpha}^{(\alpha)} y_{N_\alpha}^{(\alpha)} \hbar_\alpha \\ &= -(a_\alpha, (y_{\bar x_\alpha}^{(\alpha)})^2)_\alpha - [d_{i_\alpha}\xi_{i_{\alpha 0}}, y^{(\alpha)}]_\alpha - \beta_{-\alpha}(y_0^{(\alpha)})^2 - \beta_{+\alpha}(y_{N_\alpha}^{(\alpha)})^2,\end{aligned} \tag{11}$$

$$[\Phi^{(\alpha)}, y^{(\alpha)}]_\alpha = [\varphi^{(\alpha)}, y^{(\alpha)}]_\alpha + \mu_{-\alpha} y_0^{(\alpha)} + \mu_{+\alpha} y_{N_\alpha}^{(\alpha)}, \tag{12}$$

where $\xi_{i_{\alpha 0}} = \xi(y_{i_{\alpha 0}}, y_{i_{\alpha 0}+1})$.

Summing (11) and (12) over all $i_\beta \ne i_\alpha$, where $\beta = 1, 2, \ldots, p$, we obtain

$$\begin{aligned}[\overline{\Lambda}_\alpha y^{(\alpha)}, y^{(\alpha)}] &= \sum_{i_\beta \ne i_\alpha} \left(\sum_{i_\alpha = 0}^{N_\alpha} \overline{\Lambda}_\alpha y^{(\alpha)} y^{(\alpha)} \hbar_\alpha \right) H/\hbar_\alpha \\ &= -\sum_{i_\beta \ne i_\alpha} [(a_\alpha, (y_{\bar x_\alpha}^{(\alpha)})^2]_\alpha + [d_{i_\alpha}\xi_{i_{\alpha 0}}, y^{(\alpha)}]_\alpha + \beta_{-\alpha}(y_0^{(\alpha)})^2 + \beta_{+\alpha}(y_{N_\alpha}^{(\alpha)})^2) H/\hbar_\alpha,\end{aligned} \tag{13}$$

$$[\Phi^{(\alpha)}, y^{(\alpha)}] = \sum_{i_\beta \ne i_\alpha} ([\varphi^{(\alpha)}, y^{(\alpha)}]_\alpha + \mu_{-\alpha} y_0^{(\alpha)} + \mu_{+\alpha} y_{N_\alpha}^{(\alpha)}) H/\hbar_\alpha. \tag{14}$$

Substituting (10), (13), and (14) into (9) gives

$$\begin{aligned}&\frac{1}{2}\left(\left\|y^{(\alpha)}\right\|^2_{L_2(\overline{\omega}_h)}\right)_{\bar t} + \frac{\tau}{2}\left\|y_{\bar t}^{(\alpha)}\right\|^2_{L_2(\overline{\omega}_h)} + c_0\left\|y_{\bar x_\alpha}^{(\alpha)}\right\|^2_{L_2(\overline{\omega}_h)} \le [d_{i_\alpha}\xi_{i_{\alpha 0}}, y^{(\alpha)}] + [\varphi^{(\alpha)}, y^{(\alpha)}] \\ &- \sum_{i_\beta \ne i_\alpha} (\beta_{-\alpha}(y_0^{(\alpha)})^2 + \beta_{+\alpha}(y_{N_\alpha}^{(\alpha)})^2) H/\hbar_\alpha + \sum_{i_\beta \ne i_\alpha} (\mu_{-\alpha} y_0^{(\alpha)} + \mu_{+\alpha} y_{N_\alpha}^{(\alpha)}) H/\hbar_\alpha.\end{aligned} \tag{15}$$

Let us estimate the terms on the right-hand side of (15). By the difference analogue of the embedding theorem (see [10]), we have

$$[d_{i_\alpha}\xi_{i_{\alpha 0}}, y^{(\alpha)}] \le \sum_{i_\beta \ne i_\alpha}\left(\sum_{i_\alpha=0}^{N_\alpha} d_{i_\alpha}\left(y^{(\alpha)}_{i_{\alpha_0}}\frac{x_\alpha^{(i_{\alpha_0}+1)}-x_\alpha^\circ}{h_\alpha}+y^{(\alpha)}_{i_{\alpha_0}+1}\frac{x_\alpha^\circ-x_\alpha^{(i_{\alpha_0})}}{h_\alpha}\right)y^{(\alpha)}\hbar_\alpha\right)H/\hbar_\alpha$$
$$\le \varepsilon c_1 l_\alpha \left\|y^{(\alpha)}_{\bar{x}}\right\|^2_{L_2(\bar{\omega}_h)} + c_1 l_\alpha\left(\frac{1}{l_\alpha}+\frac{1}{\varepsilon}\right)\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}, \tag{16}$$

$$-\sum_{i_\beta\ne i_\alpha}(\beta_{-\alpha}(y_0^{(\alpha)})^2+\beta_{+\alpha}(y_{N_\alpha}^{(\alpha)})^2)H/\hbar_\alpha \le 2c_2\sum_{i_\beta\ne i_\alpha}\left\|y^{(\alpha)}\right\|^2_{C(\alpha)}H/\hbar_\alpha$$
$$\le 2c_2\sum_{i_\beta\ne i_\alpha}\left(\varepsilon\left\|y^{(\alpha)}_{\bar{x}_\alpha}\right\|^2_{L_2(\alpha)}+\left(\frac{1}{l_\alpha}+\frac{1}{\varepsilon}\right)\left\|y^{(\alpha)}\right\|^2_{L_2(\alpha)}\right)H/\hbar_\alpha = 2c_2\varepsilon\left\|y^{(\alpha)}_{\bar{x}}\right\|^2_{L_2(\bar{\omega}_h)}+2c_2\left(\frac{1}{l_\alpha}+\frac{1}{\varepsilon}\right)\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}, \tag{17}$$

where $\|\cdot\|_{C(\alpha)}$ and $\|\cdot\|_{L_2(\alpha)}$ denote norms with respect to x_α with the other variables being fixed, and

$$\|y_{\bar{x}_\alpha}]|^2_{L_2(\bar{\omega}_h)} = \sum_{i_\beta\ne i_\alpha}\left\|y^{(\alpha)}_{\bar{x}_\alpha}\right\|^2_{L_2(\alpha)}H/\hbar_\alpha, \quad \left\|y^{(\alpha)}_{\bar{x}_\alpha}\right\|^2_{L_2(\alpha)} = \sum_{i_\alpha=1}^{N_\alpha} y^2_{\bar{x}_\alpha}\hbar_\alpha,$$
$$[\varphi^{(\alpha)}, y^{(\alpha)}] \le \frac{1}{2}\left\|\varphi^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}+\frac{1}{2}\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}, \tag{18}$$

$$\sum_{i_\beta\ne i_\alpha}(\mu_{-\alpha}y_0^{(\alpha)}+\mu_{+\alpha}y_{N_\alpha}^{(\alpha)})H/\hbar_\alpha \le \frac{1}{2}\sum_{i_\beta\ne i_\alpha}(\mu^2_{-\alpha}(t_j)+\mu^2_{+\alpha}(t_j))H/\hbar_\alpha+\varepsilon\|y^{(\alpha)}_{\bar{x}_\alpha}\rceil|^2_{L_2(\bar{\omega}_h)}+\left(\frac{1}{l_\alpha}+\frac{1}{\varepsilon}\right)\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}. \tag{19}$$

Substituting (16)–(19) into (15) and using $\varepsilon \le \dfrac{c_0}{2(1+c_1l_\alpha)+2c_2}$, we obtain

$$\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}-\left\|y^{j+\frac{\alpha-1}{p}}\right\|^2_{L_2(\bar{\omega}_h)}+\frac{c_0}{2}\tau\left\|y^{j+\frac{\alpha}{p}}_{\bar{x}_\alpha}\right\rceil\Big\|^2_{L_2(\bar{\omega}_h)} \le c_3\tau\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+\frac{\tau}{2}\left\|\varphi^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}+\frac{\tau}{2}\sum_{i_\beta\ne i_\alpha}(\mu^2_{-\alpha}(t_j)+\mu^2_{+\alpha}(t_j))H/\hbar_\alpha, \tag{20}$$

where $c_3 = c_3(c_0, c_1, c_2, l_\alpha)$ are known positive constants.

Inequality (20) is summed, first, over $\alpha = \overline{1, p}$,

$$\left\|y^{j+1}\right\|^2_{L_2(\bar{\omega}_h)}-\left\|y^{j}\right\|^2_{L_2(\bar{\omega}_h)}+\tau\frac{c_0}{2}\sum_{\alpha=1}^{p}\left\|y^{j+\frac{\alpha}{p}}_{\bar{x}}\right\rceil\Big\|^2_{L_2(\bar{\omega}_h)} \le c_3\tau\sum_{\alpha=1}^{p}\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+\frac{\tau}{2}\sum_{\alpha=1}^{p}\left(\left\|\varphi^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}+\sum_{i_\beta\ne i_\alpha}(\mu^2_{-\alpha}(t_j)+\mu^2_{+\alpha}(t_j))H/\hbar_\alpha\right), \tag{21}$$

and, then, over j' from 0 to j:

$$\left\|y^{j+1}\right\|^2_{L_2(\bar{\omega}_h)} + \frac{c_0}{2}\sum_{j'=0}^{j}\tau\sum_{\alpha=}^{p}\left\|y_{\bar{x}}^{j'+\frac{\alpha}{p}}\right]\right\|^2_{L_2(\bar{\omega}_h)} \le \left\|y^0\right\|^2_{L_2(\bar{\omega}_h)} + c_3\sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left\|y^{j'+\frac{\alpha}{p}}\right]\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+\frac{1}{2}\sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left(\left\|\varphi^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{i_\beta\neq i_\alpha}(\mu^2_{-\alpha}(t_{j'}) + \mu^2_{+\alpha}(t_{j'}))H/\hbar_\alpha\right). \tag{22}$$

Inequality (22) yields

$$\left\|y^{j+1}\right\|^2_{L_2(\bar{\omega}_h)} \le c_3\sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left\|y^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + F^j, \tag{23}$$

where

$$F^j = \frac{1}{2}\left[\sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left(\left\|\varphi^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{i_\beta\neq i_\alpha}(\mu^2_{-\alpha}(t_{j'}) + \mu^2_{+\alpha}(t_{j'}))H/\hbar_\alpha\right) + \left\|y^0\right\|^2_{L_2(\bar{\omega}_h)}\right].$$

Let us show that

$$\max_{1\le\alpha\le p}\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} \le \nu_1\sum_{j'=0}^{j-1}\tau\max_{1\le\alpha\le p}\left\|y^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \nu_2 F^j,$$

where ν_1 and ν_2 are known positive constants.

Returning to (20), we write

$$\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} - \left\|y^{j+\frac{\alpha-1}{p}}\right\|^2_{L_2(\bar{\omega}_h)} \le c_3\tau\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+\frac{\tau}{2}\left\|\varphi^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \frac{\tau}{2}\sum_{i_\beta\neq i_\alpha}(\mu^2_{-\alpha}(t_j) + \mu^2_{+\alpha}(t_j))H/\hbar_\alpha. \tag{24}$$

Summing (24) over α' from 1 to α gives

$$\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} \le \left\|y^j\right\|^2_{L_2(\bar{\omega}_h)} + c_3\tau\sum_{\alpha=1}^{p}\left\|y^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+0.5\tau\sum_{\alpha=1}^{p}\left(\left\|\varphi^{(\alpha)}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{i_\beta\neq i_\alpha}(\mu^2_{-\alpha}(t_j) + \mu^2_{+\alpha}(t_j))H/\hbar_\alpha\right). \tag{25}$$

Without loss of generality, we assume that

$$\max_{1\le\alpha'\le p}\left\|y^{j+\frac{\alpha'}{p}}\right\|^2_{L_2(\bar{\omega}_h)} = \left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}.$$

Otherwise, inequality (24) is summed until α at which $\left\|y^{j+\frac{\alpha'}{p}}\right\|^2_{L_2(\bar{\omega}_h)}$ reaches its maximum. Then (25) can be rewritten as

$$\max_{1\le\alpha\le p}\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} \le \left\|y^j\right\|^2_{L_2(\bar{\omega}_h)} + pc_3\tau\max_{1\le\alpha\le p}\left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)}$$
$$+0.5\tau\sum_{\alpha=1}^{p}\left(\left\|\varphi^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{i_\beta\neq i_\alpha}(\mu^2_{-\alpha}(t_j) + \mu^2_{+\alpha}(t_j))H/\hbar_\alpha\right). \tag{26}$$

From (23), we have

$$\left\|y^j\right\|^2_{L_2(\overline{\omega}_h)} \le pc_3 \sum_{j'=0}^{j-1} \tau \max_{1 \le \alpha \le \alpha} \left\|y^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\overline{\omega}_h)} + F^j,$$

Then, for small $\tau < \tau_0 \le \dfrac{1}{2pc_3}$, we find from (26) that

$$\max_{1 \le \alpha \le p} \left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\overline{\omega}_h)} \le \nu_1 \sum_{j'=0}^{j-1} \tau \max_{1 \le \alpha \le p} \left\|y^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\overline{\omega}_h)} + \nu_2 F^j, \tag{27}$$

where ν_1 and ν_2 are known positive constants.

By introducing the notation $g_{j+1} = \max\limits_{1 \le \alpha \le p} \left\|y^{j+\frac{\alpha}{p}}\right\|^2_{L_2(\overline{\omega}_h)}$, relation (27) can be rewritten as

$$g_{j+1} \le \nu_1 \sum_{k=1}^{j} g_k \tau + \nu_2 F^j. \tag{28}$$

Taking into account inequality (28) and applying Lemma 4 from [11, p. 171], we derive from (22) the a priori estimate

$$\begin{aligned} &\left\|y^{j+1}\right\|^2_{L_2(\overline{\omega}_h)} + \sum_{j'=0}^{j} \tau \sum_{\alpha=}^{p} \left\| y_{\bar{x}_\alpha}^{j'+\frac{\alpha}{p}} \right]\right\|^2_{L_2(\overline{\omega}_h)} \\ &\le M(t)\left[\left\|y^0\right\|^2_{L_2(\overline{\omega}_h)} + \sum_{j'=0}^{j} \tau \sum_{\alpha=1}^{p} \left(\left\|\varphi^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\overline{\omega}_h)} + \sum_{i_\beta \ne i_\alpha} (\mu^2_{-\alpha}(t_{j'}) + \mu^2_{+\alpha}(t_{j'}))H/\hbar_\alpha \right)\right]. \end{aligned} \tag{29}$$

which implies the following result.

Theorem 1. *The locally one-dimensional scheme* (7), (8) *is stable with respect to the initial data and the right-hand side, so that the solution to problem* (7), (8) *satisfies estimate* (29) *for any h and* $\tau < \tau_0$.

2. APPROXIMATION ERROR

Each equation in (7) with index α does not approximate Eq. (1) separately but approximates the equation $\mathscr{L}_\alpha \vartheta_{(\alpha)} = 0$ in the usual sense. Therefore, the system of difference equations (7) is an additive scheme; i.e., $\Psi_\alpha = \overset{\circ}{\Psi}_\alpha + \Psi^*_\alpha$,

$$\overset{\circ}{\Psi}_\alpha = (\mathscr{L}_\alpha u)^{j+\frac{1}{2}}, \quad \sum_{\alpha=1}^{p} \overset{\circ}{\Psi}_\alpha = 0, \quad \overset{\circ}{\Psi}_\alpha = O(1),$$

$$\Psi^*_\alpha = O(\hbar^2_\alpha + \tau), \quad \Psi = \sum_{\alpha=1}^{p} \Psi_\alpha + \sum_{\alpha=1}^{p} \Psi^*_\alpha = O(|\hbar|^2 + \tau), \quad |h|^2 = \hbar_1^2 + \hbar_2^2 + \ldots + \hbar_p^2.$$

For the error $z^{j+\frac{\alpha}{p}} = y^{j+\frac{\alpha}{p}} - u^{j+\frac{\alpha}{p}}$, we have the problem

$$z^{(\alpha)}_{\bar{i}} = \Lambda_\alpha z^{(\alpha)} + \Psi^{j+\frac{\alpha}{p}}_\alpha, \quad \overset{\circ}{\Psi}_\alpha = \overset{\circ}{\Psi}_\alpha + \Psi^*_\alpha, \quad \overset{\circ}{\Psi}_\alpha = O(1), \quad \Psi^*_\alpha = O(\hbar^2_\alpha + \tau), \quad \alpha = \overline{1,p}, \quad x_\alpha \in \omega_h,$$

$$z^{(\alpha)}_{\bar{i},0} = \Lambda^-_\alpha z^{(\alpha)} + \Psi^{j+\frac{\alpha}{p}}_{-\alpha}, \quad \Psi_{-\alpha} = \overset{\circ}{\Psi}_{-\alpha} + \frac{\Psi^*_{-\alpha}}{0.5\hbar_\alpha}, \quad \overset{\circ}{\Psi}_{-\alpha} = O(1), \quad \Psi^*_{-\alpha} = O(\hbar^2_\alpha) + O(\hbar_\alpha \tau), \quad x_\alpha \in \gamma_{-\alpha},$$

$$z^{(\alpha)}_{\bar{i},N_\alpha} = \Lambda^+_\alpha z^{(\alpha)} + \Psi^{j+\frac{\alpha}{p}}_{+\alpha}, \quad \Psi_{+\alpha} = \overset{\circ}{\Psi}_{+\alpha} + \frac{\Psi^*_{+\alpha}}{0.5\hbar_\alpha}, \quad \overset{\circ}{\Psi}_{+\alpha} = O(1), \quad \Psi^*_{+\alpha} = O(\hbar^2_\alpha) + O(\hbar_\alpha \tau), \quad x_\alpha \in \gamma_{+\alpha},$$

$$z(x,0) = 0,$$

or

$$z_{\bar{t}}^{(\alpha)} = \overline{\Lambda}_\alpha z^{(\alpha)} + \Psi_\alpha^{j+\frac{\alpha}{p}}, \qquad z(x,0) = 0, \tag{30}$$

where

$$\overline{\Lambda}_\alpha = \begin{cases} \Lambda_\alpha, & x_\alpha \in \omega_{h_\alpha}, \\ \Lambda_\alpha^-, & x_\alpha \in \gamma_{-\alpha}, \\ \Lambda_\alpha^+, & x_\alpha \in \gamma_{+\alpha}, \end{cases} \qquad \Psi_\alpha^{j+\frac{\alpha}{p}} = \begin{cases} \Psi_\alpha, & x_\alpha \in \omega_{h_\alpha}, \\ \Psi_{-\alpha}^-, & x_\alpha \in \gamma_{-\alpha}, \\ \Psi_{+\alpha}^+, & x_\alpha \in \gamma_{+\alpha}. \end{cases}$$

3. CONVERGENCE OF THE LOCALLY ONE-DIMENSIONAL SCHEME

By analogy with [9, p. 528], the solution of (30) is represented as $z_{(\alpha)} = v_{(\alpha)} + \eta_{(\alpha)}$, $z_{(\alpha)} = z^{j+\frac{\alpha}{p}}$, where η_α is determined by the conditions

$$\frac{\eta_{(\alpha)} - \eta_{(\alpha-1)}}{\tau} = \mathring{\Psi}_\alpha \quad \text{for} \quad x \in \omega_h + \gamma_{h,\alpha}, \quad \alpha = 1, 2, \ldots, p, \quad \eta(x,0) = 0,$$

$$\mathring{\Psi}_\alpha = \begin{cases} \mathring{\Psi}_\alpha, & x_\alpha \in \omega_{\hbar_\alpha}, \\ \mathring{\Psi}_{-\alpha}, & x_\alpha \in \gamma_{-\alpha}, \\ \mathring{\Psi}_{+\alpha}, & x_\alpha \in \gamma_{+\alpha}. \end{cases} \tag{31}$$

This yields $\eta^{j+1} = \eta_{(p)} = \eta^j + \tau(\mathring{\Psi}_1 + \mathring{\Psi}_2 + \ldots + \mathring{\Psi}_p) = \eta^j = \eta^{j-1} = \eta^{j-2} = \ldots = \eta^0 = 0$, since $\eta^0 = 0$. Then, for $\eta_{(\alpha)}$, we have

$$\eta_{(\alpha)} = \tau(\mathring{\Psi}_1 + \mathring{\Psi}_2 + \ldots + \mathring{\Psi}_\alpha) = -\tau(\mathring{\Psi}_{\alpha+1} + \ldots + \mathring{\Psi}_p).$$

The function v_α is determined by the conditions

$$\frac{v_{(\alpha)} - v_{(\alpha-1)}}{\tau} = \Lambda_\alpha v_{(\alpha)} + \tilde{\Psi}_\alpha, \quad x \in \omega_h, \quad \alpha = 1, 2, \ldots, p, \tag{32}$$

$$\frac{v_{(\alpha)} - v_{(\alpha-1)}}{\tau} = \Lambda_\alpha^- v_{(\alpha)} + \frac{\tilde{\Psi}_{-\alpha}}{0.5\hbar_\alpha}, \quad x_\alpha \in \gamma_{-\alpha}, \tag{33}$$

$$\frac{v_{(\alpha)} - v_{(\alpha-1)}}{\tau} = \Lambda_\alpha^+ v_{(\alpha)} + \frac{\tilde{\Psi}_{+\alpha}}{0.5\hbar_\alpha}, \quad x_\alpha \in \gamma_{+\alpha}, \tag{34}$$

$$v(x,0) = 0, \tag{35}$$

where

$$\overline{\Psi}_\alpha = \begin{cases} \tilde{\Psi}_\alpha = \Psi_\alpha^* + \left(a_\alpha \eta_{\bar{x}_\alpha}^{j+\frac{\alpha}{p}}\right)_{x_\alpha} - d_\alpha \xi(\eta_{i_{\alpha_0}}, \eta_{i_{\alpha_0}+1}), & x_\alpha \in \omega_{h_\alpha}, \\ \dfrac{\tilde{\Psi}_{-\alpha}}{0.5\hbar_\alpha} = \dfrac{\Psi_{-\alpha}^*}{0.5\hbar_\alpha} + \dfrac{a_\alpha^{(+1_\alpha)} \eta_{x_\alpha,0}^{j+\frac{\alpha}{p}} - \beta_{-\alpha} \eta^{j+\frac{\alpha}{p}}}{0.5\hbar_\alpha} - d_\alpha \xi(\eta_{i_{\alpha_0}}, \eta_{i_{\alpha_0}+1}), & x_\alpha \in \gamma_{-\alpha}, \\ \dfrac{\tilde{\Psi}_{+\alpha}}{0.5\hbar_\alpha} = \dfrac{\Psi_{+\alpha}^*}{0.5\hbar_\alpha} + \dfrac{-a_\alpha^{(N_\alpha)} \eta_{\bar{x}_\alpha,N_\alpha}^{j+\frac{\alpha}{p}} - \beta_{+\alpha} \eta^{j+\frac{\alpha}{p}}}{0.5\hbar_\alpha} - d_\alpha \xi(\eta_{i_{\alpha_0}}, \eta_{i_{\alpha_0}+1}), & x_\alpha \in \gamma_{+\alpha}. \end{cases}$$

The solution to problem (32)–(35) is estimated with the help of Theorem 1:

$$\left\|v^{j+1}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left\|\left[v_{\bar{x}}^{j'+\frac{\alpha}{p}}\right]\right\|^2_{L_2(\bar{\omega}_h)}$$
$$\le M(t)\sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left(\left\|\tilde{\Psi}^{j'+\frac{\alpha}{p}}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{i_\beta\neq i_\alpha}(\tilde{\Psi}^2_{-\alpha}(t_{j'}) + \tilde{\Psi}^2_{+\alpha}(t_{j'}))H/\hbar_\alpha\right). \tag{36}$$

Since

$$\eta^j = 0, \quad \eta_{(\alpha)} = O(\tau), \quad \left\|z^j\right\| \le \left\|v^j\right\|,$$

estimate (36) implies the following result.

Theorem 2. *Suppose that problem* (1)–(3) *has a unique solution* $u(x, t)$ *that is continuous in* Q_T *and there are continuous* (*in* Q_T) *derivatives*

$$\frac{\partial^2 u}{\partial t^2}, \quad \frac{\partial^4 u}{\partial x_\alpha^2 \partial x_\beta^2}, \quad \frac{\partial^3 u}{\partial x_\alpha^2 \partial t}, \quad \frac{\partial^2 f}{\partial x_\alpha^2}, \quad \alpha = 1, 2, \ldots, p, \quad \alpha \neq \beta;$$

Then scheme (7), (8) converges at the rate $O(|h|^2 + \tau)$ so that

$$\left\|y^{j+1} - u^{j+1}\right\|_1^2 \le M(|h|^2 + \tau),$$
$$\left\|y^{j+1}\right\|_1^2 = \left\|y^{j+1}\right\|^2_{L_2(\bar{\omega}_h)} + \sum_{j'=0}^{j}\tau\sum_{\alpha=1}^{p}\left\|\left[y_{\bar{x}_\alpha}^{j'+\frac{\alpha}{p}}\right]\right\|^2_{L_2(\bar{\omega}_h)}.$$

The grid equations obtained after approximating loaded equations can be conveniently solved by parametric tridiagonal Gaussian elimination (see [12, p. 131]).

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 08-05-00436.

REFERENCES

1. Kh. Zh. Dikinov, A. A. Kerefov, and A. M. Nakhushev, "On a Boundary Value Problem for the Loaded Heat Equation," Differ. Uravn. **13** (1), 77–79 (1977).
2. A. V. Borodin, "An Estimate for Elliptic Equations and Its Applications to Loaded Equations," Differ. Uravn. **13** (1), 17–22 (1977).
3. A. M. Nakhushev, "Boundary Value Problems for Loaded Hyperbolic Integro-Differential Equations and Some Applications to Soil Moisture Prediction," Differ. Uravn. **15** (1), 96–105 (1979).
4. A. M. Nakhushev, "Loaded Equations," Differ. Uravn. **19** (1), 86–94 (1983).
5. Yu. A. Anokhin, A. B. Gorstko, L. Yu. Domeshek, et al., *Mathematical Models and Methods for the Control of Large-Scale Water Reservoir* (Nauka, Novosibirsk, 1987) [in Russian].
6. M. Kh. Shkhanukov, "Difference Method for Solving a Parabolic Loaded Equation," Differ. Uravn. **13** (1), 163–167 (1977).
7. V. M. Abdullaev and K. R. Aida-Zade, "On the Numerical Solution of Loaded Systems of Ordinary Differential Equations," Zh. Vychisl. Mat. Mat. Fiz. **44**, 1585–1595 (2004) [Comput. Math. Math. Phys. **44**, 1505–1515 (2004)].
8. I. V. Fryazinov, "On Difference Approximation of Boundary Conditions for the Robin Problem," Zh. Vychisl. Mat. Mat. Fiz. **4**, 1106–1112 (1964).
9. A. A. Samarskii, *The Theory of Difference Schemes* (Nauka, Moscow, 1989; Marcel Dekker, New York, 2001).
10. V. B. Andreev, "On the Convergence of Difference Schemes Approximating Neumann and Robin Boundary Value Problems for Elliptic Equations," Zh. Vychisl. Mat. Mat. Fiz. **8**, 1218–1231 (1968).
11. A. A. Samarskii and A. V. Gulin, *Stability of Difference Schemes* (Nauka, Moscow, 1973) [in Russian].
12. A. F. Voevodin and S. M. Shugrin, *Numerical Methods for Computing One-Dimensional Systems* (SO Akad. Nauk SSSR, Novosibirsk, 1981) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1175–1196. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © L.V. Perova, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1232–1254.

Propagation of Perturbations in a Two-Layer Rotating Fluid with an Interface Excited by Moving Sources

L. V. Perova

Faculty of Physics, Moscow State University, Moscow, 119992 Russia

e-mail: lada_perova@mail.ru

Received December 24, 2008

Abstract—Propagation of small perturbations in a two-layer inviscid fluid rotating at a constant angular velocity is studied. It is assumed that the lower density fluid occupies the upper unbounded half-space, while the higher density fluid occupies the lower unbounded half-space. The source of excitation is a plane wave traveling along the interface of the fluids. An explicit analytical solution to the problem is constructed, and its existence and uniqueness are proved. The long-time wave pattern developing in the fluids is analyzed.

DOI: 10.1134/S0965542509070100

Key words: stream function, rotating fluid, internal waves, surface waves.

1. STATEMENT OF THE PROBLEM

Assume that an unbounded space is filled with an inviscid two-component fluid. For the system to be in equilibrium in the unperturbed state, it is necessary that the density of the lower layer be higher than that of the upper layer. The interface between the layers at rest is a plane. Assume that both fluids rotate uniformly as a whole about an axis perpendicular to the interface of the fluids. Thus, this problem is a generalization of previously well-studied ones for a semi-infinite fluid excited by moving surface sources. We introduce a Cartesian coordinate system (x_1, x_2, x_3) tied to the two-layer fluid such that the interface of fluid components lies in the plane $x_3 = 0$ and the Ox_3 axis coincides with the rotation axis of the system. Then the angular velocity is aligned with the Ox_3 axis, and the Coriolis vector has a single nonzero component: $\alpha = \{0, 0, \alpha\}$, where α is the double frequency of fluid rotation.

We study small two-dimensional fluid motions. It is assumed that all the functions characterizing the state of the system are independent of x_2. To distinguish between two fluid layers, the variables characterizing the lower and upper fluids are denoted by superscripts (1) and (2), respectively.

The fluid dynamics equations written in the chosen coordinate system in the linear approximation (see [1, 2]) are

$$\begin{gathered}\rho^{(i)}\frac{\partial v_1^{(i)}}{\partial t} - \rho^{(i)}\alpha v_2^{(i)} + \frac{\partial p^{(i)}}{\partial x_1} = 0, \quad i = 1, 2,\\ \rho^{(i)}\frac{\partial v_2^{(i)}}{\partial t} + \rho^{(i)}\alpha v_1^{(i)} = 0,\\ \rho^{(i)}\frac{\partial v_3^{(i)}}{\partial t} - \frac{\partial p^{(i)}}{\partial x_3} = 0,\\ \frac{\partial v_1^{(i)}}{\partial x_1} + \frac{\partial v_3^{(i)}}{\partial x_3} = 0,\end{gathered} \tag{1.1}$$

where $v_j^{(i)}(x_1, x_3, t)$ $(j = 1, 2, 3)$ are the components of fluids particle velocities $\mathbf{v}^{(i)} = \{v_1^{(i)}, v_2^{(i)}, v_3^{(i)}\}$, $p^{(i)}(x_1, x_3, t)$ are the pressures, and $\rho^{(i)}$ are the densities of the fluids.

Note that the second velocity component $v_2(x_1, x_3, t)$ can be eliminated from the system, since it can be expressed from the second equation in (1.1) in terms of the first component $v_1(x_1, x_3, t)$ by the formula

$$v_2(x_1, x_3, t) = \alpha \int_0^t v_1(x_1, x_3, \tau)d\tau + v_2(x_1, x_3, 0).$$

To make system (1.1) closed, we add boundary and initial conditions. Assume that both fluids are at rest relative to the rotating coordinate system until the time $t = 0$. This means that zero initial conditions are imposed on all the functions.

Let the elevation of the interface above the unperturbed level $x_3 = 0$ be described by $x_3 = \xi(x_1, t)$. Two types of boundary conditions are set at the interface: the kinematic conditions

$$\frac{\partial \xi}{\partial t}(x_1, t) = v_3^{(1)}(x_1, x_3, t)\big|_{x_3 = 0} = v_3^{(2)}(x_1, x_3, t)\big|_{x_3 = 0}, \tag{1.2a}$$

$$\left.\frac{\partial v_3^{(1)}(x_1, x_3, t)}{\partial t}\right|_{x_3 = 0} = \left.\frac{\partial v_3^{(2)}(x_1, x_3, t)}{\partial t}\right|_{x_3 = 0} \tag{1.2b}$$

and the dynamic condition

$$p^{(1)}(x_1, x_3, t) - p^{(2)}(x_1, x_3, t) + p(x_1, x_3, t)\big|_{x_3 = 0} = \rho^{(1)} g\xi(x_1, t) - \rho^{(2)} g\xi(x_1, t)\big|_{x_3 = 0}. \tag{1.2c}$$

Here, g is the acceleration of gravity and $p(x_1, x_3, t)$ is the pressure characterizing the effect of the perturbation source on the system.

In addition to the conditions on the internal boundary of the domain, in an unbounded space, the solution has to satisfy conditions ensuring that the propagating perturbations are bounded as $x_3 \longrightarrow -\infty$ and $x_3 \longrightarrow +\infty$. Their form will be specified later.

We use a method convenient for solving fluid dynamics problems, namely, the system of first-order differential equations for several functions is reduced to an initial–boundary value problem for a single scalar function. Following [1, 2], this function is specified as the stream function $\psi(x_1, x_3, t)$, which is related to the components of the two-dimensional velocity $\mathbf{v} = \{v_1, v_3\}$ by the formula

$$\{v_1, v_3\} = \{\psi_{x_3}, -\psi_{x_1}\}. \tag{1.3}$$

Recall that, for two-dimensional motions of a rotating fluid, the second velocity component $v_2(x_1, x_3, t)$ can be found as described above.

For a two-layer fluid, $\psi(x_1, x_3, t)$ splits into two functions: $\psi^{(1)}(x_1, x_3, t)$ solves the problem in the lower half-space $\mathbb{R}^2_- \equiv \{(x_1, x_3) \in \mathbb{R}^2 : x_3 < 0\}$ and $\psi^{(2)}(x_1, x_3, t)$ is defined in the upper half-space $\mathbb{R}^2_+ \equiv \{(x_1, x_3) \in \mathbb{R}^2 : x_3 > 0\}$; i.e., $\psi(x_1, x_3, t)$ has the form

$$\psi(x_1, x_3, t) = \begin{cases} \psi^{(1)}(x_1, x_3, t), & (x_1, x_3) \in \mathbb{R}^2_- \\ \psi^{(2)}(x_1, x_3, t), & (x_1, x_3) \in \mathbb{R}^2_+. \end{cases}$$

After standard transformations of system (1.1) similar to those performed in [1–3], we obtain the fourth-order differential equation

$$\frac{\partial^2}{\partial t^2}(\psi^{(i)}_{x_1x_1} + \psi^{(i)}_{x_3x_3}) + \alpha^2 \psi^{(i)}_{x_3x_3} = 0, \quad i = 1, 2. \tag{1.4}$$

For the physical process in question, the stream function and the functions from vector system (1.1) satisfy zero initial conditions:

$$\psi^{(i)}(x_1, x_3, 0) = \psi^{(i)}_t(x_1, x_3, 0) = 0. \tag{1.5}$$

Following [3–6], we assume that the perturbation switched on at $t = 0$ is a plane periodic wave traveling along the interface $x_3 = 0$, which is hereafter denoted by Γ. Then the functions characterizing the source depend on $(x_1 - ct)$. As a result, the argument of the right-hand side of the dynamic boundary conditions for $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$ is specified as

$$\frac{\partial^2}{\partial t^2}(\rho^{(1)}\psi^{(1)}_{x_3} - \rho^{(2)}\psi^{(2)}_{x_3}) + \alpha^2(\rho^{(1)}\psi^{(1)}_{x_3} - \rho^{(2)}\psi^{(2)}_{x_3}) - g(\rho^{(1)} - \rho^{(2)})\psi^{(1,2)}_{x_1x_1}\big|_\Gamma = f(x_1 - ct)\eta(t) + C(t), \tag{1.6c}$$

where $\eta(t) \in C_0^{(2)}[0, +\infty)$, $\eta(0) = \eta'(0) = 0$, $\exists T: \eta(t) \equiv 1$ for $t > T$, the function $f(z) \in C^{(2)}(\mathbb{R}^1)$ has a period of 2π (i.e., $\forall a: f(a) = f(a + 2\pi k)$, $k \in Z$), and $C(t)$ is the function to be determined.

The kinematic boundary conditions (1.2a) and (1.2b) are transformed into the following relations for the components of the stream function:

$$\psi_{x_1}^{(1)}\Big|_\Gamma = \psi_{x_1}^{(2)}\Big|_\Gamma, \tag{1.6a}$$

$$\frac{\partial \psi_{x_1}^{(1)}}{\partial t}\Big|_\Gamma = \frac{\partial \psi_{x_1}^{(2)}}{\partial t}\Big|_\Gamma. \tag{1.6b}$$

To ensure that the solution decreases as $x_3 \longrightarrow -\infty$ and $x_3 \longrightarrow +\infty$, the functions $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$ have to satisfy the conditions

$$\begin{aligned} \left|D_t^k D_{x_j}^p \psi^{(2)}(x_1, x_3, t)\right|\Big|_{x_3 \to +\infty} &\le B(t)\exp(-\delta x_3), \\ \left|D_t^k D_{x_j}^p \psi^{(1)}(x_1, x_3, t)\right|\Big|_{x_3 \to -\infty} &\le B(t)\exp(\delta x_3), \end{aligned} \tag{1.7}$$

where $k = 0, 1, 2$; $p = 0, 1$; $j = 1, 3$; $0 < \delta < 1$; and $B(t)$ is a continuous nonnegative function.

The system of conditions for the problem is completed with the requirement that the solution be a periodic function of x_1 with the same period 2π as at the perturbation source:

$$\begin{aligned} D_t^k D_{x_j}^p \psi^{(i)}(x_1, x_3, t)\big|_{x_1 = a} &= D_t^k D_{x_j}^p \psi^{(i)}(x_1, x_3, t)\big|_{x_1 = a + 2\pi} \\ \forall a,\ k &= 0, 1, 2, \quad p = 0, 1, 2, \quad j = 1, 3. \end{aligned} \tag{1.8}$$

The complete statement of the initial–boundary value problem for $\psi(x_1, x_3, t)$ is as follows.

Problem A. For $t \ge 0$, find functions $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$ that are continuous, together with their partial derivatives involved in the conditions, in the spaces $\overline{\mathbb{R}}_-^2$ and $\overline{\mathbb{R}}_+^2$, respectively; 2π-periodic with respect to x_1; and satisfy (in the classical sense) Eqs. (1.4) in the open domains $\mathbb{R}_-^2$ and $\mathbb{R}_+^2$, boundary conditions (1.6), initial conditions (1.5), and regularity condition (1.7) at infinity. Determine $C(t)$ from boundary conditions (1.6c).

Two remarks have to be made. First, the densities $\rho^{(1)}$ and $\rho^{(2)}$, which distinguish the dynamic properties of the fluids in the upper and lower layers, are involved only in the boundary conditions but are lacking in the equations; i.e., $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$ satisfy the same equation. Second, boundary conditions (1.6) involve both unknown functions $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$. Therefore, problem A does not split into two separate problems in the upper and lower half-spaces for one of these functions.

2. CONSTRUCTION OF THE SOLUTION AND THE PROOF OF THE EXISTENCE AND UNIQUENESS THEOREMS

Since the solution to problem A is assumed to be periodic in x_1 (see (1.8)), following [1–6], we search for its components $\psi^{(1)}(x_1, x_3, t)$ and $\psi^{(2)}(x_1, x_3, t)$ by applying the Fourier method; i.e., the solution is sought in the form of the series

$$\psi^{(i)}(x_1, x_3, t) = \sum_{n=-\infty}^{+\infty} \exp(inx_1)\, v_n^{(i)}(x_3, t) \equiv \sum_{n=-\infty}^{+\infty} \psi_n^{(i)}(x_1, x_3, t), \quad i = 1, 2, \tag{2.1}$$

where $v_n^{(i)}(x_3, t)$ are as yet unknown functions defined according to their superscripts in the lower $x_3 \le 0$ and upper $x_3 \ge 0$ half-planes.

Note that, as usual, the variables are dimensionless with the old notation retained.

Problems for $v_n^{(i)}(x_3, t)$ are obtained by substituting series (2.1) into Eqs. (1.4), initial conditions (1.5), and regularity conditions (1.7). (The validity of the term-by-term differentiation of the series involved will be justified later.) As a result, we obtain a countable number of problems for $v_n^{(i)}(x_3, t)$:

$$\frac{\partial^2}{\partial t^2}\left[\left(v_n^{(i)}\right)_{x_3 x_3} - n^2 v_n^{(i)}\right] + \alpha^2 \left(v_n^{(i)}\right)_{x_3 x_3} = 0,$$

$$V_n^{(i)}\Big|_{t=0} = (V_n^{(i)})_t\Big|_{t=0} = 0,$$
$$\left|D_t^k D_{x_3}^p V_n^{(1)}(x_3, t)\right|\Big|_{x_3 \to -\infty} \le B(t)\exp(\delta x_3),$$
$$\left|D_t^k D_{x_3}^p V_n^{(2)}(x_3, t)\right|\Big|_{x_3 \to +\infty} \le B(t)\exp(-\delta x_3),$$
$$k = 0, 1, 2, \quad p = 0, 1, \quad 0 < \delta < 1.$$

Note that the solutions to two problems with $n = 0$ are immediately found, since the equations and all the corresponding conditions are satisfied only by the functions

$$V_0^{(1)}(x_3, t) = V_0^{(2)}(x_3, t) = 0. \tag{2.2}$$

The solutions to the problems with $n \neq 0$ are found using standard methods (see [3–6]), namely, by applying the Laplace transform with respect to t and solving the corresponding differential equation for the Laplace transform $\widehat{V}_n^{(i)}(x_3, p)$ with boundedness conditions at infinity. As a result, we obtain the functions

$$\widehat{V}_n^{(1)}(x_3, p) = C_n^{(1)}(p)\exp\left(|n|\frac{p}{\sqrt{p^2 + \alpha^2}}x_3\right),$$
$$\widehat{V}_n^{(2)}(x_3, p) = C_n^{(2)}(p)\exp\left(-|n|\frac{p}{\sqrt{p^2 + \alpha^2}}x_3\right),$$

where $C_n^{(i)}(p)$ are the p-dependent constants from the general solution.

Substituting these expressions into the Laplace transforms of kinematic boundary conditions (1.6), we see that $C_n^{(1)}(p) = C_n^{(2)}(p)$.

This equality is very important, since establishes the relation between the Laplace transforms of the harmonics $\widehat{\psi}_n^{(1)}(x_1, x_3, p)$ and $\widehat{\psi}_n^{(2)}(x_1, x_3, p)$ of the solution to problem A in the upper and lower half-spaces.

Taking into account $C_n^{(1)}(p) = C_n^{(2)}(p)$ and using the Laplace transforms of dynamic boundary conditions (1.6c), we obtain the Laplace transforms of the terms of series (2.1):

$$\widehat{\psi}_n^{(1)}(x, p) = \frac{\phi_n(p)\exp[inx_1 + |n|(p/\sqrt{p^2 + \alpha^2})x_3]}{(\rho^{(1)} + \rho^{(2)})p|n|\sqrt{p^2 + \alpha^2} + g(\rho^{(1)} - \rho^{(2)})n^2},$$
$$\widehat{\psi}_n^{(2)}(x, p) = \frac{\phi_n(p)\exp[inx_1 - |n|(p/\sqrt{p^2 + \alpha^2})x_3]}{(\rho^{(1)} + \rho^{(2)})p|n|\sqrt{p^2 + \alpha^2} + g(\rho^{(1)} - \rho^{(2)})n^2}, \quad n \neq 0. \tag{2.3}$$

Here and below, the argument x with no index denotes the collection of x_1 and x_3.

Note that the functions $\phi_n(p)$ used to express $C_n^{(i)}(p)$ are directly related to the Laplace transform of the right-hand side of boundary conditions (1.6c).

The solution to problem A in the original variables (x_1, x_3, t) was complicated by the fact that it did not split into two separate problems for $\psi^{(1)}(x, t)$ and $\psi^{(2)}(x, t)$ in the upper and lower half-spaces. An advantage of the transition to Laplace transforms is that this difficulty can now be overcome. Therefore, it can be perform the further actions, the pre-images of $\widehat{\psi}_n^{(i)}(x, p)$ in (2.3) can be recovered separately for each of the series $\psi^{(1)}(x, t)$ and $\psi^{(2)}(x, t)$ in (2.1.) in the half-spaces $\mathbb{R}_-^2$ - and $\mathbb{R}_+^2$, respectively.

Remark 1. Taking into account that the solutions to the problems in $\mathbb{R}_-^2$ and $\mathbb{R}_+^2$ satisfy the same equation (1.4), an analysis of the difference between $\widehat{\psi}_n^{(1)}(x, p)$ and $\widehat{\psi}_n^{(2)}(x, p)$ suggests that, given one of the corresponding solutions, for example, $\psi^{(1)}(x, t)$ in $\mathbb{R}_-^2$, the second stream-function component $\psi^{(2)}(x, t)$ in $\mathbb{R}_+^2$ can be constructed by reversing the sign of the expressions involving x_3.

This assertion can be verified in the course of constructing the solution to problem A in the lower half-space.

Based on Remark 1, we give a detailed argument only for the fluid in $\mathbb{R}^2_-$, i.e., for $\psi^{(1)}(x, t)$.

Expression (2.3) for $\widehat{\psi}_n^{(1)}(x, p)$ is transformed into a form such that the inverse Laplace transforms can be performed by applying the properties of Laplace transforms and look-up tables (see [7]). Following [1–6], this can be achieved using the integral representation of $\widehat{\psi}_n^{(1)}(x, p)$:

$$\psi_n^{(1)}(x, p) = -\frac{\exp(inx_1)}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\int_{-\infty}^{+\infty}\frac{\mu(\mu^2+1)\exp(i\mu|n|x_3)\phi_n(p)d\mu}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1][\mu^2 + p^2(p^2+\alpha^2)^{-1}]}$$

$$-\frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{g(\rho^{(1)} - \rho^{(2)})n^2\sqrt{\alpha^4 + 4Q^2g^2n^2}}\frac{p^2 + \alpha^2}{p^2 + 2^{-1}(\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2)}\phi_n(p) \tag{2.4}$$

$$\times \exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right), \quad n \neq 0.$$

Here, $Q = \dfrac{\rho^{(1)} - \rho^{(2)}}{\rho^{(1)} + \rho^{(2)}}$.

For further computations, we need the roots of the polynomial in the first square brackets in the denominator of the first term in (2.4). The root lying in the lower half-plane $\operatorname{Im}\mu < 0$ is denoted by

$$\mu^- = -i\frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg|n|}. \tag{2.5}$$

By applying the standard methods of [1–6] to the integral representation of $\widehat{\psi}_n^{(1)}(x, p)$ in (2.4), we form the sum of two products of Laplace transforms of $\phi_n(t)$ and a function of the form $\sin\beta t$, which return the convolutions of these pairs of functions in the space of preimages:

$$\psi_n^{(1)}(x, t) = -\frac{\alpha\exp(inx_1)}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\int_{-\infty}^{+\infty}\frac{\mu\exp(i\mu|n|x_3)}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2+1}}\left\{\int_0^t \sin\left[\frac{\alpha\mu(t-\tau)}{\sqrt{\mu^2+1}}\right]\phi_n(\tau)d\tau\right\}d\mu$$

$$+\frac{2\sqrt{2}Q^2g}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^4}}\exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right) \tag{2.6}$$

$$\times \int_0^t \sin\left[\frac{1}{\sqrt{2}}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}(t-\tau)\right]\phi_n(\tau)d\tau.$$

To complete the construction of the solution in $\mathbb{R}^2_-$ -, we have to specify the functions $\phi_n(t)$ in (2.6). As was noted above, they are related to $f(x_1 - ct)$ in (1.6c), which characterizes the source of perturbations. For this reason, this function is represented by the Fourier series

$$f(z) = \sum_{n=-\infty}^{+\infty} f_n\exp(inz), \tag{2.7}$$

where

$$f_n = \frac{1}{2\pi}\int_0^{2\pi} f(z)\exp(-inz)dz \tag{2.7}'$$

are the Fourier coefficients. According to the smoothness conditions imposed on $f(z)$, series (2.7) converges uniformly to this function and its coefficients f_n in (2.7)' satisfy the condition

$$\sum_{n=-\infty}^{+\infty} |f_n| < \infty. \tag{2.7"}$$

Substituting the Fourier series expansion (2.7) of $f(x_1 - ct)$ into the right-hand side of (1.6c) and the solution $\psi^{(1)}(x, t)$ to problem A in $\mathbb{R}^2_-$ constructed in the form of series (2.1) with terms (2.6) into the left-hand side of (1.6c), after setting $x_3 = 0$ on the left-hand side, we obtain

$$\sum_{n=-\infty}^{+\infty} \phi_n(t)\exp(inx_1) = \sum_{n=-\infty}^{+\infty} f_n \exp[in(x_1 - ct)]\eta(t) + C(t),$$

from which we find the unknown functions $\phi_n(t)$ with $n \neq 0$ and determine $C(t)$ from boundary conditions (1.6c):

$$\phi_n(t) = f_n \exp(-inct)\eta(t), \quad C(t) = -f_0\eta(t). \tag{2.8}$$

Thus, the solution to problem A in $\mathbb{R}^2_-$ has the form

$$\begin{aligned}
\psi^{(1)}(x, t) \equiv \sum_{n\neq 0}(u_n^{(1)}(x, t) + w_n^{(1)}(x, t)) = \sum_{n\neq 0}\Bigg(&-\frac{f_n\alpha}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\exp(inx_1) \\
&\times \int_{-\infty}^{+\infty} \frac{\mu\exp(i\mu|n|x_3)}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}}\left\{\int_0^t \sin\frac{\alpha\mu(t-\tau)}{\sqrt{\mu^2+1}}\exp(-inc\tau)\eta(\tau)d\tau\right\}d\mu \\
&+ \frac{2\sqrt{2}f_nQ^2g}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^4}} \\
&\times \exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right)\int_0^t \sin\left[\frac{1}{\sqrt{2}}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}(t-\tau)\right]\exp(-inc\tau)\eta(\tau)d\tau\Bigg).
\end{aligned} \tag{2.9}$$

According to Remark 1, the solution $\psi^{(2)}(x, t)$ to problem A in $\mathbb{R}^2_+$ can be obtained without performing computations from formula (2.9) for $\psi^{(1)}(x, t)$ by reversing the sign of the expressions containing x_3 in the exponent in the second terms:

$$\begin{aligned}
\psi^{(2)}(x, t) \equiv \sum_{n\neq 0}[u_n^{(2)}(x, t) + w_n^{(2)}(x, t)] = \sum_{n\neq 0}\Bigg(&-\frac{f_n\alpha}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\exp(inx_1) \\
&\times \int_{-\infty}^{+\infty} \frac{\mu\exp(i\mu|n|x_3)}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}}\left\{\int_0^t \sin\frac{\alpha\mu(t-\tau)}{\sqrt{\mu^2+1}}\exp(-inc\tau)\eta(\tau)d\tau\right\}d\mu \\
&+ \frac{2\sqrt{2}f_nQ^2g}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^4}} \\
&\times \exp\left(inx_1 - \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right)\int_0^t \sin\left[\frac{1}{\sqrt{2}}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}(t-\tau)\right]\exp(-inc\tau)\eta(\tau)d\tau\Bigg).
\end{aligned}$$

Note that the model of a two-layer rotating fluid differs from the model in [3] for a single rotating fluid occupying the lower half-space with a free upper boundary. Specifically, in the present problem, perturbations propagate not only in the lower half-space but also in the upper one filled with a lower density fluid rotating at the same angular velocity. For the fluid in the lower half-space, the source of perturbations in two problems is identical: a plane wave traveling along its surface. Therefore, we can expect that setting $\rho^{(2)} = 0$ in the solution $\psi^{(1)}(x, t)$ to problem A in $\mathbb{R}^2_-$ yields an explicit solution to the problem of oscilla-

tions propagating in a rotating homogeneous fluid with a free upper boundary. Setting $\rho^{(2)}$ in (2.9) equal to zero and comparing the result with the solution to the problem in [3], we find that the difference is that the denominator of the fractions in (2.9) involves the constant value $\rho^{(1)}$. This is explained by the fact that $f(x_1 - ct)$ also differs from its analogue in the problem in [3] by the same multiplier in the denominator. Thus, our conjecture is true.

Note that the derivation of formula (2.9) involves the term-by-term differentiation of series (2.1). This operation has to be justified. For this purpose, we prove the following result.

Theorem 1. *Problem A has a solution that is given by* (2.9) *and* (2.9)'.

Proof. An analysis of the differences between the components of the solution to problem A in the lower and upper half-spaces (see (2.9) and (2.9)') shows that it suffices to prove the theorem for either $\psi^{(1)}(x, t)$ or $\psi^{(2)}(x, t)$. We do it for the solution $\psi^{(1)}(x, t)$ to problem A in $\mathbb{R}^2_-$. Consider an analogue of series (2.9) in which the terms $\psi_n^{(1)}(x, t)$ are given by (2.7) with as yet unspecified functions $\phi_n(t)$.

Consider the series consisting of the first terms $u_n^{(1)}(x, t)$ of the harmonics $\psi_n^{(1)}(x, t)$ representable as $\psi_n^{(1)}(x, t) = u_n^{(1)}(x, t) + w_n^{(2)}(x, t)$, i.e., the series $u^{(1)}(x, t) = \Sigma_{n \neq 0} u_n^{(1)}(x, t)$. For the integral with respect to μ in $u_n^{(1)}(x, t)$, we introduce the function

$$I_n(x_3, t) = -\frac{\alpha}{\pi i g(\rho_0^{(1)} - \rho_0^{(2)})n^2} \int_{-\infty}^{+\infty} \frac{\mu \exp(i\mu|n|x_3)\sin[\alpha^2\mu(\mu^2+1)^{-1/2}t]}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2+1}} d\mu.$$

Such integrals can be evaluated using the contour integration method (see [7]), according to which the values of the integral and its partial derivatives with respect to x_3 and t are determined by the residues at the singular points of the integrand lying in the complex plane domain to which this function is analytically continued. For the lower half-plane $\operatorname{Im}\mu < 0$, these are the points $\mu_1 = -i$ and $\mu_2 = \mu^-$ (indicated in (2.5)). Note that they do not coincide when $\alpha \neq 0$. The function $I_n(x_3, t)$ then satisfies the estimate

$$\left|\frac{\partial^k}{\partial t^k}\frac{\partial^p}{\partial x_3^p} I_n(x_3, t)\right| \leq |n|^{p-2}\left\{C_{\mathrm{cr}}^{(1)}(\varepsilon_1)\exp[(1-\varepsilon_1)|n|x_3]\exp[C^{(2)}(\varepsilon_1)t]\right.$$
$$\left. + C_{\mathrm{cr}}^{(3)}(\varepsilon_2)\exp\left[\left(\frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg|n|} - \varepsilon_2\right)|n|x_3\right]\exp[C^{(4)}(\varepsilon_2)t]\right\}, \tag{2.10}$$

where $k, p = 0, 1, 2, \ldots$, and $C_{\mathrm{cr}}^{(1, 2, 3, 4)}(\varepsilon_{1, 2})$ are positive constants with $\varepsilon_{1, 2} > 0$.

The terms of the series $u_n^{(1)}(x, t)$ are written using $I_n(x_3, t)$ as

$$u_n^{(1)}(x, t) = e^{inx_1}\int_0^t I_n(x_3, t-\tau)\phi_n(\tau)d\tau.$$

Since the functions $\phi_n(t)$ are related to $f(z)$ and $\eta(t)$ by formula (2.8), they belong to the same smoothness class ($\phi_n(t) \in C_0^{(2)}[0, +\infty)$) and satisfy the estimate $|\phi_n(t)| \leq \|f_n\|$. These properties of $\phi_n(t)$ and estimate (2.10) imply the uniform convergence of the series with the terms

$$\frac{\partial^k}{\partial t^k}\frac{\partial^l}{\partial x_1^l}\frac{\partial^p}{\partial x_3^p} u_n^{(1)}(x, t), \quad k, p, l = 0, 1, \ldots,$$

in $\mathbb{R}^2_-$. This substantiates the term-by-term differentiation of $u^{(1)}(x, t) = \Sigma_{n \neq 0} u_n^{(1)}(x, t)$. Additionally, we have proved that $\psi^{(1)}(x, t)$, together with its partial derivatives, as many as indicated in the statement of problem A, is continuous in $\mathbb{R}^2_-$.

Consider the second terms $w_n^{(1)}(x, t)$ of the harmonics $\psi_n^{(1)}(x, t)$ in (2.6). The integrals they involve are proper. Their integrands are estimated taking into account the coefficient multiplying each integral:

$$\left|\frac{\partial^k}{\partial t^k}\frac{\partial^p}{\partial x_3^p}w_n^{(1)}(x, t)\right| \le C_{\text{cr}}\frac{(\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2)^{p+(k-1)/2}}{\sqrt{\alpha^4 + 4Q^2g^2n^2}}$$

$$\times \max_{0<t<T}|\phi_n(t)|\exp\left(\frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right), \quad x_3 \le -\delta < 0;$$

where C_{cr} is independent of n.

The properties of $\phi_n(t)$ imply that $\psi^{(1)}(x, t)$ corresponding to the series $w^{(1)}(x, t) = \Sigma_{n\neq 0}w_n^{(1)}(x, t)$ also satisfies the requirements of the definition of a classical solution to problem A. Therefore, like $u^{(1)}(x, t)$, the series $w^{(1)}(x, t)$ can be differentiated term by term in $\overline{\mathbb{R}}_-^2$.

We do not need to verify whether the solution $\psi^{(1)}(x, t)$ to problem A in $\mathbb{R}_-^2$ satisfies the equation and the boundary and initial conditions, since expression (2.9) was not presented in an explicit form but rather was directly constructed.

The theorem is proved.

Let us discuss the uniqueness of a solution to problem A. We prove this for functions possessing the following properties: in any doubly connected domain consisting of two closed subdomains $\overline{Q}_-$, and $\overline{Q}_+$, where

$$\overline{Q}_- = \{(x_1, x_3) \in \mathbb{R}_-^2 : -a \le x_3 \le -\varepsilon < 0, |x_1| < b\} \times [0, T] \equiv \overline{\Omega}_- \times [0, T],$$

$$\overline{Q}_+ = \{(x_1, x_3) \in \mathbb{R}_+^2 : 0 < \varepsilon \le x_3 \le a, |x_1| < b\} \times [0, T] \equiv \overline{\Omega}_+ \times [0, T]$$

(here, a, b, and ε are positive constants), the functions have continuous partial derivatives $D_t^k D_{x_j}^p \psi(x, t)$, where $k = 0, 1, 2$; $p = 0, 1, 2$; and $j = 1, 3$. We say that these functions belong to the class D.

Theorem 2. *Problem A has a unique solution in the smoothness class D.*

Proof. Following [1–6], we prove the theorem by the energy method. Specifically, the solutions $\psi^{(1)}(x, t)$ and $\psi^{(2)}(x, t)$ in $\mathbb{R}_-^2$ and $\mathbb{R}_+^2$ are substituted in turn into Eq. (1.4); the resulting identities are multiplied by $\rho^{(1)}\overline{\psi}_t^{(1)}(x, t)$ and $\rho^{(2)}\overline{\psi}_t^{(2)}(x, t)$, respectively; and the resulting equalities for $\rho^{(1)}\overline{\psi}_t^{(1)}(x, t)$ and $\rho^{(2)}\overline{\psi}_t^{(2)}(x, t)$ are integrated over some domains Q_- and Q_+, respectively. For the fluid in the lower half-space, we obtain

$$\int_0^t\int_{\Omega_-}\int\left(\frac{\partial^2}{\partial\tau^2}\frac{\partial^2\psi^{(1)}}{\partial x_1^2} + \frac{\partial^2}{\partial\tau^2}\frac{\partial^2\psi^{(1)}}{\partial x_3^2} + \alpha^2\frac{\partial^2\psi^{(1)}}{\partial x_3^2}\right)\rho^{(1)}\frac{\partial\overline{\psi}^{(1)}}{\partial\tau}d\Omega_-d\tau = 0.$$

By using the Ostrogradsky–Gauss formula for a two-dimensional domain and vector analysis formulas, this equation is reduced to

$$\frac{\rho^{(1)}}{2}\left\|\nabla\psi_t^{(1)}\right\|_{L_2(\Omega_-)}^2 + \frac{\rho^{(1)}\alpha^2}{2}\left\|\psi_{x_3}^{(1)}\right\|_{L_2(\Omega_-)}^2 = \int_0^t\int_{\Gamma_-}(N_{x\tau}\psi^{(1)})\rho^{(1)}\overline{\psi}_\tau^{(1)}d\Gamma_-d\tau. \tag{2.11}$$

Here, Γ_- is the boundary of Ω_- and N_{xt} denotes the differential operator

$$N_{\text{xt}}\psi = \frac{\partial^2}{\partial t^2}\frac{\partial\psi}{\partial n_x} + \alpha^2(\mathbf{n}_x, \mathbf{e}_3)\frac{\partial\psi}{\partial x_3},$$

where $\mathbf{n}_x$ is the outward normal to Γ_- at the point $(x_1, x_3) \in \Gamma_-$and $\mathbf{e}_1$ and $\mathbf{e}_3$ are the basis vectors along the Ox_1 and Ox_3 axes, respectively.

Following [3–6], the domain Ω_- is specified as a rectangle of width 2π:

$$\Pi_k^{\varepsilon a,-} = \{(x_1, x_3) \in \mathbb{R}_-^2 : -a < x_3 < -\varepsilon < 0,\ 2\pi k < x_1 < 2\pi(k+1)\},\quad k \in Z.$$

As a result, we can use the periodicity of solution (1.8). The contour integral on the right-hand side of (2.11) is written along the components of Γ_-, namely, along the sides of $\Pi_k^{\varepsilon a,-}$:

$$\begin{aligned}
&\int_{0}^{t-\varepsilon}\int_{-a}\rho^{(1)}(\psi^{(1)}_{\tau\tau x_1}\overline{\psi}^{(1)}_{\tau}\big|_{x_1=2\pi(k+1)} - \psi^{(1)}_{\tau\tau x_1}\overline{\psi}^{(1)}_{\tau}\big|_{x_1=2\pi k})dx_3 d\tau\\
&\quad+\int_{0}^{t}\int_{2\pi k}^{2\pi(k+1)}\rho^{(1)}(-\psi^{(1)}_{\tau\tau x_3} - \alpha^2\psi^{(1)}_{x_3})\overline{\psi}^{(1)}_{\tau}\big|_{x_3=-a}dx_1 d\tau\\
&\quad+\int_{0}^{t}\int_{2\pi k}^{2\pi(k+1)}\rho^{(1)}(\psi^{(1)}_{\tau\tau x_3} + \alpha^2\psi^{(1)}_{x_3})\overline{\psi}^{(1)}_{\tau}\big|_{x_3=-\varepsilon}dx_1 d\tau.
\end{aligned}$$

Since $\psi^{(1)}(x, t)$ is a periodic function of x_1 (see (1.8)), we conclude that the integrand in the first integral along the lateral sides $x_1 = 2\pi k$ and $x_1 = 2\pi(k+1)$ of $\Pi_k^{\varepsilon a,-}$ vanishes. For the integrand of the middle integral along the lower base of $\Pi_k^{\varepsilon a,-}$, regularity conditions (1.7) as $x_3 \longrightarrow -\infty$ imply the estimate

$$\rho^{(1)}(\psi^{(1)}_{ttx_3} + \alpha^2\psi^{(1)}_{x_3})\overline{\psi}^{(1)}_{t}\big|_{x_3=-a} \le \tilde{B}(t)\exp(-2\delta a),\quad 0 < \delta < 1.$$

Passage to the limit as $\varepsilon \longrightarrow 0$ and $a \longrightarrow +\infty$ in (2.11) transforms the rectangle $\Pi_k^{\varepsilon a,-}$ into the semi-infinite strip

$$\Pi_k^{-\infty} = \{(x_1, x_3) \in \overline{\mathbb{R}}_-^2 : -\infty < x_3 \le 0,\ 2\pi k < x_1 < 2\pi(k+1)\},\quad k \in Z,$$

in $\mathbb{R}_-^2$, in which the following limiting version of (2.11) holds:

$$\frac{\rho^{(1)}}{2}\left\|\nabla\psi^{(1)}_t\right\|^2_{L_2(\Pi_k^{-\infty})} + \frac{\rho^{(1)}\alpha^2}{2}\left\|\psi^{(1)}_{x_3}\right\|^2_{L_2(\Pi_k^{-\infty})} = \int_{0}^{t}\int_{2\pi k}^{2\pi(k+1)}\rho^{(1)}(\psi^{(1)}_{\tau\tau x_3} + \alpha^2\psi^{(1)}_{x_3})\overline{\psi}^{(1)}_{\tau}\big|_{x_3=0}dx_1 d\tau. \tag{2.12}$$

Equation (1.4) is integrated with respect to spatial variables in $\Pi_k^{\varepsilon a,-}$, and, next, $\Pi_k^{\varepsilon a,-}$ is stretched into the semi-infinite strip $\Pi_k^{-\infty}$. Following the same line of reasoning as before, we obtain

$$\int_{2\pi k}^{2\pi(k+1)}\rho^{(1)}(\psi^{(1)}_{ttx_3} + \alpha^2\psi^{(1)}_{x_3})\big|_{x_3=0}dx = 0. \tag{2.13}$$

Similar results can be derived for the fluid in the upper half-space $\mathbb{R}_+^2$. Specifically, for the semi-infinite strip

$$\Pi_k^{+\infty} = \{(x_1, x_3) \in \overline{\mathbb{R}}_+^2 : 0 \le x_3 < +\infty,\ 2\pi k < x_1 < 2\pi(k+1)\},\quad k \in Z,$$

we have

$$\frac{\rho^{(2)}}{2}\left\|\nabla\psi^{(2)}_t\right\|^2_{L_2(\Pi_k^{+\infty})} + \frac{\rho^{(2)}\alpha^2}{2}\left\|\psi^{(2)}_{x_3}\right\|^2_{L_2(\Pi_k^{+\infty})} = \int_{0}^{t}\int_{2\pi k}^{2\pi(k+1)}\rho^{(2)}(-\psi^{(2)}_{\tau\tau x_3} - \alpha^2\psi^{(2)}_{x_3})\overline{\psi}^{(2)}_{\tau}\big|_{x_3=0}dx_1 d\tau \tag{2.12}'$$

and

$$\int_{2\pi k}^{2\pi(k+1)}\rho^{(2)}(-\psi^{(2)}_{ttx_3} - \alpha^2\psi^{(2)}_{x_3})\big|_{x_3=0}dx = 0. \tag{2.13}'$$

Since the integrals in the pairs of formulas (2.12), (2.12)' and (2.13), (2.13)' are evaluated over identical integration domains, adding (2.12) to (2.12)' and (2.13) to (2.13)' yields, in the strip

$$\Pi_k^\infty = \{(x_1, x_3) \in \mathbb{R}^2 : -\infty < x_3 < +\infty,\ 2\pi k < x_1 < 2\pi(k+1)\}, \quad k \in Z,$$

two equalities

$$\frac{\rho^{(1)}}{2}\left\|\nabla\psi_t^{(1)}\right\|^2_{L_2(\Pi_k^{-\infty})} + \frac{\rho^{(1)}\alpha^2}{2}\left\|\psi_{x_3}^{(1)}\right\|^2_{L_2(\Pi_k^{-\infty})} + \frac{\rho^{(2)}}{2}\left\|\nabla\psi_t^{(2)}\right\|^2_{L_2(\Pi_k^{+\infty})} + \frac{\rho_0^{(2)}\alpha^2}{2}\left\|\psi_{x_3}^{(2)}\right\|^2_{L_2(\Pi_k^{+\infty})}$$
$$= \int_0^t \int_{2\pi k}^{2\pi(k+1)} [\rho^{(1)}(\psi_{ttx_3}^{(1)} + \alpha^2\psi_{x_3}^{(1)})\overline{\psi}_\tau^{(1)} - \rho^{(2)}(\psi_{ttx_3}^{(2)} + \alpha^2\psi_{x_3}^{(2)})\overline{\psi}_\tau^{(2)}]\Big|_{x_3=0} dx_1 d\tau, \tag{2.12"}$$

$$\int_{2\pi k}^{2\pi(k+1)} [\rho^{(1)}(\psi_{ttx_3}^{(1)} + \alpha^2\psi_{x_3}^{(1)}) - \rho^{(2)}(-\psi_{ttx_3}^{(2)} - \alpha^2\psi_{x_3}^{(2)})]\Big|_{x_3=0} dx = 0. \tag{2.13"}$$

Assume that problem A has two different solutions $\psi_a(x, t)$ and $\psi_b(x, t)$ that correspond to two different functions $C_a(t)$ and $C_b(t)$ in the boundary conditions. Consider the difference between these solutions $\tilde{\psi}(x, t) = \psi_a(x, t) - \psi_b(x, t)$. Since the problem is linear, this function satisfies the homogeneous equations and the conditions in problem A, but, for $\tilde{\psi}(x, t)$, we have to retain only the difference between $C_a(t)$ and $C_b(t)$ on the right-hand side of (1.6c):

$$\frac{\partial^2}{\partial t^2}(\rho^{(1)}\tilde{\psi}_{x_3}^{(1)} - \rho^{(2)}\tilde{\psi}_{x_3}^{(2)}) + \alpha^2(\rho^{(1)}\tilde{\psi}_{x_3}^{(1)} - \rho^{(2)}\tilde{\psi}_{x_3}^{(2)}) - g(\rho^{(1)} - \rho^{(2)})\tilde{\psi}_{x_1x_1}^{(1,2)}\Big|_{x_3=0} = C_a(t) - C_b(t). \tag{1.6c}$$

The function $\tilde{\psi}(x, t)$ belongs to the class D as the difference of two solutions to problem A. Therefore, formulas (2.12)" and (2.13)" hold for it. By using boundary conditions (1.6c)', the integrand in (2.13)" is replaced with an equivalent expression

$$\int_{2\pi k}^{2\pi(k+1)} [g(\rho^{(1)} - \rho^{(2)})\tilde{\psi}_{x_1x_1}^{(1,2)} + C_a(t) - C_b(t)]\Big|_{x_3=0} dx = 0.$$

Integrating this and taking into account periodicity condition (1.8), we conclude that $C_a(t) \equiv C_b(t)$.

Consider the right-hand side of formula (2.12)" for $\tilde{\psi}(x, t)$. At the interface of the fluids, due to kinematic conditions (1.6b), we have

$$\overline{\tilde{\psi}}_\tau^{(1)}\Big|_{x_3=0} = \overline{\tilde{\psi}}_\tau^{(2)}\Big|_{x_3=0}.$$

By using condition (1.6c)' and taking into account that $C_a(t)$ and $C_b(t)$ coincides identically, the right-hand side of (2.12)" is rewritten as

$$\int_0^t \int_{2\pi k}^{2\pi(k+1)} [g(\rho^{(1)} - \rho^{(2)})\tilde{\psi}_{x_1x_1}^{(1,2)}]\overline{\tilde{\psi}}_\tau^{(1,2)}\Big|_{x_3=0} dx_1 d\tau.$$

While integrating the inner integral by parts, we take into account that the solution is a periodic function of x_1 and move the remaining expression to the left-hand side to obtain

$$\frac{\rho^{(1)}}{2}\left\|\nabla\psi_t^{(1)}\right\|^2_{L_2(\Pi_k^{-\infty})} + \frac{\rho^{(1)}\alpha^2}{2}\left\|\psi_{x_3}^{(1)}\right\|^2_{L_2(\Pi_k^{-\infty})} + \frac{\rho^{(2)}}{2}\left\|\nabla\psi_t^{(2)}\right\|^2_{L_2(\Pi_k^{+\infty})}$$
$$+ \frac{\rho^{(2)}\alpha^2}{2}\left\|\psi_{x_3}^{(2)}\right\|^2_{L_2(\Pi_k^{+\infty})} + \frac{g}{2}(\rho^{(1)} - \rho^{(2)})\int_0^t \frac{\partial}{\partial\tau}\int_{2\pi k}^{2\pi(k+1)} \tilde{\psi}_{x_1}^{(1)}\overline{\tilde{\psi}}_{x_1}^{(1)}\Big|_{x_3=0} dx_1 d\tau = 0.$$

Let us analyze the result. All five terms on the right-hand side are nonnegative. Therefore, their sum vanishes only if each of them is zero. This means that $\nabla\tilde{\psi}_t^{(1)} = 0$, $\nabla\tilde{\psi}_t^{(2)} = 0$, $\tilde{\psi}_{x_3}^{(1)} = 0$, and $\tilde{\psi}_{x_3}^{(2)} = 0$ in the $\Pi_k^{+\infty}$ and $\Pi_k^{-\infty}$, respectively, and $\tilde{\psi}_{x_1}^{(1)} = \tilde{\psi}_{x_1}^{(2)} = 0$ at the interface.

Since $\tilde{\psi}^{(1)}$ and $\tilde{\psi}^{(2)}$ are regular as $x_3 \longrightarrow -\infty$ and $x_3 \longrightarrow +\infty$, we find from (1.7) that $\tilde{\psi}(x, t) \equiv 0$ in the vertical strip Π_k^∞. Since the number k is arbitrary and the entire space can be treated as the union of an infinite countable number of such strips, we conclude that $\tilde{\psi}(x, t) \equiv 0$ everywhere in the solution domain of the problem for $\tilde{\psi}(x, t)$. Therefore, $\psi_a(x, t)$ coincides with $\psi_b(x, t)$ and problem A cannot have two different solutions in the class D.

The theorem is proved.

3. LONG-TIME ASYMPTOTICS OF THE SOLUTION

Although the resulting analytical solution (2.9), (2.9)' to problem A is of interest, it has a rather complex structure, which prevents the qualitative understanding of processes proceeding in the two-component fluid. For this reason, following [3–6], we study the asymptotics of the series $\psi^{(1)}(x, t)$, $\psi^{(2)}(x, t)$ as $t \longrightarrow \infty$, which make it possible to describe the wave pattern developing in the upper and lower fluids at long times.

For this purpose, the function characterizing the perturbation source in conditions (1.6c) is represented as the sum of two functions:

$$f(x_1 - ct)\eta(t) = f(x_1 - ct)[\eta(t) - 1] + f(x_1 - ct) \equiv f(x_1 - ct)\sigma(t) + f(x_1 - ct).$$

Since problem A is linear, it can be split into two similar initial–boundary value problems with different inhomogeneous right-hand sides in (1.6c). In the first problem, the moving source is described by the function $f(x_1 - ct)\sigma(t)$. This mathematical model is interpreted as follows: since $\sigma(t)$ vanishes after completing the transient period T, the external force does not affect the system starting at this time. The problem with a compactly supported boundary function of this type is referred to as problem $\mathbf{A}^\sigma$, and the index σ appears in all the components of the solution to problem A^σ.

In the dynamic boundary condition of the second problem, the function $\eta(t)$ associated with the transient regime is formally replaced by 1. This problem models a process with an infinitely long periodic external force affecting the system. It is referred to as problem $\mathbf{A}^1$, and all the components of its solution are denoted by superscript 1.

According to the superposition principle, the solution to problem A is the sum of the solutions to problems A^σ and A^1.

The study of the asymptotics of the solution is begun with problem A^σ. According to formula (2.8) for $\phi_n(t)$, the solution $\psi^{(1),\sigma}(x, t)$ to this problem in $\mathbb{R}^2_-$ is obtained from (2.9) by replacing the factor $\eta(t)$ in the integrand with $\sigma(t)$. Moreover, since $\sigma(\tau)$ in the integrands is compactly supported, the variable upper limit of integration t can be replaced with the constant T.

The modifications made in series (2.9) in order to transform it into the solution $\psi^{(1),\sigma}(x, t)$ to problem A^σ do not change the structure of its terms. Therefore, each harmonic $\psi_n^{(1),\sigma}(x, t)$ is again the sum of two terms: $\psi_n^{(1),\sigma}(x, t) \equiv u_n^{(1),\sigma}(x, t) + w_n^{(1),\sigma}(x, t)$.

Let us analyze the series consisting of the first components $u_n^{(1),\sigma}(x, t)$:

$$u^{(1),\sigma}(x, t) \equiv \sum_{n \neq 0} u_n^{(1),\sigma}(x, t) = -\sum_{n \neq 0} \frac{f_n \alpha}{\pi i g(\rho^{(1)} - \rho^{(2)})n^2} \exp(inx_1)$$
$$\times \int_{-\infty}^{+\infty} \frac{\mu \exp(i\mu|n|x_3)}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}} \left\{ \int_0^T \sin\left[\frac{\alpha\mu(t-\tau)}{\sqrt{\mu^2+1}}\right] \exp(-inc\tau)\sigma(\tau)d\tau \right\} d\mu. \tag{3.1}$$

To apply the standard methods used to estimate such series (see [1–6]), the variable of integration in the inner integral is changed to $\xi = \alpha\mu/\sqrt{\mu^2 + 1}$. The range of the old variable μ—the entire real axis

$(-\infty, \infty)$—is then mapped for ξ into the finite interval $[-\alpha, \alpha]$ of the real axis, and each harmonic $u_n^{(1),\sigma}(x, t)$ in (3.1) splits into two terms of the same structure:

$$u_n^{(1),\sigma}(x, t) = -\frac{f_n}{\pi i g(\rho^{(1)} - \rho^{(2)})n^2}\exp(-inx_1)$$
$$\times\left\{\int_{-\alpha}^{\alpha}\frac{\exp[-i|n|\xi(\alpha^2 - \xi^2)^{-1/2}x_3]}{1 + i\xi\sqrt{\alpha^2 - \xi^2}(Qg|n|)^{-1}}\sin(\xi t)\left[\int_0^T\cos(\xi\tau)\exp(-inc\tau)\sigma(\tau)d\tau\right]d\xi\right. \tag{3.2}$$
$$\left.-\int_{-\alpha}^{\alpha}\frac{\exp[-i|n|\xi(\alpha^2 - \xi^2)^{-1/2}x_3]}{1 + i\xi\sqrt{\alpha^2 - \xi^2}(Qg|n|)^{-1}}\cos(\xi t)\left[\int_0^T\sin(\xi\tau)\exp(-inc\tau)\sigma(\tau)d\tau\right]d\xi\right\}.$$

Since their structure is identical, it suffices to study the asymptotics of only the series formed by the first terms (3.2). Let the inner integral be denoted by

$$H_n(\xi) = \frac{f_n}{g(\rho^{(1)} - \rho^{(2)})n^2}\int_0^T\cos(\xi\tau)\exp(-inc\tau)\sigma(\tau)d\tau.$$

We have the obvious estimate

$$|H_n(\xi)| \le \frac{|f_n|}{g(\rho^{(1)} - \rho^{(2)})n^2}T\max_{0 \le t \le T}|\sigma(t)| \le \frac{|f_n|}{g(\rho^{(1)} - \rho^{(2)})n^2}T. \tag{3.3}$$

By using $H_n(\xi)$,the first term $u_n^{(1),\sigma,\sin}(x, t)$ can be written in the form

$$u_n^{(1),\sigma,\sin}(x, t) = -\frac{\exp(inx_1)}{\pi i}\int_{-\alpha}^{\alpha}\frac{\exp[-i|n|\xi(\alpha^2 - \xi^2)^{-1/2}x_3]}{1 + i\xi\sqrt{\alpha^2 - \xi^2}(Qg|n|)^{-1}}H_n(\xi)\sin(\xi t)d\xi \equiv \int_{-\alpha}^{\alpha}h_n(x, \xi)\sin(\xi t)d\xi.$$

Taking into account inequality (3.3), the integrand for $u_n^{(1),\sigma,\sin}(x, t)$ is estimated as follows:

$$|h_n(x, \xi)\sin(\xi t)| \le \frac{|f_n|T}{\pi g(\rho_0^{(1)} - \rho_0^{(2)})n^2}. \tag{3.4}$$

Taking into account the properties (2.7)" of the Fourier coefficients f_n of $f(x_1 - ct)$ (see (2.7)'), we conclude that the series $\Sigma_{n\neq 0}h_n(x, \xi) = h(x, \xi)$ is absolutely converging. Moreover, its sum $h(x, \xi)$ is integrable with respect to ξ on the interval $[-\alpha, \alpha]$ for fixed x_1 and x_3. Additionally, we can write the chain of inequalities

$$\sum_{n=1}^{N}\left|u_n^{(1),\sigma}(x, t)\right| \le 2\sum_{n=1}^{N}\left|u_n^{(1),\sigma,\sin}(x, t)\right| \le \frac{2T}{\pi g(\rho^{(1)} - \rho^{(2)})}\sum_{n=1}^{N}\frac{|f_n|}{n^2}, \quad N \in \mathbb{N}.$$

Combining this with (2.7)", we conclude the convergence of series (3.1).

The argument above implies that the order of integration and summation in (3.1) can be changed (see, e.g., [1]). For the series consisting of $u_n^{(1),\sigma,\sin}(x, t)$, the result of this operation is

$$\sum_{n\neq 0}u_n^{(1),\sigma,\sin}(x, t) = \int_{-\alpha}^{\alpha}h(x, \xi)\sin(\xi t)d\xi.$$

For $h(x, \xi)$ possessing the properties described above, the asymptotics of this integral as $t \longrightarrow \infty$ is well known (see, e.g., [8]), namely, for every $x \in \mathbb{R}_-^2$, the integral tends to zero. The same is true of the series consisting of the second terms $u_n^{(1),\sigma,\cos}(x, t)$. Thus, the contribution of the series $\psi^{(1),\sigma}(x, t)$ to the solution $\Sigma_{n\neq 0}u_n^{(1),\sigma}(x, t)$ of problem A^σ tends to zero at long times:

$$\lim_{t\to\infty}\sum_{n\neq 0}u_n^{(1),\sigma}(x, t) = 0. \tag{3.5}$$

Since the first terms in the members of series (2.9) and (2.9)' are identical, the same is true for the solution $\psi^{(2),\sigma}(x, t)$ to problem A in $\mathbb{R}^2_+$.

Now, we examine the series composed of the second terms $w_n^{(1),\sigma}(x, t)$ of the harmonics $\psi_n^{(1),\sigma}(x, t)$:

$$w^{(1),\sigma}(x, t) \equiv \sum_{n\neq 0} w_n^{(1),\sigma}(x, t) = \sum_{n\neq 0} \frac{2\sqrt{2} f_n Q^2 g}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^4}}$$
$$\times \exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right)\int_0^t \sin\left[\frac{1}{\sqrt{2}}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}(t-\tau)\right]\exp(-inc\tau)\sigma(\tau)d\tau. \quad (3.6)$$

The terms of series (3.6) contain only proper integrals and satisfy the estimate

$$\left|w_n^{\sigma}(x, t)\right| \le \frac{2\sqrt{2}|f_n|Q^2 g T \max\limits_{0\le t\le T}|\alpha(t)|}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^4}}\exp\left(\frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3\right).$$

Taking into account the properties of the Fourier coefficients f_n in (2.7)", we conclude that series (3.6) converges absolutely and uniformly to a continuous function $w^{(1),\sigma}(x, t)$, but the asymptotics for the sum of the series $w^{(1),\sigma}(x, t)$ as $t \longrightarrow \infty$ are not constructed.

Let us interpret the process corresponding to the series $w^{(1)}(x, t) = \Sigma_{n\neq 0} w_n^{(1)}(x, t)$, which consists of the second terms of the harmonics in (2.9); i.e., we interpret the solution $\psi^{(1)}(x, t)$ to the original problem A with dynamic boundary condition (1.6c) without assuming a compact support in time.

By analogy with the study of perturbations propagating in a two-layer stratified fluid (see [4]) and a uniformly rotating free-surface fluid (see [3]), we can expect that the wave motion of the fluid described by this series is localized near the interface of the layers in the two-component rotating fluid. According to [3], if there is no rotation in the model of a single homogeneous free-surface fluid, the harmonics of arising perturbations damp exponentially as they propagate deep into the fluid and a nontrivial process is observed only in a thin upper layer near the excited surface, a phenomenon known as surface waves.

In the model of a two-layer fluid, the effect of rotation is described by terms involving the double frequency α of fluid rotation. Therefore, a model of a two-layer unstratified fluid without rotation can be obtained by setting $\alpha = 0$ in the original problem. The solution to the corresponding initial–boundary value problem was constructed in [4] and has the form

$$\psi_{(0)}^{(1)}(x, t) = \sum_{n\neq 0} \frac{f_n\sqrt{Q}}{\sqrt{g(\rho^{(1)} - \rho^{(2)})|n|}^{3/2}}\exp(inx_1 + |n|x_3)\int_0^t \sin[\sqrt{Qg|n|}(t-\tau)]\exp(-inc\tau)\eta(\tau)d\tau. \quad (3.7)$$

The harmonics of series (3.7) depend exponentially on x_3 ($\exp(|n|x_3)$), which suggests that the lower component of the nonrotating two-layer fluid does not contain waves propagating at infinity. In other words, the process represents oscillations similar to surface waves, which damp quickly with depth.

Returning to problem A in $\mathbb{R}^2_-$, we discuss the series consisting of the second terms $\Sigma_{n\neq 0} w_n^{(1)}(x, t)$ in (2.9). It is easy to see that setting $\alpha = 0$ gives a series exactly coinciding with (3.7). This support the hypothesis that the series $\Sigma_{n\neq 0} w_n^{(1)}(x, t)$ describes wave motions of the fluid concentrated near the excited boundary between the fluids. Specifically, this conclusion is also true for the series $w^{(1),\sigma}(x, t) = \Sigma_{n\neq 0} w_n^{(1),\sigma}(x, t)$ in the solution to problem A^{σ}.

Let us construct asymptotics of the solution to problem A^1 with dynamic boundary conditions involving $f(x_1 - ct)$. According to formula (2.8) for $\phi_n(t)$, the terms of the series $\psi^{(1),1}(x, t) = \Sigma_{n\neq 0}\psi_n^{(1),1}(x, t)$ are obtained from (2.6) by replacing $\phi_n(\tau)$ with $f_n\exp(-inc\tau)$. This means that, in contrast to the solution of problem A^{σ}, the integrands in the integrals with respect to τ lack a multiplier involving the unspecified function $\eta(\tau)$. Therefore, both integrals in $\psi_n^{(1),1}(x, t)$ can be directly evaluated.

Each harmonic $\psi_n^{(1),1}(x, t)$ of the solution to problem A^1 is decomposed into the sum of two terms: $\psi_n^{(1),1}(x, t) \equiv u_n^{(1),1}(x, t) + w_n^{(1),1}(x, t)$. The study of the asymptotics starts with the series of first terms:

$$u_n^{(1),1}(x, t) = -\frac{f_n\alpha}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\exp(-inx_1)\int_{-\infty}^{+\infty}\frac{\mu\exp(i\mu|n|x_3)}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}}$$
$$\times\left\{\int_0^t \sin\frac{\alpha\mu(t-\tau)}{\sqrt{\mu^2 + 1}}\exp(-inc\tau)d\tau\right\}d\mu.$$

After evaluating the inner integral, each harmonic $u_n^{(1),1}(x, t)$ splits into the sum of three improper integrals:

$$u_n^{(1),1}(x, t) = -\frac{f_n\alpha^2}{\pi ig(\rho^{(1)} - \rho^{(2)})n^2(\alpha^2 - n^2c^2)}\exp(inx_1 - inct)$$
$$\times\int_{-\infty}^{+\infty}\frac{\mu\exp(i\mu|n|x_3)d\mu}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1][\mu^2 - n^2c^2(\alpha^2 - n^2c^2)^{-1}]}$$
$$-\frac{f_n\alpha\exp(inx_1)}{2\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\int_{-\infty}^{+\infty}\frac{\mu\exp(i\mu|n|x_3)d\mu}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}}\frac{\exp(i\alpha\mu(\mu^2 + 1)^{-1/2}t)}{i\alpha\mu(\mu^2 + 1)^{-1/2} + nc}d\mu \quad (3.8)$$
$$-\frac{f_n\alpha\exp(inx_1)}{2\pi ig(\rho^{(1)} - \rho^{(2)})n^2}\int_{-\infty}^{+\infty}\frac{\mu\exp(i\mu|n|x_3)d\mu}{[\mu^2 + i\alpha^2(Qg|n|)^{-1}\mu + 1]\sqrt{\mu^2 + 1}}\frac{\exp(-i\alpha\mu(\mu^2 + 1)^{-1/2}t)}{i\alpha\mu(\mu^2 + 1)^{-1/2} - nc}d\mu$$
$$\equiv K_n^{(1)}(x, t) + K_n^{(2)}(x, t) + K_n^{(3)}(x, t).$$

Let us analyze (3.8), excluding the cases when $|nc| = \alpha$ for integers n. The first integral $K_n^{(1)}(x, t)$ has the simplest structure, since the singular points of its integrand (four simple poles) are easy to determine. The point μ^- in (2.5) always lies in the lower half-plane $\mathrm{Im}\,\mu < 0$. Two singular points corresponding to the roots of the second square brackets in the denominator $\mu^2 - n^2c^2(\alpha^2 - n^2c^2)^{-1}$ lie on the real axis if $0 < |nc| < \alpha$. Then the integral in $K_n^{(1)}(x, t)$ is calculated by summing the residues at both poles. For $|nc| > \alpha$, the residue only at one of the purely imaginary roots of the expression $\mu_0 = -i|n|c(n^2c^2 - \alpha^2)^{-1/2}$ lying in the domain $\mathrm{Im}\,\mu < 0$ contributes to the integral in $K_n^{(1)}(x, t)$. First, assume that this root does not coincide with μ^-. As a result, we obtain two values of $K_n^{(1)}(x, t)$:

$$K_n^{(1)}(x, t) = \frac{2f_ngQ^2(\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2)}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}[(\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2)(\alpha^2 - n^2c^2) + 2Q^2g^2n^2]}$$
$$\times\exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3 - inct\right) \quad (3.9a)$$
$$+\frac{f_nQ}{(\rho^{(1)} - \rho^{(2)})(gQ - c\sqrt{n^2c^2 - \alpha^2})n^2}\exp\left(inx_1 + \frac{n^2c}{\sqrt{n^2c^2 - \alpha^2}}x_3 - inct\right), \quad |nc| > \alpha,$$

or

$$K_n^{(1)}(x,t) = \frac{2f_n g Q^2(\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2)}{(\rho^{(1)} - \rho^{(2)})\sqrt{\alpha^4 + 4Q^2g^2n^2}[(\sqrt{\alpha^4 + 4Q^2g^2n^2 + \alpha^2})(\alpha^2 - n^2c^2) + 2Q^2g^2n^2]}$$
$$\times \exp\left(inx_1 + \frac{\sqrt{\alpha^4 + 4Q^2g^2n^2} + \alpha^2}{2Qg}x_3 - inct\right)$$
$$+ \frac{f_nQ}{2(\rho^{(1)} - \rho^{(2)})(Qg + inc|n|^{-1}\sqrt{\alpha^2 - n^2c^2})n^2}\exp\left(inx_1 + i\frac{|n|nc}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right)$$
$$+ \frac{f_nQ}{2(\rho^{(1)} - \rho^{(2)})(Qg - inc|n|^{-1}\sqrt{\alpha^2 - n^2c^2})n^2}\exp\left(inx_1 - i\frac{|n|nc}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right), \quad 0 < |nc| < \alpha. \tag{3.9b}$$

Since the structures of $K_n^{(2)}(x,t)$ and $K_n^{(3)}(x,t)$ in (3.8) are similar, we study their asymptotics simultaneously. First, we change the variable of integration to $\xi = \alpha\mu/\sqrt{\mu^2 + 1}$ (as was done in the study of problem A^σ). As a result, $K_n^{(2)}(x,t)$ and $K_n^{(3)}(x,t)$ become

$$K_n^{(2,3)}(x,t) = -\frac{f_n}{2\pi i g(\rho^{(1)} - \rho^{(2)})n^2}\exp(inx_1)\int_{-\alpha}^{\alpha}\frac{\exp[-i|n|(\xi/\sqrt{\alpha^2 - \xi^2})x_3]}{1 + i\xi\sqrt{\alpha^2 - \xi^2}(Qg|n|)^{-1}}\frac{\exp(\pm i\xi t)}{\xi \pm nc}d\xi,$$

where the upper sign in $\pm$ corresponds to $K_n^{(2)}(x,t)$, and the lower sign, to $K_n^{(3)}(x,t)$.

Note that the integrals with respect to the new variable ξ have the same structure as their analogues in [3–6]. For this reason, the asymptotics of the series $\Sigma_{n\neq 0}K_n^{(2,3)}(x,t)$ are analyzed following the scheme used in [3–6].

For the common part of the integrand in two integrals, we introduce the notation

$$q(x,\xi) = \frac{f_n\exp(inx_1)}{2g(\rho^{(1)} - \rho^{(2)})n^2}\frac{\exp[-i|n|(\xi/\sqrt{\alpha^2 - \xi^2})x_3]}{1 + i\xi\sqrt{\alpha^2 - \xi^2}(Qg|n|)^{-1}}.$$

This integrand is continuous in the interval $(-\alpha, \alpha)$. By taking into account the expressions for $K_n^{(2)}(x,t)$ and $K_n^{(3)}(x,t)$, it is rewritten in the compact form

$$K_n^{(2,3)}(x,t) = \frac{1}{\pi i}\int_{-\alpha}^{\alpha}q(x,\xi)\frac{\exp(\pm i\xi t)}{\xi \pm nc}d\xi. \tag{3.10}$$

Now, it is clear that whether or not the integrand has singular points in the integration domain $[-\alpha, \alpha]$ is determined by the relation between $|nc|$ and α. Note that the same condition separates two values of $K_n^{(1)}(x,t)$.

For harmonics with $|nc| > \alpha$, the expressions $\xi \pm nc$ do not vanish on the interval $[-\alpha, \alpha]$ and the integrands in (3.10) satisfy the estimates

$$\left|q(x,\xi)\frac{\exp(\pm i\xi t)}{\xi \pm nc}\right| \le \frac{|f_n|}{2g(\rho_0^{(1)} - \rho_0^{(2)})n^2(|nc| - \alpha)}.$$

For this solution component, using the same argument as in analyzing the asymptotics of $\Sigma_{n\neq 0}u_n^{(1),\sigma}(x, t)$ in problem A^σ, we see that

$$\lim_{t\to\infty}\sum_{|nc|>\alpha} K_n^{(2,3)}(x, t) = 0. \tag{3.11}$$

Therefore, the contribution to the asymptotics, as $t \longrightarrow \infty$, of the solution $\psi^{(1),1}(x, t)$ to problem A^1 with a constantly acting source of perturbations from small fluid motions described by $\Sigma_{n\neq 0}K_n^{(2,3)}(x, t)$ is zero.

For a finite number of remaining harmonics with small indices such that $|nc| < \alpha$, the roots of $\xi \pm nc$ (i.e., the points $\xi_0 = \mp nc$) lie in the integration interval $[-\alpha, \alpha]$ of $K_n^{(2)}(x, t)$, $K_n^{(3)}(x, t)$ and $\xi_0 = -nc$. Therefore, the singular points $\xi_0 = nc$ and $\xi_0 = -nc$, respectively, appear in the integrands.

For n at which $K_n^{(2)}(x, t)$ and $K_n^{(3)}(x, t)$ become improper integrals, their asymptotics are constructed using Lebesgue's lemma (see, e.g., [1]), according to which integrals of the form

$$\frac{1}{\pi i}\int_{-\alpha}^{\alpha} q(x, \xi)\frac{\exp(i\xi t)}{\xi - \xi_0}d\xi \equiv B(t, \xi_0)$$

satisfy the limiting relation $\lim\limits_{t\to\infty}\exp(-i\xi_0 t)B(t, \xi_0) = q(x, \xi_0)$.

The lemma yields the following limiting values of $K_n^{(2,3)}(x, t)$ for $|nc| < \alpha$:

$$K_n^{(2)}(x, t) = \begin{cases} -\dfrac{f_n Q}{2(\rho^{(1)} - \rho^{(2)})n^2(Qg + ic\sqrt{\alpha^2 - n^2c^2})}\exp\left(inx_1 + i\dfrac{n^2c}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right) + g(x, t), \\ -\alpha < nc < 0, \\ -\dfrac{f_n Q}{2(\rho^{(1)} - \rho^{(2)})n^2(Qg - ic\sqrt{\alpha^2 - n^2c^2})}\exp\left(inx_1 - i\dfrac{n^2c}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right) + g(x, t), \\ 0 < nc < \alpha, \end{cases} \tag{3.12}$$

and

$$K_n^{(3)}(x, t) = \begin{cases} \dfrac{f_n Q}{2(\rho^{(1)} - \rho^{(2)})n^2(Qg - ic\sqrt{\alpha^2 - n^2c^2})}\exp\left(inx_1 - i\dfrac{n^2c}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right) + g(x, t), \\ -\alpha < nc < 0, \\ \dfrac{f_n Q}{2(\rho^{(1)} - \rho^{(2)})n^2(Qg + ic\sqrt{\alpha^2 - n^2c^2})}\exp\left(inx_1 + i\dfrac{n^2c}{\sqrt{\alpha^2 - n^2c^2}}x_3 - inct\right) + g(x, t), \\ 0 < nc < \alpha, \end{cases} \tag{3.13}$$

where $g(x, t)$ tends to zero as $t \longrightarrow \infty$.

Note that both components of these expressions are present in formula (3.9b) for the precise value of $K_n^{(1)}(x, t)$ in the case of $|nc| < \alpha$.

Now, we discuss the series consisting of the second terms of the harmonics $\psi_n^{(1),1}(x, t)$ in the solution to problem A^1. Evaluating the proper integrals involved gives explicit expressions for $w_n^{(1),1}(x, t)$:

$$w_n^{(1),1}(x,t) = \frac{4f_nQ^2g}{(\rho^{(1)}-\rho^{(2)})\sqrt{\alpha^4+4Q^2g^2n^2}(\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2-2n^2c^2)}$$
$$\times \exp\left(inx_1+\frac{\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2}{2Qg}x_3-inct\right)$$
$$-\frac{f_n\sqrt{2}Qg}{(\rho^{(1)}-\rho^{(2)})\sqrt{\alpha^4+4Q^2g^2n^2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}}\exp\left(inx_1+\frac{\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2}{2Qg}x_3\right) \quad (3.14)$$
$$\times\left[\frac{\exp\left(i2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}t\right)}{2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}+nc}+\frac{\exp\left(-i2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}t\right)}{2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}-nc}\right].$$

Note that the exponent in the first term in $w_n^{(1),1}(x, t)$ coincides with that of the first component of $K_n^{(1)}(x, t)$ in (3.9). Moreover, this term is involved in the formulas for $K_n^{(1)}(x, t)$ for both $|nc| > \alpha$ and $|nc| < \alpha$. The case of $|nc| = \alpha$ leads to resonance and has to be analyzed separately. By using simple algebra, the coefficient multiplying the exponential in $K_n^{(1)}(x, t)$ can be reduced to a form that coincides (up to the sign) with the coefficient of $w_n^{(1),1}(x, t)$.

For three cases of relations between α and nc and the sign of n, summing the contributions to the asymptotics made by the solutions to problems A^σ and A^1, (see (3.5), (3.6), (3.9), and (3.11)–(3.14)) and collecting like terms, we obtain the basic formula for the asymptotics of the solution $\psi^{(1)}(x, t)$ to problem A in $\mathbb{R}^2_-$ as $t \longrightarrow \infty$:

$$\psi^{(1)}(x,t) = -\sum_{n\neq 0}\frac{f_n\sqrt{2}Qg}{(\rho^{(1)}-\rho^{(2)})\sqrt{\alpha^4+4Q^2g^2n^2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}}\exp\left(inx_1+\frac{\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2}{2Qg}x_3\right)$$
$$\times\left[\frac{\exp\left(i2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}t\right)}{2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}+nc}+\frac{\exp\left(-i2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}t\right)}{2^{-1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}-nc}\right]$$
$$+\sum_{|nc|>\alpha}\frac{f_nQ}{(\rho^{(1)}-\rho^{(2)})n^2(Qg-c\sqrt{n^2c^2-\alpha^2})}\exp\left(inx_1+i\frac{n^2c}{\sqrt{n^2c^2-\alpha^2}}x_3-inct\right)$$
$$+\sum_{-\alpha<nc<0}\frac{f_nQ}{(\rho^{(1)}-\rho^{(2)})n^2(Qg-ic\sqrt{\alpha^2-n^2c^2})}\exp\left(inx_1-i\frac{n^2c}{\sqrt{\alpha^2-n^2c^2}}x_3-inct\right) \quad (3.15)$$
$$+\sum_{0<nc<\alpha}\frac{f_nQ}{(\rho^{(1)}-\rho^{(2)})n^2(Qg+ic\sqrt{\alpha^2-n^2c^2})}\exp\left(inx_1+i\frac{n^2c}{\sqrt{\alpha^2-n^2c^2}}x_3-inct\right)$$
$$+\sum_{n\neq 0}\frac{2\sqrt{2}f_nQ^2g}{(\rho^{(1)}-\rho^{(2)})\sqrt{\alpha^4+4Q^2g^2n^2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^4}}\exp\left(inx_1+\frac{\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2}{2Qg}x_3\right)$$

$$\times \int_0^t \sin\left[\frac{1}{\sqrt{2}}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}(t-\tau)\right]\exp(-inc\tau)\sigma(\tau)d\tau, \quad x<0.$$

Let us analyze the form of the asymptotics for the second part of the solution to problem A, i.e., for $\psi^{(2)}(x, t)$ defined in $\mathbb{R}^2_+$. The general structure of the expression remains the same, but the following modifications appear, which are noted in Remark 1. In the terms of the first series, the plus sign before the second term in the exponent is replaced with a minus. Indeed, only this form of the series satisfies the regularity condition as $x_3 \longrightarrow +\infty$. Similar changes in the sign of the coefficients of x_3 are made in the exponents in the terms of the second and last series in (3.15). In the partial sums of the third and fourth terms of the asymptotics of $\psi^{(1)}(x, t)$, the signs are changed not only before the terms with x_3 in the exponent but also in the second brackets in the denominators of the coefficients multiplying them.

4. DISCUSSION OF THE RESULTS

Below, we interpret the constructed asymptotics of the solution to problem A with a special emphasis on the processes observed in $\mathbb{R}^2_-$. Note that formula (3.15) is the sum of three series (first, second, and fifth terms) and two partial sums (the third and fourth terms). The terms of the first four components in (3.15) depend on t as exp(-int), where the subscript n in ω_n is the index of a series term and ω_n is naturally interpreted as the frequency of the nth harmonic in the solution. Note that ω_n has no superscript, since the asymptotics $\psi^{(2)}(x, t)$ of the solution to problem A in the upper fluid layer involves the same coefficients ω_n in the exponents before the terms with t; i.e., the set of frequencies is the same for the solutions to problem A in $\mathbb{R}^2_-$ and $\mathbb{R}^2_+$.

The analysis of the asymptotics of the solution to problem A begins with the first and fifth series in (3.15). Recall that they are formed of the series with the terms $w_n^{(1),1}(x, t)$ (see (3.14)) and $w_n^{(1),\sigma}(x, t)$ (see (3.6)). It was shown above that these series of the second terms of the solutions to problems A^1 and A^σ describe oscillations localized near the interface of the fluids.

Consider the first series in (3.15) separately. The oscillation frequencies of its terms are given by a complicated function of n, namely, $\omega_n = 2^{-1/2}\sqrt{(\alpha^4 + 4Q^2g^2n^2)^{1/2} + \alpha^2}$, which differs from the set of frequencies of the perturbation source $\omega_n^0 = nc$. Note that the denominators of the series coefficients also involve ω_n and may vanish for some integers $n = n_0$, in which case formula (3.15) formally makes no sense. The asymptotics of such harmonics have to be calculated separately, i.e., according to a scheme other than that used above.

The harmonics with such indices satisfy

$$\omega_{n_0} \equiv |n_0 c| = \sqrt{\left(\frac{Qg}{c}\right)^2 + \alpha^2}. \tag{4.1}$$

For these n_0, the second terms in the solution to problem A^1 are

$$w_{n_0}^{(1)}(x, t) = \frac{f_n Q^2 g}{(\rho^{(1)} - \rho^{(2)})(2n_0^2c^2 - \alpha^2)|n_0|c}\int_0^t \sin[n_0c(t-\tau)]\exp(-in_0c\tau)d\tau.$$

Evaluating the integrals involved, we have

$$w_{n_0}^{(1)}(x, t) = \frac{f_n Q^2 g}{(\rho^{(1)} - \rho^{(2)})(2n_0^2c^2 - \alpha^2)|n_0|c}\exp\left(inx_1 + \frac{n_0^2c^2}{Qg}x_3 + i\frac{\pi}{2}\right)\left[t\exp(-in_0ct) - \frac{\sin(n_0ct)}{n_0c}\right].$$

If $\rho^{(1)}$, $\rho^{(2)}$, α, and c are such that (4.1) holds for some integer n_0 and if the spectrum of $f(z)$ contains the frequency $\omega_{n_0} = n_0c$, i.e., the Fourier coefficient $f_{n_0} \neq 0$ is nonzero, then the solution to problem A contains a harmonic whose amplitude increases linearly with time. It is called a resonance harmonic.

Note that this term in the first series in (3.15) depends on x_3 in $\mathbb{R}^2_-$ as $\exp[n_0^2c^2(Qg)^{-1}x_3]$; i.e., its absolute value decreases rapidly as the distance from the interface increases. Note that the same resonance har-

monic is present in $\mathbb{R}^2_+$, and its amplitude decreases away from the source according to the same law $\exp[-n_0^2c^2(Qg)^{-1}x_3]$. Therefore, in a two-component fluid, the resonance phenomenon corresponding to the first series in (3.15) is observed only near the interface.

Note that, in the mathematical model of perturbations propagating in a single rotating or stratified free-surface fluid occupying the lower half-space (see [3, 5]), the analogue of a resonance harmonic is a surface wave. Moreover, since $Q < 1$, the resonance frequency in the lower component of a two-layer fluid is higher than that in a one-layer rotating fluid with all other things being equal.

A comparison of the wave pattern with that observed in a stratified two-component fluid [4] reveals resemblance of the resonance effects localized near the interface, but the resonance harmonics in a stratified fluid are determined by the buoyancy frequency of each fluid.

Finally, let us compare the resonance phenomenon with that observed in a nonrotating two-layer fluid. As was mentioned earlier, the resonant frequency is determined by formally setting α equal to zero. Moreover, according to formula (4.1), the frequency of the resonance harmonic for a stationary fluid is lower than that for a uniformly rotating system.

Note that no unbounded amplitude growth as $t \longrightarrow \infty$ is observed in reality. The presence of such series terms in the asymptotics of the solution suggests that the mathematical model of the phenomenon in question is incomplete. Specifically, the fluid viscosity, which smoothes the resonance phenomena, is ignored in the model.

Consider the last series in (3.15). It is easy to see that it is formed in the course of solving problem A^σ. Therefore, the integrand of its terms contains the unspecified function $\sigma(\tau) = \eta(t) - 1$, which prevents the evaluation of the integrals. However, by using trigonometric formulas, the terms of the last series $w_n^{(1),\sigma}(x, t)$ can be transformed so that their dependence on t is explicitly given:

$$w_n^{(1),\sigma}(x,t) = \frac{2\sqrt{2}f_nQ^2g}{(\rho_0^{(1)}-\rho_0^{(2)})\sqrt{\alpha^4+4Q^2g^2n^2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}}$$

$$\times \exp\left(inx_1 + \frac{\sqrt{\alpha^4+4Q^2g^2n^2}+\alpha^2}{2Qg}x_3\right)\left[\sin\left(\frac{1}{\sqrt{2}}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}\,t\right)\right.$$

$$\times \int_0^T \cos\left(\frac{1}{\sqrt{2}}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}\,\tau\right)\exp(-inc\tau)\sigma(\tau)d\tau$$

$$\left. - \cos\left(\frac{1}{\sqrt{2}}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}\,t\right)\int_0^T \sin\left(\frac{1}{\sqrt{2}}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}\,\tau\right)\exp(-inc\tau)\sigma(\tau)d\tau\right].$$

The values of the remaining integrals are independent of t. As before, the coefficients of t in the arguments of the trigonometric functions multiplying the integrals are interpreted as the frequencies of the nth harmonics. Moreover, their values $\omega_n = 2^{1/2}\sqrt{(\alpha^4+4Q^2g^2n^2)^{1/2}+\alpha^2}$ are the same as for the first series in (3.15); i.e., the frequency spectra of the first and last series of the asymptotics in (3.15) coincide.

Now, we discuss three middle terms of the asymptotics of the solution to problem A in $\mathbb{R}^2_-$. As was noted above, they are formed from the first terms $u_n^{(1),1}(x, t)$ of the harmonics $\psi_n^{(1),1}(x, t)$. Recall the structure of the solution to the problem of perturbations propagating in a homogeneous uniformly rotating fluid occupying the upper half-space with a solid bottom along which a source of perturbations moves [6]. Since there is no free surface, the solution to the problem in [6] consists of various types of internal gyroscopic waves. Comparing the general form of the solution $\psi^{(1)}(x, t)$ to the given problem in (2.9) with its analogue in [6], we see that the solution to the problem for a single rotating fluid consists only of series similar to $u^{(1),1}(x, t)$ and $u^{(1),\sigma}(x, t)$ with no harmonics of the form $w_n^{(1),1}(x, t)$ and $w_n^{(1),\sigma}(x, t)$. This supports the interpretation of two extreme series in (3.15) as analogues of surface waves and suggests that the contribution of three middle terms to the asymptotics of the solution to problem A corresponds to purely internal gyroscopic waves in the fluid.

A characteristic feature of the harmonics describing internal waves in a half- space is that they are divided into two classes. The first contains an infinite number of harmonics forming a series whose terms decrease exponentially away from the boundary with a source of perturbations moving along it. This is explained by the fact that the expressions with x_3 involved in the exponent in such harmonics take real negative values. This part of the asymptotics of the solution to problem A in $\mathbb{R}^2_-$ is represented by the series from the second term in (3.15). Note that the denominator of its terms depends on n in such a manner that a certain combination of the parameters for some integer $n = n_0$ can lead to this expression vanishing, in which case formula (3.15) becomes invalid. Importantly, these terms of the series have the same index $n = n_0$ as that for which (4.1) becomes an equality; i.e., they correspond to resonance harmonics among the surface waves.

To avoid the contradiction that the asymptotics of the solution cannot be determined by formula (3.15) at such $n = n_0$, we note that the second series in (3.15) is related to the contribution made to the integral $K_n^{(1)}(x, t)$ by the residue of its integrand at the point $\mu_0 = -i|n_0|c(n_0^2c^2 - \alpha^2)^{-1/2}$, which is a simple pole. It can be shown that, if $|nc| > \alpha$ and (4.1) holds for some integer n_0, the roots μ_0 and μ^- of the denominator in the integrand of $K_n^{(1)}(x, t)$ in (3.8) coincides. Therefore, the integrand has one second-order pole rather than two simple poles in the half-plane $\operatorname{Im}\mu < 0$. The integral evaluated with the help of residue theory turns out to be zero.

According to mathematical computations, the contribution to the resonance effect for the harmonic $\psi_{n_0}^{(1),1}(x, t) = u_{n_0}^{(1),1}(x, t) + w_{n_0}^{(1),1}(x, t)$ is determined not by the first term $u_{n_0}^{(1),1}(x, t)$, which describes internal gyroscopic waves in the lower half-space, but rather by the second component $w_{n_0}^{(1),1}(x, t)$, which corresponds to a process localized near the interface of the fluids. A similar interpretation of the resonance harmonic was given for the interface of a two-layer stratified fluid [4] and the free surface of a one-layer rotating fluid [3] excited by a moving source.

The second class of harmonics describing internal fluid motions consists of a finite number of terms with purely imaginary coefficients of x_3 in the exponent. Therefore, their modulus does not vary with x_3 increasing in absolute value. They describe waves that originate at the interface and propagate without damping deep into the fluid in both half-spaces.

In (3.15) they correspond to the partial sums in the third and fourth terms. The frequencies of the corresponding harmonics lie in the interval $0 < |nc| < \alpha$, which coincides with the range of this component of the asymptotics for a single homogeneous semi-infinite rotating fluid occupying only the upper half-space (see [6]) or only the lower half-space (see [3]). Therefore, the number of such harmonics is determined by the angular velocity of the fluid $\alpha/2$ and by the speed of the source c. This agrees with the fact that, for a two-component fluid with different densities, the frequency range of this part of the solution is the same in the upper and lower half-spaces.

Recall that the index n_0 of a possible resonance harmonic lies beyond the frequency interval $0 < |nc| < \alpha$; i.e., the resonance harmonic corresponds to that part of the asymptotics that describes fluid oscillations rapidly damping with increasing absolute value of x_3.

Thus, we have given a physical interpretation for each of the five terms of the asymptotics of the solution to problem A. However, the explicit analytical solution (2.9) involves a few additional series that do not contribute to asymptotic formula (3.15), for example, $\Sigma_{n\neq 0}u_n^{(1),\sigma}(x, t)$, which appears in the solution of problem A^σ. This series describes perturbations propagating deep into the lower fluid layer from a source affecting the system over the finite time interval $[0, T]$.

Let us discuss in more detail the harmonics $u_n^{(1),1}(x, t)$, which correspond to waves propagating in $\mathbb{R}^2_-$. Consider a harmonic with a positive index n satisfying $0 < nc < \alpha$:

$$u_n^{(1),1}(x, t) = \frac{f_n Q \exp[inx_1 + i(n^2c/\sqrt{\alpha^2 - n^2c^2})x_3 - inct]}{(\rho^{(1)} - \rho^{(2)})n^2(Qg + ic\sqrt{\alpha^2 - n^2c^2})}.$$

More briefly, it can be written as $u_n^{(1),1}(x, t) = A_n \exp\{i[(\mathbf{k}^{(1)}, \mathbf{x}) - \omega_n t]\}$, where $\mathbf{k}^{(1)}$ is the two-dimensional wave vector

$$\mathbf{k}^{(1)} = \{k_1, k_3\} = \left\{ n, \frac{n^2 c}{\sqrt{\alpha^2 - n^2 c^2}} \right\}. \tag{4.2}$$

Note that the wave vector given by (4.2) coincides with that in the problem of perturbations propagating in a one-component homogeneous free-surface rotating fluid occupying the lower half-space (see [3]) with only difference in the amplitude of A_n, which now depends on the densities of both fluids.

Typically, the last component of the vector $\mathbf{k}^{(1)}$, i.e., the expression $k_3 = n^2 c/\sqrt{\alpha^2 - n^2 c^2}$ is positive; i.e., the wave vector is oriented so that, from this point of view, the gyroscopic waves propagate from infinity to the interface of the fluids. It is well known that the wave vector determines the direction of the phase velocity of a wave, while the direction of its propagation is determined by the group velocity rather than by the phase velocity (see [1, 2, 9]). The stream function $\psi(x, t)$ was introduced in (1.3) via the relation to the group velocity. Therefore, the group velocity for the chosen harmonic has the form

$$\mathbf{v}_{\mathrm{gr}}^{(1)} = \{v_1, v_3\} = \frac{f_n Q \exp[inx_1 + i(n^2 c/\sqrt{\alpha^2 - n^2 c^2})x_3 - inct + i\pi/2]}{(\rho^{(1)} - \rho^{(2)}) n^2 (Qg + ic\sqrt{\alpha^2 - n^2 c^2})} \left\{ \frac{n^2 c}{\sqrt{\alpha^2 - n^2 c^2}}, -n \right\}, \tag{4.3}$$

The differentiation of the asymptotics of the solution $\psi^{(1)}(x, t)$required for deriving formula (4.3) is valid.

Note that the two-dimensional group velocity is orthogonal to the phase velocity, which is aligned with $\mathbf{k}^{(1)}$. Moreover, the projection of the group velocity onto the Ox_3 axis (i.e., its last component) has a sign opposite to that of the last component of the phase velocity. Therefore, the gyroscopic waves in $\mathbb{R}_-^2$ diverge from the perturbation source, which is a wave traveling along the surface of the fluid. The same pattern, which agrees with the laws of physics, is observed for internal gyroscopic waves developing in $\mathbb{R}_+^2$.

A similar mutual orientation of the phase and group velocities of internal waves is a characteristic feature for perturbations originating at the interface of half-spaces filled with rotating or stratified fluids (see [1–6]).

To conclude the study of the problem, we note that the wave pattern developing in a two-layer rotating fluid at long times is generally similar to that observed in a two-layer stratified fluid excited by a similar source moving along the interface [4]. In a stratified fluid, however, the wave patterns in the upper and lower layers differ to a higher degree. This is explained by the fact that the solutions for $\mathbb{R}_-^2$ and $\mathbb{R}_+^2$ in [4] satisfy different equations, while the stream function equation in the model of a rotating fluid is the same in the entire space.

As a further step, it is of interest to analyze perturbations propagating in a two-layer stratified fluid that rotates uniformly about an axis aligned with the direction of stratification.

ACKNOWLEDGMENTS

The author is grateful to Professor A.G. Sveshnikov for helpful advice and remarks concerning this paper.

This work was supported by the Russian Foundation for Basic Research, project no. 08-01-0037a.

REFERENCES

1. S. A. Gabov and A. G. Sveshnikov, *Linear Problems in the Theory of Unsteady Internal Waves* (Nauka, Moscow, 1990) [in Russian].
2. S. A. Gabov, *New Problems in the Mathematical Theory of Waves* (Nauka, Moscow, 1998) [in Russian].
3. L. V. Perova, "On Oscillations of a Semi-Infinite Rotating Liquid with Its Free Surface Excited by Moving Sources," Zh. Vychisl. Mat. Mat. Fiz. **46**, 955–970 (2006) [Comput. Math. Math. Phys. **46**, 891–906 (2006)].

4. L. V. Perova, "Propagation of Perturbations in a Two-Layer Stratified Fluid with an Interface Excited by Moving Sources," Zh. Vychisl. Mat. Mat. Fiz. **48**, 1062–1086 (2008) [Comput. Math. Math. Phys. **48**, 1001–1023 (2008)].
5. L. V. Perova, "Oscillations of Semi-Infinite Stratified Fluid with Its Free Surface Excited by Moving Sources," Zh. Vychisl. Mat. Mat. Fiz. **45**, 1107–1124 (2005) [Comput. Math. Math. Phys. **45**, 1068–1085 (2005)].
6. L. V. Perova, Yu. D. Pletner, A. G. Sveshnikov, and M. O. Korpusov, "Long-Time Asymptotics of an Initial–Boundary Value Problem for the Two-Dimensional Sobolev Equation," Differ. Uravn, **35**, 1421–1425 (1999).
7. A. G. Sveshnikov and A. N. Tikhonov, *The Theory of Functions of a Complex Variable* (Nauka, Moscow, 1991) [in Russian].
8. M. F. Fedoryuk, *Asymptotics: Integrals and Series* (Nauka, Moscow, 1987) [in Russian].
9. S. A. Gabov and A. G. Sveshnikov, *Problems in Stratified Fluid Dynamics* (Nauka, Moscow, 1986) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1197–1211. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © V.A. Titarev, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1255–1270.

Numerical Method for Computing Two-Dimensional Unsteady Rarefied Gas Flows in Arbitrarily Shaped Domains

V. A. Titarev

Cranfield University, Cranfield, UK, MK43 0AL

e-mail: v.a.titarev@cranfield.ac.uk, titarev@mail.ru

Received November 5, 2008

Abstract—A high-order accurate method for analyzing two-dimensional rarefied gas flows is proposed on the basis of a nonstationary kinetic equation in arbitrarily shaped regions. The basic idea behind the method is the use of hybrid unstructured meshes in physical space. Special attention is given to the performance of the method in a wide range of Knudsen numbers and to accurate approximations of boundary conditions. Examples calculations are provided.

DOI: 10.1134/S0965542509070112

Key words: rarefied gas, *S*-model, unstructured mesh, kinetic equation, TVD scheme.

INTRODUCTION

The numerical solution of the Boltzmann kinetic equation with an exact or model collision integral is a major technique for analyzing unsteady rarefied gas flows in a wide range of Knudsen and Mach numbers. To this day, solution methods for this equation have been well developed for multidimensional problems with relatively simple geometry on structured meshes (see, for example, [1–3]). However, in domains of complex geometry, these methods are rather laborious, primarily, because of the difficulties related to the construction of meshes in physical space and the organization of computations on multiblock meshes.

The use of high-order accurate finite-volume methods on unstructured meshes in physical space seems the natural approach to simplifying the numerical solution of multidimensional problems. Although these methods are widely applied to numerical simulation in gas dynamics, they are little used to solve kinetic equations. Apparently, the only exception is [4], where a numerical method based on Cartesian unstructured meshes was developed for the Boltzmann equation with an exact collision integral.

The basic goal of this study is to develop a high-order accurate numerical method for solving the Boltzmann equation with the *S*-model collision integral (see [5, 6]), which applies to the computation of two-dimensional rarefied gas flows in arbitrarily shaped domains in a wide range of Knudsen numbers. The method consists of two components: an upwind TVD scheme of high order in space for discretizing the advection operator on hybrid unstructured meshes and a procedure for computing macroscopic gas parameters. As examples, the method is applied to a number of two-dimensional external and internal problems.

1. GOVERNING EQUATIONS

The description of two-dimensional rarefied gas flows is based on the Boltzmann kinetic equation for the molecular velocity distribution function f with the *S*-model collision integral (see [5, 6]). For plane flows, the distribution function depends on time t, spatial coordinates $(x, y) = (x_1, x_2)$, and three components of the molecular velocity $\xi = (\xi_1, \xi_2, \xi_3) = (\xi_x, \xi_y, \xi_z)$. Let n_0 and T_0 be some characteristic values of the density and temperature, and L be the characteristic length scale of the problem. To change to dimensionless variables, we use the following reference values of the spatial coordinates (x, y), time t, velocity (u_x, u_y), density n, temperature T, viscosity μ, heat flux (q_x, q_y), and distribution function f, respectively:

$$L, \quad L/\sqrt{2RT_0}, \quad \sqrt{2RT_0}, \quad n_0, \quad T_0,$$

$$\frac{5}{16}mn_0\sqrt{2\pi RT_0}\lambda_0, \quad mn_0(2RT_0)^{3/2}, \quad n_0(2RT_0)^{-3/2}.$$

Here, m is the molecular mass and λ_0 is the mean free path corresponding to n_0 and T_0. In what follows, the dimensionless variables are denoted by the same letters as the corresponding dimensional variables. Following [7], we reduce the dimension of the problem by proceeding from f to a pair of functions $\varphi = (\varphi_1, \varphi_2)^T$ according to the formulas

$$\varphi = \int (1, \xi_z^2)^T f d\xi_z.$$

In dimensionless variables, the S-model kinetic equation for φ becomes

$$\frac{\partial \varphi}{\partial t} = -\xi_x \frac{\partial \varphi}{\partial x} - \xi_y \frac{\partial \varphi}{\partial y} + \nu(\varphi^+ - \varphi), \tag{1.1}$$

where the right-hand side (model collision integral) is given by the formulas

$$\varphi_1^+ = \varphi_M[1 + \Delta(c_1^2 + c_2^2 - 2)], \quad \varphi_2^+ = \frac{1}{2} T \varphi_M[1 + \Delta(c_1^2 + c_2^2 - 1)],$$

$$\varphi_M = n(\pi T)^{-1} \exp(-c_1^2 - c_2^2), \quad \Delta = \frac{4}{5}(1 - \mathrm{Pr})(S_1 c_1 + S_2 c_2), \tag{1.2}$$

$$v_i = \xi_i - u_i, \quad c_i = \frac{v_i}{\sqrt{T}}, \quad S_i = \frac{2q_i}{nT^{3/2}}, \quad \nu = \frac{8}{5\sqrt{\pi}} \frac{nT}{\mu} \frac{1}{\mathrm{Kn}}.$$

Here, $\mathrm{Kn} = \lambda_0 / L$ is the Knudsen number. The macroscopic gas parameters involved in the model collision integral are expressed in terms of φ in the form of integrals with respect to the molecular velocity:

$$(n, nu_x, nu_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (1, \xi_x, \xi_y) \varphi_1 d\xi_x d\xi_y,$$

$$\frac{3}{2} nT + n(u_x^2 + u_y^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [(\xi_x^2 + \xi_y^2)\varphi_1 + \varphi_2] d\xi_x d\xi_y, \tag{1.3}$$

$$(q_x, q_y) = \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (v_x, v_y)[(v_x^2 + v_y^2)\varphi_1 + \varphi_2] d\xi_x d\xi_y.$$

We consider boundary conditions of four types: specular reflection from the surface of the body; diffuse reflection from the surface of the moving body with the accommodation coefficient α; a supersonic incoming flow with given constant macroscopic parameters n_∞, T_∞, $u_{x,\infty}$, $u_{y,\infty}$; and a nonreflecting boundary condition (outlet flow).

Denote by $\mathbf{n}_b$ the outward unit normal to the surface of the body, and let $\xi = (\xi_x, \xi_y)$ and $\xi_n = (\xi, \mathbf{n}_b)$. The specular reflection boundary condition for molecules moving inside the computational domain ($\xi_n < 0$) is written as

$$\varphi_w(\xi) = \varphi(\xi'), \quad \xi' = \xi - 2\xi_n \mathbf{n}_b, \tag{1.4}$$

where ξ' is the velocity of the reflected molecules. For the diffuse reflection boundary condition in the case of a surface moving at the velocity $\mathbf{u} = (u_w, v_w)$, the distribution function of reflected molecules is

$$\varphi_w = \frac{n_w}{\pi T_w} \begin{pmatrix} 1 \\ \frac{1}{2} T_w \end{pmatrix} \exp\left(-\frac{(\xi_x - u_w)^2 + (\xi_y - v_w)^2}{T_w} \right). \tag{1.5}$$

The density n_w and the temperature T_w of reflected molecules are found from the impermeability condition and the energy balance for the gas in its interaction with the wall:

$$n_w = 2\sqrt{\frac{\pi}{T_w}} N_i, \quad T_w = \alpha_w T_s + (1 - \alpha_w) \frac{E_i}{N_i}, \tag{1.6}$$

where T_s is the dimensionless temperature of the surface and N_i and E_i are the mass and energy fluxes of incident molecules given by the expressions

$$N_i = -\int\limits_{\xi_n<0} \xi_n \varphi_1 d\xi, \quad E_i = -\frac{1}{2}\int\limits_{\xi_n<0} \xi_n[(\xi_x^2+\xi_y^2)\varphi_1+\varphi_2]d\xi_x d\xi_y. \tag{1.7}$$

The accommodation coefficient α_w varies in the range $0 \le \alpha_w \le 1$. Here, $\alpha_w = 1$ corresponds to complete temperature accommodation ($T_w = T_s$), while, at $\alpha_w = 0$, the total energy flux through the surface is zero (heat-insulated surface).

The boundary conditions of the third and fourth types are used to solve external flow problems. Denote by $\mathbf{n}_{\text{out}}$ the outer unit normal to the boundary of the computational domain. For the boundary condition of the third type for molecules moving inside the computational domain ($(\xi, \mathbf{n}_{\text{out}}) < 0$), we have an unperturbed flow with given parameters n_∞, T_∞, and $(u_x, u_y)_\infty$:

$$\varphi_\infty = \frac{n}{\pi T_\infty}\begin{pmatrix} 1 \\ \frac{1}{2}T_\infty \end{pmatrix} \exp\left(-\frac{(\xi_x - u_{x,\infty})^2 + (\xi_y - u_{y,\infty}^2)}{T_\infty}\right).$$

For the boundary condition of the fourth type for molecules moving from within the computational domain ($(\xi, \mathbf{n}_{\text{out}}) > 0$), we use the Maxwellian distribution function

$$\varphi = \frac{n}{\pi T}\begin{pmatrix} 1 \\ \frac{1}{2}T \end{pmatrix} \exp\left(-\frac{(\xi_x - u_x)^2 + (\xi_y - u_y)^2}{T}\right).$$

Other types of boundary conditions (evaporation/condensation, radiation) are possible, but they are not considered in this paper.

2. GENERAL FORMULATION OF THE NUMERICAL METHOD

In the physical variables $\mathbf{x} = (x_1, x_2) = (x, y)$, we introduce a mesh consisting of polygons (cells) E_i. Each cell is determined by $L(i)$ vertices with coordinates $\mathbf{x}^{(l)} = (x^{(l)}, y^{(l)})$, $l = 1, 2, \ldots, L(i)$, $\mathbf{x}^{(L(i)+1)} \equiv \mathbf{x}^{(1)}$. The cell area is denoted by $|E_i|$. In this paper, we consider meshes consisting of triangular ($L(i) = 3$) and quadrilateral ($L(i) = 4$) cells. To compute macroscopic gas parameters, improper integrals (1.3) in the space of molecular velocities (ξ_x, ξ_y) are replaced by proper integrals over the square domain $|\xi_x|, |\xi_y| \le \xi_0$. We introduce a uniform rectangular mesh in ξ_x and ξ_y with nodes ξ_α and cell size $\Delta\xi$. For a given node ξ_α, the finite-volume scheme for kinetic equation (1.1) in the cell E_i is written as

$$\frac{\partial}{\partial t}\overline{\varphi}_{i,\alpha} = \mathbf{L}(\overline{\varphi}_{i\alpha}) = \mathbf{D}_{i\alpha} + \nu_i(\varphi^+_{i\alpha} - \overline{\varphi}_{i\alpha}), \quad \mathbf{D}_{i\alpha} = -\frac{1}{|E_i|}\sum_{l=1}^{L(i)} \Phi_{i\alpha l}. \tag{2.1}$$

The cell-averaged value $\overline{\varphi}_{i\alpha}(t)$ and the numerical flux $\Phi_{i\alpha l}$ through the side l of E_i are given by the formulas

$$\overline{\varphi}_{i\alpha} = \frac{1}{|E_i|}\int\limits_{E_i} \varphi(t, \mathbf{x}, \xi_\alpha)dxdy, \quad \Phi_{i\alpha l} = \int\limits_{\mathbf{x}^{(l)}}^{\mathbf{x}^{(l+1)}} \xi_n \varphi(t, \mathbf{x}, \xi_\alpha)dl, \tag{2.2}$$

where $\mathbf{n} = (n_x, n_y)$ is the outward unit normal to l and $\xi_n = (\xi_\alpha, \mathbf{n})$. To compute numerical fluxes, the distribution function values on the cell sides are determined from the known averaged values $\overline{\varphi}_{i\alpha}$. The use of a piecewise constant approximation of the distribution function inside each cell leads to Godunov's scheme [8] written on an unstructured mesh. In [9] a piecewise linear reconstruction with slope limiters based on the minimum derivative principle [10, 11] was used. In this paper, we use second- and higher order accurate quasi-monotone piecewise polynomial reconstructions of the distribution function in a local coordinate system. The slope limiter is specified as in [12].

To complete the description of the numerical method, we have to specify the following parts of the numerical algorithm: the method used to construct a piecewise polynomial approximation of the distribution function; the method used to compute numerical fluxes across cell sides to a prescribed order of

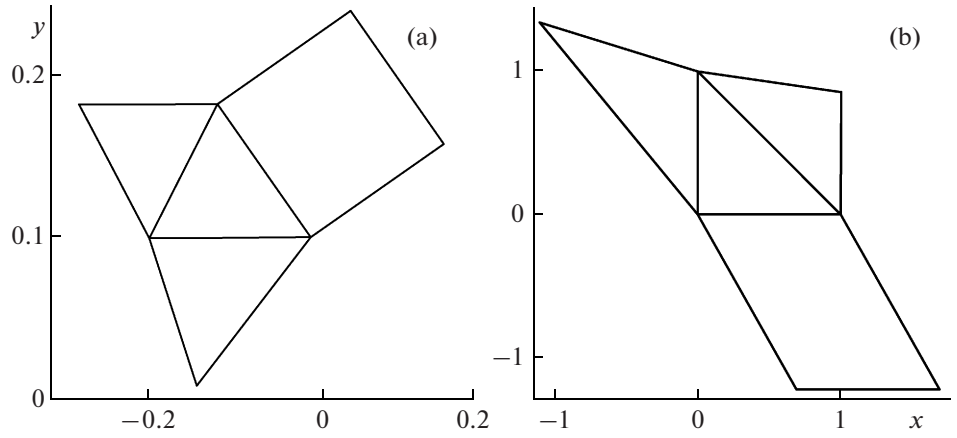


Fig. 1.

accuracy taking into account the boundary conditions; the method used to compute macroscopic gas parameters from known values $\overline{\varphi}_{i\alpha}$ with the mass, momentum, and total energy conservation laws satisfied; and a time differencing method.

3. RECONSTRUCTION ON AN UNSTRUCTURED MESH

For simplicity, the formulas below are written for a scalar function $\varphi(x, y)$ independent of ξ. In computations, the resulting expressions are used for each component $\overline{\varphi}_\alpha$. The dependence on t is dropped since the reconstruction procedure is executed at a fixed time corresponding to a step in the time integration method. In what follows, the global spatial index i is omitted and a local indexing of cells with the index m is used. In this case, the cell E_i has the index 0. The reconstruction problem is formulated as follows: construct a polynomial $p(x, y)$ of a given order r such that its cell-averaged value is equal to $\overline{\varphi}_0$:

$$\overline{\varphi}_0 = \frac{1}{|E_0|}\int_{E_0} p(x, y)dxdy. \tag{3.1}$$

For smooth solutions, $p(x, y)$ must approximate φ inside E_0 to the $(r + 1)$th order of accuracy. The polynomial is conveniently constructed in a local coordinate system $\hat{\mathbf{x}} = (\hat{x}, \hat{y})$ (see [13]), which is specified by the linear transformation

$$\mathbf{x} = \mathbf{x}^{(1)} + J_i\hat{\mathbf{x}}, \quad J_i = \begin{pmatrix} x^{(2)} - x^{(1)} & x^{(L(i))} - x^{(1)} \\ y^{(2)} - y^{(1)} & y^{(L(i))} - y^{(1)} \end{pmatrix}.$$

The reconstruction stencil is obtained by recursively adding all the direct neighbors E_m of the cell E_0 and all the neighbors of the cells already added to the stencil. Cells are added until the required number of cells M is reached in the stencil. The reconstruction stencil in the local coordinate system is obtained by applying the inverse mapping $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{x})$ and consists of the cells E_m', $m = 0, 1, ..., M$. Figure 1 shows the reconstruction stencil in the (a) global and (b) local coordinate systems for a second-order accurate scheme in space.

In the local coordinate system $\hat{\mathbf{x}}$, the reconstruction polynomial $p(\hat{x}, \hat{y})$ is written as

$$p(\hat{x}, \hat{y}) = \overline{\varphi}_0 + \sum_{k=1}^{K} \alpha_k e_k(\hat{x}, \hat{y}), \tag{3.2}$$

where the basis functions $\{e_k\}$, $k = 1, 2, \ldots$, are given by the formulas

$$e_k \equiv \psi_k - \frac{1}{|E_0'|}\int\limits_{E_0'} \psi_k d\hat{x}d\hat{y}, \quad \{\psi_k\} = \hat{x}, \hat{y}, (\hat{x})^2, \ldots. \tag{3.3}$$

The number K of unknown degrees of freedom α_k is related to the degree r of the polynomial by the formula $K = \frac{1}{2}(r+1)(r+2) - 1$. Note that conservativeness condition (3.1) for the reconstruction holds identically due to the choice of the basis functions in the form of (3.3). The unknown degrees of freedom α_k are determined by assuming that, for each cell E_m' in the stencil, the mean value of p is equal to the mean value of the solution $\bar{\varphi}_m$:

$$\int\limits_{E_m'} p(\hat{x}, \hat{y}) d\hat{x}d\hat{y} = |E_m'|\bar{\varphi}_0 + \sum_{k=1}^{K} \int\limits_{E_m'} \alpha_k e_k d\hat{x}d\hat{y} = |E_m'|\bar{\varphi}_m, \quad m = 1, 2, \ldots, M. \tag{3.4}$$

Obviously, no less than K equations of type (3.4) are required to determine K unknown degrees of freedoms. Moreover, the use of the minimum number of cells in the stencil ($M \equiv K$) leads to an unstable scheme [13, 14]. We set $M = (1.5 \ldots 2) \times K$. The resulting overdetermined system of equations for α_k is solved by the weighted least squares method [15].

To suppress the spurious oscillations in domains of rapid variations in the distribution function, we use a limiter of the coefficients of the polynomial ψ. This is equivalent to α_k replaced by the modified degrees of freedom $\tilde{\alpha}_k = \psi\alpha_k$. Then (3.2) is rewritten as

$$p(\hat{x}, \hat{y}) = \bar{\varphi}_0 + \psi\left(\sum_{k=1}^{K} \alpha_k e_k(\hat{x}, \hat{y})\right) = \bar{\varphi}_0 + \sum_{k=1}^{K} \tilde{\alpha}_k e_k(\hat{x}, \hat{y}). \tag{3.5}$$

In this paper, ψ is computed using the technique described in [12], which is a modification of the method in [14] (see also [16]).

Below, we need the mean values of the distribution function $\varphi_0^{(l)}$ along each side l. Using the expression for $p(\hat{x}, \hat{y})$ yields

$$\varphi_0^{(l)} = \bar{\varphi}_0 + \sum_{k=1}^{K} \tilde{\alpha}_k \left(\frac{1}{|\hat{x}^{(l)}\hat{x}^{(l+1)}|}\int\limits_{\hat{\mathbf{x}}^{(l)}}^{\hat{\mathbf{x}}^{(l+1)}} e_k(\hat{x}, \hat{y}) d\hat{l}\right) = \bar{\varphi}_0 + \sum_{k=1}^{K} \tilde{\alpha}_k \bar{e}_k^{(l)}, \tag{3.6}$$

where $|\hat{x}^{(l)}\hat{x}^{(l+1)}|$ is the side length of the transformed cell E_0'. The mean values of $\bar{e}_k^{(l)}$ for each side l are preliminarily found for each cell and are stored in the course of the computations.

4. COMPUTATION OF NUMERICAL FLUXES AND THE COLLISION INTEGRAL

Numerical fluxes (2.2) are computed as follows. First, for each cell E_i and all ξ_α in the velocity mesh, the reconstruction polynomial $\mathbf{p}_{i,\alpha}(x, y)$ is constructed by applying the above reconstruction method to each component of $\boldsymbol{\varphi}$. Then, for each cell side, the mean values $\boldsymbol{\varphi}_{i,\alpha}^{(l)}$ are calculated by formula (3.6).

Next, for all cell sides adjacent to the rigid boundaries, the macroscopic parameters involved in the distribution function of reflected molecules for the diffuse reflection boundary condition (1.5) are calculated by integration with respect to $\boldsymbol{\xi}$. Again, let $\mathbf{n}$ be the outward unit normal to the side l. The mass and energy

fluxes (1.7) are computed so that the impermeability condition holds identically. For this purpose, n_w and T_w are determined by the formulas

$$n_w = -\frac{N_i}{M_i}, \quad T_w = \alpha_w T_s + (1-\alpha_w)\frac{E_i}{N_i},$$

$$N_i = \sum_{\xi_n > 0} \xi_n (\varphi_1)^{(l)}_{i,\alpha} \Delta\xi_x \Delta\xi_y, \quad E_i = \sum_{\xi_n > 0} \xi_n [\xi^2 \varphi_1 + \varphi_2]^{(l)}_{i,\alpha} \Delta\xi_x \Delta\xi_y, \tag{4.1}$$

$$M_i = \sum_{\xi_n \le 0} \xi_n \frac{1}{\pi T_w} \exp\left(-\frac{(\xi_\alpha - \mathbf{u}_w)^2}{T_w}\right) \Delta\xi_x \Delta\xi_y.$$

To calculate $\Phi_{i\alpha l}$, we have to consider two basic cases. If $\xi_n = \xi_\alpha \cdot \mathbf{n} \ge 0$ for the side l in the cell E_i, the flux across the cell side is determined by the simple formula

$$\Phi_{i\alpha l} = \xi_n \varphi^{(l)}_{i,\alpha} \left| x^{(l)} x^{(l+1)} \right|, \tag{4.2}$$

where $|x^{(l)}x^{(l+1)}|$ is the cell size in x and y.

The opposite case $\xi_n < 0$ is divided into several subcases. If l is adjacent to the side l_1 in the neighboring cell indexed by i_1, then the numerical flux across the side is given by a formula similar to (4.2):

$$\Phi_{i\alpha l} = \xi_n \varphi^{(l_1)}_{i_1,\alpha} \left| x^{(l)} x^{(l+1)} \right|. \tag{4.3}$$

When there is no neighboring cell for l, for the second to fourth boundary conditions, the numerical flux is determined by the formula

$$\Phi_{i\alpha l} = \xi_n \varphi_m \left| x^{(l)} x^{(l+1)} \right|, \tag{4.4}$$

where φ_m is a locally Maxwellian function with macroscopic parameters corresponding to the boundary condition.

For the specular reflection of molecules, the velocity ξ' of reflected molecules is first found by formula (1.4), in which we set $\mathbf{n}_b \equiv -\mathbf{n}$. Then the mean value $\varphi^{(l)}_i(\xi'_\alpha)$ of the distribution function over the side l for $\xi \equiv \xi'_\alpha$ is determined by parabolic interpolation with respect to ξ. The numerical flux across the cell side is given by the formula

$$\Phi_{i\alpha l} = \xi_n \varphi^{(l)}_i(\xi'_\alpha) \left| x^{(l)} x^{(l+1)} \right|. \tag{4.5}$$

Note that it is convenient to interpolate the logarithms of the distribution function rather than its values so that interpolation is exact for locally Maxwellian functions.

To evaluate the model collision integral on the right-hand side of (2.1), we have to know six macroscopic gas parameters, namely, the density, temperature, two velocity components, and heat flux. They are determined as integrals (1.3) of the distribution function with respect to molecular velocity. Let ω_α be the weights of the composite Simpson rule, which is used to calculate the integrals with respect to molecular velocity. The direct approximation of (1.3) according to the formulas (with the index i dropped for simplicity)

$$(n, nu_x, nu_y) = \sum_\alpha (1, \xi_x, \xi_y)_\alpha \varphi_{1,\alpha} \omega_\alpha,$$

$$\frac{3}{2} nT + n(u_x^2 + u_y^2) = \sum_\alpha [(\xi_x^2 + \xi_y^2)\varphi_1 + \varphi_2]_\alpha \omega_\alpha, \tag{4.6}$$

$$(q_x, q_y) = \frac{1}{2} \sum_\alpha (v_x, v_y)_\alpha [(v_x^2 + v_y^2)\varphi_1 + \varphi_2]_\alpha \omega_\alpha$$

leads to a nonconservative numerical method, in which the discrete mass, momentum, and energy conservation laws are violated. Then the computation of flows at low Knudsen numbers becomes very difficult.

To ensure that the method is conservative with respect to the collision integral, the macroscopic gas parameters are found by solving the system of equations used to construct the model equation (see [17, 18]). Omitting, for simplicity, the index i, we introduce

$$\mathbf{I}_\alpha = (\varphi_\alpha^+ - \overline{\varphi}_\alpha).$$

For a given cell i, the vector of macroscopic values

$$\mathbf{w} = (n, u_x, u_y, T, q_x, q_y)^{\mathrm{T}}$$

is determined from the system of equations

$$\mathbf{R}(\mathbf{w}) = \mathbf{0}, \tag{4.7}$$

where

$$(R_1, R_2, R_3) = \sum_\alpha (1, \xi_x, \xi_y) I_{1,\alpha} \omega_\alpha,$$

$$R_4 = \sum_\alpha [(\xi_x^2 + \xi_y^2) I_1 + I_2]_\alpha \omega_\alpha,$$

$$R_5 = \frac{4}{3} q_x + \sum_\alpha v_x [(v_x^2 + v_y^2) I_1 + I_2]_\alpha \omega_\alpha,$$

$$R_6 = \frac{4}{3} q_y + \sum_\alpha v_y [(v_x^2 + v_y^2) I_1 + I_2]_\alpha \omega_\alpha.$$

System (4.7) is easily solved by Newton's method:

$$\mathbf{M}(\mathbf{w}^p)(\mathbf{w}^{p+1} - \mathbf{w}^p) = -\mathbf{R}(\mathbf{w}^p),$$

where $\mathbf{M}$ is the Jacobian of system (4.7), which represents discrete sums of the derivatives φ^+ with respect to the macroscopic variables $\mathbf{w}$. Values (4.6) are used as an initial approximation. The iterations are executed until

$$\nu(\mathbf{w})\mathbf{R}(\mathbf{w}) \le 10^{-6}.$$

The computation of $\mathbf{M}$ is a slow part of the solution procedure for system (4.7). In [17, 18], the value of $\mathbf{M}$ was found exactly, which guaranteed the fast convergence of iterations but considerably increased the computation time. In this paper, the method is simplified by proceeding in the expression for $\mathbf{M}$ from numerical integration to exact one. As a result, $\mathbf{M}$ is explicitly expressed in terms of macroscopic gas parameters:

$$\mathbf{M}(\mathbf{w}) \approx \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ u_x & n & 0 & 0 & 0 & 0 \\ u_y & 0 & n & 0 & 0 & 0 \\ u_x^2 + u_y^2 + \frac{3}{2}T & 2nu_x & 2nu_y & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{10}{3} - 2\mathrm{Pr} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{10}{3} - 2\mathrm{Pr} \end{pmatrix} + \mathbf{O}(\Delta\xi^r),$$

where $r = 4$ for the composite Simpson rule. Computations show that the use of the approximate formula for $\mathbf{M}$ hardly affects the convergence of the iterations, but the CPU time is reduced considerably.

Thus, the right-hand side of scheme (2.1) is completely determined.

5. TIME DIFFERENCING

In this paper, unsteady flows are computed with the use of splitting with respect to physical processes [19]. In the first-order accurate splitting method, the passage from the time t^n to the next time t^{n+1} consists of two stages. The first stage corresponds to spatially homogeneous relaxation, and the second stage, to free-molecular motion.

At the first stage, for $0 \le \tau \le \Delta t$, solving the equation

$$\frac{\partial}{\partial \tau}\overline{\varphi}_{i,\alpha} = \nu_i(\varphi^+_{i\alpha} - \overline{\varphi}_{i\alpha}) \tag{5.1}$$

with the initial condition

$$\overline{\varphi}_{i,\alpha}(0) = \overline{\varphi}^n_{i,\alpha}$$

yields an intermediate distribution function value $\overline{\varphi}^*_{i,\alpha}$. At the second stage, we solve a homogeneous equation corresponding to free-molecular motion

$$\frac{\partial}{\partial \tau}\overline{\varphi}_{i,\alpha} = \mathbf{D}_{i,\alpha} = -\frac{1}{|E_i|}\sum_{l=1}^{L(i)}\Phi_{i\alpha l} \tag{5.2}$$

with the initial condition

$$\overline{\varphi}_{i,\alpha}(0) = \overline{\varphi}^*_{i,\alpha}.$$

The computation of the free-molecular stage is completed by calculating macroscopic gas parameters.

In the *S*-model at the relaxation stage described by Eq. (5.1), all the macroscopic gas parameters (except for the heat flux) and the collision frequency are equal to their values at the time t^n. The variation of heat at the relaxation stage is given by the formula

$$q_x(\tau) = q^n_x \exp\left(-\frac{2}{3}\nu\tau\right), \quad q_y(\tau) = q^n_y \exp\left(-\frac{2}{3}\nu\tau\right),$$

so that, at the end of the relaxation stage, we have

$$q^*_x = q^n_x \exp\left(-\frac{2}{3}\nu\Delta t\right), \quad q^*_y = q^n_y \exp\left(-\frac{2}{3}\nu\Delta t\right). \tag{5.3}$$

In view of these relations for macroscopic gas parameters at the relaxation stage, Eq. (5.1) can be analytically integrated. The distribution function value φ^* at the end of the relaxation stage is given by formula (for simplicity, the mesh indices are dropped)

$$\varphi^* = \varphi^n e^{-\nu\Delta t} + \begin{pmatrix} 1 \\ \frac{1}{2}T \end{pmatrix}\varphi^n_M(1 - e^{-\nu\Delta t}) + 3\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} e^{-\frac{2}{3}\nu\Delta t}\begin{pmatrix} 1 - e^{-\frac{1}{3}\nu\Delta t} \end{pmatrix},$$

$$A_k = \frac{8}{5}(1 - \mathrm{Pr})\frac{q_x V_x + q_y V_y}{nT^2}\left(\frac{V_x^2 + V_y^2}{T} - 3 + k\right), \quad k = 1, 2.$$

It should be stressed that all the macroscopic parameters in the expressions for φ^* correspond to t^n. At the end of the relaxation stage, the values of n, u, and T are equal to those at t^n, while the heat flux values are given by formula (5.3). This completes the description of the relaxation stage.

At the free-molecular stage, we use the standard two-step second-order accurate predictor–corrector method.

The time step is specified by the formula

$$\Delta t = K_\xi \frac{\min_i d_i}{\xi_0}, \quad K_\xi \le \frac{1}{2},$$

where d_i is the diameter of the element E_i. For triangular cells, d_i is set equal to the diameter of the inscribed circle, while, for quadrilateral cells, d_i is specified as the ratio of the cell area to the longest side

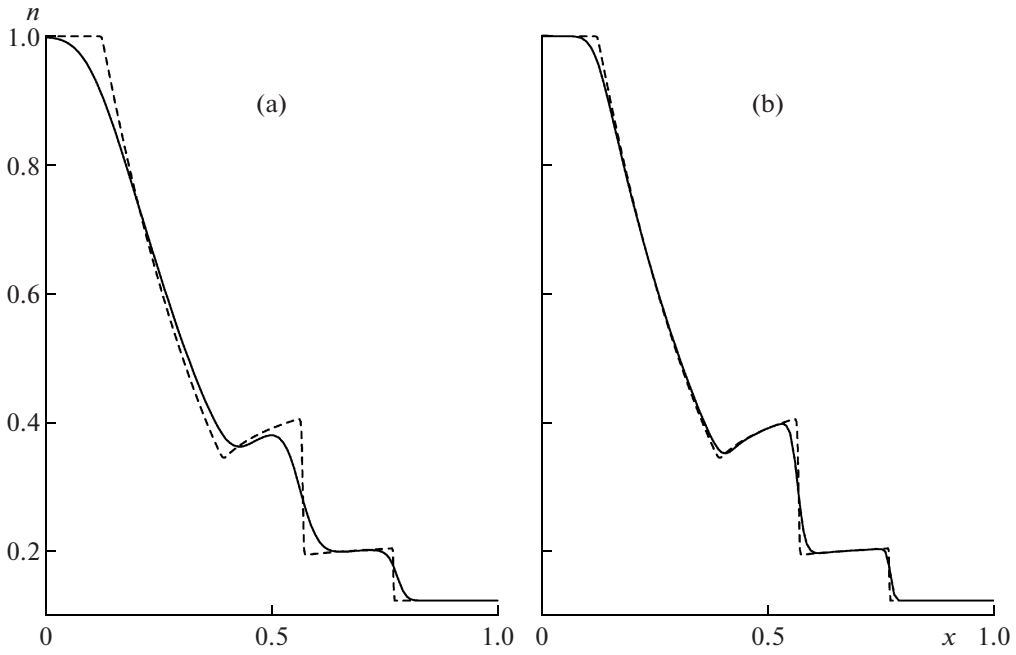


Fig. 2.

length. Note that Δt is independent of the Knudsen number, which facilitates the computation of unsteady flows with $\mathrm{Kn} \ll 1$.

Note that, in the splitting method described, the free-molecular stage is calculated to second-order accuracy, while the relaxation stage is calculated exactly. In practice, we used a second-order accurate splitting method consisting of two relaxation stages with the step size $\Delta t/2$ and one free-molecular stage with the step size Δt.

The time-marching computation of steady flows is based on the first-order accurate scheme

$$\overline{\varphi}_{i,\alpha}^{n+1} = \overline{\varphi}_{i,\alpha}^{n} + \Delta t \mathbf{L}(\overline{\varphi}_{i\alpha}^{n}). \tag{5.4}$$

The time step is specified by the formula

$$\Delta t = \min\left(K_\xi \frac{\min\limits_i d_i}{\xi_0}, \quad K_\nu \min\limits_i \nu_i\right), \quad K_\xi \le \frac{1}{2}, \quad K_\xi \le 1,$$

where the collision frequency ν is defined in (1.2). The resulting method is stable due to the use of a slope limiter in the reconstruction procedure (see [10, 11]). Convergence can be somewhat accelerated by using local time steps in (5.4).

6. NUMERICAL EXAMPLES

The numerical method was verified by computing several test problems with the application of the first- and second-order accurate schemes in space. All the results below were obtained with $\mu = \sqrt{T}$, which corresponds to the hard-sphere intermolecular interaction. The Prandtl number was $\mathrm{Pr} = 2/3$. We also used schemes of the third and higher orders of accuracy in space. For the test problems considered, the numerical results were comparable in accuracy to those produced by the second-order accurate scheme and, for this reason, are not presented here. A uniform mesh with $\Delta\xi \approx 0.4...0.5$ was used for all the computations in the molecular velocity space.

The first test problem concerns the dispersion of a cylindrical cloud of high pressure and density in the surrounding space. The initial condition for the distribution function was specified as the locally Max-

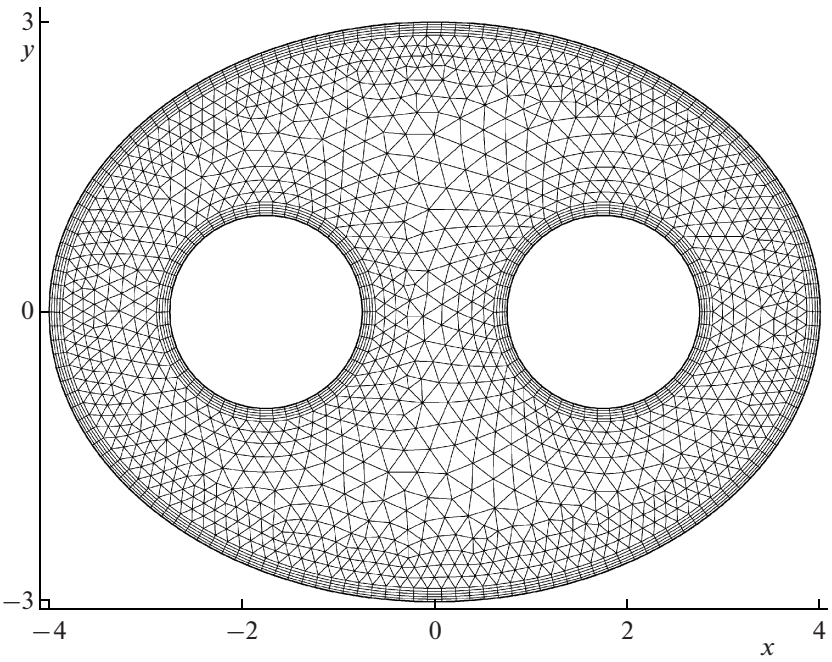


Fig. 3.

wellian function corresponding to a stationary gas with the density and pressure given by the formula (see [20])

$$(n, p) = \begin{cases} (1, 1), & r < 0.4 \\ (0.125, 0.1), & r > 0.4, \end{cases}$$

where $r = \sqrt{x^2 + y^2}$. The resulting rarefied gas flow is independent of z and is described by Eq. (1.1). The Knudsen number in the problem is a formal parameter that defines the scale of the phenomenon in question.

The computational domain in (x, y) was specified as a sector with a central angle of $22.^\circ5$ and $0 \le r \le 1$. A uniform partition with a given number N_r of nodes was specified on the radial boundaries of the sector. The computations were performed using the first- and second-order accurate schemes in space on several refined meshes at $\mathrm{Kn} = 10^{-5}$. The specular reflection condition (1.4) was set on the boundaries of the computational domain. The time differencing scheme was based on the second-order accurate splitting method with respect to physical processes.

For reasons of space, the results are presented only for the finest mesh with $N_r = 200$ (4429 triangular cells). Figure 2 displays the one-dimensional density distributions along the x axis at $t = 0.3$ for (a) first- and (b) second-order accurate schemes. The dashed line depicts the solution to the compressible Euler equations. The gasdynamic solution to the problem consists of a shock wave and a contact discontinuity that propagate in the unperturbed gas and a rarefaction wave moving to the origin. Both schemes correctly reproduce all three waves in the solution, but the second-order accurate scheme in space is noticeably more accurate for the discontinuities and the rarefaction wave. Moreover, the specular reflection conditions used on the boundaries of the computational domain do not lead to artifacts in the numerical solution.

In the second test problem, we considered the steady monatomic rarefied gas flow between three infinite cylinders. The outer elliptical cylinder is at rest and is specified by the equation

$$\left(\frac{x}{4}\right)^2 + \left(\frac{y}{3}\right)^2 = 1.$$

Two rotating circular cylinders of unit radius centered at $x = \pm 3/2$, $y = 0$ are placed inside the outer cylinder. The left inner cylinder rotates at the angular velocity $\Omega_1 = 1$, while the right cylinder moves at the lower

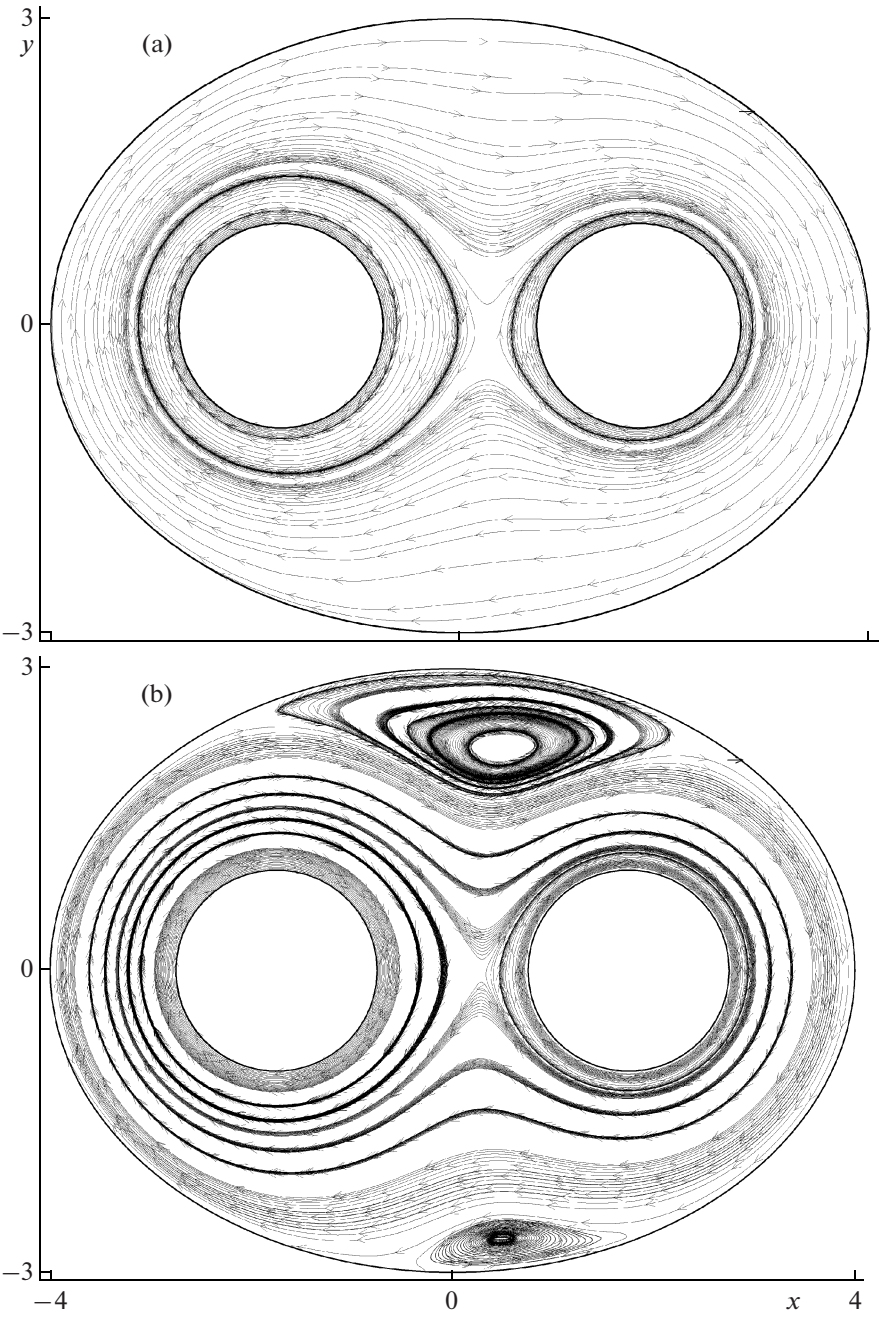


Fig. 4.

angular velocity $\Omega_2 = 0.5$. The surface of the outer cylinder is held at the constant temperature $T_w = 1$, and the inner cylinders are heated to $T_w = 1.5$. The diffuse reflection condition with complete accommodation of the momentum and energy of incident molecules is set on all the surfaces. The resulting rarefied gas flow is independent of z and is described by Eq. (1.1) with the length scale being the radius of the inner cylinder.

The computations were performed on a mixed-element unstructured mesh. In the near-wall domains, we used quadrilateral cells with a cell size of ≈ 0.025 in the normal direction to the surface. Triangular cells were used in the remaining part of the computational domain. The total number of mesh cells was 3602, of which 1250 were quadrilaterals (see Fig. 3). A steady-state solution was calculated by time marching with the help of the first-order accurate scheme (5.4) in time and a second-order accurate scheme in space.

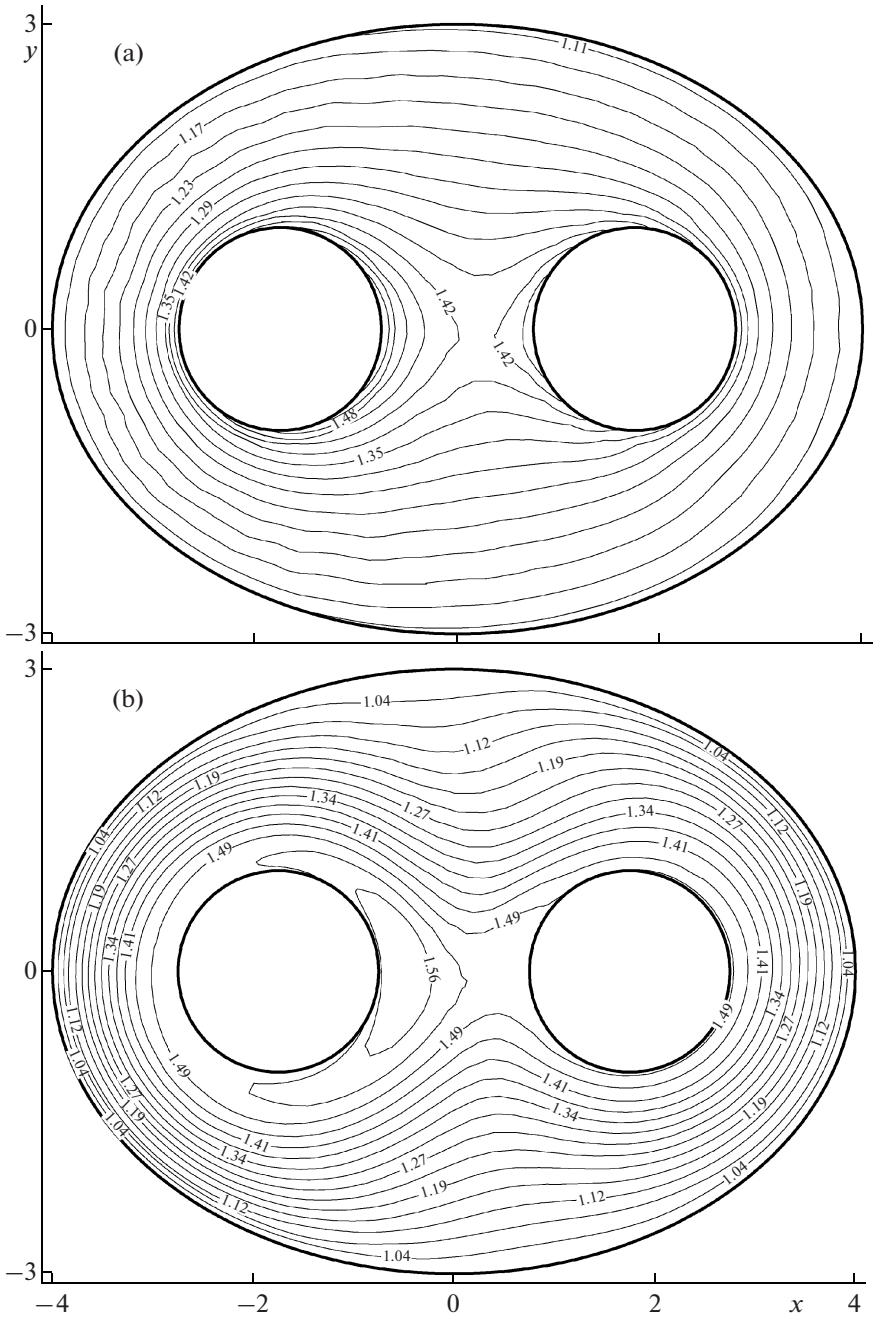


Fig. 5.

Figure 4 shows streamlines for (a) Kn = 1 and (b) Kn = 0.01. As the Knudsen number decreases, the gas velocity away from the surfaces of the inner cylinders falls rapidly. Two circulation flow regions of different sizes develop near the surface of the ellipse. Figure 5 presents contour lines of temperature at the same Knudsen numbers. It can be seen that a higher angular velocity of the left cylinder leads to a noticeably higher temperature of the gas near it as compared with the right cylinder. As the Knudsen number decreases, the temperature profile becomes more asymmetric.

In the third test problem, time marching to a steady state was used to compute the supersonic rarefied plane gas flow over a body of complex geometry. The shape of the body is shown in Fig. 6. The free-stream values of density and temperature were used as their scales. The length scale was specified by the radius of the frontal part of the body. In dimensionless variables, the free-stream velocity was $U_\infty = 2$ and the incoming flow made an angle of 15° with the x axis. The diffuse reflection condition with complete thermal

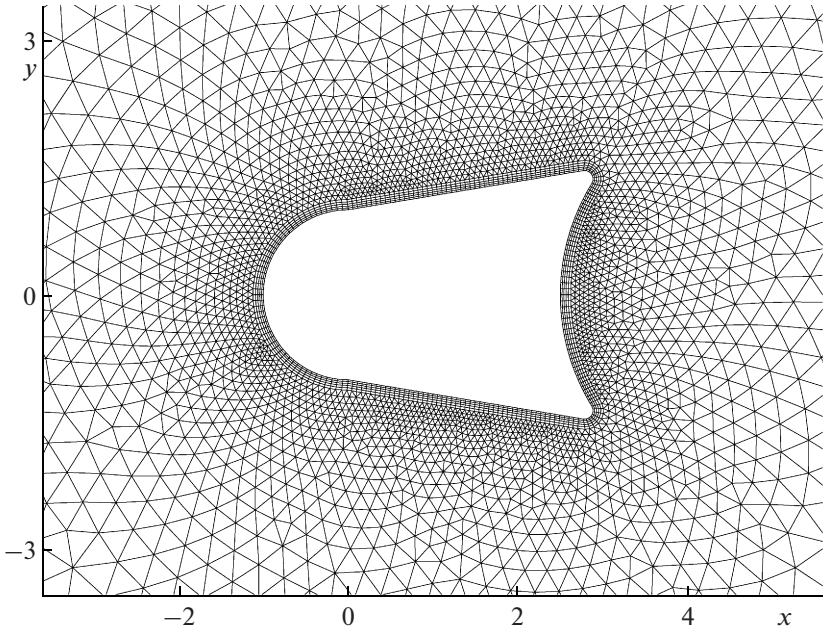


Fig. 6.

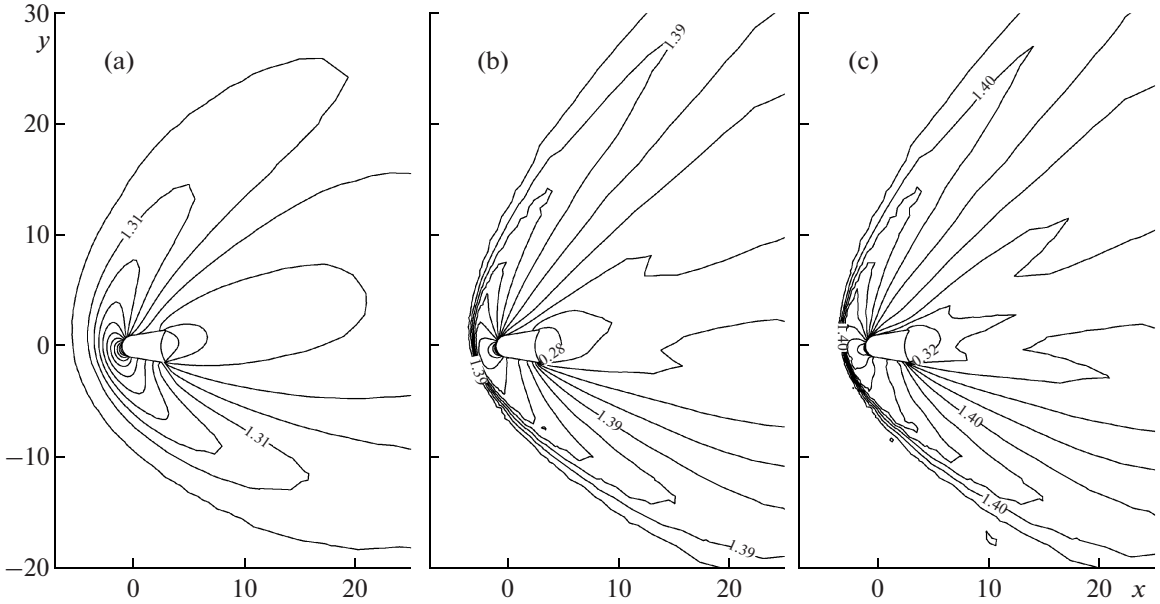


Fig. 7.

accommodation to the dimensionless surface temperature $T_w = 2$ was set on the surface of the body, which was assumed to be homogeneous. In this case, the distribution function is independent of z.

As usual, a finite range of variables ($-10 \le x \le 25$, $-35 \le y \le 60$) was chosen so that the influence of its boundaries could be neglected. In the near-wall layer, we used quadrilaterals with a small cell size of about 0.02 in the normal direction to the surface. A triangular mesh was used in the remaining part of the domain (see Fig. 6). The total number of mesh cells was 8264, of which 1070 were quadrilaterals. A steady-state solution was constructed by time marching with the help of the first-order accurate scheme (5.4) in time and a second-order accurate scheme in space.

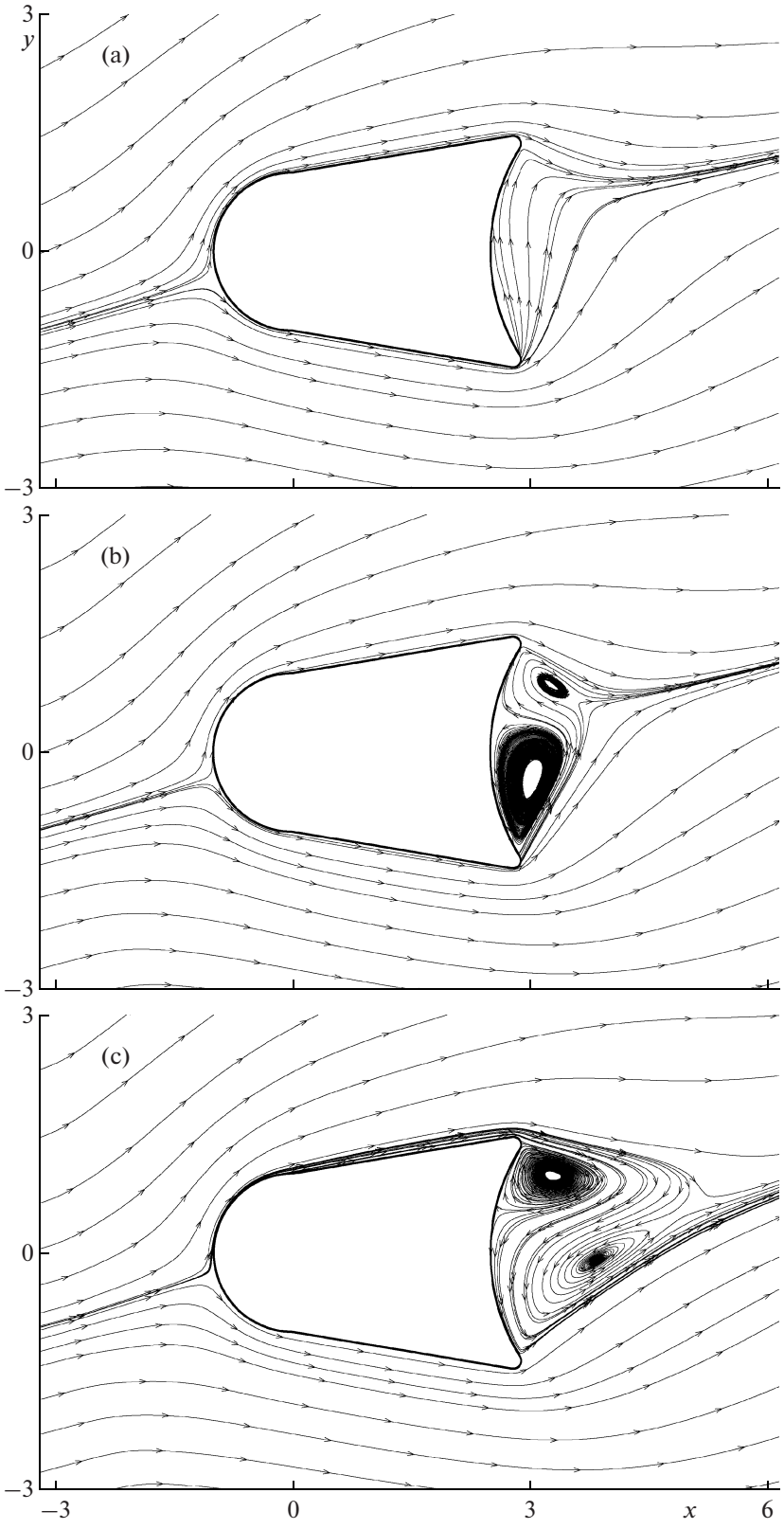


Fig. 8.

Figure 7 shows contour lines of density at Kn = 1 (a), 0.1 (b), and 0.02 (c). Figure 8 illustrates streamlines in the gas at the same Knudsen numbers. Overall, the flow pattern is typical of the supersonic flow past a blunt body. As Kn decreases, a detached shock wave is formed ahead of the body and two regions with a strong recirculation flow appear in the bottom vacuum region.

7. CONCLUSIONS

A high-order accurate numerical method for the *S*-model kinetic equation was proposed for solving two-dimensional unsteady problems in arbitrary multiply connected domains. The method makes use of hybrid unstructured meshes consisting of triangular and quadrilateral cells. The performance of the method was demonstrated by solving three test problems: the cylindrical explosion test problem, the steady flow between rotating bodies, and the steady supersonic flow past a body of complex geometry. In all the cases, the method proved robust and high accurate.

ACKNOWLEDGMENTS

The author is grateful to M. Dumbser (University of Trento, Italy), V.A. Rykov (Dorodnicyn Computing Center of the Russian Academy of Sciences), and E.M. Shakhov (Bauman State Technical University) for helpful remarks.

REFERENCES

1. A. A. Frolova and F. G. Cheremisin, "Rarefied Gas Flow around Cylindrical Bodies," Zh. Vychisl. Mat. Mat. Fiz. **38**, 2096–2102 (1998) [Comput. Math. Math. Phys. **38**, 2012–2018 (1998)].
2. Z.-H. Li and H.-X. Zhang, "Study on Gas Kinetic Unified Algorithm for Flows from Rarefied Transition to Continuum," J. Comput. Phys. **193**, 708–738 (2004).
3. I. N. Larina and V. A. Rykov, "Study of the Rarefied Gas Flow past a Circular Cylinder at Steady-State and Self-Oscillating Regimes," Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 1, 166–175 (2006).
4. V. I. Kolobov, R. R. Arslanbekov, V. V. Aristov, et al., "Unified Solver for Rarefied and Continuum Flows with Adaptive Mesh and Algorithm Refinement," J. Comput. Phys. **223**, 589–608 (2007).
5. E. M. Shakhov, "Generalization of the Krook Relaxation Kinetic Equation," Izv. Akad. Nauk SSSR. Mekh. Zhidk. Gaza, No. 5, 142–145 (1968).
6. E. M. Shakhov, *The method for Analyzing Rarefied Gas Flows* (Nauka, Moscow, 1974) [in Russian].
7. C. K. Chu, "Kinetic-Theoretic Description of the Formation of a Shock Wave," Phys. Fluids **8** (1), 12–22 (1965).
8. S. K. Godunov, "Difference Method for Computing Discontinuous Solutions of Fluid Dynamics Equations," Mat. Sb. **47** (89), 271–306 (1959).
9. N. I. Tillyaeva, "Generalization of Modified Godunov Scheme to Arbitrary Irregular Meshes," Uch. Zap. Tsentr. Aerogidrodin. Inst. **17** (2), 18–26 (1986).
10. V. P. Kolgan, "The Principle of Derivative's Minimal Values as Applied to the Construction of Finite-Difference Schemes for Computing Discontinuous Gas Flows," Uch. Zap. Tsentr. Aerogidrodin. Inst. **3** (6), 68–77 (1972).
11. V. P. Kolgan, "Numerical Method for Solving Three-Dimensional Gas Dynamics Problems and Flow Computations at Nonzero Angle of Attack," Uch. Tsentr. Aerogidrodin. Inst. **6** (2), 1–6 (1975).
12. J. J. W. van der Vegt and H. van der Ven, "Discontinuous Galerkin Finite Element Method with Anisotropic Local Mesh Refinement for Inviscid Compressible Flows," J. Comput. Phys. **141**, 46–77 (1998).
13. M. Dumbser and M. Köser, "Arbitrary High Order Nonoscillatory Finite Volume Schemes on Unstructured Meshes for Linear Hyperbolic Systems," J. Comput. Phys. **221**, 693–723 (2007).
14. T. J. Barth and P. O. Frederickson, "Higher Order Solution of the Euler Equations on Unstructured Meshes Using Quadratic Reconstruction," AIAA Paper, No. 90-0013, *28th Aerospace Sciences Meeting, 1990.*
15. C. F. Ollivier-Gooch and M. van Altena, "A High-Order-Accurate Unstructured Mesh Finite-Volume Scheme for the Advection-Diffusion Equation," J. Comput. Phys. **181**, 729–752 (2002).
16. A. G. Kulikovskii, N. V. Pogorelov, and A. Yu. Semenov, *Mathematical Aspects of Numerical Solution of Hyperbolic Systems* (Fizmatlit, Moscow, 2001; Chapman and Hall/CRC, London, 2001).
17. V. A. Titarev and E. M. Shakhov, "Numerical Calculation of the Hypersonic Rarefied Gas Transverse Flow past a Cold Flat Plate," Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 5, 139–154 (2005).
18. V. A. Titarev, "Conservative Numerical Methods for Model Kinetic Equations," Comput. Fluids **36**, 1446–1459 (2007).
19. V. A. Titarev and E. M. Shakhov, "Numerical Study of Intense Unsteady Evaporation from the Surface of a Sphere," Zh. Vychisl. Mat. Mat. Fiz. **44**, 1314–1328 (2004) [Comput. Math. Math. Phys. **44**, 1245–1258 (2004)].
20. E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics* (Springer-Verlag, Berlin, 1999).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1212–1220. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © I.V. Savenkov, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1271–1279.

Features of the Linear Stage of Development of 3D Wave Packets in a Plane Poiseuille Flow

I. V. Savenkov

Dorodnicyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119991 Russia
e-mail: isavenkov@mail.ru

Received November 17, 2008

Abstract—In the framework of the asymptotic theory of free interaction, the linear stage of the development of 3D wave packets in a plane Poiseuille flow is studied. Numerical results show the presence of "ripples" in the lateral direction in the first phase of the linear stage. The disturbances propagate within a certain angle.

DOI: 10.1134/S0965542509070124

Key words: plane Poiseuille flow, Tollmien–Schlichting waves, three-dimensional disturbances, wave packets, asymptotic expansions, free interaction theory, Navier–Stokes equations.

INTRODUCTION

The plane Poiseuille flow is a simple model that provides a basis for studying the instabilities of flows in a channel caused by a pressure gradient. The disturbed flows in channels at high Reynolds numbers can be described using the asymptotic free interaction theory approach [1–3]. Furthermore, this theory was successfully used to describe the linear instability of the Poiseuille flow when the Reynolds number tends to infinity (see [4, 5]). However, only two-dimensional disturbances were considered in these studies. In this paper, we investigate the three-dimensional wave packets generated by a low-amplitude source.

1. STATEMENT OF THE PROBLEM

Consider the flow of a viscous incompressible fluid driven by a constant pressure gradient g^* between two parallel plates. Define the orthogonal coordinate system placing its origin at the same distance from the plates and directing the axis x^* downstream (so that the equations $y^* = \pm b^*$ describe the plates); the axis y^* is orthogonal to the plates. This flow is described by the classical Poiseuille solution

$$u^* = u_0^*(y^*) = \frac{1}{2}U_0^*[1 - (y^*/b^*)^2], \tag{1.1}$$

where $U_0^* = g^*b^{*2}/(\rho^*\nu^*)$, ρ^* is the density, and ν^* is the kinematic viscosity of the fluid.

Assume that the Couette–Poiseuille flow is disturbed by a local inhomogeneity of the flow (for example, caused by a wall deformation or by an injection through an opening). We also assume that the characteristic size of the disturbed flow is such that it is described by the triple-deck theory of free interaction developed for the Poiseuille flow in [4–6]. Then, as the Reynolds number $\mathrm{R} \longrightarrow \infty (\mathrm{R} = g^*b^{*3}/(\rho^*\nu^{*2})$, the flow falls into three typical subdomains—the main core of the flow (where $y^* \sim b^*$) and two narrow near-wall layers with the thickness $\Delta y^* \sim b^*\mathrm{R}^{-2/7}$. Let us consider the flow in each of these subdomains.

Viscous near-wall layers. According to [4–6], the following asymptotic expansions hold in the near-wall domains (as $\mathrm{R} \longrightarrow \infty$) for the flow functions:

$$u^* = U_0^*(a^{1/3}\mathrm{R}^{-2/7}u_\pm + \ldots), \quad v^* = U_0^*(a^{-1/3}\mathrm{R}^{-5/7}v_\pm + \ldots), \quad w^* = U_0^*(a^{1/3}\mathrm{R}^{-2/7}w_\pm + \ldots),$$
$$p^* = p_0^* + g^*b^*ax + \rho^*U_0^{*2}(a^{2/3}\mathrm{R}^{-4/7}p_\pm + \ldots). \tag{1.2}$$

Here, the functions with the subscripts $\pm$ corresponding to the upper and the lower layers, respectively, depend on the dimensionless coordinates x, $y_\pm$, z, and the time t defined as follows:

$$t^* = \frac{b^*}{U_0^*}a^{2/3}\mathrm{R}^{3/7}t, \quad x^* = b^*ax, \quad y^* = b^*(\pm 1 + a^{1/3}\mathrm{R}^{-2/7}y_\pm), \quad z^* = b^*az.$$

The numerical value of the constant parameter a will be determined later.

Substituting (1.2) in the Navier–Stokes equations, we obtain the Prandtl set of equations for the boundary layer:

$$\begin{gathered}\frac{\partial u_\pm}{\partial x}+\frac{\partial v_\pm}{\partial y_\pm}+\frac{\partial w_\pm}{\partial z}=0,\quad \frac{\partial p_\pm}{\partial y_\pm}=0,\\ \frac{\partial u_\pm}{\partial t}+u_\pm\frac{\partial u_\pm}{\partial x}+v_\pm\frac{\partial u_\pm}{\partial y_\pm}+w_\pm\frac{\partial u_\pm}{\partial z}=-\frac{\partial p_\pm}{\partial x}+\frac{\partial^2 u_\pm}{\partial y_\pm^2},\\ \frac{\partial w_\pm}{\partial t}+u_\pm\frac{\partial w_\pm}{\partial x}+v_\pm\frac{\partial w_\pm}{\partial y_\pm}+w_\pm\frac{\partial w_\pm}{\partial z}=-\frac{\partial p_\pm}{\partial z}+\frac{\partial^2 w_\pm}{\partial y_\pm^2}.\end{gathered}\tag{1.3}$$

However, in distinction from the classical Prandtl statement, the pressure $p_\pm$ is self-induced and is determined along with the other functions of the flow.

The main core of the flow. In this domain, the main Couette–Poiseuille flow (1.1) is subject to small perturbations determined by the following asymptotic expansions (see [4–6]):

$$\begin{gathered}u^*=U_0^*[U_0(y)+\mathrm{R}^{-2/7}a^{1/3}u+\ldots],\quad w^*=U_0^*(a^{2/3}\mathrm{R}^{-4/7}w+\ldots),\\ v^*=U_0^*(a^{-2/3}\mathrm{R}^{-3/7}v+\ldots),\quad p^*=p_0^*+g^*b^*ax+\rho^*U_0^{*2}(a^{2/3}\mathrm{R}^{-4/7}p+\ldots),\end{gathered}\tag{1.4}$$

here, the functions depend on the variables t, x, and $y=y^*/b^*$, and $U_0(y)=\frac{1}{2}(1-y^2)$ describes the Poiseuille flow.

Substituting expansions (1.4), we obtain the set of equations

$$\begin{gathered}\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0,\quad U_0(y)\frac{\partial u}{\partial x}+v\frac{dU_0}{dy}=0,\\ \frac{\partial p}{\partial y}=-a^{-7/3}U_0(y)\frac{\partial v}{\partial x},\quad U_0(y)\frac{\partial w}{\partial x}=-\frac{\partial p}{\partial z}.\end{gathered}\tag{1.5}$$

The integration of the first two equations in (1.5) yields the solution

$$u=A(t,x,z)\frac{dU_0}{dy},\quad v=-U_0(y)\frac{\partial A}{\partial x}\tag{1.6}$$

with the yet unknown function $A(t,x,z)$ that describes the instantaneous displacement of the flow lines in the main core of the flow. The substitution of solution (1.6) in the third equation in (1.5) allows us to find the pressure

$$p=p(t,x,-1,z)+a^{-7/3}\frac{\partial^2 A}{\partial x^2}\int_{-1}^{y}U_0^2(\xi)d\xi,$$

which can be used to find the pressure difference between the upper and the lower walls:

$$p(t,x,1,z)-p(t,x,-1,z)=2\frac{\partial^2 A}{\partial x^2}.\tag{1.7}$$

Finally, we have found the numerical value of the constant $a=\left(\int_{-1}^{1}U_0^2(\xi)d\xi\right)^{3/7}$, which makes it possible to reduce (1.7) to the canonical form (see [4–6]).

Matching the asymptotic solutions. Matching asymptotic expansions (1.2) and (1.4) with consideration of (1.6) and (1.7) yields the following boundary conditions (see [4–6]):

$$\begin{gathered} p_+(t,x,z) - p_-(t,x,z) = 2\frac{\partial^2 A}{\partial x^2}, \\ u_+ + y_+ \longrightarrow -A(t,x,z) \quad \text{as} \quad y_+ \longrightarrow -\infty, \\ u_- - y_- \longrightarrow A(t,x,z) \quad \text{as} \quad y_- \longrightarrow +\infty, \\ w_\pm \longrightarrow 0 \quad \text{as} \quad y_\pm \longrightarrow \mp\infty. \end{gathered} \tag{1.8}$$

If we also set initial conditions and set boundary conditions on the walls, we obtain closed problem (1.3), (1.8) that enables us to find all the functions of the flow.

2. DISPERSION RELATION

Let us investigate the free oscillations of the flow by setting the homogeneous boundary conditions $u_\pm = v_\pm = w_\pm = 0$ for $y_\pm = 0$ and linearizing system (1.3), (1.8) in the small parameter $\delta \longrightarrow 0$:

$$(u_+ + y_+, u_- - y_-, v_\pm, w_\pm, p_\pm, A) = \delta(u'_+, u'_-, v'_\pm, w'_\pm, p'_\pm, A'). \tag{2.1}$$

Extracting the harmonic dependence

$$(u'_\pm, v'_\pm, w'_\pm, p'_\pm, A') = (\bar{u}_\pm, \bar{v}_\pm, \bar{w}_\pm, \bar{p}_\pm, \bar{A})\exp(\omega t + ikx + imz),$$

in the explicit form, we obtain a system of ordinary differential equations that can be reduced to the Airy equation (see [4–6]). Its solution yields the dispersion relation

$$\begin{gathered} \Phi(\Omega) = Q(k,m), \\ \Phi(\Omega) = \frac{d\mathrm{Ai}(\zeta)}{d\zeta}\left[\int_\Omega^\infty \mathrm{Ai}(\zeta)d\zeta\right]^{-1}, \quad \Omega = \omega(ik)^{-2/3}, \quad Q = (ik)^{1/3}(k^2+m^2), \end{gathered} \tag{2.2}$$

which relates the complex frequency ω to the wave numbers (k, m) of the free oscillations of the flow. Here, $\mathrm{Ai}(\zeta)$ is the Airy function, which exponentially damps in the sector $|\arg\zeta| < \pi/3$.

When $m = 0$, dispersion relation (2.2) coincides with the dispersion relation for the two-dimensional waves obtained in [5]:

$$\Phi(\Omega) = Q_0(k), \quad Q_0 = (ik)^{1/3}k^2. \tag{2.3}$$

Furthermore, making the change of variable $K = k^{1/7}(k^2 + m^2)^{3/7}$, we can reduce (2.2) to form (2.3), namely, $\Phi(\Omega) = Q_0(K)$; from this, we derive the relation between the frequencies of the three-dimensional (ω) and two-dimensional (ω_0) oscillations:

$$\omega = \omega_0(K)(k/K)^{2/3}, \quad K = k^{1/7}(k^2+m^2)^{3/7}. \tag{2.4}$$

Thus, the problem on the three-dimensional oscillations is reduced to the well-known problem on the two-dimensional oscillations (see [5, 7]). It was shown in [5, 7] that dispersion relation (2.3) has a countable number of roots $\omega_{0,n}$; among them, only the first root $\sigma_0(k) = \mathrm{Re}\,\omega_{0,1}(k) > 0$ for the values of the wave number $k > k_{0,*} = 1.005$ is unstable. The numerical analysis (see [5]) shows that the maximum of $\sigma_0(k)$ is $\sigma_{\max} = 0.937$ at $k = k_{\max} = 1.74$. Since $k \le K$, we have $\sigma(k, m) = \mathrm{Re}\,\omega_1(k, m) \le \sigma_0(K)$; therefore, the plane wave with $(k, m) = (k_{\max}, 0)$ is the most unstable one. The complete picture of the waves' increment of growth is shown in Fig. 1a. Note that the isolines are clearly scythe-shaped and bend backwards; due to this fact, for small k (say, in a neighborhood of $k \approx k_{0,*} \approx 1$), the direct waves (with $m = 0$) are almost not growing or even damp, while the oblique waves (with $m \neq 0$) can have a considerable increment of growth $\sigma \approx 0.5$ for m in a neighborhood of $m \approx 1.5$. The isolines $\mathrm{Im}\,\omega_1(k, m)$ shown in Fig. 1b also carry important

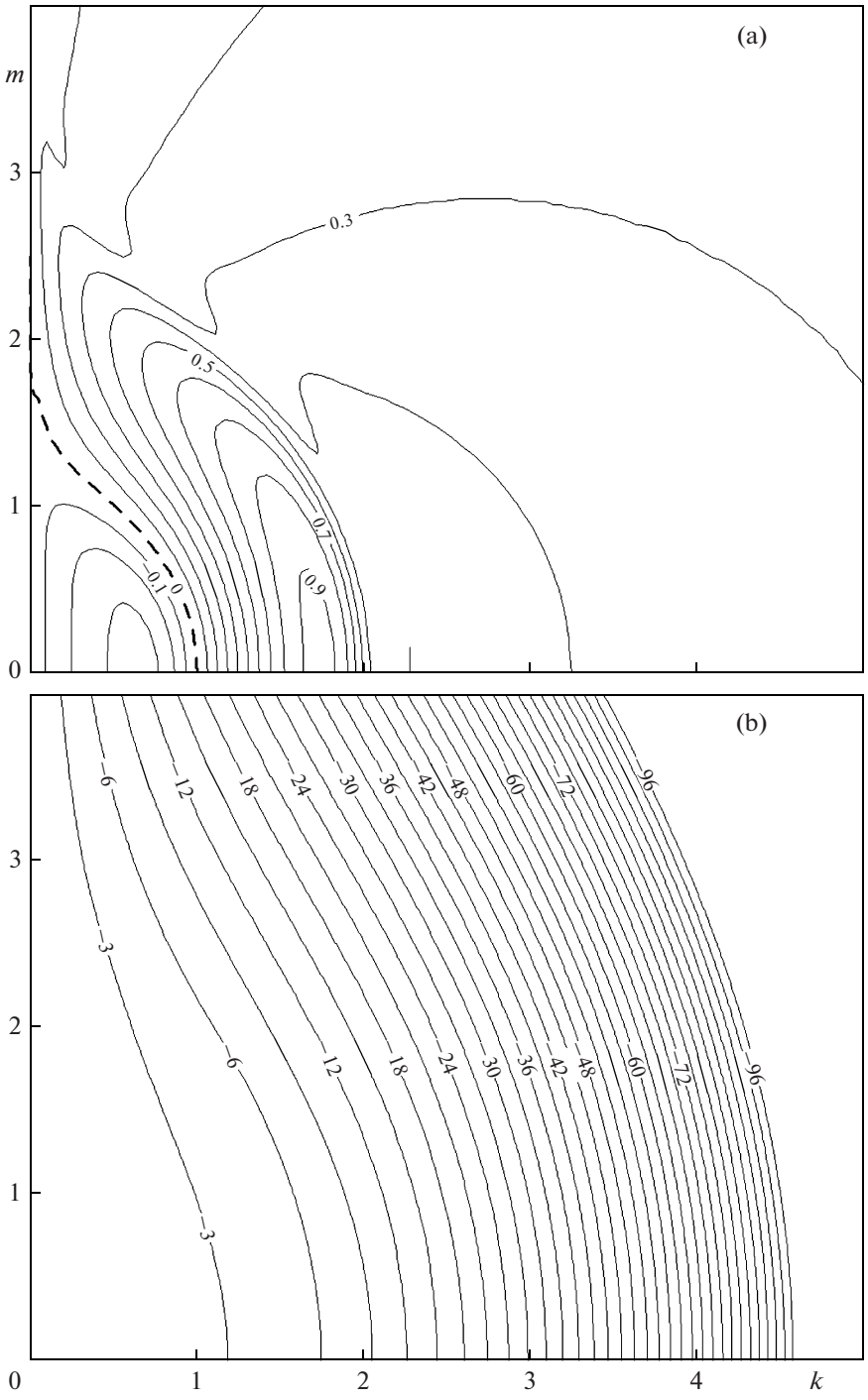


Fig. 1.

information. The normal to an isoline drawn at a point (k, m) shows the direction of the propagation of a group of waves from a neighborhood of (k, m).

3. WAVE PACKET

Consider disturbances generated by a pulse source. We assume that the source of disturbances is the injection through an opening in the lower wall. Then, the boundary conditions are

$$u_\pm = w_\pm = 0, \quad v_- = \delta v_0(t, x, z), \quad v_+ = 0 \quad \text{as} \quad y_\pm = 0. \tag{3.1}$$

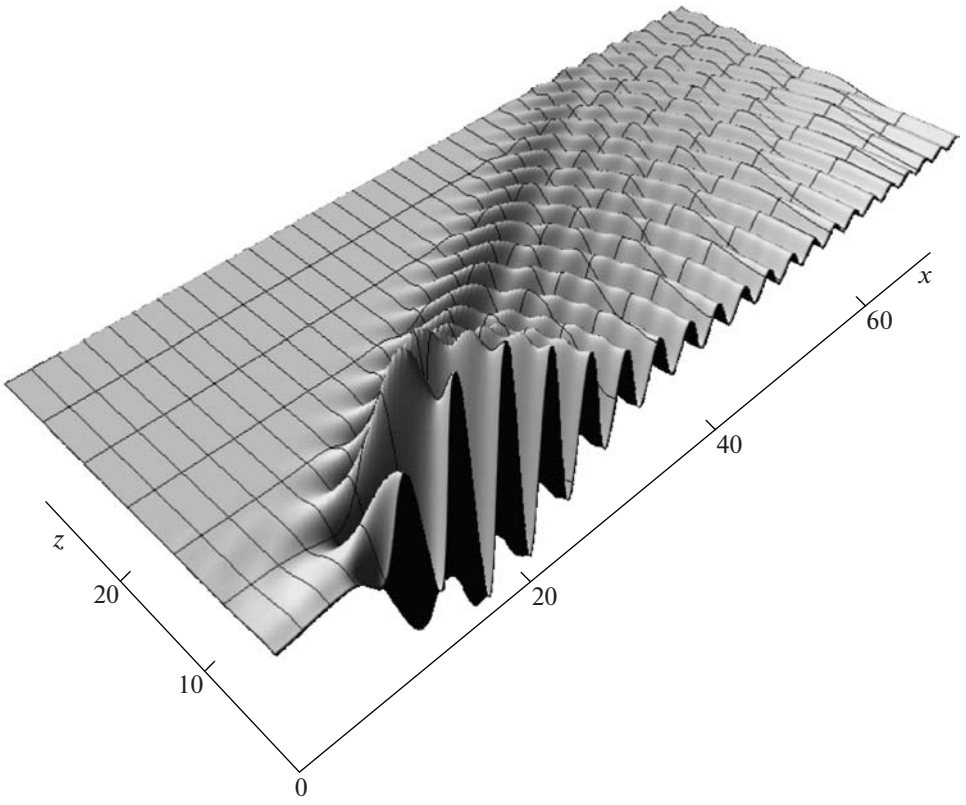


Fig. 2.

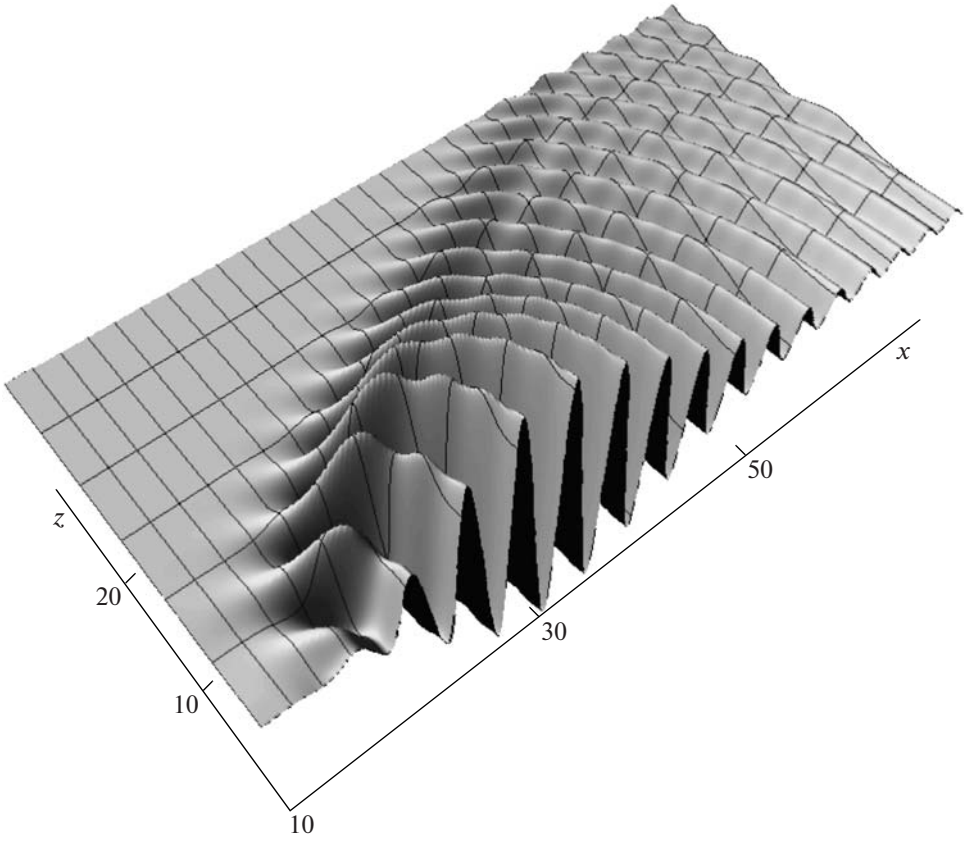


Fig. 3.

Assuming that the disturbances are small, we linearize problem (1.3), (1.8), (3.1) in the amplitude parameter $\delta \longrightarrow 0$ using (2.1); then, we expand the unknown functions (marked by a prime) in the Laplace–Fourier integrals. For example, the expansion for the function A is

$$A'(t, x, z) = 2\mathrm{Re}\left(\frac{1}{2\pi i}\int_{-\infty}^{\infty} dm \exp(imz)\int_{0}^{\infty} dk \int_{b-i\infty}^{b+i\infty} \bar{A}(\omega, k, m)\exp(\omega t + ikx)d\omega\right). \tag{3.2}$$

Hence, for the images of the unknown functions, we have a system of differential equations (in Y) that reduces to solving the Airy equation. Finally, we have

$$\bar{A} = -(ik)^{-2/3}\frac{\bar{v}_0(\omega, k, m)}{\Phi(\Omega) - Q(k, m)},$$

where

$$\Phi(\Omega) = \frac{d\mathrm{Ai}(\zeta)}{d\zeta}\left[\int_{\Omega}^{\infty}\mathrm{Ai}(\zeta)d\zeta\right]^{-1}, \quad \Omega = \omega(ik)^{-2/3}, \quad Q = (ik)^{1/3}(k^2 + m^2).$$

For simplicity, we assume that $v_0(t, x, z) = \delta(t)v_{00}(x, z)$, where $\delta(t)$ is the Dirac delta function and $v_{00}(x, z) = \exp(-x^2 - z^2)$. Using the residual theorem for the inner integral in (3.2), we obtain

$$A'(t, x, z) = 2\mathrm{Re}\left(\int_{-\infty}^{\infty} dm \int_{0}^{\infty} dk \sum_{n=1}^{\infty} \frac{\bar{v}_{00}(k, m)}{\frac{d\Phi}{d\Omega}(\Omega_n)}\exp[\omega_n(k, m)t + ikx + imz]\right),$$

where ω_n are the roots of dispersion relation (2.2) mentioned above. Neglecting the contribution of all the damping roots, we obtain

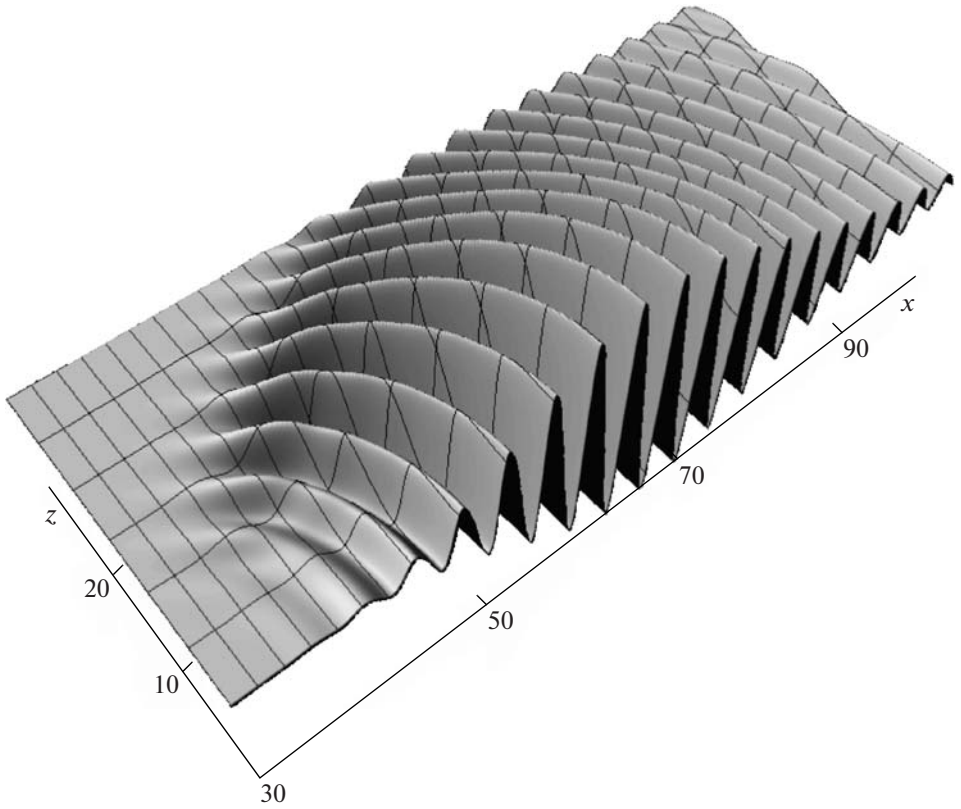


Fig. 4.

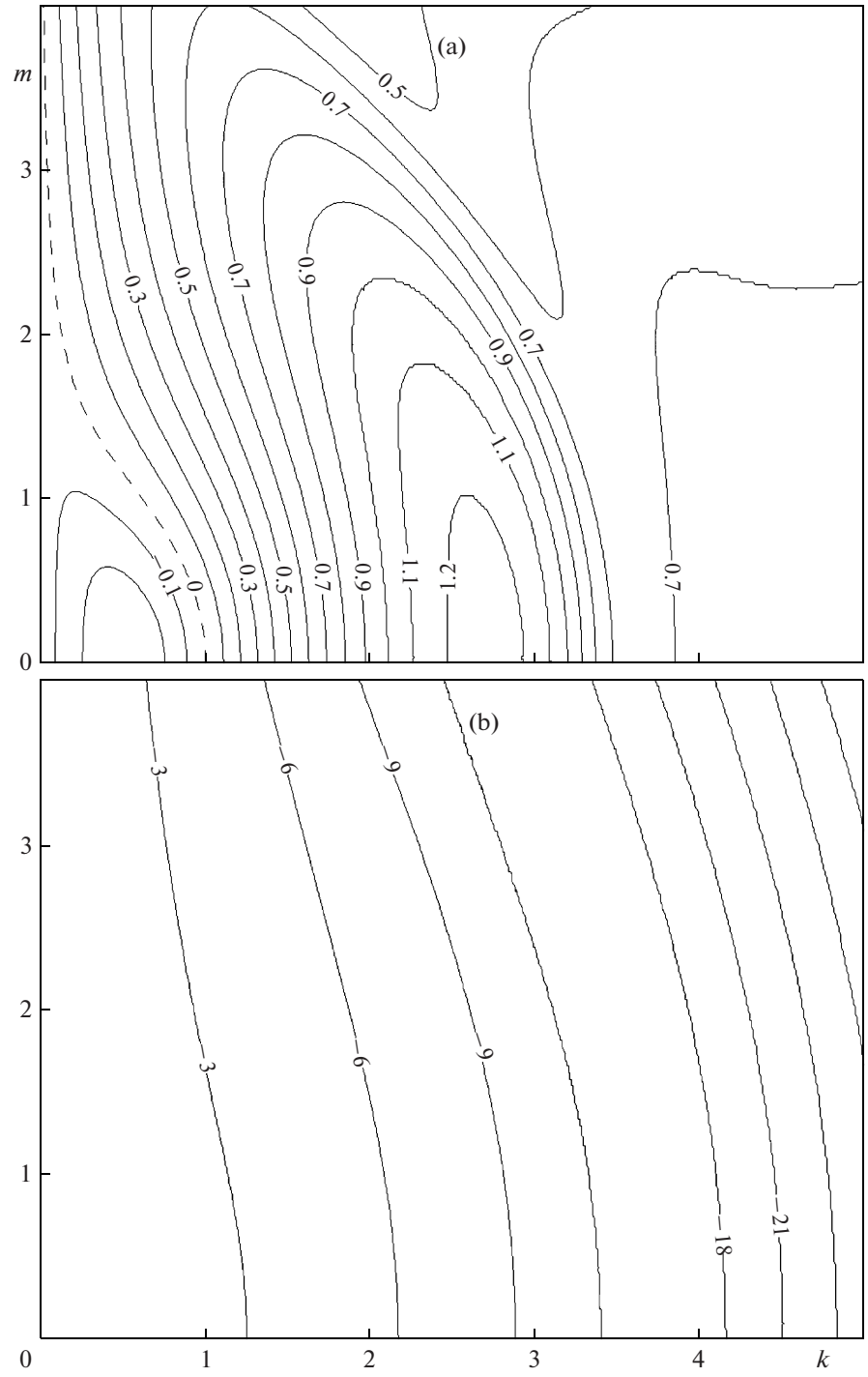


Fig. 5.

$$A'(t,x,z) = 2\mathrm{Re}\left(\int_{-\infty}^{\infty} dm \int_{0}^{\infty} dk \frac{\overline{V}_{00}(k,m)}{\dfrac{d\Phi}{d\Omega}(\Omega_1)} \exp[\omega_1(k,m)t + ikx + imz]\right). \quad (3.3)$$

Integral (3.3) was calculated using the trapezoid rule with the steps $\Delta k = 0.01$ and $\Delta m = 0.02$, which provided an accuracy of 1%. The results are shown in Figs. 2, 3, and 4 for the times $t = 3$, $t = 5$, and $t = 10$,

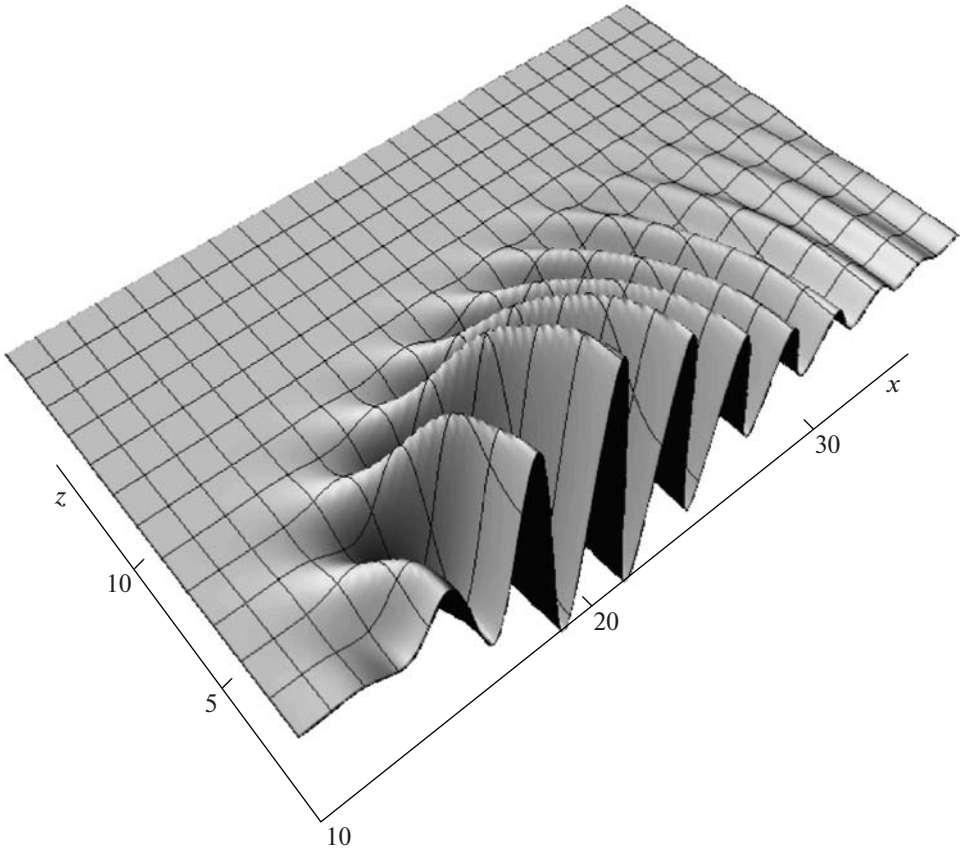


Fig. 6.

respectively. In these figures, the development of the wave packet formed by the most unstable waves from a neighborhood of $k = k_{\max} = 1.738$ (see Fig. 1a) and propagating at the group velocity $U = -\partial \mathrm{Im}(\omega_1(k = k_{\max}; m = 0))/\partial k = 7.02$ downstream is clearly visible. Such a pattern of one packet distinguished on the general background of the perturbations is typical for many shear perturbations (in particular, for the boundary layer [7–9]).

We also observe a specific feature of the propagation of the perturbations in the lateral direction (z). In Figs. 2 and 3, the ripples are clearly visible, which recede when the wave packet acquires a fully developed shape by the time $t = 10$ (Fig. 4). These ripples are characteristic of the initial phase of the packet development when the most unstable waves from the neighborhood of $(k, m) = (k_{\max}, 0)$ still do not completely define the shape of the wave packet. For t in the range from 3 to 5, the oblique waves in the neighborhood of $(k, m) \approx (1, 1.5)$, which have a considerable increment of growth $\sigma \approx 0.5$, compete with the most unstable waves. It is precisely these oblique waves that cause the ripples.

Another feature is clearly observed in Figs. 2 and 3; namely, there is an angle within which the perturbations propagate. The measure of this angle can be estimated as 30° judging from these figures, which is in agreement with the shape of the isolines in Fig. 1b (the normal to an isoline drawn at a point (k, m) shows the direction of the propagation of a group of waves from a neighborhood of (k, m)). Moreover, we can give an asymptotic estimate of the measure of this angle using the asymptotics found in [7, 8]: $\mathrm{Im}\,\omega_1(k, m) \approx -ik(k^2 + m^2)$ for sufficiently large k and m. Then,

$$\tan\alpha = -\frac{\partial \mathrm{Im}\,\omega_1(k, m)/\partial m}{\partial \mathrm{Im}\,\omega_1(k, m)/\partial k} \approx \frac{2(m/k)}{3 + (m/k)^2} \leq \frac{\sqrt{3}}{3},$$

hence, we obtain an asymptotic estimate for the maximal angle of the propagation of the perturbations $\alpha \approx \arctan(\sqrt{3}/3) = 30°$, which is an excellent agreement.

To emphasize these features, we make a comparison with the wave packet in the boundary layer on the flat plate (see [8, 9]). For the case of the boundary layer on the flat plate, the dispersion relation has the

same form (2.2); however, $Q = (ik)^{1/3}\sqrt{(k^2 + m^2)}$ in this case. The change in the topology of the isolines of the waves' increment of growth $\sigma(k, m) = \mathrm{Re}\omega_1(k, m)$ and their frequency $\mathrm{Im}\omega_1(k, m)$ can be seen in Figs. 5a and 5b, respectively. It is seen that the "scythes" in Fig. 5a do not bend backwards as sharply, and there is no clear dominance of the increments of growth of the oblique waves over the plane waves in the range $k \approx [1, 1.5]$; therefore, there are no ripples on the wave packet in Fig. 6. Furthermore, the normals to the isolines in Fig. 5b are less tilted to the axis k; as a result, the wave packet does not spread in the lateral direction (z) as is seen in Figs. 2–4.

4. CONCLUSIONS

In the framework of the asymptotic free interaction theory, it is shown that ripples in the lateral direction can occur in the plane Poiseuille flow at the initial stage of the development of a 3D wave packet; the ripples are caused by unstable oblique waves. It is established that the perturbations propagate within a 30° angle.

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research, project no. 07-01-00589.

REFERENCES

1. V. Ya. Neiland, "On the Theory of Laminar Boundary Layer in a Supersonic Flow," Izv. Akad. Nauk SSSR, Mekhan. Zhidkosti Gaza, No. 4, 53–58 (1969).
2. K. Stewartson and P. G. Williams, "Self-Induced Separation," Proc. R. Soc. London, Ser. A **312** (1509), 181–206 (1969).
3. A. F. Messiter, "Boundary-Layer Flow near the Trailing Edge of a Flat Plate," SIAM J. Appl. Math. **18** (1), 241–257 (1970).
4. V. I. Zhuk and O. S. Ryzhov, "On the Free Interaction of Near-Wall Layers the Core of a Poiseuille Flow," Dokl. Akad. Nauk SSSR **257** (1), 55–59 (1981).
5. E. V. Bogdanova and O. S. Ryzhov, "Oscillation Excited by a Harmonic Oscillator in a Poiseuille Flow," Dokl. Akad. Nauk SSSR **257**, 837–841 (1981).
6. F. T. Smith, "On the High Reynolds Number Theory of Laminar Flows," IMA J. Apl. Math., 1981, vol. 27, pp. 133–175. **28** (3), 207–281 (1982).
7. O. S. Ryzhov and E. D. Terent'ev, "On the Transient Mode Characterizing the Launch of a Vibrator in the Subsonic Boundary Layer on a Plate," Prikl. Mat. Mekh. **50**, 974–986 (1986).
8. O. S. Ryzhov and I. V. Savenkov, "Asymptotic Theory of Wave Packets in the Boundary Layer on a Plate," Prikl. Mat. Mekh. **51**, 820–828 (1987).
9. O. S. Ryzhov and I. V. Savenkov, "Asymptotic Approach to the Theory of Hydrodynamic Stability," Mat. Modelir. **1** (4), 61–86 (1989).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1221–1234. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © A.D. Savel'ev, A.I. Tolstykh, D.A. Shirobokov, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1280–1294.

Application of Compact and Multioperator Schemes to the Numerical Simulation of Acoustic Fields Generated by Instability Waves in Supersonic Jets

A. D. Savel'ev, A. I. Tolstykh, and D. A. Shirobokov

Dorodnicyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119991 Russia

e-mail: savel@ccas.ru, tol@ccas.ru

Received January 22, 2009

Abstract—Acoustic fields generated by instability waves in supersonic jets were numerically simulated. A seventh-order multioperator scheme was used to solve the Euler equations linearized about the mean flow field in an axisymmetric turbulent jet. The mean field was computed using fifth-order compact approximations of the convective terms under conditions similar to experimental data. The numerical results were found to agree well with the experiment.

DOI: 10.1134/S0965542509070136

Key words: compact and multioperator schemes, numerical simulation of acoustic fields, instability waves of supersonic jets.

1. INTRODUCTION

An area of research into sources of noise generated by nozzle jets is the study of instability waves. This concept was introduced in [1] for the case of supersonic jets. According to [1], the emission of sound in the course of instability development proceeds as follows.

The shear layers at the boundary of the initial jet segment are characterized by high transverse gradients of velocity, which lead to instability and, accordingly, to the growth of harmonic amplitudes. As a result, the amplitude of a simple harmonic wave generated by an external source at the nozzle exit increases as the wave propagates along the jet axis. However, the thickness of the shear layers increases downstream, which, combined with decreasing transverse gradients, leads to the stabilization and subsequent damping of the wave. An increase and then a decrease in the wave amplitude leads to a broadband wavenumber spectrum. Its short-wave range may include harmonics with a supersonic phase velocity relative to the surrounding medium, which are the source of sound.

The following procedure for analyzing this phenomenon was proposed in [1]. The steady mean field pattern is assumed to be known. The Euler equations written in cylindrical coordinates are linearized about this field. The resulting linear system is solved by the method of matched asymptotic expansions, where the small parameter is that characterizing the slow expansion of the jet. In [1] the theory was verified using the experiments from [2], in which the flow parameters of a supersonic axisymmetric jet were measured together with acoustic characteristics. An approximation of the experimental data was used in [1] as input data in the numerical implementation of analytical solutions. The theory thus constructed was found to agree well with the experiment.

A similar approach was used in [3], where the theory from [1] was further developed and was supported by experimental data.

The results suggesting that instability waves are a source of sound in supersonic jets motivate attempts to solve the three-dimensional nonstationary Euler equations linearized about the mean field of the jet. This opens up the possibility of studying acoustic fields generated by jets outflowing from nozzles of various shapes.

The basic difficulties in the numerical simulation of acoustic fields are associated with phase and amplitude errors (mainly in the short-wave range of harmonics resolved by actual schemes), which are characteristic for most traditional methods. In the numerical integration of nonstationary equations, these errors can accumulate and noticeably distort the solutions.

A conventional approach is to use high-order accurate schemes. This ensures small errors in the long wavelength approximation, when the dimensionless wave number kh (h is the mesh size) tends to zero.

However, high-order schemes do not guarantee that the errors are small in the shortwave range, when $kh \geq \pi/2$ (note that the wavelength in the case of $kh = \pi/2$ is equal to $4h$).

The idea of optimizing schemes for an aeroacoustic problem was proposed in [4]. Specifically, by partially sacrificing the high order in the approximation of derivatives on a multipoint stencil, we can try to extend the range of wave numbers for which the phase errors are small, for example, in the rms norm by choosing suitable free parameters. For dissipative schemes, the same approach can be used to minimize the amplitude error (or both errors simultaneously).

This paper consists of two parts. The first describes a multioperator scheme for which the construction principles were presented in [5, 6] and some application results can be found in [7, 8]. This scheme was used in this study to solve the linearized Euler equations numerically. Seventh-order accurate in space, this scheme is characterized by rather wide ranges of wave numbers with small phase and amplitude errors. It is based on a linear combination of upwind compact approximation operators other than those used in [5, 6]. The application of these operators to the discretization of convective terms with an arbitrarily high order was described in [9].

The other part presents numerical results obtained for acoustic fields generated by instability waves in supersonic jets. Preliminarily, the mean turbulent flow in an axisymmetric jet was computed under conditions close to experimental data. The resulting fields of gasdynamic parameters were used in the linearized Euler equations, which were solved by applying a multioperator scheme.

2. STATEMENT OF THE PROBLEM AND DIFFERENCE SCHEMES FOR ITS SOLUTION

1. The statement of the problem corresponds to the strategy proposed in [1] for studying acoustic fields in the case of supersonic jets. Specifically, the Euler equations linearized about a known mean flow field are used in the presence of acoustic disturbances at the inlet boundary. Two problems arise in the numerical implementation of this strategy. First, we have to compute the steady mean jet flow described by the Navier–Stokes equations with a turbulence model. The resulting numerical solution is used as a base in the linearization of the Euler equations. Let the mean-flow parameters be denoted by the index 0. Then, assuming that the excitation is nonstationary with respect to perturbations in the density ρ, velocity components u, v, and w, and internal energy e, the resulting three-dimensional system has the form

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} + \frac{\partial \mathbf{H}}{\partial z} = 0, \tag{1}$$

where

$$\mathbf{W} = \begin{bmatrix} \rho \\ u_0\rho + \rho_0 u \\ v_0\rho + \rho_0 v \\ w_0\rho + \rho_0 w \\ (e_0 + U_2)\rho + \rho_0(e + U_1) \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} u_0\rho + \rho_0 u \\ u_0^2\rho + 2\rho_0 u_0 u + P_1 \\ u_0 v_0\rho + \rho_0 v_0 u + \rho_0 u_0 v \\ u_0 w_0\rho + \rho_0 w_0 u + \rho_0 u_0 w \\ E_2(u_0\rho + \rho_0 u) + \rho_0 u_0 E_1 \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} v_0\rho + \rho_0 v \\ u_0 v_0\rho + \rho_0 v_0 u + \rho_0 u_0 v \\ v_0^2\rho + 2\rho_0 v_0 v + P_1 \\ v_0 w_0\rho + \rho_0 w_0 v + \rho_0 v_0 w \\ E_2(v_0\rho + \rho_0 v) + \rho_0 v_0 E_1 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} w_0\rho + \rho_0 w \\ u_0 w_0\rho + \rho_0 w_0 u + \rho_0 u_0 w \\ v_0 w_0\rho + \rho_0 w_0 v + \rho_0 v_0 w \\ w_0^2\rho + 2\rho_0 w_0 w + P_1 \\ E_2(w_0\rho + \rho_0 w) + \rho_0 w_0 E_1 \end{bmatrix}.$$

Here, $U_2 = (u_0^2 + v_0^2 + w_0^2)/2$, $U_1 = u_0 u + v_0 v + w_0 w$, $E_1 = \gamma e + U_1$, $E_2 = \gamma e_0 + U_2$, and $P_1 = (\gamma - 1)(e_0\rho + \rho_0 e)$.

This system describes perturbations of gasdynamic parameters in the three-dimensional case and applies to jets outflowing from nozzles with variously shaped exit cross sections. In this case, the mean field was axisymmetric and corresponded to the experimental conditions in [2], but the computer code did not take into account this symmetry.

In what follows, it is assumed that the x axis coincides with the jet axis and all the sizes are divided by the radius r of the jet's initial cross section. In the jet's initial cross section $x = 0$, $r = \sqrt{y^2 + z^2} < 2$, we set $v = w = 0$, specified a perturbation of the pressure $p = (\gamma - 1)(e\rho_0 + \rho e_0)$, and used the adiabatic condition. Another condition was that the perturbation in the Riemann invariant was transferred from interior to boundary nodes for a one-dimensional flow along the x axis. In the supersonic part of the outlet boundary $x = L$ of the computational domain, all the unknown functions were transferred. In the subsonic part of this boundary, we used a condition for Riemann-invariant perturbations (the perturbation for the "entering" invariant was set equal to zero) and the entropy conservation condition. Moreover, the perturbations in v and w were transferred to the boundary. Similar conditions were used on the lateral boundaries.

To prevent the influence of reflected perturbations on acoustic fields, the term $-\beta W$, where $\beta = \text{const} < 0$, was added to the right-hand side of Eqs. (1) near the outlet and lateral boundaries. As a result, perturbations reaching the boundaries had smaller amplitudes.

A similar statement of the problem concerning the mean supersonic turbulent flow in a jet was presented in [10], where an axisymmetric jet outflowing from a nozzle was considered at given distributions of flow parameters in a cross section inside the nozzle. In this paper, input data of the problem were chosen so that the flow parameters were similar to those obtained experimentally in [2]. We started from the Reynolds-averaged Navier–Stokes equations and the modified two-parameter turbulence model $q - \nu$ (see [11]).

The first problem was solved using a fifth-order upwind difference scheme with compact approximations. It was described in detail in [12].

2. Consider a multioperator scheme of the seventh order in space with upwind compact approximations other than those used in [5–8]. We applied this scheme to the linearized Euler equations, but it can also be applied to various problems in fluid dynamics.

Following the notation used in [6], consider three-point operators defined on the grid $\omega_h = \{x_j = ih, j = \pm 1, \pm 2, \ldots, h = \text{const}\}$ in the form of linear combinations of the identity operator I and three-point central differences Δ_0 and Δ_2, where

$$\Delta_0 = T_1 - T_{-1}, \quad \Delta_2 = T_1 - 2I + T_{-1}, \quad T_{\pm 1} v_j = v(jh \pm h).$$

To construct multioperators, the original one-parameter family of compact approximations defined on ω_h has the form

$$\left[\frac{\partial u}{\partial x}\right]_j = L_5(s)u_j + O(h^5),$$
$$L_5(s) = [\Delta(s) + sR_1^{-1}(s)Q_1(s)\Delta_2]/h, \tag{2}$$

where $\Delta(s) = 0.5(\Delta_0 - s\Delta_2)$, s is a parameter, and $Q_1 = \tilde{Q}_1 (I + \Delta_2/12)^{-1}$. The three-point operators R_1 and $\tilde{Q}_1$ are given by

$$R_1(s) = I + \frac{1}{6s}\Delta_0 + \frac{1}{5}\Delta_2, \quad \tilde{Q}_1(s) = I + \left(\frac{17}{60} - \frac{1}{9s^2}\right)\Delta_2.$$

The action of the operator $L_5(s) = [\Delta(s) + sR_1^{-1}(s)Q_1\Delta_2]/h$ on some grid function u_j ($j = 0, 1, \ldots, N$) is calculated via the following operations:

(i) Calculate $v_j = (I + \Delta_2/12)^{-1}\Delta_2 u_j$ by inverting the tridiagonal matrix $(I + \Delta_2/12)$.

(ii) Calculate $w_j = R_1^{-1}(\tilde{Q}_1 v_j)$ again by converting a tridiagonal matrix.

(iii) Calculate $[\Delta(s)u_j + w_j]/h$.

While inverting a matrix, we always assume that, depending on the statement of the problem, some boundary conditions are set at $j = 0$ and $j = N$.

By using the one-parameter family of operators $L_5(s)$, the first derivatives can be approximated to an arbitrarily high order in the form of linear combinations of the basis operators $L_5(s_i)$, $i = 1, 2, \ldots, M$.

Following [6], the action of $L_5(s)[u]_h$ on a sufficiently smooth function $u(x)$ considered at grid points is represented as a series in powers of h:

$$L_5(s)[u]_h = u_x^{(1)} + p_1(s)u_x^{(6)}h^5 + p_2(s)u_x^{(7)}h^6 + \ldots + p_m(s)u_x^{(m+5)}h^{m+4} + O(h^{m+5}), \tag{3}$$

where $p_k(s)$ are polynomials of degree k. Specifically,

$$p_1(s) = \frac{1}{135s} - \frac{s}{120}, \quad p_2(s) = \frac{41}{12600} - \frac{1}{405s^2}, \quad p_3(s) = \frac{160s^{-3} - 291s^{-1} + 54s}{194400}.$$

Restricting our consideration to seventh-order multioperators, we fix three values s_1, s_2, and s_3 of s and substitute them into (3). The resulting equalities are multiplied by the coefficients γ_1, γ_2, and γ_3, where $\gamma_1 + \gamma_2 + \gamma_3 = 1$, and are added up. Since the coefficients of h^5 and h^6 vanish, taking into account the equality for their sum, which guarantees that the coefficient of the first derivative is 1, we obtain the following system for determining γ_1, γ_2, and γ_3:

$$\sum_{i=1}^{3}\gamma_i = 1, \quad \sum_{i=1}^{3}\gamma_i p_1(s_i) = 0, \quad \sum_{i=1}^{3}\gamma_i p_2(s_i) = 0. \tag{4}$$

In contrast to systems for other basis operators [5–8], this system is not reduced to one with a Vandermonde matrix and the existence of its solution for any s_1, s_2, and s_3 is not guaranteed. However, direct calculations show that the determinant of its matrix is nonzero if s_1, s_2, and s_3 are not related by the formula

$$s_1s_2 + s_1s_3 + s_2s_3 + 8/9 = 0. \tag{5}$$

Specifically, equality (5) is not satisfied if

$$s_2 = \frac{s_1 + s_3}{2}.$$

Although the systems for γ generated by the family $L_5(s)$ may have nonsingular matrices, their doubtless advantage is that the condition number does not increase sharply with increasing dimension, which is typical of Vandermonde matrices.

Consider some properties of the operator $L_M(s_1, s_2, s_3)$ that follow from the properties of basis operators.

Skew-Symmetric and Self-Adjoint Components

Meaning the Hilbert space of square summable grid functions defined on ω_h, we write the operator $L_5(s)$ in the form of the sum of its skew-symmetric and self-adjoint parts (denoted by indices (1) and (0), respectively):

$$L_5(s) = L_5^{(1)}(s) + L_5^{(0)}(s).$$

Accordingly, L_M can be represented as

$$L_M(s_1, s_2, s_3) = \sum_{i=1}^{3}\gamma_i L_5^{(1)}(s_i) + \sum_{i=1}^{3}\gamma_i L_5^{(0)}(s_i).$$

Algebraic rearrangements using the permutability of Δ_0 and Δ_2 yield

$$L_5^{(1)}(s) = L_5^{(1)}(-s), \quad L_5^{(0)}(s) = -L_5^{(0)}(-s). \tag{6}$$

Consider $L_M(-s_1, -s_2, -s_3)$. It follows from (4) that the solution γ_i ($i = 1, 2, 3$) does not change after the substitution $s_i = -s_i$. This means that, by virtue of (6),

$$L_m(-s_1, -s_2, -s_3) = \sum_{i=1}^{3}\gamma_i L_5^{(1)}(s_i) - \sum_{i=1}^{3}\gamma_i L_5^{(0)}(s_i),$$

i.e., the skew-symmetric part of L_M is invariant under a change in the sign of the parameters, while the self-adjoint part reverses its sign. Thus, if the multioperator is positive for some parameter values, it becomes negative after changing their signs.

Representability of the Action of Multioperators as the Difference of "Fluxes"

The action of L_m on a grid function can be represented as

$$L_M u_j = q_{M,j+1/2} - q_{M,j-1/2},$$

where $q_{M,j+1/2}$ ($j = 0, \pm1, \pm2, \ldots$) are "fluxes" through the imaginary faces $x_{j\pm1/2}$ of the control cell. This property is derived from the easy-to-establish equality

$$L_5(s)u_j = q_{j+1/2}(s) - q_{j-1/2}(s)$$

by summation

$$q_{M,j+1/2} = \sum_{i=1}^{3} \gamma_i q_{j+1/2}(s_i).$$

This property suggests that conservative schemes can be constructed.

In what follows, for convenience, we use a uniform distribution of the parameters such that s_2 is the half-sum of s_1 and s_3. Then the multioperator becomes dependent on two parameters s_1 and s_3.

Dispersion and Dissipative Properties

Denote by $\mu_M(\alpha; s_1, s_3)$ ($\alpha = kh$) the Fourier transform of the operator L_M with respect to k. It can be obtained by making the formal substitutions

$$\Delta_2 \longrightarrow e^{i\alpha} - 2 + e^{-i\alpha}, \quad \Delta_0 \longrightarrow e^{i\alpha} - e^{-i\alpha}.$$

Obviously, considering the space of bounded grid functions defined on ω_h, we conclude that μ_M is an eigenvalue of L_M with the eigenfunction $e^{i\alpha j}$:

$$L_M e^{i\alpha j} = \mu_M e^{i\alpha j}, \quad j = 0, \pm 1, \pm 2, \ldots.$$

Consider a multioperator approximation of the x-derivative for the Cauchy problem

$$u_t + a u_x = 0, \quad u(0, x) = e^{ikx}$$

without discretizing the time derivative

$$(u_j)_t + a L_M(s_1, s_3) u_j = 0, \quad u(0, x_j) = e^{i\alpha j},$$

where $\alpha = kh$ is the dimensionless wave number. The solution to this difference problem has the form

$$u_j(t) = e^{-\mathrm{Re}\,\mu_M t} e^{ik(x_j - a_* t)}, \tag{7}$$

where $a_* = a\,\mathrm{Im}\,\bar{\mu}_M(\alpha, s_1, s_3)/\alpha$ is the "numerical" phase velocity and $\bar{\mu}_M = h\mu_m$. In what follows, it is always assumed that s_1 and s_3 are such that $\mathrm{Re}\,\mu_M(\alpha; s_1, s_3) > 0$ for $0 < \alpha \le \pi$. This means that the multioperator is positive for these parameter values, which is important in the design of robust schemes for fluid dynamics equations. The domains of "admissible" values of s_1 and s_3 in the plane $\Omega = (s_1, s_3)$ can be found by numerically solving the problem of determining domains where a function of two variables is positive.

To describe acoustic fields accurately, it is important that the phase and amplitude errors be close to zero and that numerical solution (7) be different from the exact one in a possibly wider range of wave numbers $\alpha = kh$ in the interval $[0, \pi]$ supported by a grid with the mesh size h. This means that the damping factor of the harmonic $\mathrm{Re}\,\hat{L}_M$ is small for this range (i.e., the amplitude errors are small) and a_* differs little from a (i.e., the phase errors are small). From the point of view of acoustics, the dissipative mechanism defined by $\mathrm{Re}\,\hat{L}_M$ is an enforced measure dictated by the stability of the computations and by the filtration of spurious short-wave oscillations.

Small phase and amplitude errors for $0 \le kh \le \pi/2$ are ensured, to some extent, by high-order schemes. However, in aeroacoustic problems, it is desirable to extend this range. To do this, we can use the fact that the multioperator parameters $s_1, s_3 \in \Omega$ are still arbitrary. Figure 1 shows $d = \mathrm{Re}\,\mu_M(\alpha)$ and $r = a_*/a = \mathrm{Im}\,\mu_M(\alpha)/a$ as functions of the wave number $\alpha = kh$ at $s_1 = 0.82$ and $s_3 = 10$.

These parameters were chosen by searching through admissible values, and they are not necessarily optimal. However, Fig. 1 reveals that $\mathrm{Re}\,\mu_M(\alpha)$ differs from zero and a_*/a differs from unity in a rather wide range of α.

3. When multioperator schemes are constructed for equations with convective terms, it is convenient to use flux splitting. Let the operator $L_5(s_1, s_3)$ be positive. Denote it by L^+. Then the operator $L^- = L_5(-s_1, -s_3)$ is negative. In the case of the nonlinear equation

$$u_t + f(u)_x = 0,$$

the spatial derivative is approximated up to $O(h^7)$ as follows:

$$f(u)_x \approx \frac{1}{2}[L^+(f(u) + Cu) + L^-(f(u) - Cu)] = \frac{1}{2}(L^+ + L^-)f(u) + \frac{1}{2}C(L^+ - L^-)u, \tag{8}$$

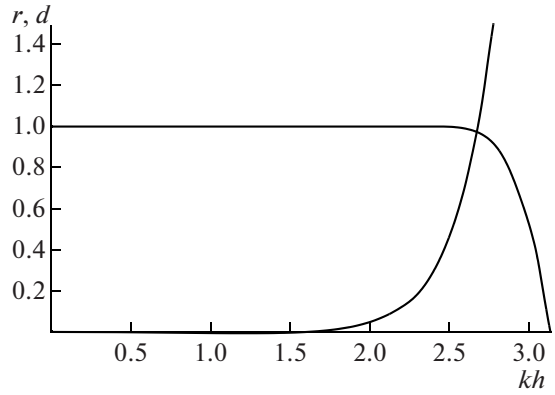


Fig. 1. Phase velocity and dissipation of the seventh-order accurate multioperator scheme.

where $C > 0$ is a parameter. In the case of the linear problem $f(u) = au$ with $a = \text{const}$, the approximating operator is positive, since $L^+ - L^- = L_M^{(0)}(s_1, s_3) > 0$. The skew-symmetric operator $(L^+ + L^-)/2$ is used to approximate the derivatives, while the second term on the right-hand side of (8) serves as a dissipative addition. Approximation (8) is easily extended to vector functions u and f: it is sufficient to replace the constant C with the matrix CE, where E is the identity matrix. In all the cases, C can be used as a parameter governing the level of dissipation. Specifically, when $C = 0$, we have a nondissipative symmetric approximation.

In equations with several spatial variables, the derivative with respect to each variable is approximated in a similar manner.

To compute the action of the multioperator in the case of a single processor, the action of basis operators has to be calculated three times by setting $s = s_1$, $s = s_2$, and $s = s_3$. In the case of a parallel computer system, these computations can be executed in parallel. Moreover, if the time required for data exchange is ignored, the actions of the multioperator are computed in the same time as the actions of the basis operators.

The accuracy of the multioperator scheme can be estimated by applying it to the widely used test problem for Burgers' equation (see, e.g., [13])

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\frac{u^2}{2} = 0, \qquad (9)$$
$$u(0, x) = 1 + 0.5\sin(\pi x), \quad -1 \le x \le 1,$$

with periodicity conditions. The exact solution to problem (9) is smooth up to some time and can be obtained by the method of characteristics. Its value at nodes is denoted by u_j^{ref}.

The computations were based on the fourth-order Runge–Kutta method. To eliminate the influence of the time step on accuracy, we used relatively low Courant numbers. The accuracy and its order were estimated as follows:

$$E_c(N) = \max_j \left|u_j - u_j^{\text{ref}}\right|, \quad k_c = \log_2 \frac{E_c(N)}{E_c(2N)},$$

where N is the number of grid points. The table presents E_c and k_c for various N in the case of a smooth exact solution ($t = 0.3$) for the multioperator scheme (the multioperator is denoted by L_{57}). The table also lists the results obtained in [8] for the scheme with the basis operator L_5 and for the fifth-order WENO scheme. The results in the table suggest that, for coarse grids ($N = 8, 16$), the error in the multioperator scheme is only several times less than the errors in the fifth-order schemes. However, for $N \ge 32$, the accuracy of the former increases rapidly, exceeding that of the latter by several orders of magnitude (roughly 4 orders of magnitude for $N = 256$).

Table

Schemes \ N		8	16	32	64	128	256
WENO-5	E_c	6.72e-2	1.25e-2	1.20e-3	9.50e-5	3.31e-6	8.66e-8
	k_c		2.4	3.7	3.7	4.8	5.3
L_5	E_c	3.35e-2	6.24e-3	4.19e-4	3.22e-6	4.13e-7	1.22e-8
	k_c		2.71	3.81	4.74	5.00	5.03
L_{57}	E_c	1.82e-2	2.14e-3	3.61e-5	2.21e-7	8.03e-10	3.38e-12
	k_c		3.1	5.9	7.3	8.1	7.9

The principle described was used to approximate the spatial derivatives in the linear system. Although the base flow was axisymmetric, acoustic fields were computed with a three-dimensional code with the purpose of possible generalizations to three dimensions.

3. NUMERICAL RESULTS

Mean Flow Computation

The computations were performed in the experimental conditions from [2]. We considered a supersonic jet outflowing from an axisymmetric nozzle with distributions of supersonic flow parameters specified in some internal nozzle cross section. The boundary conditions were chosen so that the flow was close to that studied in [2] starting at some distance from the nozzle exit. The computational domain and the grid used are shown in Fig. 2. The averaged Navier–Stokes equations were supplemented by two equations of the modified turbulence model $q - \nu$ (see [11]). The convective terms of the equations were discretized by applying fifth-order accurate upwind compact approximations, while the diffusion terms were approximated by fourth-order centered differences. Time marching was based on the scheme described in [12].

In Figs. 3 and 4, the numerical results (solid lines) are compared with the experimental data from [2] (markers). Figure 3 shows the Mach number distribution along the jet axis. Since the nozzle contour was not intended for ideal flow expansion (as occurred in the experiment), nonmonotone variations in the Mach number, which suggest the presence of rolls, are observed near the nozzle exit in Fig. 3. Further downstream, the computed and experimental distributions agree fairly well.

Figure 4 displays the Mach number distributions in several cross sections where measurements were conducted. It can be seen that the numerical results also agree with the experimental data.

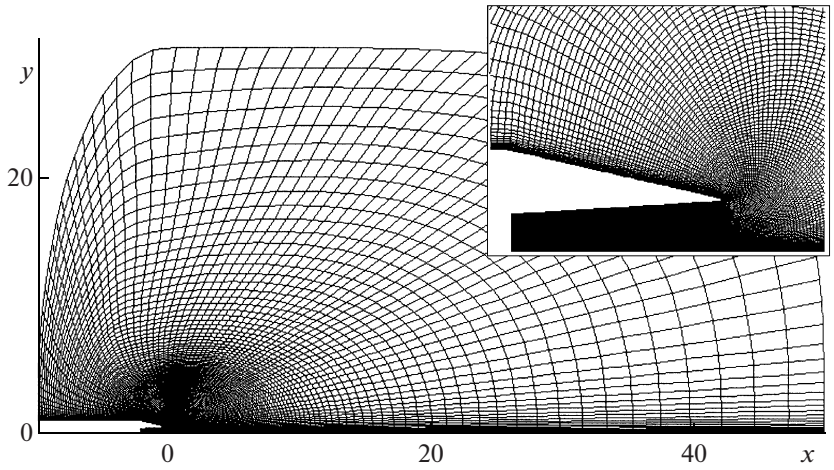


Fig. 2.

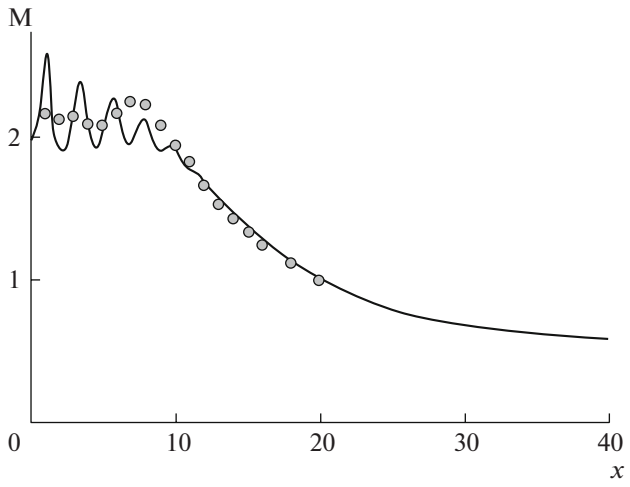


Fig. 3.

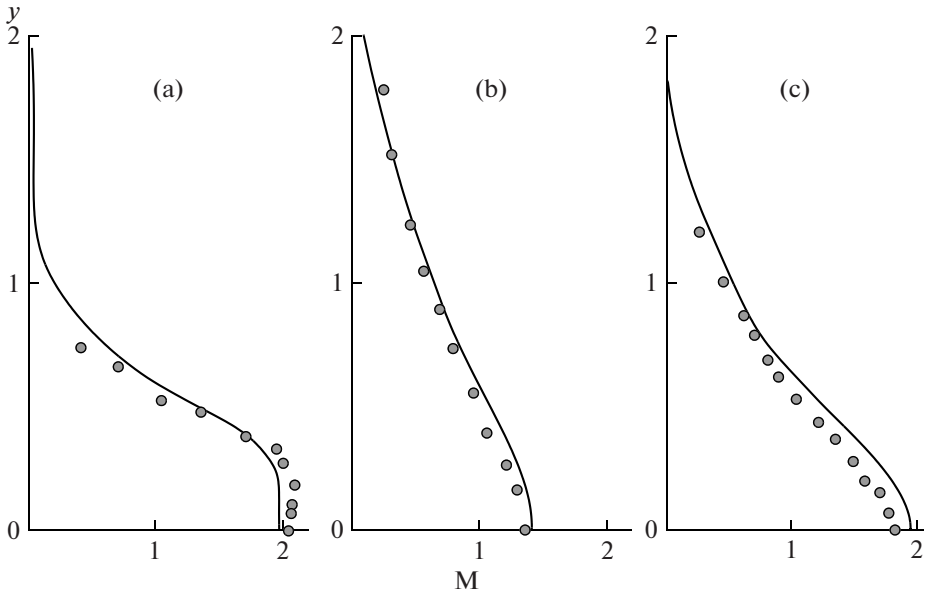


Fig. 4. (a) $x/D = 5$, (b) $x/D = 10$, and (c) $x/D = 15$.

Computation of Acoustic Fields

The turbulent flow parameters obtained behind the shock region near the nozzle exit were used in system (1). Obviously, this linear system has the trivial solution in the absence of perturbation sources. A periodic perturbation (e.g., of density or pressure) specified in the initial cross section propagates downstream and, under certain conditions, generates an acoustic field outside the jet.

In theoretical studies [2, 3], an approximate solution to system (1) was constructed by the method of matched asymptotic expansions, in which the small parameter was that characterizing the slow expansion of the supersonic jet. Due to this approach, not only acoustic field parameters similar to experimental data were obtained in the plane and axisymmetric cases but also the subtle mechanism of acoustic field generation was described. The numerical solution to problem (1) gives no information on this mechanism, but it makes no use of any assumptions, except the only one that the acoustic field is described by linearized Euler equations. It can be assumed that, in this approximation, high-accuracy numerical methods produce fairly accurate acoustic field parameters in the general three-dimensional case for various shapes and characteristics of the jet and for various initial perturbations. At the same time, the above linear model

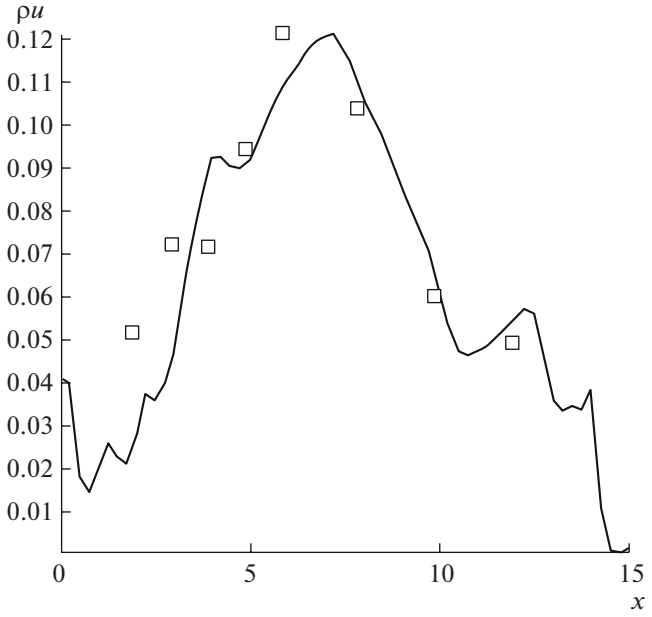


Fig. 5.

eliminates certain effects of the nonlinear interaction between waves themselves and between waves and the mean flow.

The above-described seventh-order multioperator approximations and the fourth-order Runge–Kutta method for time stepping were used to develop a computer code for the general three-dimensional problem. In this study, the code was used to test the performance of the method by comparing it with the acoustic fields measured in [2].

Let us discuss some details of the numerical implementation of the scheme. In contrast to the mesh used for the mean flow, the mesh for solving the linear system was constructed in Cartesian coordinates x, y, z with the x axis coinciding with the jet axis. As a result, jets outflowing from nozzles with arbitrary cross-sectional shapes could be considered. Near the jet boundary, the mesh was refined by transforming y and z with uniform meshes in the new coordinates. The base flow fields on these meshes were calculated by interpolating grid functions defined on the grid shown in Fig. 2.

In the experiment in [2], the perturbation at the nozzle exit was either specified as white noise or was produced by a glow discharge generated periodically at a fixed frequency. In this paper, in at the given Strouhal number $\mathrm{St} = \omega D/(2\pi U)$ (U is the velocity on the jet axis in the initial cross section and D is the jet diameter), the perturbation the initial cross section was specified as

$$p(r, t) = Af(r)\sin(\omega t + n\varphi), \quad r = (x^2 + y^2)^{1/2}, \quad A = \text{const},$$

where n is the index of an azimuthal mode, φ is the azimuthal angle, ω is the frequency, and $f(r)$ is a smooth function that is nonzero inside the jet. White noise perturbations were produced as linear combinations of harmonics with random phases. Thus, the solution of the linear problem was determined up to a constant A. Since the perturbation amplitudes in [2] were unknown, this constant was specified, following [1], by using experimental and numerical data coinciding at some fixed point of the domain.

The fluctuation mass flux distribution ρu in a shear layer was presented in [2]. Figure 5 compares this function with its values computed for the axisymmetric mode ($n = 0$) at $\mathrm{St} = 0.2$. Since these values were determined up to a constant, they were normalized so that their maxima coincided with those in the experimental data. This technique was used in [1] in the construction of analytical solutions. It can be seen that the shapes of both distributions and their maxima are fairly similar to each other. A maximum suggests that the amplitudes of instability waves grow and then decrease because of the increased thickness of the shear layer at the jet boundary. The location of these maxima was interpreted in [2] as an indication of a domain where sound is generated. Figure 6 presents a similar comparison in the case of a white noise perturbation specified in the initial cross section. This corresponds to the experiment in [2] with no artificial glow discharge excitation. It can be seen that the maxima have noticeably different locations, but the shapes of the curves remain similar. Finally, Fig. 7 displays the computed distribution ρu along the jet axis

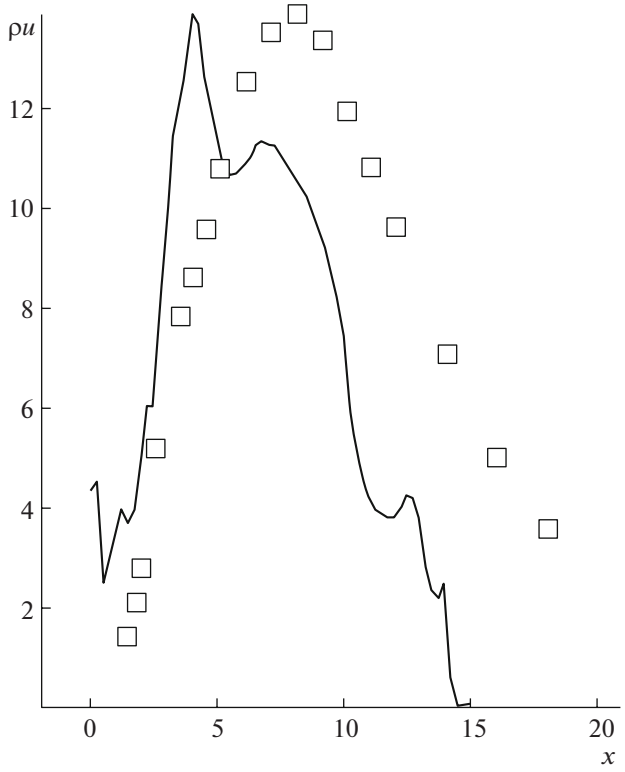


Fig. 6.

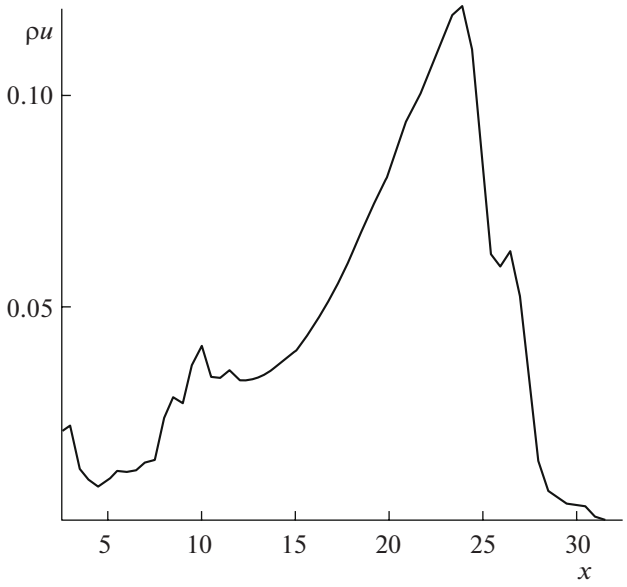


Fig. 7.

for the first azimuthal mode ($n = 1$) at the same perturbation frequency (St = 0.2). No experimental data are available for this case in [2]. It can be seen that the maximum of the curve shifts considerably downstream.

Consider the acoustic fields emitted by the jet. Figures 8–10 show the instantaneous acoustic pressure patterns in a meridional plane for three above cases, respectively: St = 0.2 and $n = 0$, St = 0.2 and $n = 1$,

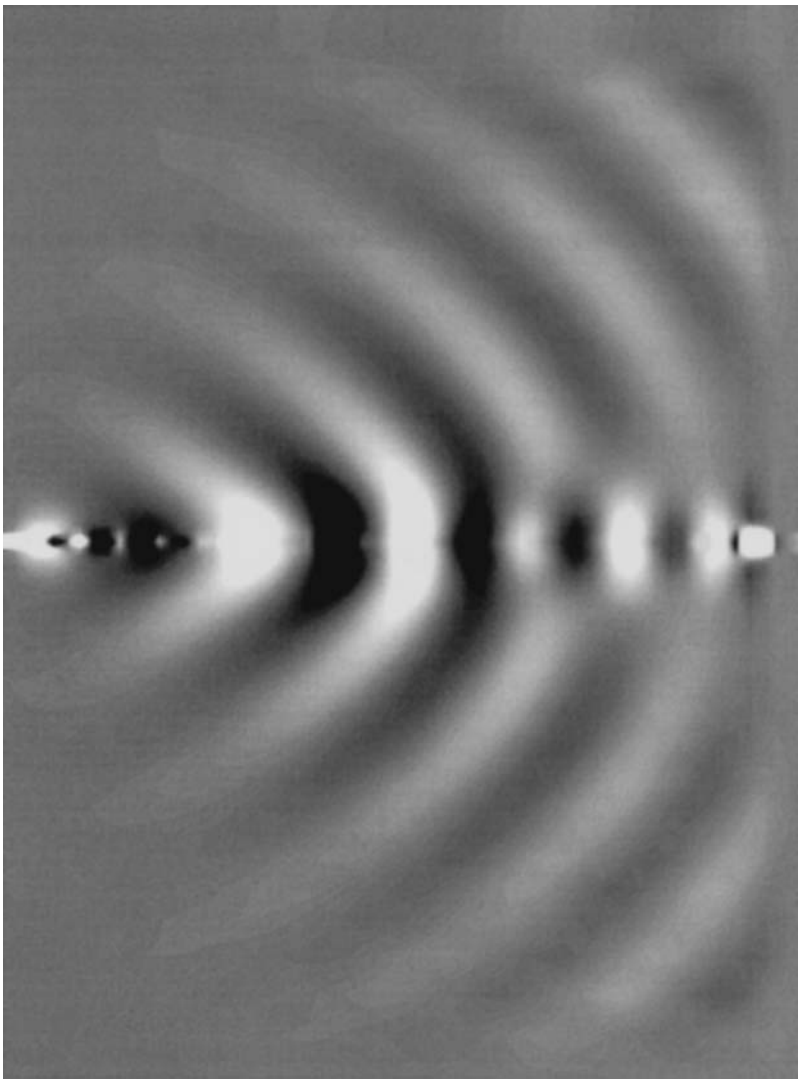


Fig. 8.

and white noise in the case of an axisymmetric initial perturbation ($n = 0$). In Fig. 8, the contrasts of light and dark grades increase and then decrease along the jet axis. This corresponds to a maximum in the fluctuation mass flux distribution presented in Fig. 5 and is in complete agreement with the evolution of instability waves as the shear layers at the jet boundary become thicker (see [1]). The near-field pattern of periodic contrasts characterizes sound emitted by the zero mode downstream at some angle to the jet axis. In the case of the first azimuthal mode, the pressure pattern (Fig. 9) is entirely different. First, the areas of maximum and minimum pressures inside the jet alternate in a staggered manner. Second, the fluctuation intensity in the near field is noticeably lower than that inside the jet. This means that a considerably lower fraction of energy is emitted as sound by the azimuthal mode. This conclusion agrees with the experimental data in [3], which suggest that, for a white noise initial perturbation, the azimuthal modes contribute insignificantly to the intensity of emitted sound away from the nozzle exit. The pressure field in the case of initial white noise is shown in Fig. 10. The plot suggests that the sound propagates approximately in the same direction as in the case of $n = 0$ represented in Fig. 8. This also agrees with the conclusions in [2] that the directional diagram depends weakly on the character of perturbations generated at the nozzle exit and, therefore, the mechanism of sound emission is the same in all the cases. In contrast to Fig. 8, the field in Fig. 10 lacks periodic locations of maxima and minima, which suggests the superposition of many harmonics. Finally, Fig. 11 displays the computed directional diagram for the meridional mode at $\mathrm{St} = 0.2$ (emitted sound

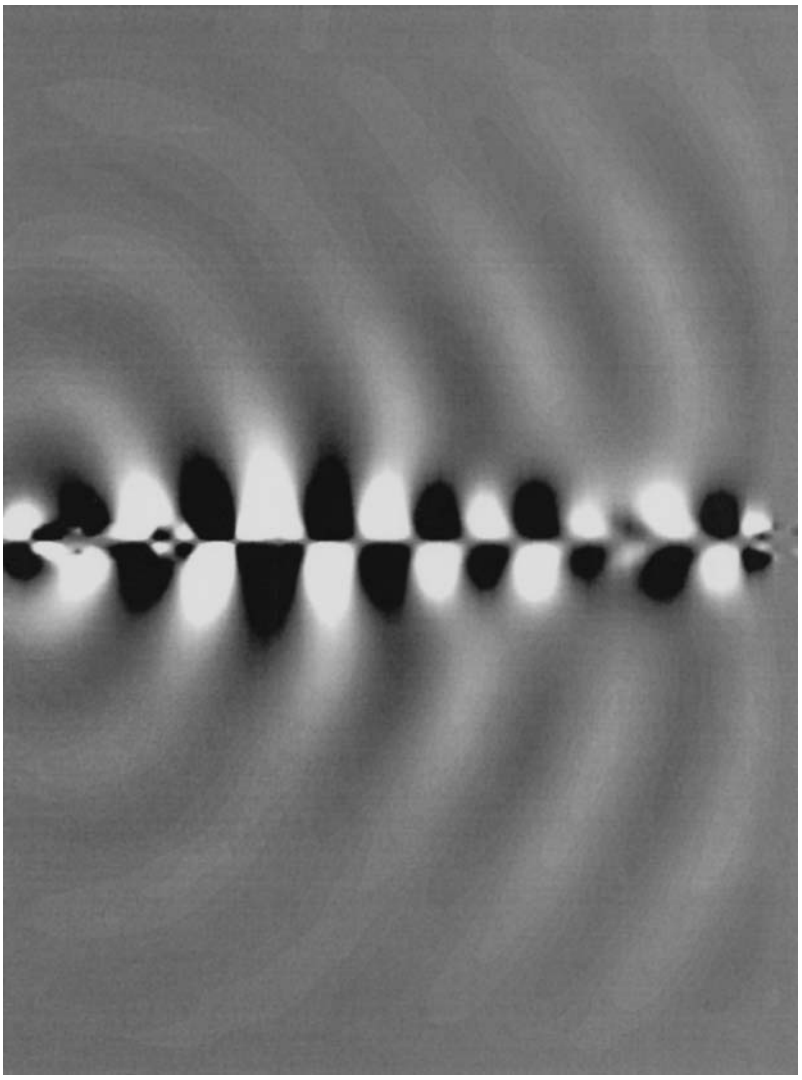


Fig. 9.

levels are presented in decibels). Figure 11 also shows the experimental data from [2], which are similar to the numerical results (markers).

Below are some conclusions.

1. The numerical results agree well with the experimental data, which suggests that linearized Euler equations can be used to simulate acoustic fields for more general supersonic jets.

2. The numerical results also agree with the mechanism of sound emission by instability waves described in [1], which is associated with an increase and then a decrease in the wave amplitudes in the course of thickening shear layers. Accordingly, agreement is also observed with the results produced by the method of matched asymptotic expansions in the axisymmetric case for experimental conditions.

3. A comparison of the linearized model with direct numerical simulation of jet flows suggests the following conclusions. In principle, direct simulation can produce rather complete information on the near-field flow, including acoustic fields. Specifically, this information can include sound self-generation under certain conditions due to shear layer instability, the nonlinear interaction of waves, and data on the role of turbulent fluctuations. However, concerning the complete resolution of all flow scales, the applicability of direct numerical simulation is now restricted to rather low Reynolds numbers even in the case of high-order accurate schemes with a high resolution.

The linearized model has a considerably smaller domain of application. For example, it can hardly be used for subsonic jets. Moreover, it can be treated only as a mechanism of gaining and emitting already

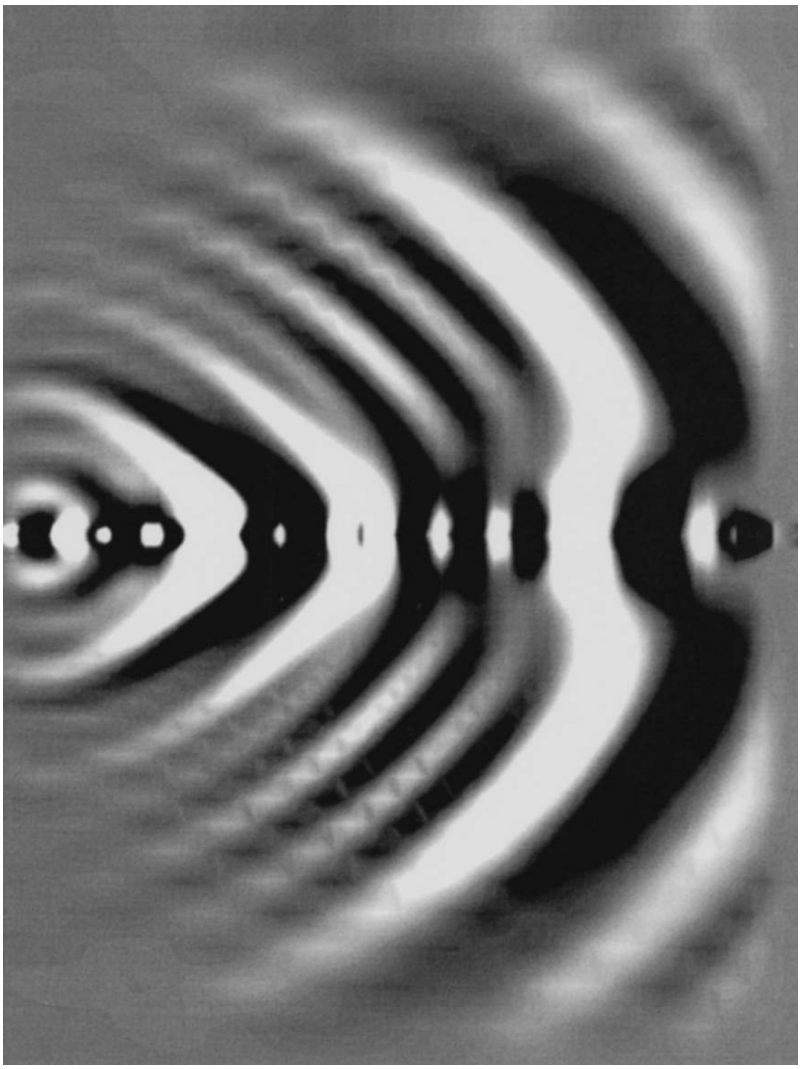


Fig. 10.

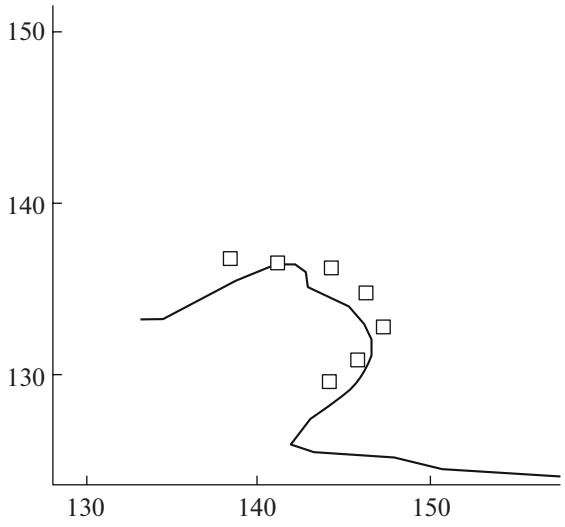


Fig. 11.

available perturbations in the framework of a broadband filter. At the same time, the linearized model is easy to implement in the general three-dimensional case with the use of high-order schemes, since it does not require that the smallest scales be resolved. Moreover, turbulence can be roughly taken into account using the easy-to-implement model of averaged Navier–Stokes equations (with a suitable semiempirical turbulence model), which is used to describe the "base" flow in the jet.

Thus, the described method of application of the high-accuracy schemes for some problems of an aeroacoustics is represented reasonable enough.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 08-01-00354, 06-08-01628) and by the Department of Mathematical Sciences of the Russian Academy of Sciences (project no. 3).

REFERENCES

1. C. K. W. Tam and D. E. Burton, "Sound Generated by Instability Waves of Supersonic Flows: Part 2. Axisymmetric Jets," J. Fluid Mech. **138**, 273–295 (1984).
2. T. R. Troutt and D. K. McLouphlin, "Experiments on the Flow and Acoustic Properties of a Moderate-Reynolds-Number Supersonic Jet," J. Fluid Mech. **116**, 123–156 (1982).
3. V. F. Kopiev, S. A. Cherneshev, M. Yu. Zaitsev, and V. M. Kusznetsov, "Experimental Validation of Instability Wave Theory for Round Supersonic Jet," AIAA Paper, 2006–2595 (2006).
4. C. K. W. Tam and J. C. Webb, "Dispersion Relation Preserving Finite Difference Schemes for Computational Acoustics," J. Comput. Phys. **194**, 194–214 (2004).
5. A. I. Tolstykh, "Multioperator Schemes of Arbitrary Order Based on Noncentered Compact Approximations," Dokl. Akad. Nauk **366**, 319–322 (1999) [Dokl. Math. **59**, 409–412 (1999)].
6. A. I. Tolstykh, "Construction of Schemes of Prescribed Order of Accuracy with Linear Combinations of Operators," Zh. Vychisl. Mat. Mat. Fiz. **40**, 1206–1220 (2000) [Comput. Math. Math. Phys. **40**, 1159–1172 (2000)].
7. M. V. Lipavskii and A. I. Tolstykh, "Fifth- and Seventh-Order Accurate Multioperator Compact Schemes," Zh. Vychisl. Mat. Mat. Fiz. **43**, 1018–1034 (2003) [Comput. Math. Math. Phys. **43**, 975–990 (2003)].
8. M. V. Lipavskii, A. I. Tolstykh, and E. N. Chigarev, "A Parallel Computational Scheme with Ninth-Order Multioperator Approximations and Its Application to Direct Numerical Simulation," Zh. Vychisl. Mat. Mat. Fiz. **46**, 1433–1452 (2006) [Comput. Math. Math. Phys. **46**, 1359–1377 (2006)].
9. A. I. Tolstykh, "Development of Arbitrary-Order Multioperator-Based Schemes for Parallel Calculations. 1: Higher-Than-Fifth Order Approximations to Convection Terms," J. Comput. Phys. **225**, 2333–2353 (2007).
10. A. D. Savel'ev, "Numerical Simulation of Axisymmetric Afterbody Flows with Jet Exhaust," Zh. Vychisl. Mat. Mat. Fiz. **47**, 310–320 (2007) [Comput. Math. Math. Phys. **47**, 301–310 (2007)].
11. A. D. Savel'ev, "Computations of Viscous Gas Flows Based on the $q - \nu$ Turbulence Model," Zh. Vychisl. Mat. Mat. Fiz. **43**, 589–600 (2003) [Comput. Math. Math. Phys. **43**, 564–574 (2003)].
12. A. D. Savel'ev, *High-Order Accurate Difference Schemes with Composite Stabilizing Additions* (Vychisl. Tsentr Ross. Akad. Nauk, Moscow, 2003) [in Russian].
13. N. A. Adams and K. A. Sharif, "A High-Resolution Compact-ENO Scheme for Shock-Turbulence Interaction Problems," J. Comput. Phys. **127**, 27–51 (1996).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1235–1244. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © V.N. Diesperov, G.L. Korolev, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1295–1305.

Triple-Deck Analysis of Formation of Supersonic and Local Separation Regions in Transonic Steady Flow over a Roughness Element on the Surface of a Body of Revolution

V. N. Diesperov[a] and G. L. Korolev[b]

[a] *Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudnyi, Moscow oblast, 141700 Russia*
e-mail: diesvn@list.ru

[b] *Zhukovsky Central Institute of Aerohydrodynamics, Zhukovskii, Moscow oblast, 140160 Russia*
e-mail: glk777@mail.ru

Received December 26, 2008

Abstract—Transonic axisymmetric flow over a body of rotation with a small roughness element located on its surface is considered. The body is manly cylindrical. The roughness height is assumed to be much smaller than the radius of the cylinder and such that a triple-deck flow is induced in its neighborhood. The goal of the work is to study the effect of the cylinder radius and the roughness shape on the triple-deck flow when the cylinder radius is of the same order as the transverse size of the triple-deck interaction region. In this case, the effect of three-dimensionality of the flow is exhibited even in the first approximation. Special attention is given to the structure of supersonic regions and closing shock waves arising in the outer potential region, as well as to local separation regions if they develop in the lower viscous boundary sublayer. Specifically, it is shown that, as the radius of the cylinder increases at a fixed roughness height, the shock intensity grows considerably, whereas the position of the main shock varies little.

DOI: 10.1134/S0965542509070148

Key words: transonic steady flow problems, supersonic region formation, local separation regions, unbounded layer problem.

INTRODUCTION

Experimental studies [1] have shown that the interaction of an incident shock wave with a laminar boundary layer is of a complicated nature and goes beyond the classical view of the boundary layer. It was also shown that the boundary layer has a large effect on the formation of a transonic flow. Further studies confirmed these conclusions (see [2]). The triple-deck theory with equations describing the interaction mechanism of a shock wave with a boundary layer was developed in [3, 4]. In the case of a transonic steady outer flow and a laminar boundary layer, the triple-deck equations were derived in [5]. Their numerical solution [6–9] was used to analyze the behavior of skin friction. The conclusions drawn in [1, 3, 4] were in complete agreement with the results of experiments in [1].

The triple-deck studies of the influence exerted on a supersonic flow by a small roughness surface element located in the lower viscous sublayer were begun in [10, 11].

In the case of an incompressible flow, those studies were continued in [12–15]. Axisymmetric supersonic flows were analyzed in [16]. A review of the results can be found in [17]. The flow pattern arising in the flow over a three-dimensional roughness element located on a body of revolution was also investigated in [17].

A triple-deck analysis of transonic steady flow over a small roughness element was performed in [18, 19]. It was found that a closed supersonic region with two shock waves is formed in the outer potential region. These results agree well with experiments [1]. For some relations between the flow parameters, a local separation region is also formed in the lower viscous sublayer.

The formation and development of supersonic regions, closing shock waves, and local separation regions in unsteady transonic flow over the same small roughness elements as in [18, 19] were studied in [20].

STATEMENT OF THE PROBLEM

We analyze a transonic laminar axisymmetric gas flow over a small roughness element located on a body of revolution at a distance L from its leading edge. It is assumed that the incoming flow is homogeneous and that the body has an extended cylindrical segment that is much longer than the roughness element. The radius of the cylindrical part of the body is denoted by R_w. It is assumed that $\delta = R_w/L \ll 1$. The Reynolds number $\mathrm{Re} = \rho_\infty U_\infty L/\mu_\infty$ is determined by the gasdynamic characteristics in the incoming flow (parallel to the axis of rotation) and is assumed to tend to infinity. Here, ρ_∞, U_∞, and μ_∞ are the free-stream density, velocity, and viscosity, respectively. Denote by x^* and r^* cylindrical coordinates with the origin at the nose of the body.

It is assumed that δ is small but much larger than the boundary layer thickness; i.e., $\delta \gg O(\mathrm{Re}^{-1/2})$.

As a result, the incoming flow is weakly perturbed and the boundary layer in the first approximation is described by the two-dimensional boundary layer equations and by the Blasius solution. The roughness height is assumed to be $O(L\mathrm{Re}^{-3/5})$. It is also assumed that the free-stream Mach number M_∞ approaches unity as $\mathrm{Re} \longrightarrow \infty$ and $\mathrm{M}_\infty^2 = 1 + O(\mathrm{Re}^{-1/5})$.

Under these assumptions, the roughness element induces a viscous-inviscid interaction between the incoming transonic flow and the boundary layer. A triple-deck region develops, whose longitudinal size is $O(\mathrm{Re}^{-3/10})$ (see [5]). It consists of the lower boundary sublayer adjacent to the body surface, the main part of the boundary layer, and the outer sublayer adjacent to the free stream and the boundary layer.

Following the ideas of [5], we briefly describe the derivation of the triple-deck equations and the boundary and initial conditions for a roughness element located on the surface of a body of revolution placed in the lower viscous sublayer. The variables are expanded in the form

$$x^* = L(1 + \mathrm{Re}^{-3/10}x), \quad r^* = L(\delta + \mathrm{Re}^{-3/5}y^\circ).$$

Assume that the shape of the roughness element in dimensionless Cartesian coordinates (x, y°) is specified by the equation $y^\circ = f(x)$, where $f(x)$ is a smooth function. The boundary layer equations and boundary conditions in the lower viscous sublayer are most conveniently analyzed when expressed in variables x and y, where $y = y^\circ - f(x)$. In this case, the flow variables are expanded in the form

$$u^* = U_\infty \mathrm{Re}^{-1/10}u + \ldots, \quad v^* = U_\infty \mathrm{Re}^{-2/5}v + \ldots,$$

$$p^* = p_\infty + \rho_\infty U_\infty^2 \mathrm{Re}^{-1/5}p + \ldots, \quad \rho^* = \rho_\infty\rho + \ldots.$$

Here, p^* and ρ^* are the pressure and density of the gas and u^* and v^* are the flow velocity components in coordinates (x^*, r^*). For simplicity, we assume that the body surface is heat-insulated, the Prandtl number is $\mathrm{Pr} = 1$, and the first viscosity is $\mu = CT$ (where C is the Chapman constant). Then, according to Crocco's relation [21], the temperature on the body surface is a constant. Denote by $R(0)$ the dimensionless density on the body surface. In the approximation used, the equation of state implies that $\rho = R(0)$. Substituting the above expansions in the lower viscous sublayer into the Navier–Stokes equations, in the first approximation, we obtain the Prandtl equation for an incompressible boundary layer:

$$\begin{gathered}\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad \frac{\partial p}{\partial y} = 0,\\ u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{R(0)}\frac{\partial p}{\partial x} + \frac{C}{R^2(0)}\frac{\partial^2 u}{\partial y^2}.\end{gathered} \tag{1.1}$$

For $x \longrightarrow -\infty$, the solution of system (1.1) in the lower viscous sublayer is matched with the Blasius solution. At $y = 0$, the velocity components satisfy the no-slip condition $u = v = 0$.

Thus, in the lower viscous sublayer, where the roughness element is situated, the flow in the first approximation is formed by viscous stresses. The effect of heat conduction can be neglected, since the gas is nearly incompressible at low velocities.

In the main part of the boundary layer, the flow is rotational, although dissipative factors in the first approximation are no longer substantial. The expansions of the variables and flow parameters in this sublayer have the form

$$\begin{gathered}x^* = L(1 + \mathrm{Re}^{-3/10}x), \quad r^* = L(\delta + \mathrm{Re}^{-1/2}Y),\\ u^* = U_\infty(U(Y) + \mathrm{Re}^{-1/10}\bar{u}(Y) + \ldots), \quad v^* = U_\infty \mathrm{Re}^{-3/10}\bar{v}(Y) + \ldots,\\ p^* = p_\infty + \rho_\infty U_\infty^2 \mathrm{Re}^{-1/5}\bar{p} + \ldots, \quad \rho^* = \rho_\infty(R(Y) + \mathrm{Re}^{-1/10}\bar{\rho} + \ldots).\end{gathered} \tag{1.2}$$

The functions $U(Y)$ and $R(Y)$ are described by the Blasius solution and determine the state of the boundary layer ahead of the interaction region. They behave as follows: $U(Y) \longrightarrow O(Y)$ and $R(Y) \longrightarrow R(0)$ as $Y \longrightarrow 0$, while $U(Y) \longrightarrow 1$ and $R(Y) \longrightarrow 1$ as $Y \longrightarrow \infty$. Substituting (1.2) into the Navier–Stokes equations yields ordinary differential equations for determining $\bar{u}$, $\bar{v}$, $\bar{p}$, and $\bar{\rho}$. Their solutions are

$$\bar{u} = \bar{A}(x)\frac{dU}{dY}, \quad \bar{v} = \frac{d\bar{A}}{dx}U(Y), \quad \bar{\rho} = \bar{A}(x)\frac{dR}{dY}, \quad \bar{p} = \bar{p}(x). \tag{1.3}$$

Here, $\bar{A}(x)$ is an arbitrary function such that $\bar{A}(x) \longrightarrow 0$ as $x \longrightarrow -\infty$. Solutions (1.3) mean that, in the first approximation, the displacement of streamlines in the main part of the boundary layer is completely determined by the roughness element and the displacement thickness due to influence exerted by the flow in the lower viscous sublayer on the entire triple-deck flow.

In the outer sublayer, the flow is potential and the variables and flow parameters are expanded according to the formulas

$$\begin{gathered} x^* = L(1 + \mathrm{Re}^{-3/10}x), \quad r^* = L\mathrm{Re}^{-1/5}r, \\ u^* = U_\infty(1 + \mathrm{Re}^{-1/5}\tilde{u} + \ldots), \quad v^* = U_\infty\mathrm{Re}^{-3/10}\tilde{v} + \ldots, \\ p^* = p_\infty + \rho_\infty U_\infty^2\mathrm{Re}^{-1/5}\tilde{p} + \ldots, \quad \rho^* = \rho_\infty(1 + \mathrm{Re}^{-1/5}\tilde{\rho} + \ldots). \end{gathered} \tag{1.4}$$

Here, we assumed that $\delta = \mathrm{Re}^{-1/5}r_w$ where r_w is on the order of 1. In this case, the effect of the cylinder curvature is exhibited even in the first approximation. Substituting (1.4) into the Navier–Stokes equations, we find that the flow in the upper inviscid region is described by the Karman–Guderley equation for the perturbed velocity potential $\phi(x, r)$ (see [22]):

$$\begin{gathered} (K_\infty - (\gamma + 1)\phi_x)\phi_{xx} + \phi_{rr} + \phi_r/r = 0, \quad K_\infty = \frac{1 - \mathrm{M}_\infty^2}{\mathrm{Re}^{-1/5}}, \\ \tilde{u} = \phi_x, \quad \tilde{v} = \phi_r. \end{gathered} \tag{1.5}$$

Here, K_∞ is the transonic similarity parameter. The solution to Eq. (1.5) must satisfy the boundary conditions

$$\phi \longrightarrow 0, \quad x^2 + r^2 \longrightarrow \infty, \tag{1.6}$$

$$p(x) = -\phi_x(x, r_w). \tag{1.7}$$

The first condition in (1.6) is determined by matching the triple-deck flow parameters with the free-stream ones, while the second condition in (1.6) shows the relation between the perturbed velocity potential and the pressure at the outer edge of the boundary layer.

The interaction conditions are derived by matching the solution in the main part of the boundary layer with that in the outer sublayer and with that in the lower boundary sublayer. These conditions are

$$\frac{\partial\phi}{\partial r}(x, r_w) = \frac{df}{dx} - \frac{dA}{dx}, \tag{1.8}$$

$$\frac{dA}{dx} \longrightarrow -\frac{v(x, y)}{u(x, y)}, \quad y \longrightarrow \infty, \tag{1.9}$$

respectively. Here, $A(x) = \bar{A}(x) + f(x)$.

NUMERICAL COMPUTATIONS

By applying affine transformations similar to those used in [18], the equations and the boundary and initial conditions describing the transonic flow in the triple-deck region induced by the small roughness element are reduced to canonical form. In numerical computations, it is convenient to use the boundary layer equations written for the vorticity $\omega = \partial u/\partial y$. For the transformed variables and functions involved in the equations and the initial and boundary conditions, we retain the previous notation. As a result, in the new variables, the flow in the lower viscous sublayer is described by the boundary layer equation for ω with the following initial and boundary conditions:

$$\left(\int_0^y \omega dy\right)\frac{\partial\omega}{\partial x} + \left(\int_0^y y\frac{\partial\omega}{\partial x}dy - y\int_0^y \frac{\partial\omega}{\partial x}dy\right)\frac{\partial\omega}{\partial y} = \frac{\partial^2\omega}{\partial y^2}, \tag{2.1}$$

$$\omega \longrightarrow 1, \quad x \longrightarrow -\infty, \quad y \longrightarrow \infty, \tag{2.2}$$

$$\frac{\partial \omega}{\partial y}(x, 0) = \frac{dp}{dx}. \tag{2.3}$$

Conditions (2.2) describe the flow in the regions to the left of and above the viscous lower sublayer. Condition (2.3) specifies the relation between the vorticity and the pressure gradient on the body surface.

In the outer potential region, the flow is governed by the Karman–Guderley equation for ϕ:

$$(K_\infty - (\gamma + 1)\phi_x)\phi_{xx} + \phi_{rr} + \phi_r/r = 0. \tag{2.4}$$

Here, K_∞ is a new transonic similarity parameter appearing after the affine transformations. The boundary conditions for Eq. (2.4) are given by (1.6) and (1.7).

The interaction conditions are given by (1.8) and (1.9). As before, $A(x) = \bar{A}(x) + f(x)$.

In the computations, we used a roughness element whose shape was specified by the equation

$$f(x) = \frac{h}{1 + x^2}. \tag{2.5}$$

The solution to problem (2.1)–(2.5) depends on three parameters: K_∞, h, and r_w.

The problem was solved numerically by a direct method developed for solving the transonic triple-deck equations (see [23]). Two nonuniform grids were introduced in the viscous sublayer and the outer inviscid region. At interior grid points, the Prandtl equations were approximated by second-order accurate finite-difference schemes, where the first x- and y-derivatives involved in the nonlinear terms of the equations were approximated by finite differences depending on the sign of the longitudinal and vertical velocity components, respectively.

Similarly, depending on the sign of the nonlinear term, the Karman–Guderley equation was approximated by the finite-difference scheme represented in [22].

The resulting nonlinear system of finite-difference equations was solved by the Newton–Kantorovich method with block Gaussian elimination used to improve the approximation of the unknown functions. The block Gaussian elimination algorithm as applied to triple-deck problems was described in detail in [23].

The numerical results presented below were obtained primarily on a nonuniform grid with the maximum number of points $200 \times (80 + 120)$, with the minimum mesh size in x ($\Delta x = 0.02$), and with the minimum mesh size in y ($\Delta y = \Delta r = 0.02$) in the boundary layer and the outer transonic flow. To verify the results, we doubled the number of grid points ($400 \times (160 + 240)$) and halved the minimum mesh sizes in x, y, and r: $\Delta x = \Delta y = \Delta r = 0.01$. The maximum discrepancy between the characteristics computed in these cases was found in the near-wall area of steepest pressure gradients and did not exceed 4%.

Before analyzing the numerical results, we discuss the experiments conducted in [1]. Specifically, symmetric double-convex airfoils composed of circular arcs are placed in a transonic flow with a laminar boundary layer at $\mathrm{Re} \approx 10^6$. A closed supersonic region with two or more shock waves inclined upstream is formed in the experiments. The shock waves divide the supersonic region into several subregions. The first shock and the supersonic region ahead of are called basic. Behind the basic shock, the pressure increases, while the Mach number decreases, but the flow remains supersonic. Thus, a closed supersonic region much smaller than the basic one is formed downstream of the basic shock. The boundary layer becomes thicker at the place where the basic shock impinges. The shape of the thickening has a small radius of curvature, and the flow past it gives rise to a rarefaction wave, which leads to lower pressure. As a result, another shock wave develops in this supersonic region. If the flow behind it is supersonic, the process repeats until the flow velocity behind the newly formed shock is subsonic. The pressure distribution on the surface of the boundary layer is continuous.

Figure 1 shows contour lines of $w = (1 + \gamma)(\partial\phi/\partial x) - K_\infty$ (which is proportional to the deviation of the squared local Mach number from unity) at $K_\infty = 2$, $h = 5$, and $r_w = 1$. It can be clearly seen that, due to the displacing effect of the boundary layer, a supersonic zone with two shock waves inclined upstream is formed in the outer potential region. The bases of the shock waves are smeared, as usually occurs when they impinge on the boundary layer (see [2]), and the pressure distribution on the body surface becomes continuous. This can be inferred from the behavior of $w_s = w(r_w, x)$ and the nondimensionalized skin friction $\tau = \omega(0, x)$ plotted in Fig. 2 ($K_\infty = 2$, $h = 5$, $r_w = 1$, solid lines).

The skin friction computations also show that a flow separation occurs at the above values of K_∞, h, and r_w. The streamline pattern in the viscous boundary sublayer is displayed in Fig. 3.

In the transonic steady triple-deck theory, disturbances in the outer region are caused by the displacement thickness of the boundary layer in the lower near-wall region and by the roughness of the surface.

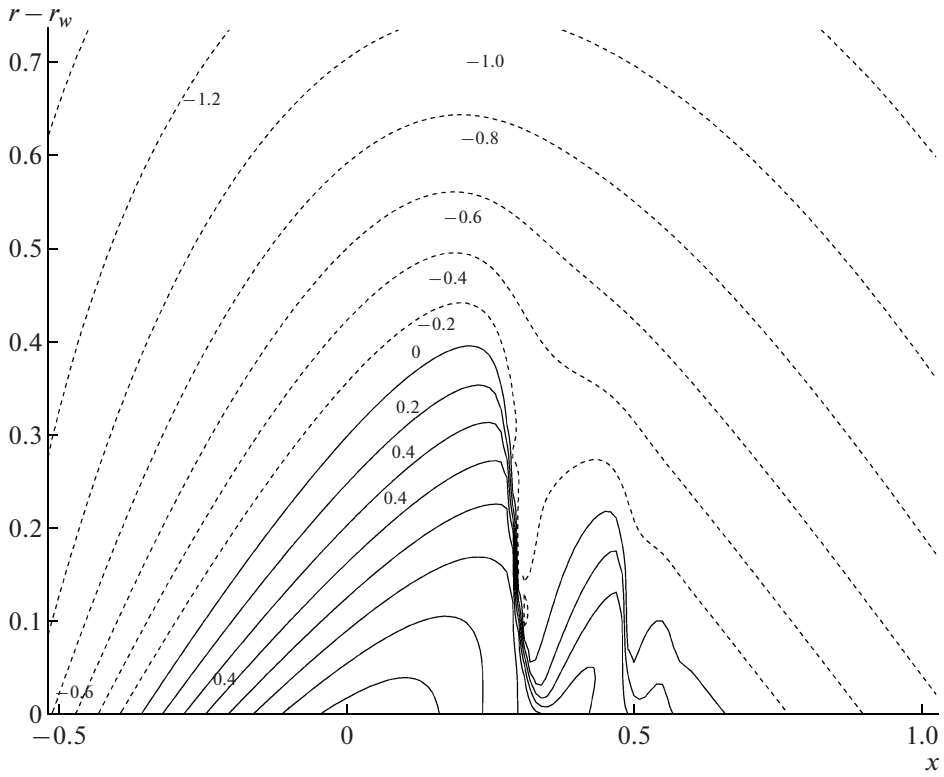

$r - r_w$
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−0.5
0
0.5
1.0
x
−1.0
−1.2
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.4
0.4
−0.6

Fig. 1.

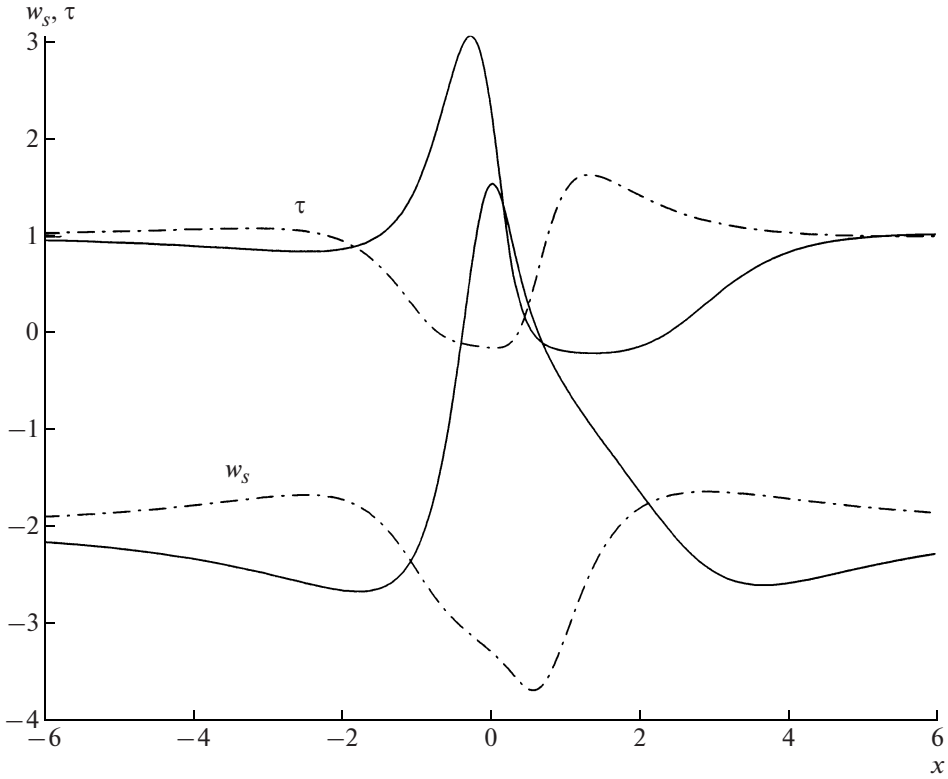

w_s, τ
3
2
1
0
−1
−2
−3
−4
−6
−4
−2
0
2
4
6
x
τ
w_s

Fig. 2.

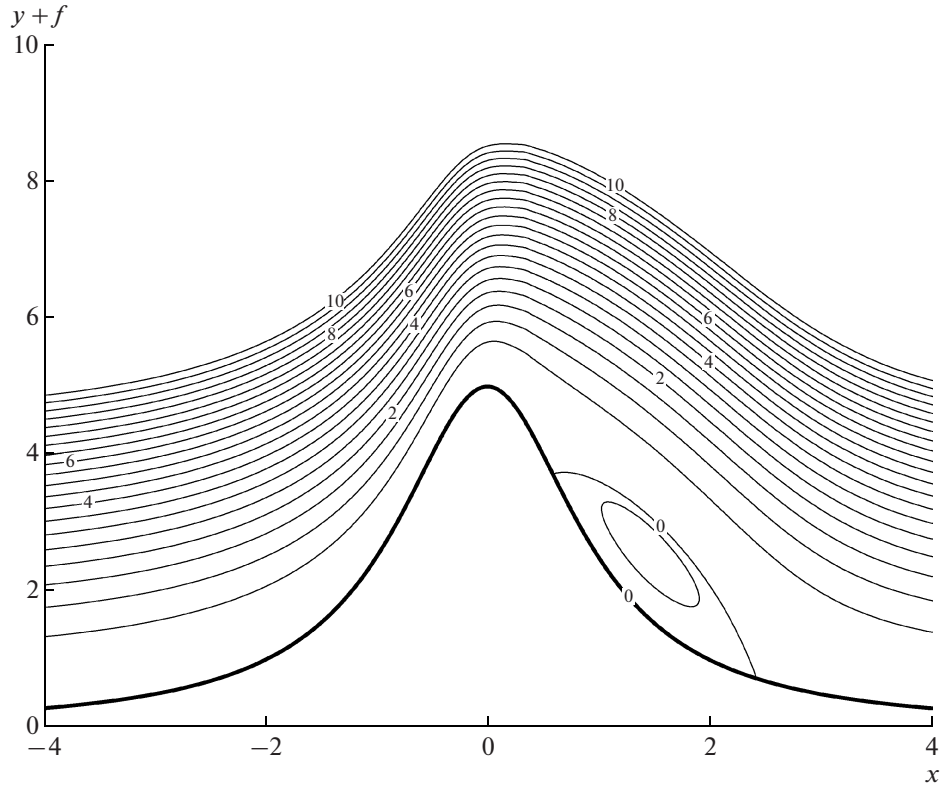


Fig. 3.

The main part of the boundary layer transmits these disturbances without changes. Because of the interaction, the lower boundary in the outer potential region contains an extended segment with a convex smooth profile, on which a supersonic region is formed in the same way as in the transonic flow past an airfoil (see [2]). In transonic flows past convex airfoils, supersonic regions are generally closed by shock waves (see [2, 24, 26]), and a shock wave is absent only for a specially chosen airfoil shape. A review concerning the design of shock-free airfoils in a certain range of Mach numbers can be found in [24]. However, even small variations in the airfoil shape or the free-stream Mach number eventually lead to a shock wave (see [25]). Of course, a unique situation is possible when, depending on the parameters of the problem, the displacement thickness of the boundary layer has a shape, at which no shock wave arises. However, the numerical computations performed for the chosen hump shapes showed that this situation did not occur. The shape of the supersonic regions qualitatively agrees with that observed in experiments [1]. The gas velocity behind the basic shock is supersonic (see Figs. 1, 2). Therefore, there is another small supersonic region behind it. Eventually, this supersonic region is closed by a second shock wave, the velocity behind which is subsonic. Compared with the experiments in [1], the supersonic regions have a more pronounced double-hump structure.

In addition to supersonic regions and shock waves in the outer potential region, local separation zones can develop in the lower viscous boundary sublayer, depending on the values of K_∞, h, and r_w. Computations show that they may coexist or form separately.

The effect of r_w on the flow characteristics can be seen in Figs. 4 and 5, which present the distributions of w and τ, respectively, at $K_\infty = 1$ and $h = 2$ for (a) $r_w = \infty$, (b) $r_w = 5$, (c) $r_w = 1$, and (d) $r_w = 0.1$. Evidently, for fixed K_∞ and h, the flow characteristics in a plane flow ($r_w = \infty$) are much more affected by the roughness element than in a flow with a large effect of the radius of curvature of the body. There are no local separation regions in the latter case. However, the results obtained with $K_\infty = 2$ and $h = 5$ (see Fig. 6) show that local separation regions develop in the flow when $r_w = 1$ or $r_w = 0.5$ (curves a and b, respectively). As r_w decreases, they disappear. This is well seen in the plot of the skin friction at $r_w = 0.1$ (curve c).

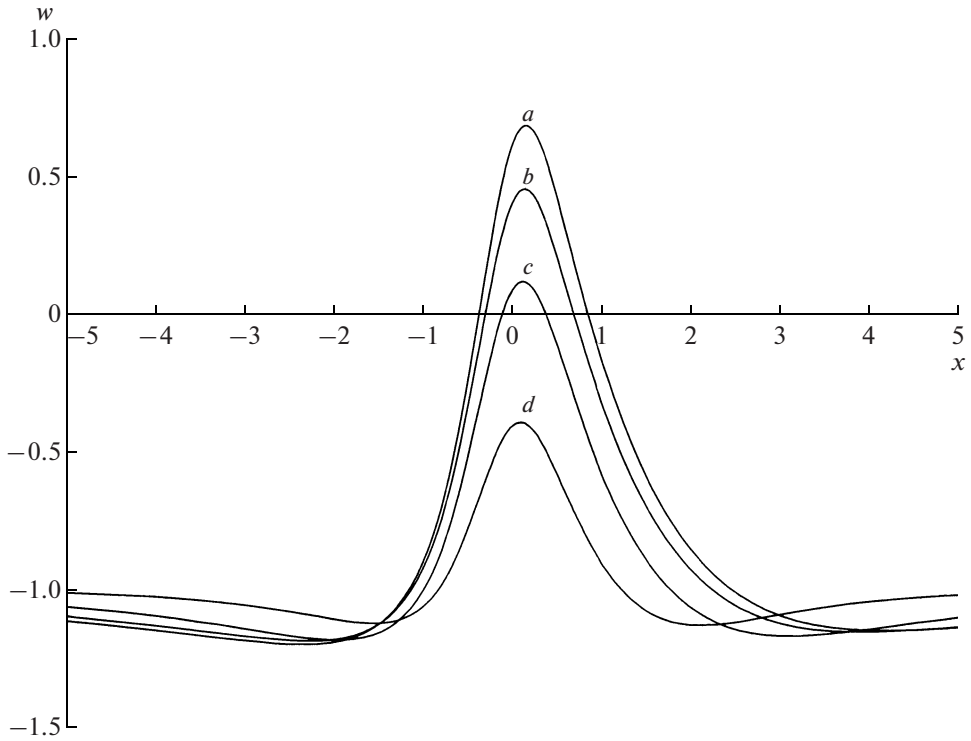

w
1.0
0.5
0
−0.5
−1.0
−1.5
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
a
b
c
d

Fig. 4.

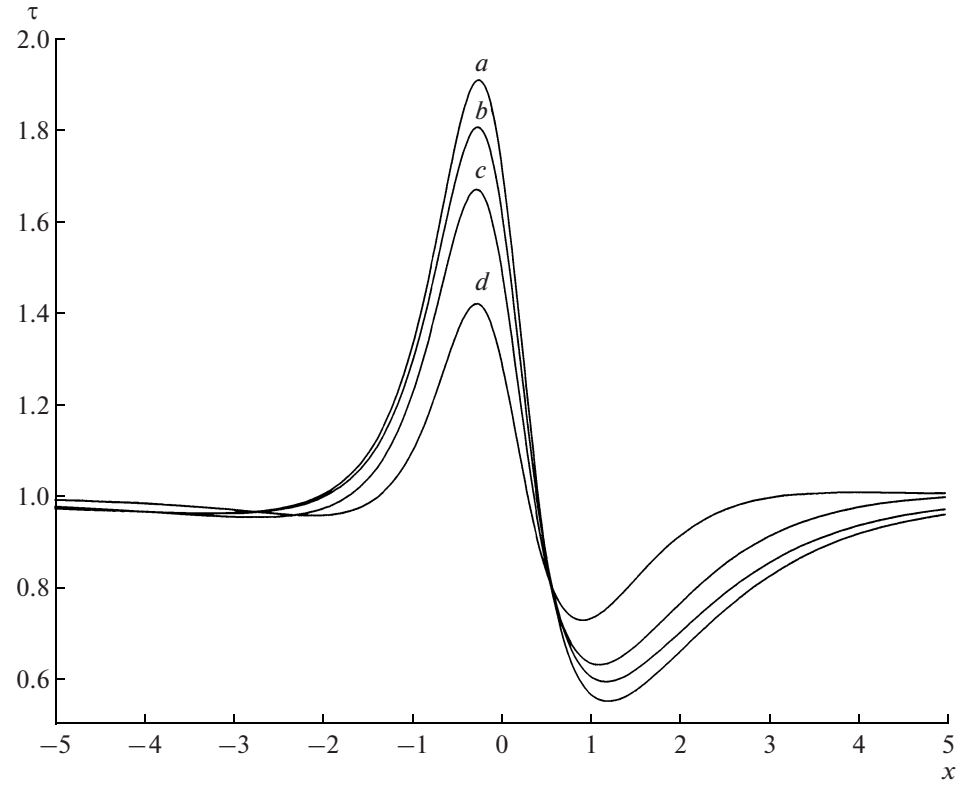

τ
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
a
b
c
d

Fig. 5.

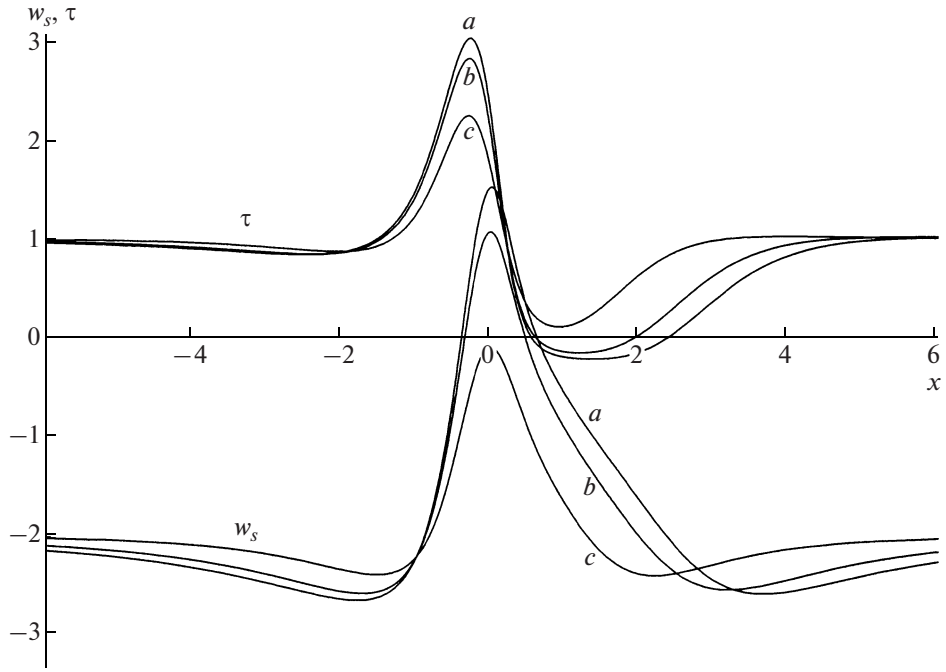

w_s, τ
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
6
x
a
b
c
τ
a
b
c
w_s

Fig. 6.

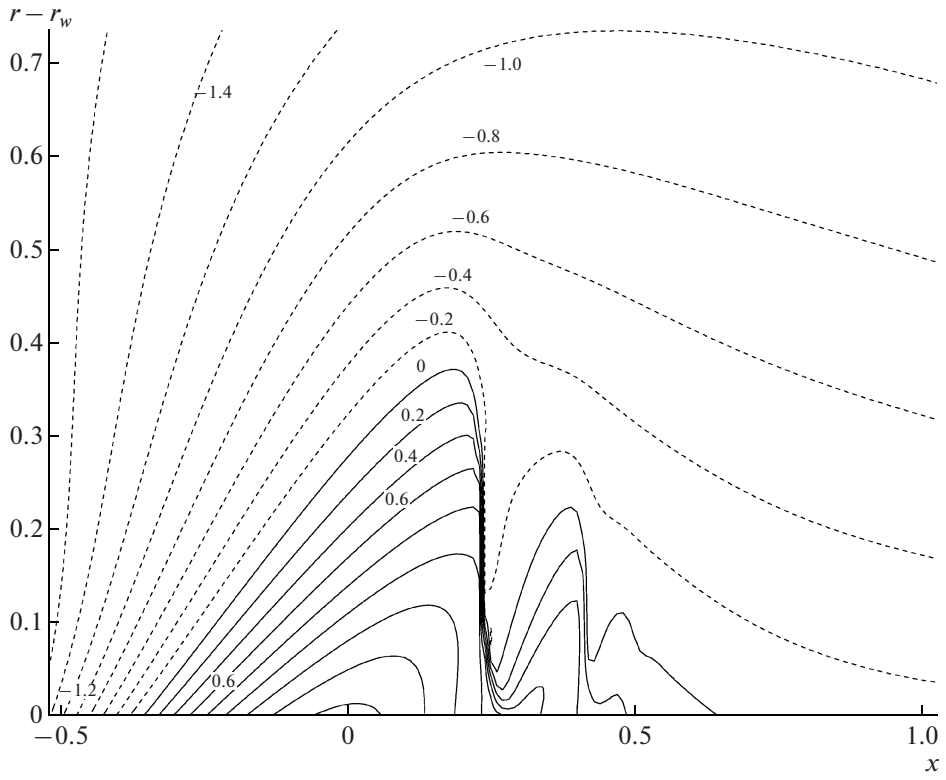

$r - r_w$
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−0.5
0
0.5
1.0
x
−1.0
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
−1.4
−1.2
0.6

Fig. 7.

Figure 2 displays the numerical results obtained for the flow past a cavity at $K_\infty = 2$, $h = -5$, and $r_w = 1$. We can see local separation regions but supersonic regions are absent (the plots of τ and w are depicted by dash-dotted lines). The solid curves show the behavior of τ and w_s at $K_\infty = 2$, $h = 5$, and $r_w = 1$.

As the shape of the hump varies, the qualitative flow pattern remains the same. This is well seen in Fig. 7, which shows contour lines of w corresponding to the flow over a hump given by $f(x) = \cos(\pi x/2)$, $x \in (-1, +1)$ at $K_\infty = 2$, $h = 5$, and $r_w = 1$.

CONCLUSIONS

The basic difference between the axisymmetric and plane cases in the triple-deck problem is that the former case involves the additional similarity parameter r_w and that the Karman–Guderley equation (1.5) contains an additional term caused by the fact that the flow is three-dimensional and axisymmetric. This term takes into account the curvature of the body-of-revolution surface. The qualitative flow pattern remains the same as in the plane case. For an axisymmetric flow, the intensity of shock waves and the sizes of local separation regions grow with increasing r_w, while the position of the basic shock wave at fixed K_∞ varies little. Note that, if we make the substitution $r = r_w + n$, then, in the limit as $r_w \longrightarrow \infty$, the axisymmetric transonic triple-deck equations turn into the plane transonic ones.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 07-01-00999) and by the Program "Development of the Scientific Potential of Higher Educational Institutions" (project no. 2.1.1/500).

REFERENCES

1. H. F. Liepmann, "The Interaction between Boundary Layer and Shock Waves in Transonic Flow," Acronaut. Sci. **13**, 623–638 (1946).
2. T. H. Moulden, *Fundamentals of Transonic Flow* (Willey, New York, 1984).
3. V. Ya. Neiland, "On the Theory of Laminar Boundary Layer Separation in Supersonic Flow," Izv. Akad. Nauk SSSR. Mekh. Zhidk. Gaza, No. 4, 53–57 (1969).
4. K. Stefardson and P. G. Williams, "Self-Induced Separation," Proc. R. Soc. London, Ser. A **312** (1509), 181–206 (1969).
5. A. F. Messiter, A. Feo, and R. E. Melnic, "Shock-Ware Strength for Separation of Laminar Boundary Layer at Transonic Speeds," AIAA J. **9**, 1197–1198 (1971).
6. H. M. Brilliant and T. C. Adamson, Jr., "Shock Wave Boundary-Layer Interactions in Laminar Transonic Flow," AIAA. J. **12**, 323–329 (1974).
7. R. J. Bodonyi and A. Kluwick, "Supercritical Transonic Trailing-Edge Flow," Quart. J. Mech. Appl. Math. **35**, 265–277 (1982).
8. G. L. Korolev, "Viscous Transonic Flow near the Trailing Edge of a Plate," Izv. Akad. Nauk SSSR. Mekh. Zhidk. Gaza **11** (3), 23–28 (1983).
9. G. L. Korolev, "Asymptotic Theory of Viscous Transonic Flow over a Slender Body," Uch. Zap. Tsentr. Aerogidrodin. Inst. **14** (4), 105–109 (1983).
10. V. V. Bogolepov and V. Ya. Neiland, in *Analysis of Local Disturbances in Viscous Supersonic Flow* (Nauka, Moscow, 1976), pp. 104–118 [in Russian].
11. V. V. Bogolepov, "Computation of the Interaction of a Supersonic Boundary Layer with a Thin Obstruction," Uch. Zap. Tsentr. Aerogidrodin. Inst. **5** (6), 30–38 (1974).
12. V. N. Trigub, "Outer Flow/Boundary Layer Interaction in Flows past Thin Axisymmetric Bodies," Uch. Zap. Tsentr. Aerogidrodin. Inst. **14** (6), 8–17 (1983).
13. P. W. Duck, "The Effect of a Surface Discontinuity on an Axisymmetric Boundary Layer," Quart. J. Mech. Appl. Math. **37** (1), 57–74 (1984).
14. S. N. Timoshin, "Laminar Flow near the Break Point of the Surface of a Prolate Body of Revolution," Uch. Zap. Tsentr. Aerogidrodin. Inst. **16** (5), 10–21 (1985).
15. S. N. Timoshin, "Outer Flow/Boundary Layer Interaction in the Longitude Flow past a Prolate Body of Rotation," Uch. Zap. Tsentr. Aerogidrodin. Inst. **17 (**2), 33–41 (1986).
16. A. Kluwick, P. Gittler, and R. Bodonyi, "Viscous-Inviscid Interactions on Axisymmetric Bodies of Revolution in Supersonic Flow," J. Fluid Mech. **140** (1984).
17. V. V. Sychev, "Three-Dimensional Flows over Roughness Elements on the Surface of an Axisymmetric Body," Uch. Zap. Tsentr. Aerogidrodin. Inst. **24**, 12–28 (1993).

18. V. N. Diesperov and G. L. Korolev, "Formation of Supersonic Zones and Local Separation Zones in Transonic Steady-State Flow Past Surface Roughness in the Free Interaction Regime," Izv. Akad. Nauk SSSR. Mekh. Zhidk. Gaza, No. 1, 50–59 (2003).
19. A. N. Bogdanov, V. N. Diesperov, V. I. Zhuk, et al., "Asymptotic Models of Viscous Transonic Flows," in *Abstracts of Papers of the All-Russia Conference on Theoretical and Applied Mechanics, August 23–29, 2003* (Perm, 2003), pp. 106–107.
20. V. N. Diesperov and G. L. Korolev, "Triple-Deck Analysis of Formation and Evolution of Supersonic Zones and Local Separation Zones in Unsteady Transonic Flow over a Surface Roughness Element," Zh. Vychisl. Mat. Mat. Fiz. **45**, 536–544 (2005) [Comput. Math. Math. Phys. **45**, 516–523 (2005)].
21. H. W. Liepmann and A. Roshko, *Elements of Gasdynamics* (Wiley, New York, 1957; Inostrannaya Literatura, Moscow, 1960).
22. J. D. Cole and L. P. Cook, *Transonic Aerodynamics* (North-Holland, Amsterdam, 1986; Mir, Moscow, 1989).
23. G. L. Korolev, "A Method for Solving Problems in the Asymptotic Theory of the Interaction of the boundary Layer with Outer Flow," Zh. Vychisl. Mat. Mat. Fiz. **27**, 1224–1232 (1987).
24. G. Y. Nienwland and B. M. Spee, "Transonic Airfoils: Recent Developments in Theory, Experiment, and Design," Ann. Rev. Fluid Mech. **5**, 119–150 (1973).
25. L. Bers, *Mathematical Aspects of Subsonic and Transonic Gas Dynamics* (New York Univ., New York, 1952; Inostrannaya Literatura, Moscow, 1961).

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1245–1256. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © V.N. Koterov, Yu.S. Yurezanskaya, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1306–1318.

Simulation of Suspended Substance Dispersion on the Ocean Shelf: Effective Hydraulic Coarseness of Polydisperse Suspension

V. N. Koterov and Yu. S. Yurezanskaya

Dorodnitsyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia
e-mail: koterov@ccas.ru, july@ccas.ru

Received October 20, 2008

Abstract—Suspended substance dispersion in a water body is simulated in the case when the spread area is considerably larger than the depth of the water body. It is shown that, even if vertical turbulent exchange has a large effect, the problem of computing the evolution of polydisperse suspended pollutants produced by an instantaneous point source can be reduced to the integration of a few one-dimensional evolution problems and to the solution of one two-dimensional problem. This result can be used to design efficient solution methods for the practically important problem of computing the dispersion of a suspension produced by a time-continuous and/or spatially distributed source of water pollution.

DOI: 10.1134/S096554250907015X

Key words: numerical methods, advection–diffusion equation, turbulent exchange, shelf, pollutants, polydisperse suspensions.

1. INTRODUCTION

Recently, interest in the computation of suspended substance dispersion in water bodies has increased as motivated by the necessity of performing assessments of various anthropogenic impacts on the environment. Such assessments are required, for example, in the construction of underwater pipelines and drilling platforms on the oceanic shelf, in dredging operations, etc. In principle, available regulations impose rather severe requirements on the quality of mathematical models used for such assessments. For example, according to these regulations, the total concentration of mineral suspension at control points located about 250–500 m away from a pollution source must not exceed 1 mg/l, while near the source this characteristic usually amounts to 10 g/l and more.

In the description of suspended substance transport, we can distinguish two qualitatively different areas, namely, a near-field region, whose length scale correlates with the size of a pollutant source (e.g., a water outlet structure, a dredger, etc.), and a far-field region (with control points deployed), whose size considerably exceeds the characteristic length of the near-field region.

In the near-field region, the concentrations of suspended substances are high and the simulation of their transport generally has to be based on nonlinear equations of multiphase media dynamics (see, e.g., [1]). In the far-field region, which is the object of study in this paper, the substance concentrations decrease considerably due to turbulent mixing and solid deposition. The suspended substances undergo passive dispersion (see, e.g., [2]) and can be treated as a passive scalar whose transport is determined only by the flow velocity and the intensity of turbulent diffusion. Moreover, the superposition principle applies in the far-field region. This means that the spread of a suspended substance can be represented as the motion of a collection of individual noninteracting "clouds" produced by instantaneous point sources of pollutants. These clouds move through the water under the action of local currents and possibly deposit on the bottom. In the course of motion, they increase in size due to horizontal turbulent diffusion, while the concentrations of suspended substances in them decrease. The suspension concentration C at an arbitrary point $\mathbf{r}$ of the water body is represented as the sum of the passive scalar concentrations C^j in individual clouds including this point at a given time. For example, for a time-continuous stationary point source that releases a polydisperse suspension starting at $t = 0$, we have

$$C(\mathbf{r}, t) = \int_0^t \sum_{j=1}^{N} C^j(\mathbf{r}, t - t_0, t_0)\, dt_0, \quad t > 0,$$

where j is the index of a pollutant fraction and N is the number of fractions.

In many practically interesting cases, the three-dimensional numerical simulation of pollutant transport is, at least, unjustified or difficult to perform due to the following factors: the size of the spread area considerably exceeds the depth of the water body; the number of various substance fractions is large; the deposition rates of these fractions may differ by many orders of magnitude; and the control-point concentrations, which have to be reliably computed by the numerical model, differ by four and more orders of magnitude from the suspension concentration at the pollution source.

Two-dimensional (depth-averaged) models are frequently considered in practice. They are based on the integral relation

$$\frac{\partial H\bar{C}^j}{\partial t} + \frac{\partial}{\partial x_i}[H(\overline{u_iC^j} + \bar{J}_i^j)] + J_z^j(H) = 0,$$
$$\bar{C}^j = \frac{1}{H}\int_0^H C^j dz, \quad \bar{J}_i^j = \frac{1}{H}\int_0^H J_i^j dz, \quad \overline{u_iC^j} = \frac{1}{H}\int_0^H u_iC^j dz. \tag{1.1}$$

Here and below, $\mathbf{x} = (x_1, x_2)$ are horizontal Cartesian coordinates; z is a vertical coordinate computed from the water surface to the bottom; $H = H(\mathbf{x})$ is the local depth of the water body; $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) = (u_1, u_2)$ is the horizontal current velocity; J_i^j is the horizontal diffusive suspension flux due to turbulent exchange; $J_z^j(H)$ is the flux of particles toward the bottom; the overbar denotes depth-averaged quantities; and summation is implied over the repeated index $i = 1, 2$.

The total depth-averaged concentration $\bar{C}$ of suspended particles and the suspension mass m per unit area deposited on the bottom are determined by the expressions

$$\bar{C} = \sum_{j=1}^N \bar{C}^j, \quad \frac{\partial m}{\partial t} = \sum_{j=1}^N J_z^j(H).$$

Integral relation (1.1) is an exact consequence of the suspension mass conservation law. In computations, one usually uses the approximate expressions $\overline{u_iC^j} = \bar{u}_i\bar{C}^j$, and $J_z^j(H) = W_j\bar{C}^j$, where W_j is the hydraulic coarseness of the jth fraction (the deposition rate of the fraction in stationary water with no vertical turbulent exchange). Then we have the following two-dimensional "depth-averaged" advection–diffusion equation for suspended substances (see, e.g., [3]):

$$\frac{\partial H\bar{C}^j}{\partial t} + \frac{\partial}{\partial x_i}[H(\bar{u}_i\bar{C}^j + \bar{J}_i^j)] + W_j\bar{C}^j = 0, \quad \bar{u}_i = \frac{1}{H}\int_0^H u_i dz. \tag{1.2}$$

However, this approach is not always acceptable. Specifically, Eq. (1.2) ignores vertical turbulent exchange, which is substantial for particles with small values of W_j. The interaction of deposited particles with the bottom and the effect of a particular vertical location of the pollution source are not taken into account.

In this paper, for an instantaneous point source of suspension, a technique that is free from the indicated shortcomings is proposed and justified for averaging the three-dimensional advection–diffusion equation. The technique is based on the concept of the time-dependent effective hydraulic coarseness of a polydisperse suspension, which is determined by solving only N one-dimensional evolution problems. The depth-averaged suspension concentration is then found by solving the two-dimensional advection–diffusion equation.

We do not consider any solution methods for this equation. Note only that the cloud method and the stochastic discrete particle method (see, e.g., [4–6]) can be used within the approach proposed below.

2. FORMULATION OF THE PROBLEM IN THE CASE OF AN INSTANTANEOUS POINT SOURCE OF SUSPENSION

Assuming that turbulent mixing can be divided into horizontal turbulent exchange and vertical turbulent diffusion (see, e.g., [7]), the three-dimensional equation and the initial condition for determining the

far-field evolution of the jth fraction concentration in a suspension produced by an instantaneous point source can be written as

$$\frac{\partial C^j}{\partial t} + \frac{\partial u_i C^j}{\partial x_i} + \frac{\partial (u_z + W_j) C^j}{\partial z} + \frac{\partial J_i^j}{\partial x_i} - \frac{\partial}{\partial z} K_z \frac{\partial C^j}{\partial z} = 0, \tag{2.1}$$

$$C^j = M_j \delta(\mathbf{x}) \delta(z - z_0) \quad \text{at} \quad t = 0, \quad C^j \longrightarrow 0 \quad \text{as} \quad |\mathbf{x}| \longrightarrow \infty.$$

Here, u_z is the vertical current velocity, J_i^j are the components of the horizontal turbulent suspension flux, K_z is the vertical turbulent diffusivity, M_j is the initial mass of the suspension, δ is the Dirac delta function, and z_0 is the vertical coordinate of the suspension source. It is assumed that the suspension source is located at the point $\mathbf{x} = 0$ and starts operating at the time $t = 0$.

At the water surface, there is no suspension flux, so

$$W_i C^j - K_z \partial C^j / \partial z = 0 \quad \text{at} \quad z = 0.$$

The boundary condition at the bottom of the reservoir depends on the interaction of the deposited substance with the bottom surface. In the general case, this condition is written as

$$W_j C^j - K_z \partial C^j / \partial z = W_j \beta_j C^j \quad \text{at} \quad z = H(\mathbf{x}), \tag{2.2}$$

where β_j is a parameter that, following the terminology in [8], can be referred to as a dimensionless mass transfer coefficient to the bottom.

The values of β_j depend on the adsorbing properties of the bottom surface. For a completely adsorbing surface, $\beta_j = \infty$ (i.e., the conditions $C^j = 0$ must be set at the bottom instead of (2.2)). No diffusive flux ($\partial C^j / \partial z = 0$) is sometimes specified at the bottom (see [9]), which corresponds to $\beta_j = 1$. The limiting case is $\beta_j = 0$, when there is no suspension flux to the bottom (the bottom is a totally nonadsorbing surface). Note that $\beta_j < 0$ corresponds to the regime of disturbing the deposits at the bottom. It is not considered in this paper.

In what follows, we use the following assumptions.

Assumption 1. The depth of the water body varies relatively slowly:

$$|\partial H / \partial x_i| \ll 1, \quad i = 1, 2.$$

Assumption 2. The shallow water approximation is used, and the vertical turbulent diffusivity K_z in the entire considered domain can be represented as $K_z = u_* H K(\xi)$ with $\xi = z/H$, where the constant u_* is the characteristic vertical diffusion rate, which can sometimes be identified with the friction velocity in the bottom boundary layer; ξ is the dimensionless vertical coordinate; and $K(\xi)$ is the nondimensional vertical turbulent diffusivity.

Assumption 3. The components u_i of the horizontal velocity $\mathbf{u}$ are independent of z; i.e., they are represented as $u_i = \bar{u}_i(\mathbf{x}, t)$.

Assumption 4. The components J_i^j of the horizontal turbulent suspension flux as a function of C^j are given by linear operators independent of z: $J_i^j = \bar{J}_i[C^j]$. We use the hypothesis of gradient horizontal turbulent transfer, which is traditional for the problem in question: $J_i^j = -\bar{A}(\mathbf{x}, t) \partial C^j / \partial x_i$, where $\bar{A}$ is the vertically averaged horizontal turbulent diffusivity. In the case of free horizontal turbulence, however, this hypothesis may be physically unjustified, and the operators $\bar{J}_i[C^j]$, though linear, may have a more complex structure (see [5, 6]).

For a time-dependent current velocity $\mathbf{u}$, Assumption 2, which states that the friction velocity in the boundary layer at the bottom of the reservoir is constant in the entire computational domain, is rather restrictive. This assumption holds, for example, if the characteristic time of varying the horizontal current speed exceeds the time required for scattering a pollution cloud. It can be assumed that this assumption also holds approximately on the oceanic shelf when the current is caused by tidal processes, the eccentricity of the current velocity hodograph is not high, and the current speed does not vary too strongly.

Assumptions 3 and 4 seem relatively severe. However, as a rule, there is no detailed information on the spatial distribution of the current velocity in practice. Moreover, the velocity field $\mathbf{u}$ required for computing suspension transport is frequently obtained by numerical simulation based on the two-dimensional shallow water equations. The corresponding numerical results, which, strictly speaking, are also based on

Assumption 1, reproduce only vertically averaged distributions of the horizontal current velocity in the area of water.

Due to the assumptions made, u_z in (2.1) can be neglected in the first approximation and the problem can be represented in a form similar to (1.1) and (1.2):

$$\frac{\partial H(\mathbf{x})C^j}{\partial t} + \frac{\partial}{\partial x_i}\{H(\mathbf{x})(\bar{u}_i(\mathbf{x}, t)C^j + \bar{J}_i[C^j])\} + u_* \frac{\partial}{\partial \xi}\left[\varepsilon_j C^j - K(\xi)\frac{\partial C^j}{\partial \xi}\right] = 0, \quad \varepsilon_j = \frac{W_j}{u_*},$$

$$C^j = \frac{M_j}{H(0)}\delta(\mathbf{x})\delta(\xi - \xi_0) \quad \text{at} \quad t = 0,$$

$$\varepsilon_j C^j - K(\xi)\frac{\partial C^j}{\partial \xi} = 0 \quad \text{at} \quad \xi = 0 \quad \text{(on the water surface)}, \tag{2.3}$$

$$\varepsilon_j C^j - K(\xi)\frac{\partial C^j}{\partial \xi} = \varepsilon_j \beta_j C^j \quad \text{at} \quad \xi = 1 \quad \text{(on the bottom)}.$$

Here, $H(0)$ is the depth of the water body at the discharge point, ξ_0 is the dimensionless vertical coordinate of the suspension source, and ε_j is a dimensionless parameter equal to the ratio of the particle deposition rate to the characteristic velocity of vertical turbulent diffusion. In practice, the values of this parameter can vary widely. The limit $\varepsilon_j \longrightarrow 0$ corresponds to the case of a fine-particle nondepositing (so-called conservative) suspension. When $\varepsilon_j \longrightarrow \infty$ (large-particle suspension), vertical turbulent mixing can be ignored.

The operator of averaging concentrations over the depth is defined in the usual manner as

$$\bar{C}^j = \int_0^1 C^j(\mathbf{x}, t, \xi)G(\xi)d\xi, \quad \int_0^1 G(\xi)d\xi = 1. \tag{2.4}$$

Here, the nonnegative function $G(\xi)$ is the averaging kernel. Specifically, if $G \equiv 1$, then $\bar{C}^j$ are the depth-averaged concentrations used in (1.1), (1.2).

Below, we formulate the equations and initial conditions for the total concentration $\bar{C} = \Sigma_{j=1}^N \bar{C}^j$ if the function G does not vanish on the interval [0, 1].

3. EXPANSION OVER VERTICAL DIFFUSION MODES

Consider the eigenvalue problem

$$\frac{d}{d\xi}\left[\varepsilon_j Z_n^j - K(\xi)\frac{dZ_n^j}{d\xi}\right] = \lambda_n^j Z_n^j, \tag{3.1a}$$

$$\varepsilon_j Z_n^j - K(0)\frac{dZ_n^j}{d\xi} = 0 \quad \text{at} \quad \xi = 0, \quad \varepsilon_j Z_n^j - K(1)\frac{dZ_n^j}{d\xi} = \varepsilon_j \beta_j Z_n^j \quad \text{at} \quad \xi = 1. \tag{3.1b}$$

By multiplying both sides of Eq. (3.1a) by

$$\rho^j(\xi) = \exp\left[-\varepsilon_j \int_0^\xi K^{-1}(\xi')d\xi'\right],$$

problem (3.1) can be reduced to the classical Sturm–Liouville eigenvalue problem for the equation

$$-\frac{d}{d\xi}\left[p^j(\xi)\frac{dZ_n^j}{d\xi}\right] = \lambda_n^j \rho^j(\xi)Z_n^j, \quad p^j(\xi) = \rho^j(\xi)K(\xi),$$

with boundary conditions (3.1b).

It is well known (see, e.g., [10]) that, for $p^j(\xi) \geq \text{const} > 0$,[1] this problem has a countable set of simple nonnegative eigenvalues $\lambda_n^j = \lambda_n^j(\varepsilon_j, \beta_j)$, $(n = 0, 1, \ldots)$ that depend on the parameters ε_j and β_j (it is

[1] In some cases, the coefficients may have singularities (specifically, vanish) at the endpoints of the interval $0 < \xi < 1$.

assumed below that λ_n^j are arranged in increasing order). Moreover, the eigenfunctions $Z_n^j = Z_n^j(\xi; \varepsilon_j, \beta_j)$, which are referred to hereafter as vertical diffusion modes, form a complete system and satisfy the orthogonality relation

$$(Z_n^j, Z_m^j) = \int_0^1 Z_n^j(\xi) Z_m^j(\xi) \rho^j(\xi) d\xi = \delta_{nm}, \tag{3.2}$$

where δ_{nm} is the Kronecker delta.

A solution to problem (2.3) is sought in the form of the series

$$C^j = \sum_{n=0}^{\infty} C_n^j(\mathbf{x}, t) Z_n^j(\xi), \quad C_n^j = (C^j, Z_n^j). \tag{3.3}$$

Substituting (3.3) into (2.3) and using orthogonality relation (3.2), in view of (3.1a), we easily obtain equations for the expansion coefficients C_n^j:

$$\begin{aligned} &\frac{\partial H C_n^j}{\partial t} + \frac{\partial}{\partial x_i}\{H(\bar{u}_i C_n^j + \bar{J}_i[C_n^j])\} + u_* \lambda_n^j C_n^j = 0, \\ &C_n^j = \frac{M_j}{H(0)} \delta(\mathbf{x}) Z_n^j(\xi_0) \rho^j(\xi_0) \quad \text{at} \quad t = 0. \end{aligned} \tag{3.4}$$

According to (3.3), the depth-averaged total concentration distribution $\bar{C}$ can be found as follows:

$$\bar{C} = \sum_{j=1}^{N} \sum_{n=0}^{\infty} C_n^j(\mathbf{x}, t) \bar{Z}_n^j, \quad \bar{Z}_n^j = \int_0^1 Z_n^j(\xi; \varepsilon_j, \beta_j) G(\xi) d\xi. \tag{3.5}$$

4. LONG-TIME ASYMPTOTICS

It can be seen that the solutions to problems (3.4) tend exponentially to zero as $t \longrightarrow \infty$. The slowest decaying components C_0^j in (3.3) correspond to the minimum eigenvalues λ_0^j, which can be called the dimensionless hydraulic coarseness of fraction j with allowance for vertical turbulent exchange (see (1.2)). In other words, as $t \longrightarrow \infty$, the leading term in (3.5) is that with the fraction index $j = j_m$, for which λ_0^j is minimal. Note that the index j_m not necessarily corresponds to the fraction with the minimum hydraulic coarseness (i.e., with the minimum value of ε_j) since λ_0^j also depends on β_j, which takes into account the interaction of the depositing suspension with the bottom.

In the general case, the vertical diffusivity $K(\xi)$ has an inhomogeneous distribution over depth and eigenvalue problem (3.1) can be solved only by numerical methods. However, when $K = 1$, the problem has the analytical solution

$$\lambda_n^j = \frac{\varepsilon_j^2}{4} + (\omega_n^j)^2, \quad Z_n^j(\xi) = \frac{\varepsilon_j \sin(\omega_n^j \xi) + 2\omega_n^j \cos(\omega_n^j \xi)}{(D_n^j)^{1/2}} \exp\left(\frac{\varepsilon_j \xi}{2}\right),$$

$$D_n^j = \frac{\varepsilon_j^2}{2} + 2(\omega_n^j)^2 + \frac{2}{\omega_n^j}\left[(\omega_n^j)^2 - \frac{\varepsilon_j^2}{4} \sin \omega_n^j \cos \omega_n^j\right] + 2\varepsilon_j^2 \sin^2 \omega_n^j.$$

Here, ω_n^j are the positive roots of the equation

$$\frac{4\beta_j \omega_n^j \varepsilon_j}{4(\omega_n^j)^2 - (2\beta_j - 1)\varepsilon_j^2} = \tan \omega_n^j.$$

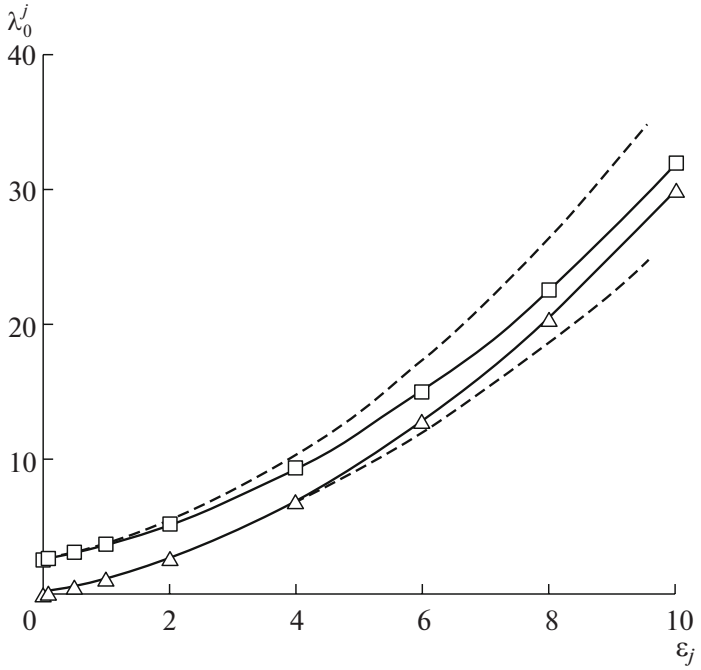


Fig. 1.

For $\beta_j > 0$, this equation has a countable set of positive roots separated roughly by π with the minimum positive root ω_0^j lying in the range $0 < \omega_0^j < \pi$.

For fixed β_j and $\varepsilon_j \longrightarrow 0$, the principal eigenvalue λ_0^j has the asymptotic expansion

$$\lambda_0^j = \beta_j\varepsilon_j + \frac{1}{6}\beta_j(3 - 2\beta_j)\varepsilon_j^2 + O(\varepsilon_j^3), \tag{4.1}$$

which is not regular as $\beta_j \longrightarrow \infty$. Without performing a detailed analysis, we indicate that, for $\beta_j = \infty$ (a completely adsorbing bottom surface) and $\varepsilon_j \longrightarrow 0$, we have

$$\lambda_0^j = \left(\frac{\pi}{2}\right)^2 + \varepsilon_j + \frac{\varepsilon_j^2}{4} + O(\varepsilon_j^3). \tag{4.2}$$

Figure 1 shows the plots of $\lambda_0^j(\varepsilon_j)$ for $\beta_j = 1$ (curve with triangles) and $\beta_j = \infty$ (curve with squares) and asymptotics (4.1) and (4.2) (dashed curves).

In view of the above analysis, in some cases, the shape of the pollution plume at a sufficiently long time after a pollutant discharge can be found by the numerical integration of the two-dimensional advection–diffusion equation (3.4) for the fraction that minimizes λ_0^j. However, if we need to compute the total amount of suspension deposited on the bottom and the suspension concentrations over the entire time interval of plume existence, then series (3.3) and (3.5) have to be summed. An analysis of the case of $K = 1$ shows that, for $\varepsilon_j < 1$, series (3.3) and the inner series in (3.5) are well summable, so, for small ε_j, it is sufficient to retain only the principal diffusion mode Z_0^j in these expansions. However, the convergence of the series degrades tremendously with increasing ε_j and decreasing t. This is not surprising, since problem (2.3) degenerates in the limit as $\varepsilon_j \longrightarrow \infty$. Fortunately, there is a technique for overcoming this difficulty.

5. EQUATION FOR THE DEPTH-AVERAGED CONCENTRATION OF SUSPENDED SUBSTANCES IN A CONSTANT-DEPTH WATER BODY

Applying summation operator (3.5) to Eqs. (3.4) yields the following relations for the depth-averaged suspension concentration $\bar{C}$:[2]

$$\frac{\partial H\bar{C}}{\partial t} + \frac{\partial}{\partial x_i}\{H(\bar{u}_i\bar{C} + \bar{J}_i[\bar{C}])\} + W\bar{C} = 0, \quad \bar{C} = \frac{M}{H(0)}\delta(\mathbf{x})G(\xi_0) \quad \text{при} \quad t = 0, \tag{5.1a}$$

$$W = u_*w, \quad w = \sum_{j=1}^{N}\sum_{n=0}^{\infty}\lambda_n^j C_n^j \bar{Z}_n^j \Big/ \sum_{j=1}^{N}\sum_{n=0}^{\infty} C_n^j \bar{Z}_n^j, \quad M = \sum_j M_j. \tag{5.1b}$$

Here, M is the total initial mass of pollutants and W can be called the effective hydraulic coarseness of the suspension (w is the dimensionless effective hydraulic coarseness).

For a constant-depth water body with $H = \text{const}$, we show that w is independent of $\mathbf{x}$ but is a function of t and describe a technique for determining w.

In the case under study, it is easy to see that the solutions to problems (3.4) can be represented as

$$C_n^j = \frac{M_j}{H} Z_n^j(\xi_0)\rho^j(\xi_0)\mu^0(\mathbf{x}, t)\exp(-u_*\lambda_n^j t/H). \tag{5.2}$$

Here, the function C^0 independent of j or n describes the conservative dispersion of a cloud of unit mass and satisfies the equation

$$\frac{\partial \mu^0}{\partial t} + \frac{\partial}{\partial x_i}(\bar{u}_i\mu^0 + \bar{J}_i[\mu^0]) = 0, \quad \mu^0 = \delta(\mathbf{x}) \quad \text{at} \quad t = 0. \tag{5.3}$$

Substituting (5.2) into (5.1b) gives

$$w \equiv w(t) = \frac{\sum_{j=1}^{N} M_j \sum_{n=0}^{\infty} \lambda_n^j Z_n^j(\xi_0)\rho^j(\xi_0)\exp(-u_*\lambda_n^j t/H)\bar{Z}_n^j}{\sum_{j=1}^{N} M_j \sum_{n=0}^{\infty} Z_n^j(\xi_0)\rho^j(\xi_0)\exp(-u_*\lambda_n^j t/H)\bar{Z}_n^j}. \tag{5.4}$$

As was mentioned above, for arbitrary ε_j and t, the inner series in (5.4) are poorly summable and attempts to determine $w(t)$ by solving eigenvalue problem (3.1) and applying summation (5.4) fail. However, this function is independent of the current velocity $\mathbf{u}$ or the horizontal diffusive flux components $\bar{J}_i$. Therefore, it can be found by integrating the following collection of one-dimensional evolution problems (see (2.3)), averaging their solutions over ξ, summing up the results over the fraction index j, and computing the logarithmic derivative:

$$\frac{\partial Hc^j}{\partial t} + u_*\frac{\partial}{\partial \xi}\left[\varepsilon_j c^j - K(\xi)\frac{\partial c^j}{\partial \xi}\right] = 0, \quad c^j = \frac{M_j}{H}\delta(\xi - \xi_0) \quad \text{at} \quad t = 0,$$

$$\varepsilon_j c^j - K(\xi)\frac{\partial c^j}{\partial \xi} = 0 \quad \text{at} \quad \xi = 0, \quad \varepsilon_j c^j - K(\xi)\frac{\partial c^j}{\partial \xi} = \varepsilon_j\beta_j c^j \quad \text{at} \quad \xi = 1, \tag{5.5}$$

$$\bar{c} = \sum_{j=1}^{N}\int_0^1 c^j G(\xi)d\xi, \quad w(t) = -\frac{H}{u_*}\frac{d\ln\bar{c}}{dt}.$$

It is easy to see that w thus determined depends on the initial disperse composition of the suspension determined by ε_j and M_j, on the nondimensional mass release coefficient β_j, and on the parameter ξ_0 specifying the location of the suspension source over the bottom. Moreover, w also depends on the concentration-averaging method in the vertical direction (i.e., on the chosen form of $G(\xi)$).

Thus, for a constant-depth water body, the computation of the vertically averaged dispersion of a polydisperse suspension with N fractions is reduced to solving N one-dimensional evolution problems (5.5)

[2] The initial condition for Eq. (5.1a) was derived using the delta function expansion $\delta(\xi - \xi_0) = \sum_{n=0}^{\infty} Z_n^j(\xi_0)\rho^j(\xi_0)Z_n^j(\xi)$.

and, then, one two-dimensional evolution problem (5.1a). The solution of the latter is given by the formula

$$\bar{C} = \frac{M}{H}G(\xi_0)\mu^0(\mathbf{x}, t)\exp\left(-\frac{u_*}{H}\int_0^t w(t')dt'\right).$$

Importantly, for a time-continuous and/or spatially distributed source with a constant disperse composition, evolution problems (5.5) can be solved only once.

6. TRAJECTORY APPROXIMATION FOR THE CASE OF A VARIABLE-DEPTH WATER BODY

In the general case, when the depth H is not a constant, the solutions to problems (3.4) cannot be represented in the form of (5.2). Therefore, strictly speaking, the above approach does not apply. However, we can propose an approximate method associated with the concept of the trajectory of a suspension cloud. This method is applicable assuming that the depth of the water body varies little at distances on the order of the characteristic length of the evolving cloud.

The trajectory of motion of the cloud center $\mathbf{x}_0(t)$ is defined in the usual manner as $d\mathbf{x}_0/dt = \mathbf{u}(\mathbf{x}_0, t)$, $\mathbf{x}_0(0) = 0$. Let $H_0(t) = H(\mathbf{x}_0(t))$ be the time-varying depth of the water body at the center of the cloud.

Under the assumption formulated above, the solutions to problems (3.4) can be approximately represented as

$$C_n^j = \frac{M_j}{H}Z_n^j(\xi_0)\rho^j(\xi_0)\mu^0(\mathbf{x}, t)\exp\left(-u_*\lambda_n^j\int_0^t H_0^{-1}(t')dt'\right), \tag{6.1}$$

where μ^0 is a compactly supported function satisfying Eq. (5.3).

Indeed, substituting (6.1) into (3.4), we obtain

$$\frac{\partial\mu^0}{\partial t} + \frac{\partial}{\partial x_i}(\bar{u}_i\mu^0 + \bar{J}_i[\mu^0]) = u_*\lambda_n^j\left(\frac{1}{H_0(t)} - \frac{1}{H(\mathbf{x})}\right)\mu^0. \tag{6.2}$$

Let the characteristic half-width of μ^0 be $\sigma(t)$ and be small in comparison with the characteristic length scale of variations in the reservoir depth. Then the right-hand side of (6.2) is estimated by the small value

$$\frac{\sigma^2(t)}{H_0(t)}\max_i\left|\partial H/\partial x_i\right|_{\mathbf{x} = \mathbf{x}_0(t)}\left| \ll 1.\right.$$

If approximate representation (6.1) can be used, then all the arguments presented in Section 5 are valid. Specifically, we have Eqs. (5.1a) and (5.5) with H replaced by $H_0(t)$.

In this case, $w(t)$ becomes depending on the trajectory of the suspension cloud. However, as before, for a time-continuous and/or spatially distributed source, evolution problems (5.5) are solved only once. Indeed, making in (5.5) the substitutions

$$\mu^j = H_0(t)c^j, \quad \tau = u_*\int_0^t H_0^{-1}(t')dt',$$

we obtain

$$\frac{\partial\mu^j}{\partial\tau} + \frac{\partial}{\partial\xi}\left[\varepsilon_j\mu^j - K(\xi)\frac{\partial\mu^j}{\partial\xi}\right] = 0, \quad \mu^j = M_j\delta(\xi - \xi_0) \quad \text{at} \quad t = 0,$$

$$\varepsilon_j\mu^j - K(\xi)\frac{\partial\mu^j}{\partial\xi} = 0 \quad \text{at} \quad \xi = 0, \quad \varepsilon_j\mu^j - K(\xi)\frac{\partial\mu^j}{\partial\xi} = \varepsilon_j\beta_j\mu^j \quad \text{at} \quad \xi = 1, \tag{6.3}$$

$$\bar{\mu} = \sum_{j=1}^N\int_0^1\mu^j G(\xi)d\xi, \quad w(\tau) = -\frac{d\ln\bar{\mu}}{d\tau}.$$

The standard function $w(\tau)$ thus defined is independent of the trajectory of the suspension cloud.

Table

Fraction index j	Particle size range, mm	Mean diameter D_j, mm	Mass fraction in discharge M_j, %	Hydraulic coarseness W_j, m/s	ε_j
	>40		0.00		
1	40–20	30	1.65	9.845E–01	1.969E+01
2	20–10	15	1.28	6.955E–01	1.391E+01
3	10.0–5.0	7.5	2.17	4.906E–01	9.812E+00
4	5.0–2.0	3.5	2.28	3.323E–01	6.647E+00
5	2.0–1.0	1.5	2.56	2.109E–01	4.217E+00
6	1.0–0.5	0.75	4.58	1.377E–01	2.754E+00
7	0.5–0.25	0.375	8.87	7.823E–02	1.565E+00
8	0.25–0.1	0.175	10.25	2.921E–02	5.841E–01
9	0.1–0.05	0.075	11.56	6.216E–03	1.243E–01
10	0.05–0.01	0.03	21.25	1.010E–03	2.019E–02
11	0.01–0.005	0.0075	15.10	6.316E–05	1.263E–03
12	<0.005		18.45	0.000E+00	0.000E+00

To conclude, we note that, when vertical turbulent mixing can be neglected, W in Eq. (5.1a) must be calculated by the formula

$$W(t) = \sum_{j=1}^{N} M_j W_j \exp\left(-W_j \int_0^t H_0^{-1}(t')dt'\right)\left[\sum_{j=1}^{N} M_j \exp\left(-W_j \int_0^t H_0^{-1}(t')dt'\right)\right]^{-1}.$$

7. EXAMPLES OF COMPUTING THE EFFECTIVE HYDRAULIC COARSENESS OF A POLYDISPERSE SUSPENSION

As an illustration, we computed the effective hydraulic coarseness of an actual polydisperse suspension, namely, light sandy loam, which is met in dredging operations. Its initial disperse composition, the

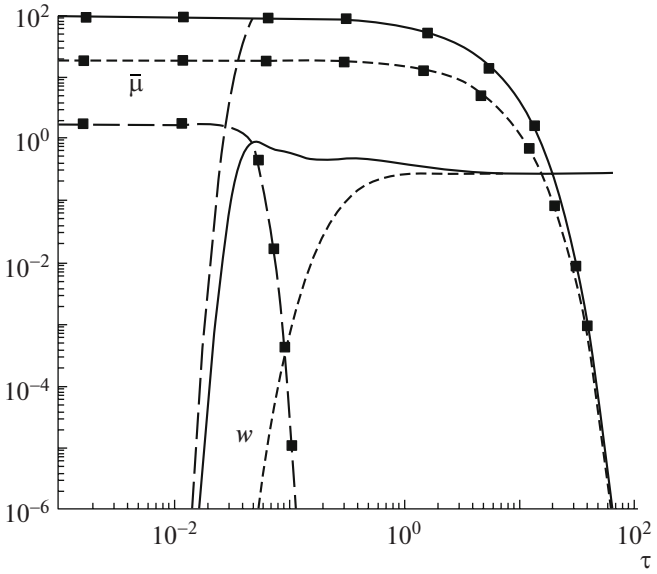


Fig. 2.

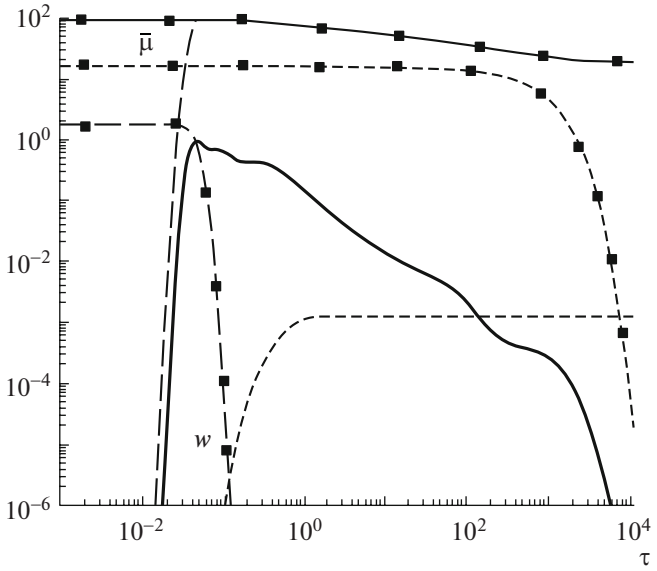


Fig. 3.

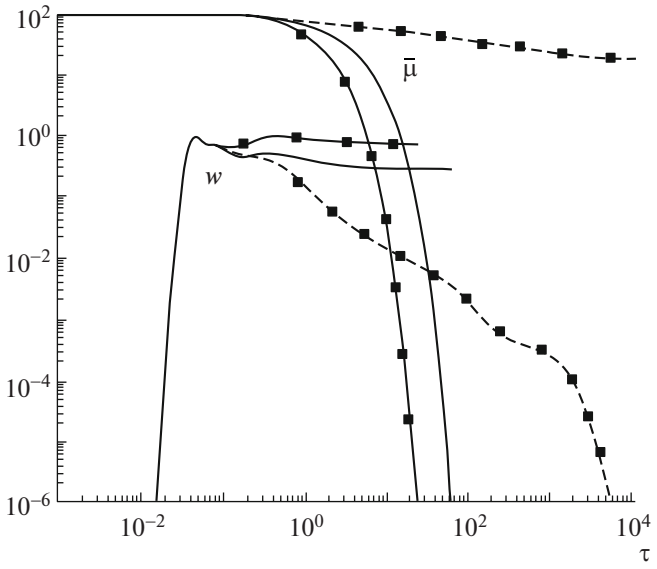


Fig. 4.

hydraulic coarseness of the components, and the values of ε_j are presented in the table. To compute ε_j, the friction velocity at the bottom was specified as $u_* = 0.05$ m/s. The dispersion of the finest fraction with the particle diameter $D_j < 0.005$ mm was treated as conservative; i.e., we set $W_j = 0$.

The dimensionless profile of the vertical turbulent diffusivity was specified by the following expression (see, e.g., [11, 12])[3]:

$$K(\xi) = (1 - \xi + \delta)(0.4 + 0.6\xi). \tag{7.1}$$

Here, δ is a small parameter equal to the ratio of the dimensionless bottom roughness to the depth.

[3] Note that the vertical coordinate z in [11, 12] was measured from the bottom to the reservoir surface.

In the computations, we set $\delta = 0.01$. The dimensionless vertical coordinate of the suspension source was $\xi_0 = 0.1$. We determined the effective hydraulic coarseness of the suspension corresponding to the vertically averaged concentration of suspended substances ($G(\xi) = 1$).

Problem (6.3) was numerically integrated using an implicit conservative difference scheme, which was solved by tridiagonal Gaussian elimination. The computations were performed for the following two cases:

(i) The bottom is completely adsorbing; i.e., $\beta_j = \infty$ for all the suspension fractions (see Fig. 2).

(ii) There are no diffusive fluxes at the bottom; i.e., $\beta_j = 1$ for all the fractions (see Fig. 3).

Both figures show the calculated dynamics of $\bar{\mu}(\tau)$ (curves with squares) and $w(\tau)$. The solid curves correspond to the polydisperse suspension presented in the table. The dashed curves depict the results obtained for the monodisperse suspension corresponding to the largest fraction in the table ($j = 1$). The dotted curves correspond to a monodisperse conservative suspension ($j = 12$ in the table) in Fig. 2. In Fig. 3, they depict the results for the monodisperse suspension corresponding to the finest nonconservative fraction ($j = 11$ in the table).

It can be seen that, at the initial times τ, the mean mass $\bar{\mu}(\tau)$ of suspended particles does not vary, since some time is required for a compact suspension cloud to achieve the bottom of the water body. Therefore, $w(\tau) = 0$. The subsequent evolution of $\bar{\mu}(\tau)$ and $w(\tau)$ is associated with the consecutive deposition of various suspension fractions on the bottom. In the case of a completely adsorbing bottom (see Fig. 2) as $\tau \longrightarrow \infty$, the evolution of these functions is determined by the conservative suspension component adsorbed by the bottom. In the case of no diffusive fluxes to the bottom (Fig. 3), the conservative component is not adsorbed and this fraction always remains in a suspended state, so that $w(\tau) = 0$ for sufficiently long times τ.

For test purposes, we also considered the case of a constant vertical turbulent diffusivity

$$\bar{K} = \int_0^1 K(\xi)d\xi = 0.3 + 0.7\delta \approx 0.3. \tag{7.2}$$

The numerical results are presented in Fig. 4. The solid curves correspond to $\beta_j = \infty$, and the dashed curves, to $\beta_j = 1$. The case of the vertical diffusivity given by (7.1) is shown by curves without markers, while the curves with markers correspond to the averaged vertical diffusivity (7.2). For $\beta_j = 1$, the curves nearly coincide. As was expected, noticeable differences are observed only when the evolution of $\bar{\mu}(\tau)$ and $w(\tau)$ is determined by the diffusion of the suspension to the bottom ($\beta_j = \infty$). In the case of no diffusive fluxes, the curves coincide, which can be explained as follows: on the one hand, vertical diffusion is negligible for large fractions; on the other hand, in the case of fine fractions at $\beta_j = 1$, this process ensures intense turbulent mixing of the suspension regardless of the vertical diffusivity profile.

8. CONCLUSIONS

It was shown that, even if vertical turbulent exchange has a large effect, the problem of computing the evolution of polydisperse suspended pollutants produced by an instantaneous point source in shallow water bodies can be reduced to the integration of N one-dimensional evolution problems (N is the number of suspension fractions) and to the solution of one two-dimensional problem. This result can be used to design efficient numerical methods intended for the mathematical modeling of the dispersion of a suspension produced by a time-continuous and/or spatially distributed source of water pollution.

The concept of effective hydraulic coarseness, which takes into account the adsorption properties of the bottom, can be used to compute pollutant dispersion in rivers, where the flow is frequently simulated using the one-dimensional Saint-Venant equations (see, e.g., [13]). The approach described can be useful for the development of prognostic mathematical models for urban air quality dynamics based on the advection–diffusion equation when detailed three-dimensional simulation (see, e.g., [14, 15]) is impossible.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 08-07-00118, 08-01-00435) and by Basic Research Program no. 3 of the Department of Mathematical Sciences of the Russian Academy of Sciences.

REFERENCES

1. R. I. Nigmatulin, *Dynamics of Multiphase Media* (Nauka, Moscow, 1987) [in Russian].
2. A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics* (MIT Press, Cambridge, Mass., 1971; Nauka, Moscow, 1965, 1967), Vols. 1, 2.
3. J. Nihoul, "Models of Passive Substance Dispersion," in *Modeling of Marine Systems* (Gidrometeoizdat, Leningrad, 1978) [in Russian].
4. B. V. Arkhipov, V. N. Koterov, A. S. Kocherova, et al., "Calculating Sediment Transport in the Coastal Zone of the Sea," Vodn. Resur. **31**, 1–8 (2004) [Water Resour. **31**, 27–34 (2004)].
5. B. V. Arkhipov, V. N. Koterov, V. V. Solbakov, et al., "Numerical Simulation of Pollution and Oil Spill Spreading by the Stochastic Discrete Particle Method," Zh. Vychisl. Mat. Mat. Fiz. **47**, 288–301 (2007) [Comput. Math. Math. Phys. **47**, 280–292 (2007)].
6. B. Arkhipov, V. Koterov, V. Solbakov, et al., "Numerical Modeling of Pollutant Dispersion and Oil Spreading by the Stochastic Discrete Particles Method," Studies in Appl. Math. **120** (1), 87–104 (2008).
7. R. V. Ozmidov, *Diffusion of Tracers in the Ocean* (Gidrometeoizdat, Leningrad, 1986) [in Russian].
8. D. A. Frank-Kamenetskii, *Diffusion and Heat Transfer in Chemical Kinetics* (Nauka, Moscow, 1987) [in Russian].
9. Bao Shi-Shiau and Jia Jung-Juang, "Numerical Study on the Far Field Diffusion of Ocean Dumping for Liquid Waste," *Proceedings of the Eighth International Offshore and Polar Engineering Conference, May 24–29, 1998, Canada* (Int. Soc. of Offshore and Polar Engineers, Cupertino, Canada, 1998), Vol. 2, pp. 327–334.
10. A. N. Tikhonov, A. B. Vasil'eva, and A. G. Sveshnikov, *Differential Equations* (Nauka, Moscow, 1985) [in Russian].
11. E. Wolanski, T. Asaeda, A. Tanaka, et al., "Three-Dimensional Island Wakes in the Field, Laboratory Experiments and Numerical Models," Continental Shelf Res. **16**, 1437–1452 (1996).
12. S. Blaise, E. Deleersnijder, L. White, et al., "Influence of the Turbulence Closure Scheme on the Finite-Element Simulation of the Upwelling in the Wake of Shallow-Water Island," Continental Shelf Res. **27**, 2329–2345 (2007).
13. A. V. Koldoba, Yu. A. Poveshchenko, E. A. Samarskaya, et al., *Methods of Mathematical Modeling of the Environment* (Nauka, Moscow, 2000) [in Russian].
14. I. V. Belov, M. S. Bespalov, L. V. Klochkova, et al., "Transport Model of Gas Tracer Dispersion in the Urban Atmosphere," Mat. Model. **12** (5), 38–46 (2000).
15. N. N. Kalitkin, N. V. Karpenko, A. P. Mikhailov, et al., *Mathematical Models of Nature and Society* (Fizmatlit, Moscow, 2005) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1257–1263. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © V.V. Gorskii, V.A. Sysenko, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1319–1326.

Effective Method for Numerical Integration of the Equations Describing the Flow of High-Temperature Multicomponent Gas Mixtures in Thermochemical Equilibrium

V. V. Gorskii and V. A. Sysenko

Research and Production Association of Machine Constructing, ul. Gagarina 33, Reutov, Moscow oblast, 143966 Russia
e-mail: vpc@npomash.ru

Received June 26, 2008

Abstract—A new stable iterative method is described for computing the flow of a high-temperature multicomponent gas mixture in thermochemical equilibrium. The method is illustrated by computing the thermochemical destruction of a carbon material in a high-temperature airflow.

DOI: 10.1134/S0965542509070161

Key words: numerical method, iterative process, diffusion, thermochemical equilibrium, convergence, boundary layer.

INTRODUCTION

The numerical solution of the equations describing the flow of high-temperature multicomponent gas mixtures usually faces a number of serious problems. The first is related to the computation of concentration profiles for chemical elements and is easy to solve only in the case of diffusion mass transfer in the binary approximation. In more rigorous approaches to the description of diffusion, the convergence of the iterative process used to determine the concentrations of chemical elements is a nontrivial problem (it has to be solved even in the case of air boundary layers, because of the diffusion separation of chemical elements [1]).

Another problem is associated with the convergence of the iterations used between the solution of the momentum and energy conservation equations, on the one hand, and the equations for determining the chemical composition of the gas mixture, on the other hand. The fact is that these systems of equations are highly coupled, since the composition of the gas mixture is uniquely determined by the gas temperature, which, in turn, is considerably affected by the heat release associated with chemical reactions proceeding between the mixture components and diffusion heat transfer.

For these problems, we develop effective solution approaches that can be for wide applications.

It should be noted that examples of solving such problems can be found in numerous publications (see, e.g., [2, 3]). However, the methodology of deriving a solution has received little attention.

1. METHOD FOR COMPUTING CONCENTRATIONS OF CHEMICAL ELEMENTS

The approach proposed in this paper is illustrated by solving the self-similar problem of the laminar flow of an unionized multicomponent gas mixture near the stagnation point of a spherical body.

Diffusion mass transfer in the gas is described by the Stefan–Maxwell equations (see [1, 4]):

$$\kappa_{i,\eta} = \frac{M}{2\rho^2\Psi u_{e,x}} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(\frac{\kappa_i J_j}{M_j D_{i,j}} - \frac{\kappa_j J_i}{M_i D_{i,j}} \right), \quad i = \overline{1, n-1}, \tag{1.1}$$

$$\Psi = 1/\sqrt{\rho_e \mu_e u_{e,x}}.$$

Here, κ_i, J_i, and M_i are the molar concentration, diffusive mass flux, and molar mass of the ith species in the mixture; $D_{i,j}$ is the binary diffusivity of the ith species in the jth species; ρ_e, μ_e, and $u_{e,x}$ are the density, dynamic viscosity, and velocity gradient on the outer edge of the boundary layer; M and ρ are the molar mass and density of the mixture; and n is the number of species in the mixture.

The index η denotes differentiation with respect to the Lees–Dorodnicyn self-similar coordinate η, and the nth species is understood as a neutral component (for example, argon).

Then, eliminating κ_n and J_n from (1.1) with the help of the Dalton equation and the condition that the diffusive mass fluxes of the components add up to zero, we use the equation of state and introduce the notation

$$W_{i,j} = \begin{cases} S_i\kappa_j\left(\dfrac{M_n}{M_j}\dfrac{D_{i,n}}{D_{i,j}} - 1\right), & j \neq i \\ -S_i\left[\displaystyle\sum_{\substack{j_1 = 1 \\ j_1 \neq i}}^{n-1} \kappa_{j_1}\frac{M_n}{M_i}\left(\frac{D_{i,n}}{D_{i,j_1}} - 1\right) + \kappa_i + \frac{M_n}{M_i}(1-\kappa_i)\right], & j = i, \end{cases} \tag{1.2}$$

$$S_i = \frac{M}{2\rho^2\Psi u_{e,x}M_nD_{i,n}} = \frac{R_{\text{un}}^2T^2}{2P^2M\Psi u_{e,x}M_nD_{i,n}},$$

to obtain

$$\kappa_{i,\eta} = \sum_{j=1}^{n-1} W_{i,j}J_j, \quad i = \overline{1, n-1}. \tag{1.3}$$

Here, R_{un} is the universal gas constant.

In turn, the chemical composition of the mixture in thermochemical equilibrium is uniquely determined by the temperature T, pressure P, and concentrations C_k^+ of the chemical elements, where $k = \overline{1, m-1}$, and m is the number of chemical elements forming the species in the mixture. Here, the mth chemical element is that forming the nth neutral component.

Then we can write $\kappa_i(\eta) = \kappa_i[C_1^+(\eta), \ldots, C_{m-1}^+(\eta), T(\eta), P(\eta)]$ and

$$\kappa_{i,\eta} = \sum_{k=1}^{m-1} \kappa_{i,C_k}C_{k,\eta}^+ + \kappa_{i,T}T_\eta + \kappa_{i,P}P_\eta, \quad i = \overline{1, n-1}. \tag{1.4}$$

Here, the indices C_k, T, and P denote the partial derivatives of corresponding functions with respect to C_k^+, T, and P.

Substituting (1.4) into (1.3) yields a system of algebraic equations for determining the diffusive mass fluxes of the species as functions of the concentration derivatives with respect to η:

$$\sum_{j=1}^{n-1} W_{i,j}J_j = \sum_{k=1}^{m-1} \kappa_{i,C_k}C_{k,\eta}^+ + \kappa_{i,T}T_\eta + \kappa_{i,P}P_\eta, \quad i = \overline{1, n-1}. \tag{1.5}$$

In problems related to the boundary layer theory, the derivative P_η is usually neglected, but the method in question is intended for a wider class of problems.

The computational method proposed is based on a specialized thermodynamic software code that (in contrast to similar standard codes) determines not only the species concentrations but also their partial (first, second, and mixed) derivatives with respect to all the arguments (i.e., with respect to the temperature, pressure, and concentrations of the chemical elements).

By applying this code, system (1.5) can be linearized to the zeroth order about the initial approximation for the concentrations of the chemical elements. The linearized system of equations is solved, for example, by Gaussian elimination, and the resulting solution is written as

$$J_i = \sum_{k=1}^{m-1} \xi_{i,k}C_{k,\eta}^+ + \xi_{i,m}, \quad i = \overline{1, n-1}.$$

As a result, the diffusive mass fluxes J_l^+ of the chemical elements can be written as functions of the concentration derivatives:

$$J_l^+ = \sum_{i=1}^{n-1} J_i \nu_{l,i} = \sum_{k=1}^{m-1} C_{k,\eta}^+ \sum_{i=1}^{n-1} \xi_{i,k} \nu_{l,i} + \sum_{i=1}^{n-1} \xi_{i,m} \nu_{l,i}, \quad l = \overline{1, m-1}. \tag{1.6}$$

Here, $\nu_{l,i}$ is the mass content of the lth chemical element in the ith species.

Solving the system of linear algebraic equations (1.6) yields the concentration derivatives of the chemical elements as functions of their diffusive mass fluxes:

$$C_{k,\eta}^+ = F_k = \sum_{l=1}^{m-1} \zeta_{k,l} J_l^+ + \zeta_{k,m}, \quad k = \overline{1, m-1}. \tag{1.7}$$

Relations (1.7) are supplemented by the mass conservation equations for the chemical elements:

$$J_{k,\eta}^+ = F_{k+m-1} = C_{k,\eta}^+ f/\Psi = F_k f/\Psi, \quad k = \overline{1, m-1},$$

where f is the dimensionless stream function. As a result, for determining the iterative values of the concentrations and diffusive mass fluxes of the chemical elements, we obtain a boundary value problem based on a system of first-order ordinary differential equations solved for the derivatives. This problem can be solved by any standard technique. We used the method of fundamental functions (see, e.g., [5]) combined with the fourth-order Runge–Kutta method [6], in which the prescribed accuracy of the desired solution was ensured by an automatic step size choice procedure.

It will be shown below that, for boundary layer problems, this technique ensures the quadratic convergence of the iterative process for determining the concentration profiles and diffusive mass fluxes corresponding to a given function $T(\eta)$. We can hope that a similar pattern is observed for nonzero derivatives of $P(\eta)$.

2. COMBINED SOLUTION OF THE ENERGY CONSERVATION EQUATIONS AND THE CHEMICAL COMPOSITION EQUATIONS

The momentum, energy, and mass conservation equations for the mixture species can be written as

$$\left(\frac{\rho\mu}{\rho_e\mu_e} f_{\eta\eta}\right)_\eta = -ff_{\eta\eta} - 0.5\left(\frac{\rho_e}{\rho} - f_\eta^2\right), \tag{2.1}$$

$$\begin{gathered}\left(\frac{\rho\lambda}{\rho_e\mu_e} T_\eta\right)_\eta = -fc_p T_\eta + \Psi T_\eta \sum_{i=1}^{n-1} (c_{p,i} - c_{p,n}) J_i + \sum_{i=1}^{n-1} \frac{\omega_i h_i}{2\rho_{u_{e,x}}}, \\ \frac{\omega_i}{2\rho u_{e,x}} = \Psi J_{i\eta} - fC_{i,\eta}, \quad i = \overline{1, n-1}.\end{gathered} \tag{2.2}$$

Here, C_i, $c_{p,i}$, ω_i, and h_i are the mass concentration, specific heat capacity at constant pressure, the mass production rate due to chemical reactions, and the enthalpy of the ith species; and c_p, μ, and λ are the specific heat capacity at constant pressure, dynamic viscosity, and the molecular thermal conductivity of the mixture.

Then the intensity of heat absorption due to chemical reactions can be written as

$$Q_{\text{chem}} = \sum_{i=1}^{n-1} \frac{\omega_i h_i}{2\rho u_{e,x}} = \sum_{i=1}^{n-1} (\Psi J_{i,\eta} - fC_{i,\eta}) h_i. \tag{2.3}$$

The iterative process between consecutive solutions of the momentum and energy conservation equations, on the one hand, and the chemical composition equations, on the other hand, is organized so that the temperature derivatives are extracted from J_i, $J_{i,\eta}$ and $C_{i,\eta}$; i.e., the latter are represented as

$$J_{i,\eta} = \alpha_i T_{\eta\eta} + \beta_i T_\eta^2 + \gamma_i T_\eta + \delta_i, \quad i = \overline{1, n-1}, \tag{2.4}$$

$$C_{i,\eta} = \gamma_{1,i} T_\eta + \delta_{1,i}, \quad i = \overline{1, n-1}, \tag{2.5}$$

$$J_i = \gamma_{2,i} T_\eta + \delta_{2,i}, \quad i = \overline{1, n-1}. \tag{2.6}$$

This problem is solved as follows.

The derivatives of the diffusive mass fluxes of the mixture species are determined by differentiating system (1.5) with respect to η. Then, representing $\kappa_{i,\eta\eta}$ in the form

$$\kappa_{i,\eta\eta} = \kappa_{i,T}T_{\eta\eta} + \kappa_{i,TT}T_\eta^2 + q_iT_\eta + r_i, \quad i = \overline{1, n-1},$$

where

$$q_i = 2\sum_{k=1}^{m-1}\kappa_{i,T,C_k}C^+_{k,\eta} + 2\kappa_{i,T,P}P_\eta,$$

$$r_i = \sum_{k=1}^{m-1}\left(\kappa_{i,C_k}C^+_{k,\eta\eta} + C^+_{k,\eta}\sum_{l=1}^{m-1}\kappa_{i,C_k,C_l}C^+_{l,\eta} + 2P_\eta C^+_{k,\eta}\kappa_{i,P,C_k}\right) + \kappa_{i,P}P_{\eta\eta} + \kappa_{i,PP}P_\eta^2,$$

we obtain

$$\sum_{j=1}^{n-1}W_{i,j}J_{j,\eta} = \kappa_{i,T}T_{\eta\eta} + \kappa_{i,TT}T_\eta^2 + q_iT_\eta + r_i - \sum_{j=1}^{n-1}W_{i,j,\eta}J_j, \quad i = \overline{1, n-1}. \tag{2.7}$$

The concentration profiles of the chemical elements are determined by the iterative procedure described above, while the molar concentrations of the species and all their partial derivatives are calculated using the thermodynamic code.

In turn, according to (1.2), if $j \neq i$, we have

$$W_{i,j,\eta} = W_{i,j}\frac{S_{i,\eta}}{S_i} + S_i\kappa_{i,\eta}\left(\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}} - 1\right) + S_i\kappa_i\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}}\left(\frac{D_{i,n,\eta}}{D_{i,n}} - \frac{D_{i,j,\eta}}{D_{i,j}}\right),$$

The binary diffusivities of the species depend only on temperature and pressure (see [4]), and

$$W_{i,i,\eta} = W_{i,i}\frac{S_{i,\eta}}{S_i} - S_i\sum_{\substack{j=1\\ j\neq i}}^{n-1}\left[\kappa_{j,\eta}\frac{M_n}{M_i}\left(\frac{D_{i,n}}{D_{i,j}} - 1\right) + \kappa_j\frac{M_n}{M_i}\frac{D_{i,n}}{D_{i,j}}\left(\frac{D_{i,n,\eta}}{D_{i,n}} - \frac{D_{i,j,\eta}}{D_{i,j}}\right)\right] - S_i\kappa_{i,\eta}\left(1 - \frac{M_n}{M_i}\right).$$

$$\frac{S_{i,\eta}}{S_i} = 2\frac{T_\eta}{T} - \frac{M_\eta}{M} - \frac{D_{i,n,\eta}}{D_{i,n}} - 2\frac{P_\eta}{P} = T_\eta\left(\frac{2}{T} - \frac{M_T}{M} - \frac{D_{i,n,T}}{D_{i,n}}\right) - \sum_{k=1}^{m-1}C^+_{k,\eta}\frac{M_{C_k}}{M} - P_\eta\left(\frac{2}{P} + \frac{M_P}{M} + \frac{D_{i,n,P}}{D_{i,n}}\right).$$

Then, by using (1.4), the general formula for computing $W_{i,j,\eta}$ can be rewritten as

$$W_{i,j,\eta} = U_{i,j}T_\eta + V_{i,j},$$

where

$$U_{i,j} = W_{i,j}\left(\frac{2}{T} - \frac{M_T}{M} - \frac{D_{i,n,T}}{D_{i,n}}\right) + S_i\kappa_{i,T}\left(\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}} - 1\right) + S_i\kappa_i\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}}\left(\frac{D_{i,n,T}}{D_{i,n}} - \frac{D_{i,j,T}}{D_{i,j}}\right), \quad j \neq i,$$

$$V_{i,j} = -W_{i,j}\sum_{k=1}^{m-1}\frac{M_{C_k}}{M}C^+_{k,\eta} + S_i\left(\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}} - 1\right)\sum_{k=1}^{m-1}\kappa_{i,C_k}C^+_{k,\eta} - P_\eta\left[W_{i,j}\left(\frac{2}{P} + \frac{M_P}{M} + \frac{D_{i,n,P}}{D_{i,n}}\right)\right.$$
$$\left. - S_i\kappa_{i,P}\left(\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}} - 1\right) - S_i\kappa_i\frac{M_n}{M_j}\frac{D_{i,n}}{D_{i,j}}\left(\frac{D_{i,n,P}}{D_{i,n}} - \frac{D_{i,j,P}}{D_{i,j}}\right)\right], j \neq i,$$

$$U_{i,i} = W_{i,i}\left(\frac{2}{T} - \frac{M_T}{T} - \frac{D_{i,n,T}}{D_{i,n}}\right) - S_i\kappa_{i,T}\left(1 - \frac{M_n}{M_i}\right)$$
$$- S_i\frac{M_n}{M_i}\sum_{\substack{j_1=1\\ j_1\neq i}}^{n-1}\left[\kappa_{j_1,T}\left(\frac{D_{i,n}}{D_{i,j_1}} - 1\right) + \kappa_{j_1}\frac{D_{i,n}}{D_{i,j_1}}\left(\frac{D_{i,n,T}}{D_{i,n}} - \frac{D_{i,j_1,T}}{D_{i,j_1}}\right)\right],$$

$$V_{i,i} = -\sum_{k=1}^{m-1} C^+_{k,\eta}\left[W_{i,i}\frac{M_{C_k}}{M} + S_i\kappa_{i,C_K}\left(1-\frac{M_n}{M_i}\right) + S_i\frac{M_n}{M_i}\sum_{\substack{j_1=1\\ j_1\neq i}}^{n-1}\kappa_{j_1,C_k}\left(\frac{D_{i,n}}{D_{i,j_1}}-1\right)\right]$$

$$-P_\eta\left\{S_i\kappa_{i,P}\left(1-\frac{M_n}{M_i}\right) + S_i\frac{M_n}{M_i}\sum_{\substack{j_1=1\\ j_1\neq i}}^{n-1}\left[\kappa_{j_1,P}\left(\frac{D_{i,n}}{D_{i,j_1}}-1\right) + \kappa_{j_1}\frac{D_{i,n}}{D_{i,j_1}}\left(\frac{D_{i,n,P}}{D_{i,n}} - \frac{D_{i,j_1,P}}{D_{i,j_1}}\right)\right]\right\} - P_\eta W_{i,i}\left(\frac{2}{P}+\frac{M_P}{M}+\frac{D_{i,n,P}}{D_{i,n}}\right).$$

In view of the notation introduced, system (2.7) becomes

$$\sum_{j=1}^{n-1} W_{i,j}J_{j,\eta} = T_{\eta\eta}\kappa_{i,T} + T_\eta^2\kappa_{i,TT} + T_\eta\left(q_i - \sum_{j=1}^{n-1}U_{i,j}J_j\right) + r_i - \sum_{j=1}^{n-1}V_{i,j}J_j, \quad i = \overline{1,n-1}.$$

Solving these set of systems of linear algebraic equations by any standard method (for example, by Gaussian elimination), we uniquely determine the coefficients α_i, β_i, γ_i, and δ_i in (2.4).

By analogy with (1.4), we can write

$$C_{i,\eta} = C_{i,T}T_\eta + \sum_{k=1}^{m-1} C_{i,C_k}C^+_{k,\eta}, \quad i = \overline{1,n-1}. \tag{2.8}$$

A comparison of (2.5) with (2.8) gives the formulas

$$\gamma_{1,i} = C_{i,T}, \quad \delta_{1,i} = \sum_{k=1}^{m-1} C_{i,C_k}C^+_{k,\eta}.$$

The coefficients in (2.6) are uniquely determined by solving the set of systems of linear algebraic equations (1.5), for example, by applying Gaussian elimination.

Substituting expansions (2.4) and (2.5) into (2.3) yields

$$Q_{\text{chem}} = T_{\eta\eta}\Psi\sum_{i=1}^{n-1}\alpha_i h_i + T_\eta^2\Psi\sum_{i=1}^{n-1}\beta_i h_i + T_\eta\sum_{i=1}^{n-1}(\gamma_i\Psi - \gamma_{1,i}f)h_i + \sum_{i=1}^{n-1}(\delta_i\Psi - \delta_{1,i}f)h_i.$$

By combining this expression (for the heat release intensity due to chemical reactions) with (2.6), Eq. (2.2) takes the final form

$$\begin{aligned} T_{\eta\eta}\left(\frac{\rho\mu}{\rho_e\mu_e} - \Psi\sum_{i=1}^{n-1}\alpha_i h_i\right) &= T_\eta^2\Psi\sum_{i=1}^{n-1}[\beta_i h_i + \gamma_{2,i}(c_{p,i}-c_{p,n})] + \sum_{i=1}^{n-1}(\delta_i\Psi - \delta_{1,i}f)h_i \\ &+ T_\eta\left\{\sum_{i=1}^{n-1}[(\gamma_i\Psi - \gamma_{1,i}f)h_i + \Psi\delta_{2,i}(c_{p,i}-c_{p,n})] - fc_p - \left(\frac{\rho\mu}{\rho_e\mu_e}\right)_\eta\right\}. \end{aligned} \tag{2.9}$$

Given boundary conditions on the surface of the body, at every outer iteration, the initial approximations to the unknown functions $f(\eta)$ and $T(\eta)$ are used to calculate the concentrations of the chemical elements. Next, we find the functions characterizing the chemical composition of the mixture and all the coefficients of Eq. (2.9) depending on it. Finally, the iterative values of $f_{it}(\eta)$ and $T_{it}(\eta)$ are determined by solving Eqs. (2.1) and (2.9).

In the transition from the kth iteration to the next one, the outer process makes use of the weighted scheme

$$\begin{aligned} f^{(k+1)}(\eta) &= f_{it}^{(k)}, \\ T^{(k+1)}(\eta) &= T^{(k)}(\eta)(1-W) + T_{it}^{(k)}(\eta)W, \\ W &= W_{\max}\min(1, D_{\max}/D). \end{aligned} \tag{2.10}$$

Here, W is the current value of the weighting coefficient; $W_{\max}$ is its maximum possible value; D is the current mismatch of the initial approximation; $D_{\max}$ is its maximum possible value, at which the maximum value of W is used; and $\min(a, b)$ denotes the minimum among a and b.

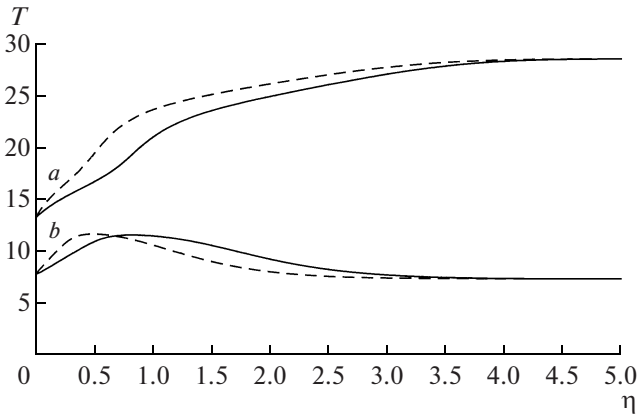


Fig. 1.

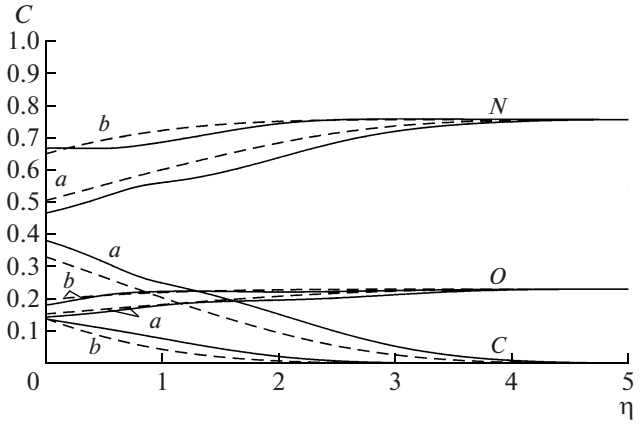


Fig. 2.

3. EXAMPLE OF CONVERGENCE OF ITERATIVE PROCESSES

The convergence of the iterative processes described above is illustrated by computing the airflow past a carbon sphere of radius 0.1 m at various Mach numbers M_∞. The pressure in the boundary layer was 1 atm, and the carbon ablation was based on the complete thermochemical destruction model described in [7].

First, the results showed that the convergence of the iterative processes degrades substantially when the squared temperature derivative with respect to η is extracted in (2.4). Accordingly, all the results presented below were obtained in the case of (2.4) replaced by the decomposition

$$J_{i,\eta} = \alpha_i T_{\eta\eta} + \gamma_i' T_\eta + \delta_i, \quad i = \overline{1, n-1},$$

$$\gamma_i' = \gamma_{1,i} T_\eta + \delta_{1,i}.$$

Figures 1 and 2 show the initial approximations of $T(\eta)$ and $C_k^+(\eta)$ (dashed lines) (which were constructed using the analogy between heat and mass transfer) and the corresponding numerical solutions (solid lines) for (a) $M_\infty = 25$ and (b) $M_\infty = 6$.

Figure 3 demonstrates the convergence of the iterative process used to compute the concentrations of the chemical elements. The convergence of the outer iterative process between the computation of dynamic and thermal flow characteristics, on the one hand, and the chemical gas composition, on the other hand, at the same Mach numbers is shown in Fig. 4. This figure also presents the weighting coefficient used to compute the initial approximation to the temperature profile at the next iteration by formula (2.10) depending on the iteration number.

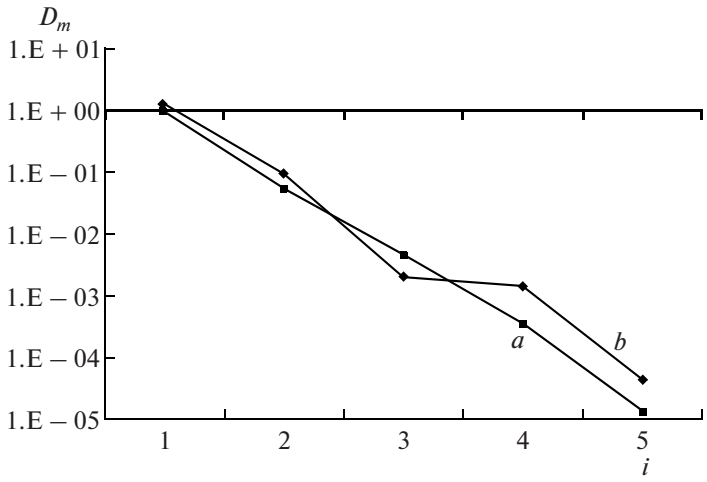


Fig. 3.

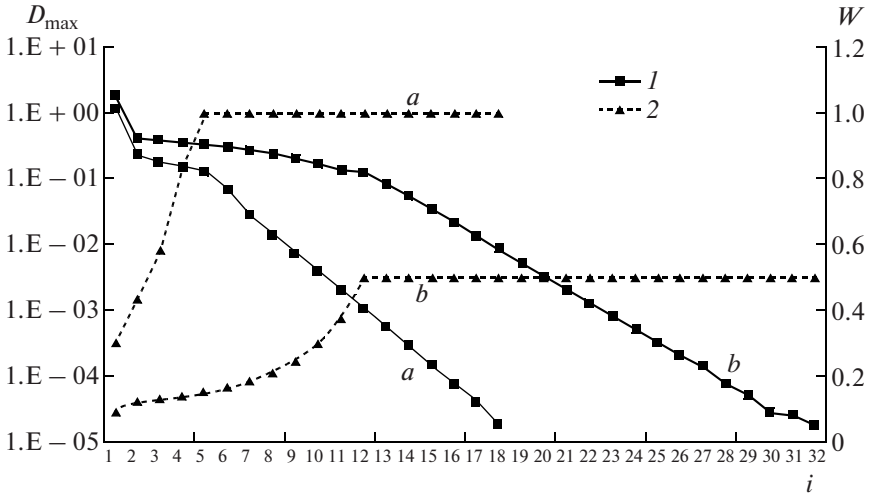


Fig. 4.

Here, i is the iteration number, D_m is the maximum relative error in the computed concentrations of the chemical elements, D_{max} is the maximum relative error in the unknown functions computed in the outer iterative process (line *1*), and W is the current weighting coefficient (line *2*).

The numerical results suggest that the approach proposed for solving the equilibrium boundary layer equations ensures the fast convergence of the corresponding iterative processes.

REFERENCES

1. N. A. Anfimov, "On Some Effects Associated with the Multicomponent Character of Gas Mixture," Izv. Akad. Nauk SSSR. Otd. Tekh. Nauk. Mekh. Mashinostr., No. 5, 117–123 (1963).
2. S. M. Scala and L. M. Gilbert, "Sublimation of Graphite at Hypersonic Speeds," AIAA J., No. 9, 87–100 (1965).
3. F. S. Zavelevich, "Graphite Combustion in a Chemically Equilibrium Boundary Layer," Izv. Akad. Nauk SSSR. Mekh. Zhidk. Gaza, No. 1, 161–167 (1966).
4. J. O. Hirschfelder, C. F. Curtiss, and R. B. Bird, *Molecular Theory of Gases and Liquids* (Wiley, New York, 1954; Inostrannaya Literatura, Moscow, 1961).
5. V. V. Stepanov, *A Course of Differential Equations* (Gostekhteorizdat, Moscow, 1950) [in Russian].
6. I. S. Berezin and N. P. Zhidkov, *Computational Methods* (Fizmatgiz, Moscow, 1962), Vol. 2 [in Russian].
7. V. V. Gorskii and Yu. V. Polezhaev, *Combustion Theory* (Energomash, Moscow, 2006) [in Russian].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1264–1275. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © Yu.V. Maksimov, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1327–1339.

Correct Algebras over Estimation Algorithms in the Set of Regular Recognition Problems with Nonoverlapping Classes

Yu. V. Maksimov

Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudnyi, Moscow oblast, 141700 Russia
e-mail: yuramaksimov@gmail.com

Received May 23, 2008; in final form, December 16, 2008

Abstract—Algebras over estimation algorithms in the set of regular problems with nonoverlapping classes are considered. A correctness criterion for the arbitrary degree algebraic closure of the model of estimation algorithms in the classification problems of this type is proposed; this criterion can be efficiently verified. An estimate of the minimal degree of the algebraic closure that is sufficient for constructing a correct classifier in an arbitrary regular problem with nonoverlapping classes is found.

DOI: 10.1134/S0965542509070173

Key words: classification problems, algebraic approach, estimation algorithms, correct algorithms.

INTRODUCTION

In the framework of the algebraic approach to pattern recognition proposed by Zhuravlev, it was shown that simple algebraic operations over the algorithms based on incorrect (heuristic) algorithms make it possible to construct correct algorithmic compositions. The base algorithms can be described in the model of estimation algorithms proposed by Zhuravlev. For the case of the problems with nonoverlapping classes, we study the question concerning the sufficiency of the compositions that can be obtained by an arbitrary degree algebraic closure of the model of estimation algorithms (EAs) for the correct classification of a test sample for a given set of training objects. The case of the linear closure of the EA model is considered separately. A sharp estimate of the degree of the algebraic closure that is sufficient for the construction of a correct classifier in an arbitrary regular problem is obtained.

1. STATEMENT OF THE PROBLEM: BASIC DEFINITIONS AND REPRESENTATION OF RECOGNIZING OPERATORS

Consider the pattern recognition problem with two nonoverlapping classes in the classical statement [1, 2]. The set of admissible objects is a union of nonempty nonoverlapping classes

$$M = \bigcup_{i=1}^{l} K_i \subseteq M_1 \times \ldots \times M_n, \quad K_i \cap K_j = \varnothing, \tag{1}$$

where M_i is the set of values of feature i equipped with the semimetric ρ_i. Every admissible object is defined by its standard feature description $I(S) = (a_1(S), ..., a_n(S))$, where $a_i \in M_i$, $i \in \{1, 2, ..., n\}$. Each admissible object S is assigned a genuine Boolean information vector $\alpha(S) = (\alpha_1(S), ..., \alpha_l(S))$, where $\alpha_i(S)$ are the values of the predicate $S \in K_i$ for $i \in \{1, 2, ..., l\}$. The genuine standard information is defined as the collection of the sets $(I(S_1), ..., I(S_m))$ and $(\alpha(S_1), ..., \alpha(S_m))$ of the complete standard descriptions of the training objects and their genuine information vectors.

The pattern recognition problem is to construct an algorithm A such that

$$A(I_0(K_1, \ldots, K_l), I(S)) = \alpha^A(S) \equiv (\alpha_1^A(S), \ldots, \alpha_l^A(S)), \quad \text{where} \quad \alpha^A(S) = \alpha(S), \quad S \in M,$$

for the genuine standard information $I_0(K_1, ..., K_l)$.

The algorithms in the EA model we want to examine are a superposition of a recognizing operator and a decision rule. The recognizing operator calculates estimates of the distances from the objects to the classes, and the decision rule classifies the objects on the basis of these estimates.

The recognizing operator B in the EA model calculates an estimate of the distance from the object i to the class j using the following formula proposed in [2]:

$$\Gamma_{ij}[B] = \frac{x_1}{N_1(j)} \sum_{\Omega \in \Omega_A} \sum_{S^t \in \tilde{S}^m \cap K_j} \omega^t \omega(\Omega) B_\Omega^{\tilde{\varepsilon}}(S^t, S_i) + \frac{x_0}{N_0(j)} \sum_{\Omega \in \Omega_A} \sum_{r \neq j} \sum_{S^t \in \tilde{S}^m \cap K_r} \omega^t \omega(\Omega)(1 - B_\Omega^{\tilde{\varepsilon}}(S^t, S_i)),$$

here, $x_0, x_1 \in \{0, 1\}$; $N_0(j)$ and $N_1(j)$ are normalizing coefficients; $\{S^1, \ldots, S^m\} = \tilde{S}^m$ is the set of training objects; $\{S_1, \ldots, S_q\} = \tilde{S}_q$ is the test sample; $K_1, \ldots, K_l$ are the classes of objects defined in (1); Ω_A is the system of support sets (the set of subsets of $\{1, 2, \ldots, n\}$); $\omega(\Omega) \in \mathbb{Q}^+$ is the weight of the support set formed by the weights of the features belonging to Ω; ω^t are the weights of the admissible objects; and $B_\Omega^{\tilde{\varepsilon}}(S^t, S_i)$ is the binary proximity function of which the value is determined on the basis of the analysis of the similarity between the descriptions of the objects S^t and S_i.

The proximity function (see [2] for details) is determined by the parameters $\tilde{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$, where $\varepsilon_i \in \mathbb{R}^+$ for any $i \in \{1, 2, \ldots, n\}$, which indicate the accuracy of measuring the features; in the most general case, it also depends on the integer nonnegative parameters q_1 and q_2:

$$B_\Omega^{\tilde{\varepsilon}, q_1, q_2}(S^t, S) = 1 \Leftrightarrow \begin{cases} \left|\{i \in \Omega : \rho_i(a_i(S^t), a_i(S)) \leq \varepsilon_i\}\right| \geq q_1, \\ \left|\{i \in \Omega : \rho_i(a_i(S^t), a_i(S)) > \varepsilon_i\}\right| \leq q_2. \end{cases} \quad (2)$$

It was proved in [1] that an arbitrary proximity function defined by Eq. (2) is a linear combination of binary proximity functions

$$B_{\{1, 2, \ldots, n\}}^{\tilde{\eta}, n, 0} = 1 \Leftrightarrow \left|\{i \in \{1, \ldots, n\} : \rho_i(a_i, b_i) \leq \eta_i\}\right| = n, \quad (3)$$

when the parameters $\tilde{\eta} = \tilde{\eta}(\tilde{\varepsilon}, \Omega)$ are chosen appropriately. The functions defined by Eq. (3) are denoted by $B[\tilde{\eta}] : B[\tilde{\eta}] = B_{\{1, 2, \ldots, n\}}^{\tilde{\eta}, n, 0}$. Therefore, in the analysis of the closures of the EA model, one can consider only the proximity functions of this type.

Consider the function

$$\Delta_\delta[\tilde{\varepsilon}] = \sum_{(\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n} B[(\varepsilon_1 + \sigma_1\delta_1, \ldots, \varepsilon_n + \sigma_n\delta_n)] \prod_{i=1}^n \sigma_i. \quad (4)$$

It was proved in [1] that, for an arbitrary $\tilde{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ and an arbitrary finite set of admissible objects, a small vector $\tilde{\delta}'$ ' with positive components can be found such that, for every $\delta < \delta'$, every term on the right-hand side of (4) is independent of δ. Therefore, the function $\Delta_\delta[\tilde{\varepsilon}]$ is also independent of small values of δ. It is called a labeling function and is denoted by $\Delta[\tilde{\varepsilon}]$:

$$\Delta[\tilde{\varepsilon}] = \sum_{(\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n} B[(\varepsilon_1 + \sigma_1\delta_1, \ldots, \varepsilon_n + \sigma_n\delta_n)] \prod_{i=1}^n \sigma_i.$$

Let $E_t(\tilde{\varepsilon}) = \{S \in M : \tilde{\rho}(S^t, S) = \tilde{\varepsilon}\}$. In the same book [1], the following propositions are proved.

Theorem 1. *For any finite set of admissible objects $\tilde{S}_q$, there exists a real (rational) vector $(\delta_1, \ldots, \delta_n)$ $(\delta_i > 0$, $i \in \{1, 2, \ldots, n\})$ such that, for any object, we have $S \in \tilde{S}_q : \Delta[\tilde{\varepsilon}](S^t, S) = 1 \Leftrightarrow S \in E_t(\tilde{\varepsilon})$.*

Theorem 2. *Let M be a finite set. Then,*

$$B[(\varepsilon_1, \ldots, \varepsilon_n)](S^t, \cdot) = \sum_{\tilde{\rho} \in P} \Delta[\tilde{\rho}](S^t, \cdot),$$

where $P = \{\tilde{\rho}(S^t, S) : S \in M, \rho_i(a_i(S^t), a_i(S)) \leq \varepsilon_i \ \forall i \in \{1, 2, \ldots, n\}\}$.

Define the set of labeling operators $D^* = \{D_{t,d}[\tilde{\varepsilon}]\}_{t,\tilde{\varepsilon},\tilde{N}_d,d}$ as follows (the correctness of this definition was proved in [1]):

$$\Gamma_{ij}(D_{t,d}[\tilde{\varepsilon}]) = \frac{1}{N_d(j)}\begin{cases}\Delta[\tilde{\varepsilon}](S^t, S_i), & S^t \in K_j^d \\ 0, & S^t \in K_j^{1-d},\end{cases} \tag{5}$$

$$\Gamma(D_{t,d}[\tilde{\varepsilon}]) = \Delta^t[\tilde{\varepsilon}] \otimes (\tilde{\alpha}(S^t)^d \circ \tilde{N}_d) \in \mathbb{Q}^{q\times l}. \tag{6}$$

By the product of two vectors $\alpha \otimes \beta = \alpha^{\mathrm{T}} \otimes \beta$, where $\alpha = (\alpha_1, \ldots, \alpha_l)^{\mathrm{T}}$, and $\beta = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$, we mean the matrix $A = \|a_{ij}\| \in \mathbb{R}^{ql}: a_{ij} = \alpha_i\beta_j$. By the product of two matrices $A \circ B$, where $A = \|a_{ij}\| \in \mathbb{R}^{ql}$ and $B = \|b_{ij}\| \in \mathbb{R}^{ql}$, we mean the matrix $C = \|c_{ij}\| \in \mathbb{R}^{ql}: c_{ij} = a_{ij}b_{ij}$. The properties of these operations are thoroughly considered in [1]. In this paper, the properties of the operations $\otimes$ and $\circ$ are used without explanations.

In formulas (5) and (6), we use the following notation:

$$K_j^0 = S^m \backslash K_j, \quad K_j^1 = K_j, \quad \tilde{N}_d = \left(\frac{1}{N_d(1)}, \ldots, \frac{1}{N_d(l)}\right)$$

$$\text{for} \quad d \in \{0, 1\}, \quad \Delta^t[\tilde{\varepsilon}] = (\Delta[\tilde{\varepsilon}](S^t, S_1)\ldots\Delta[\tilde{\varepsilon}](S^t, S_q))^{\mathrm{T}}.$$

Let $P(t) = \{\tilde{\rho}(S^t, S_i)\}_{i=1}^q = \{\tilde{\rho}_1^t, \ldots, \tilde{\rho}_{p(t)}^t\}$ and $|P(t)| = p(t)$. Assume that $p = \sum_{t=1}^m p(t)$. From the binary vectors $\Delta^t[\tilde{\varepsilon}]$, we compose the matrices

$$\Theta^t = [\Delta^t[\tilde{\rho}_1^t]\ldots\Delta^t[\tilde{\rho}_{p(t)}^t]] \in \mathbb{Q}^{q\times p(t)}, \tag{7}$$

$$\Theta = [\Theta^1\ldots\Theta^m] \in \mathbb{Q}^{q\times p}. \tag{8}$$

Theorem 3. *The set of the estimation matrices for the operators in $L(B^*)$ coincides with the set of estimation matrices in $L(D^*)$; that is, it coincides with the set of linear combinations of the matrices $\tilde{\theta} \otimes (\tilde{\alpha}(S^t)^d \circ \tilde{N}_d)$ in which $\tilde{\theta}$ is a column in the matrix Θ^t defined by Eq.* (7); *here, $t \in \{1, 2, \ldots, m\}$ and $d \in \{0, 1\}$.*

Below, we only consider the operators with fixed weight parameters $\omega(\Omega)$, ω^t, and normalizations $\tilde{N}_d(j)$. To simplify certain implications, we assume that all these parameters are equal to unity. However, all the results proved below remain valid for arbitrary fixed weights and normalizations.

Definition 1. The model of recognizing operators R^* is said to be *correct* for the recognition problem $Z = (I_0, \tilde{S}^q)$ if

$$\forall\Gamma \in \mathbb{Q}^{q\times l}\exists B \in R^*: \Gamma[B] = \Gamma.$$

Definition 2. The recognition problem $\tilde{Z}(I_0, \tilde{S}_q)$ is said to be *regular* if the following conditions are satisfied:

$$\tilde{S}^m \cap \tilde{S}_q = \varnothing, \tag{9}$$

$$|\{K_1, \ldots, K_l\}| = l, \tag{10}$$

$$|\{(\tilde{\rho}(S^1, S_i), \ldots, \tilde{\rho}(S^m, S_i))\}_{i=1}^q| = q. \tag{11}$$

Below, we only consider regular recognition problems when the linear and algebraic closures of the EA model of an arbitrary degree are examined. Note that, due to (1), condition (10) is automatically fulfilled for the problems with nonempty nonoverlapping classes.

2. LINEAR CLOSURE OF THE MODEL

Consider the linear closure of the EA model. Because this is the simplest closure, it is interesting to find out if it is sufficient for constructing a correct algorithm for the given set of training and test objects. This question was considered in many studies including [1–5]. However, the correctness conditions obtained in those studies only reduce the problem to checking other, maybe more descriptive conditions. Unfortunately, such conditions cannot be checked efficiently in most practical problems.

The main result obtained in this paper is the correctness criterion for the linear closure of the EA model that can be checked in polynomial time.

Let us define some concepts that are convenient for the further presentation.

Definition 3. The set of columns $\Delta[\tilde{\varepsilon}](S^t, S_i)$ of the matrix Θ defined by Eq. (8) for which $S^t \in K_j$ is called the representation *with respect to class* K_j. The representation with respect to class K_j is denoted by Θ^{K_j}.

Where there is no source of confusion, we will interpret the representation as the matrix consisting of the corresponding columns. The linear hull of a representation is defined as the linear closure of the corresponding set of vectors.

Definition 4. A representation is said to be *complete* if the vectors included in the representation form a basis.

Definition 5. The union of the vectors included in the representations with respect to the classes $K_1, \ldots, K_t$ is called the representations with respect to the set of the classes $K_1, \ldots, K_t$. It is denoted by $\Theta^{K_1} \cup \ldots \cup \Theta^{K_t}$.

Theorem 4. *The linear closure of the EA model is correct if and only if the condition*

$$\forall i \neq j : rank(\Theta^{K_i} \cup \Theta^{K_j}) = q, \quad i, j = \overline{1, l},$$

is fulfilled in the regular classification problem with nonoverlapping classes.

Proof. *Necessity.* Consider the representations with respect to the arbitrary classes K_{i_0}, K_{j_0}. For definiteness, assume that $j_0 > i_0$. Assume that the linear closure of the EA model is correct and the condition of the theorem is not satisfied; that is, assume that

$$rank\left(\Theta^{K_{i_0}} \cup \Theta^{K_{j_0}}\right) < q \Rightarrow \exists \hat{\mathbf{l}} = (\hat{l}_1, \ldots, \hat{l}_q) \in \mathbb{Q}^q \forall \theta \in \Theta^{K_i} \cup \Theta^{K_j} : (\hat{\mathbf{l}}, \theta) = 0, \quad \hat{\mathbf{l}} \neq 0. \tag{12}$$

Consider the following function of the vector $\hat{l}$ and of the columns i_0 and j_0 of the matrix Γ:

$$f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma) = \sum_{k=1}^{q} (\Gamma_{ki_0} - \Gamma_{kj_0}, \hat{\mathbf{l}}_k).$$

Under the assumptions of the theorem, any estimation matrix can be represented as a linear combination of the matrices $\theta_{jt} \otimes e_j^d$:

$$\forall \Gamma \exists \{c_{jtd}\} : \Gamma = \sum_{j=0}^{l} \sum_{\theta_{jt} \in \Theta^{K_j}} \sum_{d \in \{0,1\}} c_{jtd} (\theta_{jt} \otimes e_j^d)$$

$$= \sum_{j \in \{i_0, j_0\}} \sum_{\theta_{jt} \in \Theta^{K_j}} \sum_{d \in \{0,1\}} c_{jtd} (\theta_{jt} \otimes e_j^d) + \sum_{j \in \{1, 2, \ldots, l\} \setminus \{i_0, j_0\}} \sum_{\theta_{jt} \in \Theta^{K_j}} \sum_{d \in \{0,1\}} c_{jtd} (\theta_{jt} \otimes e_j^d),$$

$$\Gamma_1 = \sum_{j \in \{i_0, j_0\}} \sum_{\theta_{jt} \in \Theta^{K_j}} \sum_{d \in \{0,1\}} c_{jtd} (\theta_{jt} \otimes e_j^d), \quad \Gamma_2 = \sum_{j \in \{1, 2, \ldots, l\} \setminus \{i_0, j_0\}} \sum_{\theta_{jt} \in \Theta^{K_j}} \sum_{d \in \{0,1\}} c_{jtd} (\theta_{jt} \otimes e_j^d).$$

Note that the columns i_0 and j_0 in the matrix Γ_2 are identical. Therefore, $f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma_2) = 0$. Now, consider the terms of the sum constituting the matrix Γ_1. Every column of the matrix $\theta \otimes \alpha$, where α is an arbitrary binary vector, either coincides with θ or is zero; therefore, condition (12) implies that $\forall \theta' \in \mathrm{col}(\theta' \otimes \alpha) : (\hat{\mathbf{l}}, \theta') = 0$, where $\theta \in \Theta^{K_i} \cup \Theta^{K_j}$ and $\alpha \in \{e_{i_0}, e_{j_0}, \bar{e}_{i_0}, \bar{e}_{j_0}\}$. (For an arbitrary matrix M, $\mathrm{col}(M)$ denotes the set of its columns.) Therefore, the function $f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma_1) = 0$ due to the linearity in the last argument. Hence, we have the following equality for the matrix Γ:

$$\forall \Gamma : f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma) = f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma_1) + f(i_0, j_0, (\hat{l}_1, \ldots, \hat{l}_q), \Gamma_2) = 0.$$

For Γ, we take an arbitrary matrix that contains the only unity in the column i_0 in the first row i for which $\hat{l}_i \neq 0$ (the vector $\hat{\mathbf{l}}$ is nonzero by condition). Then, we arrive at the contradiction.

Sufficiency. **Case 1.** Assume that all the representations Θ^{K_i}, $i \in \{1, 2, ..., l\}$ have a non-full rank. Denote by $\{l^{K_i}\}$ the basis of the orthogonal complement of the subspace spanned by the vectors included in Θ^{K_i}:

$$\forall \mathbf{l} \in \{l^{K_i}\} \quad \forall \theta \in \Theta^{K_i} : (\mathbf{l}, \theta) = 0, \quad \left|\{l^{K_i}\}\right| = \operatorname{rank}(\{l^{K_i}\}) = q - \operatorname{rank}(\Theta^{K_i}). \tag{13}$$

The representation Θ^{K_i} is of no full rank; therefore, $\operatorname{rank}(\{l^{K_i}\}) \geq 1$ and, in addition, for the linear hulls spanned by the corresponding sets of vectors, we have $L(\Theta^{K_i}) \oplus L(\{l^{K_i}\}) = \mathbb{Q}^q$ for $(i \in \{1, 2, ..., l\}$. We show that any vector that is orthogonal to all the vectors in the representation Θ^{K_i} is a linear combination of the vectors of the representation with respect to the class Θ^{K_j} $\forall j \neq i$. By the assumptions of the theorem, we have

$$\forall i = 1, 2, \ldots, l : \operatorname{rank}(\Theta^{K_1} \cup \Theta^{K_i}) = \ldots = \operatorname{rank}(\Theta^{K_l} \cup \Theta^{K_i}) = q;$$

that is,

$$L(\Theta^{K_1}) + L(\Theta^{K_i}) = \ldots = L(\Theta^{K_l}) + L(\Theta^{K_i}) = L(\Theta^{K_i}) \oplus L(\{l^{K_i}\}) = \mathbb{Q}^q,$$

From this, we obtain

$$\forall j \neq i : L(\{l^{K_i}\}) \subseteq L(\Theta^{K_j}) \Rightarrow L(\{l^{K_i}\}) \subseteq \bigcap_{j \neq i} L(\Theta^{K_j}) \Rightarrow \forall \mathbf{v} \in L(\{l^{K_i}\}) : \mathbf{v} \in \bigcap_{j \neq i} L(\Theta^{K_j}). \tag{14}$$

Now, we describe a method for constructing an algorithm producing an arbitrary matrix of estimates $\Gamma = \|\gamma_{ij}\| \in \mathbb{Q}^{q \times l}$. The proof consists of four steps:

Step 1. Consider the basis $\{l^{K_1}\}$ defined by Eq. (13). For the matrix Γ, calculate the products of the vector $\tilde{\gamma}_1 = \{\tilde{\Gamma}_{11}, ..., \tilde{\Gamma}_{q1}$ by each of the vectors of the basis $\{l^{K_1}\}$:

$$(\tilde{\gamma}_1, l_1^{K_1}) = r_1^{K_1},$$
$$(\tilde{\gamma}_1, l_2^{K_1}) = r_2^{K_1},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\left(\tilde{\gamma}_1, l^{K_1}_{\operatorname{rank}\left(\left\{l^{K_1}\right\}\right)}\right) = r^{K_1}_{\operatorname{rank}\left(\left\{l^{K_1}\right\}\right)}.$$

Assume that at least one of these products is distinct from zero. Otherwise, if all these products are zero, the vector $\tilde{\gamma}_1$ belongs to the space $L(\Theta^{K_1})$, and steps (2) and (3) of the proof may be omitted.

Step 2. It follows from (14) that

$$\exists \tilde{\mathbf{v}} \in \bigcap_{j \neq 1} L(\Theta^{K_j}) : \forall i \in \{1, \ldots, \operatorname{rank}(\{l^{K_1}\})\} : (\tilde{\mathbf{v}}, l_i^{K_1}) = r_i^{K_1};$$

moreover,

$$\forall j \neq 1 \exists \{\mathbf{c}_{jt}\} : \tilde{\mathbf{v}} = \sum_t c_{jt}\theta_{jt}, \quad \text{where} \quad \theta_{jt} \in \Theta^{K_j};$$

Step 3. Using the same notation as in the preceding step, we consider the matrix

$$\Gamma^1 = \sum_t c_{kt}(\theta_{kt} \otimes \mathbf{e}_k) + \sum_t c_{kt}(\theta_{kt} \otimes \bar{\mathbf{e}}_k) - \sum_{j \neq 1} \sum_t c_{jt}(\theta_{jt} \otimes \mathbf{e}_j),$$

where k is an arbitrary class index distinct from unity. Note that all the columns in Γ^1, except for the first one, consist of zeros, and the first column equals the vector $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_q)$ determined at the preceding step.

Step 4. Repeating the reasoning used at steps 1–3 for the other classes, we obtain the set of matrices $\Gamma^1, \ldots, \Gamma^l$ such that only the ith column in Γ^i, $i \in \{1, 2, \ldots, l\}$ is distinct from zero, and it holds that

$$\forall i \in \{1, \ldots, l\} \forall j \in \{1, \ldots, \operatorname{rank}(\{l^{K_i}\})\} : (\tilde{\gamma}_i^i, \mathbf{l}_j^{K_i}) = (\tilde{\gamma}_i, \mathbf{l}_j^{K_i}). \tag{15}$$

Consider the matrix $\Gamma^* = \Gamma - \sum_{i=1}^{l} \Gamma^i$, and denote by $\tilde{\gamma}_i^*$ the ith column of Γ^*. Due to (15), we have

$$\forall i \in \{1, \ldots, l\} \forall j \in \{1, \ldots, \operatorname{rank}(\{l^{K_i}\})\} : (\tilde{\gamma}_i^*, \mathbf{l}_j^{K_i}) = 0.$$

Therefore, $\forall i \in \{1, 2, \ldots, l\}$, we have $\tilde{\gamma}_i^* \in L(\Theta^{K_i})$. Hence,

$$\forall i \in \{1, 2, \ldots, l\} \exists \{b_{it}\} : \tilde{\gamma}_i^* = \sum_{t=0}^{t_i} b_{it}\theta_{it}, \quad \text{where} \quad \theta_{it} \in \Theta^{K_i}. \tag{16}$$

Consider the matrices Γ^{*i} defined by Eq. (16):

$$\Gamma^{*i} = \tilde{\gamma}_i^* \otimes \mathbf{e}_i = \sum_{t=0}^{t_i} b_{it}(\theta_{it} \otimes \mathbf{e}_i), \quad i \in \{1, 2, \ldots, l\}.$$

Note that each of the matrices $\{\Gamma^{*i}\}_{i=0}^{l}$ and $\{\Gamma^i\}_{i=0}^{l}$ can be constructed in the linear closure of the EA model; therefore, the matrices

$$\Gamma^* = \sum_{i=0}^{l} \Gamma^{*i}, \quad \Gamma = \Gamma^* + \sum_{i=0}^{l} \Gamma^i = \sum_{i=0}^{l} \Gamma^i + \sum_{i=0}^{l} \Gamma^{*i},$$

being the sums of matrices belonging to the linear closure, can be constructed in this closure. Because Γ was chosen arbitrarily, case 1 is completely proved.

Case 2. Some of the representations are of full rank. Let the representations with respect to the classes $K_{i_0}, \ldots, K_{i_t}$ have the full rank. It is clear that, in the linear closure of the EA model, an arbitrary classification that does not change the entries of the estimation matrix in the other columns can be realized in the columns $i_0, \ldots, i_t$. Therefore, case 2 is naturally reduced to case 1 for the problem with a smaller number of classes. It remains to note that the validity of the theorem is obvious for the case of the problem with two nonoverlapping classes and for the case when all the representations have the full rank. This completes the proof of Theorem 4.

Remark 1. Note that a necessary condition for correctness is the full rank of the matrix Θ defined by (8). In the case of two classes, this condition is also sufficient.

Remark 2. The proposed criterion can be efficiently checked in a time not exceeding $O(mq^2l^2)$ in the worst case. Here, m is the length of the training sample, q is the amount of test data, and l is the number of classes. Indeed, the time needed to construct Θ defined by (8) is of order $O(qm)$, and the check of the criterion for the given pair of classes (i, j) does not exceed $O(mq^2)$. Since the number of pairs (i, j) is $O(l^2)$, the total time does not exceed $O(qm) + O(l^2)O(mq^2)$.

We illustrate the use of the proposed criterion by the following simple example.

Example 1. Consider the following problem defined by the set of training objects $\tilde{S}_m$ and the set of test objects $\tilde{S}_q$

$\tilde{S}_m$:

	ω_1	ω_2	ω_3
S^1	0	1	1
S^2	2	2	1
S^3	0	0	0

$\tilde{S}^q$:

	ω_1	ω_2	ω_3
S_1	1	2	3
S_2	3	0	3
S_3	3	2	3
S_4	0	1	2
S_5	0	1	0

Furthermore, let the object S_1 belong to the class K_1, S_2 to the class K_2, and S_3 to the class K_3. In addition, we assume that the metric in each feature space (the set of values of each feature is the set of integers) be defined by $\rho(x, y) = |x - y|$. Let us calculate the sets of pairwise distances $P(t)$, which define the matrix Θ by formulas (7) and (8): $P(1) = \{(1, 1, 2), (3, 1, 2), (0, 0, 1)\}$, $P(2) = \{(1, 0, 2), (1, 2, 2), (2, 1, 1)\}$ and $P(3) = (1, 2, 3), (3, 0, 3), (3, 2, 3), (0, 1, 2), (0, 1, 0)\}$.

Construct the matrix Θ, and find the representations with respect to each of the classes:

$$\Theta = \left[\begin{array}{ccc|ccc|ccccc} 1&0&0&1&0&0&1&0&0&0&0\\ 0&1&0&0&1&0&0&1&0&0&0\\ 0&1&0&1&0&0&0&0&1&0&0\\ 0&0&1&0&0&1&0&0&0&1&0\\ 0&0&1&0&0&1&0&0&0&0&1 \end{array}\right],$$

$$\Theta^{K_1} = \{(1, 0, 0, 0, 0)^{\mathrm{T}}, (0, 1, 1, 0, 0)^{\mathrm{T}}, (0, 0, 0, 1, 1)^{\mathrm{T}}\},$$

$$\Theta^{K_2} = \{(1, 0, 1, 0, 0)^{\mathrm{T}}, (0, 1, 0, 0, 0)^{\mathrm{T}}, (0, 0, 0, 1, 1)^{\mathrm{T}}\},$$

$$\Theta^{K_3} = \{(1, 0, 0, 0, 0)^{\mathrm{T}}, (0, 1, 0, 0, 0)^{\mathrm{T}}, (0, 0, 1, 0, 0)^{\mathrm{T}}, (0, 0, 0, 1, 0)^{\mathrm{T}}, (0, 0, 0, 0, 1)^{\mathrm{T}}\}.$$

Consider the representation with respect to the classes K_1 and K_2. Note that the vector $l = (0, 0, 0, 1, -1)$ is orthogonal to all the vectors both in the representation with respect to the class K_1 and with respect to K_2. Therefore, $(\mathrm{rank}(\Theta^{K_1} \cup \Theta^{K_2}) < q$; by the correctness criterion proved above, the linear closure of the EA model is incorrect. Note that the second-degree closure for this problem is correct as will be proved in the next section.

Assume that a new training object $S^4 = (2, 1, 0)$ such that $S^4 \in K_2$ is added. The set of the distances $P(4)$ defining the matrix Θ^4 by formula (7) is $P(4) = \{(1, 1, 3), (2, 0, 2), (2, 0, 0)\}$. The matrix Θ and the representation with respect to K_2 (the representations with respect to the other classes do not change) take the form

$$\Theta = \left[\begin{array}{ccc|ccc|ccccc|ccc} 1&0&0&1&0&0&1&0&0&0&0&1&0&0\\ 0&1&0&0&1&0&0&1&0&0&0&1&0&0\\ 0&1&0&1&0&0&0&0&1&0&0&1&0&0\\ 0&0&1&0&0&1&0&0&0&1&0&0&1&0\\ 0&0&1&0&0&1&0&0&0&0&1&0&0&1 \end{array}\right],$$

$$\Theta^{K_2} = \{(1, 0, 1, 0, 0)^{\mathrm{T}}, (0, 1, 0, 0, 0)^{\mathrm{T}}, (0, 0, 0, 1, 1)^{\mathrm{T}}, (1, 1, 1, 0, 0)^{\mathrm{T}}, (0, 0, 0, 1, 0)^{\mathrm{T}}, (0, 0, 0, 0, 1)^{\mathrm{T}}\}.$$

The equalities $\mathrm{rank}(\Theta^{K_1} \cup \Theta^{K_3}) = \mathrm{rank}(\Theta^{K_2} \cup \Theta^{K_3}) = 5$ are obvious. It remains to verify the equality $\mathrm{rank}(\Theta^{K_1} \cup \Theta^{K_2}) = 5$. The vectors $\{\mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_5\}$ belong to Θ^{K_2}, and the vector $\{\mathbf{e}_1\}$ belongs to Θ^{K_1}. The linear hull of each of the representations contains the vector $\tilde{1} = (1, 1, 1, 1, 1)$; therefore, $\mathrm{rank}(\Theta^{K_1} \cup \Theta^{K_2}) = q = 5$, and the linear closure of the EA model is correct by Theorem 4.

3. ALGEBRAIC CLOSURE OF THE EA MODEL

The operation of the elementwise multiplication defined in Section 1 and applied to the matrices of the operator estimates in the EA model naturally induces operations on recognizing operators. If $B^* = (B_1, \ldots, B_r)$ is the set of recognizing operators in the EA model, then the set of operators $U^k(B^*)$ of the algebraic closure of the EA model of degree k is defined by

$$U^k(B^*) = L(\{B_1 \cdot \ldots \cdot B_t \,|\, B_1, \ldots, B_t \in B^*, 1 \le t \le k\}). \tag{17}$$

By the set of recognizing operators of the algebraic closure $U(B^*)$ of the EA model, we mean the union of the operators belonging to the algebraic closure of the EA model of an arbitrary degree n, where $n \in \mathbb{N}$:

$$U(B^*) = \bigcup_{n \in \mathbb{N}} U^n(B^*).$$

As we established in the preceding section, the linear closure of the EA model is incorrect for certain training and test samples, which opens the question of the possibility to construct a classification in the algebraic closures of the EA model. For the problems with nonoverlapping classes, we propose and prove a criterion for the correctness of the algebraic closure of the EA model that completely solves this problem.

Note the following consequence of the correctness criterion of the EA model obtained in [2].

Theorem 5. *The algebraic closure of the EA model in the regular problem* $Z = (I_0, \tilde{S}_q)$ *with nonoverlapping classes is correct.*

As was mentioned in Section 1, only regular problems are considered when the EA model and its closures are examined. Because the linear closures of the labeling operators $D^* = \{D_{t,d}\}$ and the operators B^* of the EA model are identical, formula (17) can be rewritten in the form

$$U^k(B^*) = U^k(D^*) = L(\{D_1 \cdot \ldots \cdot D_t \mid D_1, \ldots, D_t \in D^*, 1 \le t \le k\}).$$

In [1], the matrix $A = \begin{bmatrix} A^1 \\ \ldots \\ A^m \end{bmatrix}$ was constructed. Here, $A^t = [1...1]_{1 \times l}$, for $\alpha(S^t) \in \{(0, \ldots, 0)_{1 \times l}, (1, \ldots, 1)_{1 \times l}\}$; otherwise, $A^t = \begin{bmatrix} \tilde{\alpha}(S^t) \\ \overline{\tilde{\alpha}(S^t)} \end{bmatrix}_{2 \times l}$, where $\tilde{\alpha}(S^t) = (\alpha_1, \ldots, \alpha_l)$, and $\overline{\tilde{\alpha}(S^t)} = (1 - \alpha_1, \ldots, 1 - \alpha_l)$, $t \in \{1, 2, \ldots, m\}$).

Set

$$V = \bigcup_{t=1}^{m} \bigcup_{\theta \in \Theta^t} \bigcup_{\alpha \in A^t} (\theta \otimes \alpha), \tag{18}$$

where the matrix Θ^t is defined by (7). Note that, for the types of problems under examination, equality (18) can be rewritten in terms of the representations as

$$V = \bigcup_{t=1}^{l} \bigcup_{\theta \in \Theta^{K_t}} \bigcup_{\sigma \in \{0,1\}} (\theta \otimes \mathbf{e}_t^{\sigma}). \tag{19}$$

In [1], the following theorem was proved.

Theorem 6. *The set of matrices of operator estimates belonging to* $U^k(B^*)$ *is the set of linear combinations of the matrices* $(\tilde{\theta}_1 \circ \ldots \circ \tilde{\theta}_k) \otimes (\tilde{\alpha}_1 \circ \ldots \circ \tilde{\alpha}_k)$, *where* $(\tilde{\theta}_1^{\mathrm{T}}, \tilde{\alpha}_1), \ldots, (\tilde{\theta}_k^{\mathrm{T}}, \tilde{\alpha}_k) \in V$.

Note that $(\tilde{\theta}_1 \circ \ldots \circ \tilde{\theta}_k) \otimes (\tilde{\alpha}_1 \circ \ldots \circ \tilde{\alpha}_k) = \Gamma[D_1 \cdot \ldots \cdot D_k]$, where $\Gamma[D_i] = \theta_i \otimes \alpha_i \; \forall i \in \{1, 2, \ldots, k\}$.

Now, we give some definitions, and then formulate the correctness criterion for the algebraic closure of degree k in terms of these definitions.

Definition 6. 1-closure $[\Theta]$ of the representation Θ is defined as the union of the set of vectors in Θ and the vector of an appropriate length consisting of unities. The operation $[\cdot]$ of the 1-closure is the completion of a representation to its 1-closure.

Definition 7. A representation is said to be 1-closed (1-representation) if its 1-closure coincides with the representation itself.

Definition 8. The product of the representations (1-representations) $\Theta_1 \circ \Theta_2$ is defined if the lengths of the vectors constituting the representations are identical; in this case, the product is the set of vectors U defined as

$$U = \{\theta : \theta = \theta_1 \circ \theta_2, \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}. \tag{20}$$

Definition 9. The degree k ($k > 0$) of the representation (1-representation) Θ is the set of vectors U defined as

$$U = \{\theta : \theta = \theta_1 \circ \ldots \circ \theta_k, \theta_1 \in \Theta, \ldots, \theta_k \in \Theta\}. \tag{21}$$

It is denoted by Θ^k.

The following theorems immediately follows from Definitions 6–8.

Theorem 7. *If the representations* Θ_1 *and* Θ_2 *are* 1*-representations*, *then the set U defined by* (20) *is a* 1*-representation*.

Theorem 8. *If* Θ *is a 1-representation*, *then the set* Θ^k *defined by* (21) *also is a 1-representation*.

The union of the representation $\Theta_1 \cup \Theta_2$ is the representation with the set of vectors formed by the union of the sets of vectors Θ_1 and Θ_2 without taking into account the multiplicities of their occurrences.

Theorem 9. *For an arbitrary regular problem with nonoverlapping classes, the set of estimation matrices of the operators from* $U^k(B^*)$ *is the set of linear combinations of the matrices of the form* $(\tilde{\theta}_1 \circ \ldots \circ \tilde{\theta}_k) \otimes \tilde{\alpha}_t^{\sigma_t}$, *where* $(\tilde{\theta}_1^{\mathrm{T}}, \tilde{\alpha}_1), \ldots, (\tilde{\theta}_k^{\mathrm{T}}, \tilde{\alpha}_k) \in V$, $\sigma_t \in \{0, 1\}$, and $t \in \{1, 2, \ldots, k\}$.

Proof. It follows from the equality $\bar{\mathbf{e}}_{i_1} \circ \ldots \circ \bar{\mathbf{e}}_{i_k} = \bar{\mathbf{e}}_{i_1} \circ \ldots \circ \bar{\mathbf{e}}_{i_{k-1}} - \bar{\mathbf{e}}_{i_k} \circ \bar{\mathbf{e}}_{i_1} \circ \ldots \circ \bar{\mathbf{e}}_{i_{k-1}}$ that any product of the form $\mathbf{e}_{i_1}^{\sigma_{i_1}} \circ \ldots \circ \mathbf{e}_{i_k}^{\sigma_{i_k}}$ can be represented as a linear combination of the vectors $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}, \bar{\mathbf{e}}_{i_1}, \ldots, \bar{\mathbf{e}}_{i_k}$. The application of Theorem 6 with regard for Eq. (19) completes the proof.

Theorem 10. *For an arbitrary regular classification problem* $Z = (I_0, \tilde{S}_q)$ *with nonoverlapping classes*, *the algebraic closure of the EA model of degree k is correct if and only if*

$$\forall i \neq j \quad \operatorname{rank}([\Theta^{K_i} \cup \Theta^{K_j}] \circ [\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}) = q.$$

Proof. Consider the set of vectors of the matrices defined in Theorem 9 of the form $\theta \times \mathbf{e}_i^{\sigma_i}$, where $\sigma_i \in \{0, 1\}$. Due to Theorem 9 and Eq. (19), the set of vectors $\{\theta\}$ forming these matrices in the algebraic closure of degree k is $\Theta^{K_i} \circ [\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}$.

Therefore, the problem is reduced to the question of correctness of the linear closure in the recognition problem $Z' = (I_0', \tilde{S}^q)$, which should be considered as a problem with nonoverlapping classes (due to Theorem 9). Note that, in the proof of the correctness of the linear closure on the basis of regularity conditions (9)–(11), only condition (10) was used. Moreover, any number of vectors in the representations with respect to different classes were allowed to coincide. Condition (10) is satisfied due to the regularity of the problem Z, and the representations of this problem are used to form the representations Θ'^{K_i} of the problem Z' by the formula $\Theta'^{K_i} = \Theta^{K_i} \circ [\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}$ $\forall i \in \{1, 2, \ldots, l\}$. Application of Theorem 4 to the problem Z' completes the proof.

Corollary 1. It follows from Theorem 10 that a necessary condition for the correctness of the algebraic closure of degree k is the equality $\operatorname{rank}([\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^k) = \operatorname{rank}([\Theta]^k) = q$, while $\operatorname{rank}([\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}) = \operatorname{rank}([\Theta]^{k-1}) = q$ is a sufficient condition. In particular, the algebraic closure of degree two is correct if $\operatorname{rank}([\Theta]) = \operatorname{rank}(\Theta) = q$.

Proof. We prove that the condition $\operatorname{rank}([\Theta]^{k-1}) = q$ is sufficient for the correctness of the algebraic closure of degree k. Indeed, by the definition of the 1-representation, the 1-representation $[\Theta^{K_i} \cup \Theta^{K_j}]$ contains a vector consisting of unities for any i and j; therefore, $\operatorname{rank}([\Theta^{K_i} \cup \Theta^{K_j}] \circ [\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}) \geq \operatorname{rank}([\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1}) = q$, which implies the validity of the assertion.

Now, we prove the necessity of the condition $\operatorname{rank}([\Theta]^k) = q$ for the correctness of the algebraic closure of degree k. Note that $\forall i, j$, we have $\operatorname{rank}([\Theta]^k) = \operatorname{rank}([\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^k) \geq \operatorname{rank}([\Theta^{K_i} \cup \Theta^{K_j}] \circ [\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{k-1})$. Therefore, if the representation $\operatorname{rank}([\Theta]^k)$has no full rank, the algebraic closure of degree k is incorrect due to the correctness criterion.

Return to Example 1 considered in the preceding section; this example involves three training objects. It was shown that the linear closure of the EA model is incorrect in this example. However, it follows from Corollary 1 that the algebraic closure of degree two is correct.

4. ESTIMATE OF THE DEGREE OF THE CORRECT ALGORITHM IN THE EA MODEL

We left open the question of the degree of the polynomial that is sufficient for the construction of a correct algorithm for an arbitrary regular problem $Z = (I_0, \tilde{S}_q)$ with nonoverlapping classes. In the case of an arbitrary regular problem for l classes (without restrictions on the relationships between the classes), it was shown in [1] that the unimprovable degree of the polynomial is exactly $[\log_2 q] + [\log_2 l]$.

In this paper, we prove that, in the case of nonoverlapping classes, the degree of the correct polynomial does not exceed $I_0(l = 2)[\log_2 q] + I_0(l > 2)[\log_2 q]$ and, in addition, there exists a regular recognition problem in which the polynomial of a lower degree is insufficient for obtaining an arbitrary classification. Here, I_0 is the function defined by the equation

$$I_0(condition) = \begin{cases} 1, & \text{if the } condition \text{ is true,} \\ 0, & \text{if the } condition \text{ is false.} \end{cases}$$

Definition 10. The binary matrix $H_{q \times p}$ is said to be *mD-labelable* if it can be represented as a set of matrices $H^1_{q \times p_1}, \ldots, H^m_{q \times p_m}$: such that $H = [H^1 \ldots H^m]$, where each of the matrices H^i ($i \in \{1, 2, \ldots, m\}$) does not contain zero columns and each of its rows contains exactly one unity.

Note the following result obtained in [1], which plays an important role in the proof of Theorem 12.

Theorem 11. *For an arbitrary mD-labelable matrix H, there exists a regular recognition problem in which $\Theta = H$ (Θ is determined by Eq.* (8)).

In [1], the following lemma was proved.

Lemma 1. *Let all the rows in an mD-labelable matrix Θ be different. Then, any q-dimensional rational (real) vector can be represented by the linear combination*

$$\sum_{u \in X} c_u \tilde{\theta}_{u,1} \circ \ldots \circ \tilde{\theta}_{u, r(u)},$$

$\{\tilde{\theta}^{\mathrm{T}}_{u,z}\}^{r(u)}_{z=1,\, u \in X} \subseteq \mathrm{col}(\Theta)$, $1 \le r(u) \le [\log_2 q]$ for $u \in X$, where $|X| < \infty$. *The estimate for $r(u)$ is sharp.*

Consider the family of matrices $\{X_n\} : X_n = [x_1 \ldots x_n] \in \mathbb{Q}^{2^n \times n}$ $\forall n \in \mathbb{N}$ with the columns $x_1, \ldots, x_n$ in which every row is the binary representation of its index. The matrix

$$\Theta^*_n = [x_1 \bar{x}_1 \ldots x_n \bar{x}_n] \tag{22}$$

is *nD*-labelable; therefore, by Theorem 11, it is the estimation matrix for the labeling operators of a certain regular problem. It was proved in [1] that, in the problems in which the matrix Θ^*_n $\forall n \in \mathbb{N}$ is the matrix defined by Eq. (8), the estimate of Lemma 1 is sharp. The following result is a consequence of Lemma 1.

Lemma 2. *Let the binary matrix Θ with all the rows different be mD-labelable. Let $[\Theta]$ be the 1-closure of the representation consisting of the columns of Θ. Then, we have rank$([\Theta]^{[\log_2 q]}) = q$ and, in addition,* $\forall q \exists \Theta_q : \mathrm{rank}([\Theta_q]^{[\log_2 q] - 1}) < q$.

Theorem 12. *For an arbitrary regular problem with nonoverlapping classes, the algebraic closure $U^{I_0(l=2)[\log_2 q] + I_0(l>2)[\log_2 q]}$ of the EA model is correct, and this estimate is sharp.*

Proof. *Case 1.* Let the number of classes be two. The algebraic closure of degree $[\log_2 q]$ is correct by Theorem 10 if and only if

$$\mathrm{rank}([\Theta^{K_1} \cup \Theta^{K_2}] \circ [\Theta^{K_1} \cup \Theta^{K_2}]^{[\log_2 q] - 1}) = \mathrm{rank}([\Theta^{K_1} \cup \Theta^{K_2}]^{[\log_2 q]}) = q.$$

This equality holds true due to Lemma 2. The estimate is sharp due to the same lemma and Theorem 11.

Case 2. Let the number of classes be $l > 2$ and $\exists n \in \mathbb{N}$ such that $2^n + 1 \le q < 2^{n+1}$.

Sufficiency. In Case 2, we have $[\log_2 q] = [\log_2 q] + 1 = n + 1$. It follows from Lemma 2 that $\mathrm{rank}([\Theta]^{[\log_2 q]}) = n$ and, by Corollary 1, this equality is sufficient for the correctness of the closure of degree $n + 1$.

Necessity. Consider the case of three classes. Let $\Theta_n^* = (\theta_{ij}^n)_{i=1,j=1}^{i=2^n,j=2n}$ be the nD-labelable matrix defined by Eq. (22) for which the degree $r(u) = [\log_2 q] - 1 = n$ is insufficient for the construction of a basis in the space of columns (see Lemma 1). Consider the binary matrix $A = (a_{ij})_{i,j=1}^{i,j=2^n+1,2n+3}$, where

$$a_{ij} = \begin{cases} \theta_{ij}^n, & i \le 2^n, j \le 2n \\ 0, & i = 2^n + 1, j \le 2n, j \text{ is even}, \\ 1, & i = 2^n + 1, j \le 2n, j \text{ is odd}, \\ \delta_{i,2^n+1}, & j = 2n + 1 \\ 1 - \delta_{i,2^n+1}, & j = 2n + 2 \\ 1, & j = 2n + 3. \end{cases}$$

Observe that this matrix is $(n + 2)$-labelable and all its rows are distinct; therefore, due to Theorem 11, there exists a regular recognition problem in which A plays the role of the matrix Θ defined by Eq. (8). Let the objects $S^1, \ldots, S^n$ determining the first $2n$ columns of A belong to the class K_1; the object S^{n+1} determining the columns $2n + 1$ and $2n + 2$ belong to the class K_2; and the training object S^{n+2} determining the remaining submatrix belongs to the class K_3. Therefore, the representation Θ^{K_2} with respect to K_2 contains only two vectors $\mathbf{e}_{2^n+1}$ and $\bar{\mathbf{e}}_{2^n+1}$. The representation $\Theta^{K_3} = \{(1, \ldots, 1)\}$.

By Theorem 10, for the correctness of the algebraic closure of degree $[\log q]$, the following equality must hold:

$$\operatorname{rank}([\Theta^{K_2} \cup \Theta^{K_3}] \circ [\Theta^{K_1} \cup \Theta^{K_2} \cup \Theta^{K_3}]^{[\log q]-1}) = q.$$

It follows from Lemma 2 that the representation $[\Theta_n^*]^{[\log q]-1}$ is of non-full rank; therefore, there exists a nonzero vector $\tilde{\mathbf{l}} = (\tilde{l}_1, \ldots, \tilde{l}_q)$ that is orthogonal to all the vectors of this representation. This implies that the vector $\mathbf{l} = (\tilde{l}_1, \ldots, \tilde{l}_q, 0)$ is orthogonal to all the vectors of the representation $[\Theta^{K_1} \cup \Theta^{K_2} \cup \Theta^{K_3}]^{[\log q]-1}$.

Consider an arbitrary vector $\theta \in [\Theta^{K_1} \cup \Theta^{K_2} \cup \Theta^{K_3}]^{[\log q]-1}$ that has zero as its last component. It is clear that the product of the vectors $\theta \circ \theta_1$, where $\theta_1 \in \Theta^{K_2} \cup \Theta^{K_3}$, is either a zero vector or a vector that is identical to θ. Therefore, $(\theta \circ \theta_1, \mathbf{l}) = 0$. Moreover, $(\mathbf{e}_{2^n+1} \circ \theta_1, \mathbf{l}) = 0$ for $\theta_1 \in \Theta^{K_2} \cup \Theta^{K_3}$; this implies that $\mathbf{l}$ is orthogonal to any vector in $[\Theta^{K_2} \cup \Theta^{K_3}] \circ [\Theta^{K_1} \cup \Theta^{K_2} \cup \Theta^{K_3}]^{[\log q]-1}$, which completes the proof for Case 2. Note that this construct can be easily generalized for the case of an arbitrary number of nonoverlapping classes for which there exists a regular problem. Indeed, it is sufficient that the matrix A be a submatrix of the matrix $\Theta \in \mathbb{Q}^{q \times k}$, where $2^n < q < 2^{n+1}$, the classes K_2 and K_3 consist of a single object each, and the description of this object be induced by the description of the objects S^{n+1} and S^{n+2}, respectively.

Case 3. Let the number of clusters $l > 2$ and $\exists n \in \mathbb{N} \cup \{0\}$ such that $q = 2^n$.

In this case, we have $[\log_2 q] = [\log_2 2^n] = n$.

Necessity. The necessity of the condition under examination follows from Lemma 2 and the necessary condition for the correctness of the algebraic closure given in Corollary 1.

Sufficiency. Note that, if the number of training objects is 2^n ($n \in \mathbb{N}$) and the algebraic closure of degree $n - 1$ is incorrect, the rank of the subspace that is orthogonal to any vector in $[\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{[\log q]-1}$ is one. Indeed, removing one test object, we obtain that, for the sample of length $2^n - 1$, a basis in the closure of degree $[\log_2 2^n - 1] = n - 1$ can be obtained (by Lemma 2). Therefore, $\operatorname{rank}[\Theta^{K_1} \cup \ldots \cup \Theta^{K_l}]^{[\log q]-1} \ge q$. Note that, for any nonzero vector inducing an orthogonal subspace, none of its components is zero (indeed, if a component is zero, the corresponding test object can be arbitrarily classified).

Thus, if the algebraic closure of degree n is incorrect for the regular recognition problem under examination, the set of vectors that are orthogonal to every vector in the representation $[\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{n-1}$ forms a one-dimensional subspace and $\operatorname{rank}([\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{n-1}) = q - 1$.

If the closure of degree n is correct, then we have, for any $\forall i, j$: $\operatorname{rank}([\Theta^{K_i} \cup \Theta^{K_j}] \circ [\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{n-1}) = n$. Consider arbitrary classes K_{i_0} and K_{j_0}. The problem is regular; therefore, $\Theta^{K_{i_0}} \neq \Theta^{K_{j_0}}$. Consequently, $\Theta^{K_{i_0}} \cup \Theta^{K_{j_0}}$ contains a vector that is distinct from $\tilde{\mathbf{1}} = (1, ..., 1)$ and $0\ \tilde{\mathbf{0}} = (0, ..., 0)$. Denote it by $\tilde{\alpha}$. Without loss of generality, set $|\operatorname{Ind}(\tilde{\alpha})| \leq 2^{n-1}$ (otherwise, its negation can be used as $\tilde{\alpha}$). Rearrange the components of the vector $\tilde{\alpha}$ in such a way that the first $|\operatorname{Ind}(\tilde{\alpha})| \leq 2^{n-1}$ components contain unities and the other components are zero. Rearrange the rows in the matrix Θ in the same fashion. (The length of $\tilde{\alpha}$ is the same as the number of rows in the matrix Θ.) Below, Θ denotes the matrix with the rearranged rows. It is clear that the linear closure of the set of vectors of the representation $[\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{n-1}$ contains a vector $\tilde{\mathbf{e}} = (1, 0, ..., 0, x)$, where x is a number determined by the vector $\hat{\mathbf{l}}$. It remains to note that $(\tilde{\alpha} \circ \tilde{\mathbf{e}}, \tilde{\mathbf{l}}) = (\tilde{\mathbf{e}}_1, \tilde{\mathbf{l}}) \neq 0$; therefore, the representation $[\Theta^{K_{i_0}} \cup \Theta^{K_{j_0}}] \circ [\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{[\log q]-1}\ \forall i_0, j_0 \in \{1, ..., l\}$ has a greater rank than the representation $[\Theta^{K_1} \cup ... \cup \Theta^{K_l}]^{[\log q]-1}$; that is, it is complete, which completes the proof.

5. CONCLUSIONS

The correctness criteria for the linear closure and the algebraic closure of an arbitrary degree of the family of EA algorithms with arbitrary fixed weights of the feasible objects, weights of the features and normalizations (defined in [1]) are obtained. These criteria make it possible to efficiently find the minimal sufficient degree of the algorithmic polynomial that can correctly classify the given test sample. In addition, it is shown that the degree of the closure that makes it possible to construct a correct classifier for an arbitrary regular problem with nonoverlapping classes can be considerably reduced compared with the general case (which is described, for example, in [1]).

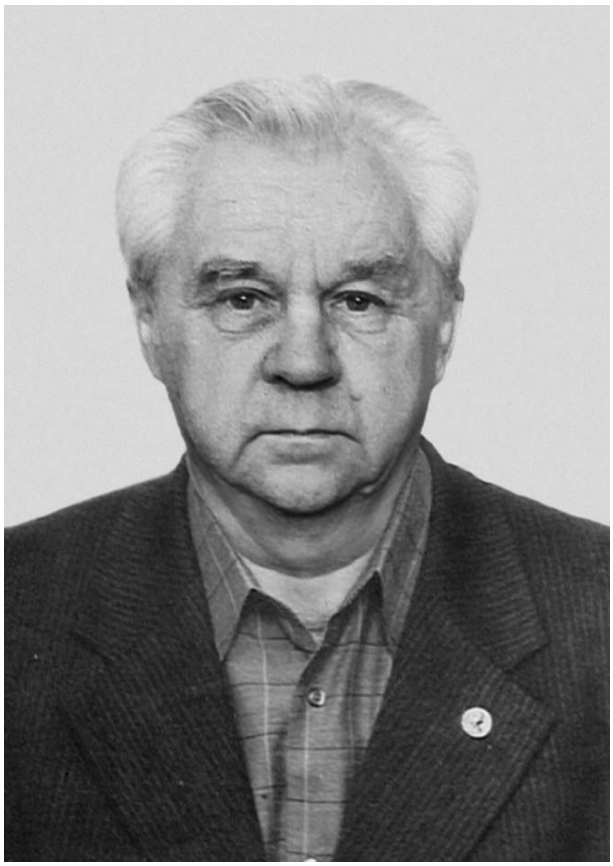
REFERENCES

1. A. G. D'yakonov, *Algebras over Estimation Algorithms* (Mosk. Gos. Univ., Moscow, 2006) [in Russian].
2. Yu. I. Zhuravlev, "An Algebraic Approach to Recognition and Classification Problems," in *Problems of Cybernetics* issue 33 (Nauka, Moscow, 1978; Hafner, 1986), pp. 5–68.
3. V. L. Matrosov, *Correct Algebras of Bounded Capacity over the Set of Estimation Algorithms*, Doctoral Dissertation in Mathematics and Physics (Computing Center, Russian Academy of Sciences, Moscow, 1986).
4. K. V. Rudakov, *On the Algebraic Theory of Universal and Local Constraints for Classification Problems*, Doctoral Dissertation in Mathematics and Physics (Computing Center, Russian Academy of Sciences, Moscow, 1992).
5. A. G. D'yakonov, "Algebra over Estimation Algorithms: Normalization and Division," Zh. Vychisl. Mat. Mat. Fiz. **47** 1099–1109 (2007) [Comput. Math. Math. Phys. **47**, 1050–1060 (2007)].

ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2009, Vol. 49, No. 7, pp. 1276–1280. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © E.L. Akim, A.L. Afendikov, K.V. Vrushlinskii, S.K. Godunov, G.V. Dolgoleva, A.B. Zhizhchenko, V.T. Zhukov, M.K. Kerimov, A.O. Latsis, A.E. Lutskii, M.V. Maslennikov, Yu.P. Popov, G.P. Prokopov, V.S. Ryaben'kii, and B.N. Chetverushkin, 2009, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2009, Vol. 49, No. 7, pp. 1340–1344.

Aleksei Valerievich Zabrodin (1933–2008)

DOI: 10.1134/S0965542509070185



December 14, 2008, marked the 75th birthday of Aleksei Valerievich Zabrodin, a corresponding member of the Russian Academy of Sciences, deputy director and a department head at the Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences, and the head of the Department of Computational Mechanics at the Faculty of Mechanics and Mathematics of Moscow State University. To our deep regret, he suddenly passed away on September 28 only a short time before his birthday anniversary, and this anniversary article has become an obituary. The Russian science has lost a leading expert in computational mathematics and its applications in mechanics and physics.

Zabrodin was born in Moscow on December 14, 1933, into a family of working intellectuals. His father was an electrician engineer in the railway transportation system, and his mother was a schoolteacher. In 1951, he finished high school no. 317 in Moscow and entered the Faculty of Mechanics and Mathematics of Moscow State University (MSU). Under the guidance of S.K. Godunov, he studied self-similar solutions of gasdynamic equations, which are of interest as asymptotics in many applications. It is pertinent to recall that the discussions of Zabrodin's final study at a seminar of Academician L.I. Sedov led to the coinage of a new term—transcendental self-similarity characteristic—which was related to the then poorly studied class of self-similar problems.

In 1954, while a twenty-year student, Zabrodin joined the staff of the Keldysh Institute of Applied Mathematics (IPM) (then it was the Department of Applied Mathematics at the Steklov Institute of Mathematics of the USSR Academy of Sciences), and this step had predetermined his future career. Begun as a student trainee study, his research concerning the numerical solution of continuum mechanics problems absorbed his interest and became the business of his entire life. After graduating from the uni-

versity in 1956, Zabrodin became a full-time employee at IPM, where he worked until his last days. The institution became his second home. It comprised a team of young enthusiastic researchers who solved important problems related to our country's defense and, hence, independence. The scientific leaders were outstanding mathematicians M.V. Keldysh, A.N. Tikhonov, and I.M. Gelfand. Problems were suggested by famous physicists, such as I.V. Kurchatov, Yu.B. Khariton, Ya.B. Zeldovich, E.I. Zababakhin, and A.D. Sakharov, who regularly attended IPM. In fact, the accomplishment of government tasks contributed to the creation of a new research discipline, namely, computational mathematics in its modern sense. The central place was occupied by the theory of finite difference schemes designed for solving nonlinear problems of mathematical physics. Simultaneously, electronic computers and programming systems were created and developed. In these circumstances, young employees quickly became experienced professionals. One of them was Zabrodin, who had deep interest in scientific problems and their details, huge diligence, and organizational grasp. He was interested in everything that happened at the institution: organizational issues; public life, in which he took an active part; and, primarily, people with whom he was acquainted, made friends, and spent his short leisure time.

Zabrodin's first activities at IPM were related to the development of a software code for two-dimensional gasdynamic computations. Its first version was launched in 1958. Specifically, this code served as a prototype of another one for computing the supersonic steady flow past a sphere. The solution to this problem was based on I.G. Petrovskii's idea of using time marching and was published by a team of authors (including Zabrodin) in 1961.

After completing his graduate study under the supervision of K.I. Babenko, in 1965 Zabrodin defended his candidate's dissertation. In 1978, he became a doctor of sciences. In 2000, he was elected a corresponding member of the Russian Academy of Sciences.

Zabrodin's science management abilities were early recognized. He guided a team of researchers who solved a series of complicated applied problems. In 1966, Zabrodin was appointed the scientific secretary of IPM and successfully worked in this challenging position for more than 10 years. Simultaneously, in 1969, he headed the hard-working Department of Computational Gas Dynamics, which was guided by Godunov before his move to Novosibirsk.

An important role in stating applied computational problems at IPM was played by its cooperation with the Sukhoi Research and Design Bureau performed by a group of departments under the guidance of Babenko. Scientific and friendly relationships between these departments have been preserved until now.

In his later life, Zabrodin's influence and the range of responsibilities extended substantially: he led a large range of activities performed by several IPM departments. In 1994, he was appointed deputy director of IPM. As scientific secretary of the Commission at the Presidium of the Russian Academy of Sciences and the director of Keldysh's memorial museum at IPM, he put much effort in perpetuating the memory of Keldysh.

Gradually, Zabrodin scientific interests embraced problems going beyond direct activities of IPM. The solution of key problems concerning scientific and technological progress in the various areas (nuclear power engineering, aerospace technologies, molecular biology, etc.) required fast development of high-performance computers. A promising direction was the creation of multiprocessor computer systems with parallel computations and data processing. In the USSR, works in this direction were initiated in the 1980s by Babenko, Zabrodin, and A.N. Myamlin. Later, the Complex Program of Developing Multiprocessor Supercomputers and Parallel Computer Technologies was organized. The program is under development as based on studies performed for justifying this initiative. The joint efforts of the "Kvant" Research Institute headed by V.K. Levin, the Keldysh Institute of Applied Mathematics, the Institute of Mathematics and Mechanics of the Russian Academy of Sciences, and other organizations have led to the creation and launching of MVS-100, MVS-1000, and MVS-1000M supercomputers. By combining the features of computational algorithms and structural engineering solutions, performance of 10^{12} operations per second has been achieved and a series of previously inaccessible fundamental and applied problems have been solved. Zabrodin was one of the leaders of a program at the Presidium of the Russian Academy of Sciences that joined supercomputer users from many institutions in different cities and ensured the financing of works on parallel computations.

Zabrodin participated in publishing activities in areas close to his interests. He was deputy editor-in-chief of the series *Mathematical Simulation of Physical Processes* in the journal *Problems of Atomic Science and Technology* and served on the editorial board of the journal *Mathematical Modeling*.

Zabrodin's pedagogical activities can be characterized, first, by his students who became candidates and doctors of sciences under his supervision. Second, he worked as a professor at the Department of Computational Mechanics created in 2000 at the MSU Faculty of Mechanics and Mathematics. In 2004, after the death of its founder, V.P. Myasnikov, he was appointed the head of this department. Finally, for

more than 20 years, Zabrodin supervised the All-Russia School–Conference "Theoretical Foundations and Design of Numerical Algorithms for Problems of Mathematical Physics" dedicated to the memory of Babenko. It was held every second year and provided the opportunity for skilled and authoritative experts to share their knowledge and experience with young students. Even in the most difficult years, the organizing committee managed to attract a sufficient number of young scientists. The XVII Conference at Abrau-Dyurso was the last business in Zabrodin's life. During the conference, he was stricken by an illness that led to his premature death on September 28, 2008.

Zabrodin's scientific interests, activities, and heritage were related to computational mathematics and its applications to modern scientific and engineering problems. First, this is concerned with the design of numerical algorithms for complicated problems in physics and mechanics, their analysis, and implementation in particular problems. More specifically, we mean two- and three-dimensional nonlinear problems in fluid and plasma dynamics. Due to the specific character of IPM works at the first stage, open publications were frequently not allowed. For this reason, only journal papers containing separate results on difference schemes could be found in the scientific literature of those years. In a more complete form, these results and their generalizations were later summarized in the monograph *Numerical Solution of Multidimensional Gas Dynamic Problems* (1976), which Zabrodin coauthored with Godunov, M.Ya. Ivanov, A.N. Kraiko, and G.P. Prokopov. Some results can also be found in the book *Theoretical Foundations and Design of Numerical Algorithms for Problems in Mathematical Physics* (1979) edited by Babenko.

It should be noted that a complete theoretical study of nonlinear problems, including convergence analysis, is possible only in extremely rare cases. Accordingly, a major role is played by the intuition of researchers and by deep knowledge of the physics of the phenomena under study.

Due to Zabrodin's determination, active attitude, and huge amount of effort, the efficiency of the principles and ideas that underlie algorithms and are implemented in software packages was confirmed in a large number of various problems. Moreover, many of these problems could not be solved by any other method despite much effort by various scientific teams.

Much of Zabrodin and his students' work was devoted to computational aerodynamics, i.e., to the numerical solution of flow problems, which are now a necessary element in the design of new planes and other airborne vehicles. According to Zabrodin, one of the most important requirements for modern techniques was the adaptability of algorithms to the features of the analyzed flow. Depending on particular conditions, i.e., even in nonstationary problems, it is reasonable to choose a computational strategy based not only on shock-capturing explicit schemes but also, in some cases, on implicit ones. We mean situations when the stability condition imposes too restrictive constraints on the time step (for example, in the computation of entropy layers or heat-conducting gas flows). At the same time, explicit shock-capturing difference schemes are generally preferable as applied to computing flow subdomains containing numerous unfitted strong discontinuities.

Since the early 1990s, studies related to the development of a cylindrical layered thermonuclear target irradiated with heavy ions have been conducted at the Institute of Theoretical and Experimental Physics and IPM. Zabrodin headed the mathematical simulation of physical processes accompanying the combustion and compression of the target. A nontrivial continuum medium model with various temperatures of ions, electrons, and photons was implemented. Special attention was given to the fundamental gas dynamics problem of shock-free compression. It is well known that the compression of a substance by high pressure generally leads to shock waves, at each of which the density can increase only a finite number of times. An interesting idea is to produce more intense compression (in the limit, unbounded). Developing the works by Zababakhin, A.F. Sidorov, etc., Zabrodin and G.V. Dolgoleva showed that a special regime can be found for supplying energy to the system so that shock-free compression of the target with intense parameters is possible. This result was summarized in the small book by Dolgoleva and Zabrodin *Energy Cumulation in Layered Systems and Implementation of Shock-Free Compression* (2004).

After the above-mentioned high-performance system of MVS computers was created at IPM, much effort was required for organizing its operation and further improvements and for developing new solution approaches based on parallel computations. These works were also initiated by Zabrodin. His greatest achievement was that he realized that the development of parallel computer systems was an extremely complex interdisciplinary task requiring close daily interactions throughout the entire technological process from engineers to mathematicians to users. In this form, the task was set up and solved in the 1990s long before parallel supercomputers became widespread and popular and the techniques for their construction and applications became conventional due to the cluster technology of multiprocessor design. The fact that, by the beginning of the cluster boom in the 2000s, our country had a school of parallel computations with practically justified preliminary supercomputer technologies was doubtlessly due to Zabrodin. Until his last days, he kept track of the development of fundamentally new supercomputer architec-

tures with primary attention given to works in this direction. In a sense, we can say that supercomputer technologies are constantly at the stage of the 1950s; i.e., they change fundamentally and are constructed anew. In these conditions, Zabrodin was an experienced professional and all of us learned much from him.

Zabrodin's activities and contributions to science were highly appreciated by the scientific community and were marked by government awards. Specifically, he was a recipient of the State Prize of the USSR (1972), the State Prize of the Russian Federation (2004), the Order of the Honor Badge (1980), and several medals. He was also awarded the title of the Merited Scientist of the Russian Federation (1999).

We are deeply grieving over the death of Zabrodin, who is not with us on his anniversary birthday. Let his life in science and his achievements serve as a role model for future generations that begin to work in his favorite discipline.

E.L. Akim, A.L. Afendikov, K.V. Vrushlinskii, S.K. Godunov, G.V. Dolgoleva, A.B. Zhizhchenko, V.T. Zhukov, M.K. Kerimov, A.O. Latsis, A.E. Lutskii, M.V. Maslennikov, Yu.P. Popov, G.P. Prokopov, V.S. Ryaben'kii, and B.N. Chetverushkin

A.V. ZABRODIN'S SELECTED PUBLICATIONS

1. S. K. Godunov, A. V. Zabrodin, and G. P. Prokopov, "Difference Scheme for Two-Dimensional Problems in Gas Dynamics and Computation of Flow with a Detached Shock Wave," Zh. Vychisl. Mat. Mat. Fiz. **1**, 1020–1050 (1961).
2. S. K. Godunov and A. V. Zabrodin, "On Second-Order Accurate Difference Schemes for Multidimensional Problems," Zh. Vychisl. Mat. Mat. Fiz. **2**, 706–708 (1962).
3. S. K. Godunov, A. A. Deribas, A. V. Zabrodin, and N. S. Kozin, "Hydrodynamics Effects in Colliding Solids," J. Comput. Phys. **5**, 517–530 (1970).
4. S. K. Godunov, A. V. Zabrodin, M. Ya. Ivanov, et al., *Numerical Solution of Multidimensional Problems in Gas Dynamics* (Nauka, Moscow, 1976) [in Russian].
5. S. K. Godunov et al., *Resolution numerique des problems multidimensionnels de la dynamiques des gas* (Mir, Moscow, 1979).
6. A. V. Zabrodin, G. P. Prokopov, and V. A. Cherkashin, "Self-Adapted Algorithms of Gas Dynamics," Lect. Notes Phys. **90**, 587–593 (1979).
7. A. V. Zabrodin and S. B. Pekarchuk, "On the Effect of Equation's Nonlinearity and Problem's Dimension on the Stability of Difference Schemes," VANT, Ser. Metod. Program. Chisl. Reshen. Zadach Mat. Fiz., No. 2 (4), 23–30 (1979).
8. N. N. Anuchina, A. V. Zabrodin, et al., "Hydrodynamic Stability Analysis at the Interface," in *Computer Study of Hydrodynamic Stability* (Inst. Prikl. Mat., Akad. Nauk SSSR, Moscow, 1981) [in Russian].
9. A. V. Zabrodin and S. B. Pekarchuk, "A Numerical Method for Solving the Nonlinear Heat Equation on," VANT, Ser. Metod. Program. Chisl. Reshen. Zadach Mat. Fiz. **2** (10), 14–22 (1982).
10. A. N. Andrianov, K. I. Babenko, A. V. Zabrodin, et al., "On the Structure of a Solver for Flow Problems: Complex Approach to Design," in *Computational Processes and Systems* (Nauka, Moscow, 1985), No. 2, pp. 13–62 [in Russian].
11. A. V. Zabrodin, Ya. M. Kazhdan, T. Yu. Luchkaya, et al., "Adaptive Method for Numerical Integration of the Equations of Motion of a Heat-Conducting Medium," VANT, Ser. Metod. Program. Chisl. Reshen. Zadach Mat. Fiz., No. 1, 3–12 (1985).
12. G. B. Alalykin, A. V. Zabrodin, V. S. Imshennik, et al., "Numerical Modeling of Substance Compression and Heating by Heavy Ion Beams," VANT, Ser. Metod. Program. Chisl. Reshen. Zadach Mat. Fiz., No. 1, 26–42 (1985).
13. A. V. Zabrodin, "O Problems on Numerical Simulation of Gasdynamic Flows with a Complex Structure," in *Design of Algorithms and the Solution of Problems in Mathematical Physics* (Inst. Prikl. Mat., Akad. Nauk SSSR, Moscow, 1987) [in Russian].
14. A. V. Zabrodin, I. D. Sofronov, and N. N. Chentsov, "Adaptive Difference Methods for Mathematical Modeling of Nonstationary Gasdynamic Flows (Review)," VANT, Metod. Program. Chisl. Reshen. Zadach Mat. Fiz., No. 4, 3–22 (1988).
15. A. V. Zabrodin and N. N. Chentsov, "Approximate Solutions to Lagrange Ballistic Problem and Its Generalizations," in *Design of Algorithms and Solving Problems of Mathematical Physics* (Inst. Prikl. Mat., Akad. Nauk SSSR, Moscow, 1989), pp. 187–192 [in Russian].
16. A. V. Zabrodin and N. N. Chentsov, "Approximation of Solution to the Lagrange Ballistic Problem," VANT, Ser. Mat. Model. Fiz. Protsessov, No. 3, 15–17 (1990).
17. V. T. Zhukov, A. V. Zabrodin, and O. B. Feodoritova, "Method for Solving Two-Dimensional Equations of Heat-Conducting Gas Dynamics in Domains of Complex Geometry," Zh. Vychisl. Mat. Mat. Fiz. **33**, 1240–1250 (1993).

18. V. T. Zhukov, A. V. Zabrodin, V. S. Imshennik, and O. B. Feodoritova, "Numerical Study of Compression of a Nonspherical Target for Heavy Ion Thermonuclear Fusion in the Approximation of Heat-Conducting Gas Dynamics," VANT, Ser. Mat. Model. Fiz. Protsessov, No. 1, 8–18 (1994).
19. M. D. Brodetskii, A. V. Zabrodin, A. E. Luchkii, and E. K. Derunov, "Comparison of Numerical and Experimental Studies of the Supersonic Flow past a Combination of Two Bodies of Revolution," Teplofiz. Aeromekh. **2** (2), 97–102 (1995).
20. A. V. Zabrodin and G. P. Prokopov, "Method for the Numerical Modeling of Plane Unsteady Flows of Heat-Conducting Gas in Three-Temperature Approximation," VANT, Ser. Mat. Model. Fiz. Protsessov, No. 3, 3–10 (1998).
21. N. P. Adamov, M. D. Brodetskii, A. V. Zabrodin, et al., "Numerical and Physical Modeling of the Supersonic Flow past Separating Cruise Missiles," Teplofiz. Aeromekh. **7**, 1–12 (2000).
22. A. V. Zabrodin, A. E. Lutskii, K. Kh. Marbashev, and L. G. Chernov, "Numerical Study of the Flow past Airborne Vehicles and Their Elements in Actual Flight Conditions," Obshcheros. Nauchno-Tekh. Zh. Polet, No. 7, 21–29 (2000).
23. G. V. Dolgoleva and A. V. Zabrodin, *Cumulation of Energy in Layered Systems and Implementation of Shock-Free Compression* (Fizmatlit, Moscow, 2004) [in Russian].
24. M. V. Keldysh, *Creative Portrait from Memoirs of Contemporaries,* Ed. by A. V. Zabrodin (Nauka, Moscow, 2002) [in Russian].
25. E. M. Galimov, A. V. Zabrodin, A. M. Krivtsov, et al., "Dynamic Models of the Formation of the Earth–Moon System," in *Geochemistry* (Nauka, Moscow, 2005), pp. 1139–1150 [in Russian].
26. A. V. Zabrodin and V. V. Ogneva, Preprint No. 21, IPM AN SSSR (Keldysh Inst. of Applied Mathematics, USSR Academy of Sciences, Moscow, 1973).
27. K. I. Babenko, A. V. Zabrodin, I. B. Zadykhailo, and A. N. Myamlin, Preprint No. 7, OVM (VINITI, Moscow, 1981).
28. G. P. Voskresenskii and A. V. Zabrodin, Preprint No. 83, IPM AN SSSR (Keldysh Inst. of Applied Mathematics, USSR Academy of Sciences, Moscow, 1986).
29. R. N. Antonova, A. V. Zabrodin, G. F. Kopytov, et al., Preprint No. 44, IPM AN SSSR (Keldysh Inst. of Applied Mathematics, USSR Academy of Sciences, Moscow, 1987).
30. G. V. Dolgoleva and A. V. Zabrodin, Preprint No. 53, IPM RAN (Keldysh Inst. of Applied Mathematics, Russian Academy of Sciences, Moscow, 1999).
31. A. V. Zabrodin, Preprint No. 71, IPM RAN (Keldysh Inst. of Applied Mathematics, Russian Academy of Sciences, Moscow, 1999).
32. A. V. Zabrodin, V. K. Smirnov, and V. S. Shtarkman, *On the Memory of A.N. Myamlin Dedicated to His 75th Birthday* (Keldysh Inst. of Applied Mathematics, Russian Academy of Sciences, Moscow, 2000) [in Russian].