Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1075–1092.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1203–1221.

GRAVITATION, ASTROPHYSICS

Discrete Gravity, the Problem of Fermion State Doubling, and Quantum Anomalies

S. N. Vergeles

Landau Institute for Theoretical Physics, Russian Academy of Science, Chernogolovka, Moscow oblast, 142432 Russia

**e-mail: vergeles@itp.ac.ru*

Received June 19, 2003

Abstract—The problem of doubling of fermion states is studied in the framework of the theory of discrete gravitation. Examples of amorphous lattices (simplicial two-, three-, and four-dimensional complexes) free of doubling of fermion states are given. Possible consequences of this fact, such as the absence of quantum anomalies in divergence of axial currents, are considered. On the basis of the absence of axial anomalies and the finiteness of the number of physical degrees of freedom in the model of discrete quantum gravity proposed in [1] and of the continuum theory of gravitation constructed with the help of the dynamic quantization method [2], the following conclusion has been drawn: discrete quantum gravity [1] in the continuum limit is transformed into the theory of gravitation constructed in accordance with the algorithm of the dynamic quantization method [2]. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In the recent publication [1], a new version of quantum gravity on a lattice was proposed. In addition to variables describing the gravitational degrees of freedom, this theory also includes fermion degrees of freedom, the fermion action being local and γ^5-invariant. In addition, in the naive continuum limit, the lattice action is transformed into the Hilbert action plus the action of a massless Dirac field, which is connected with the gravitational field to the minimal extent. Since this action will be used here, we describe it in detail.

Let $\mathfrak{K}$ be a four-dimensional simplicial complex permitting a geometrical realization. The definition and required properties of simplicial complexes can be found in [1]. A detailed theory of simplicial complexes is given, for example, in [3, 4]. Instead of the word combination "simplicial complex," we will henceforth use just the term "complex" and will regard as synonyms the following pairs of concepts: a 0-simplex and a vertex, a 1-simplex and an edge, a 2-simplex and a triangle, and a 3-simplex and a tetrahedron. Of special interest are finite complexes with the 4-disk topology. Such complexes have a boundary $\partial\mathfrak{K}$, where $\partial\mathfrak{K}$ is a three-dimensional complex with a topology of sphere S^3. We denote by α_q, $q = 0, 1, 2, 3, 4$, the number of q simplices of complex $\mathfrak{K}$. Indices i, j, k, and l enumerate the vertices of the complex: a_i, a_j, and so on. Two vertices will be referred to as neighboring if these vertices are the boundary vertices of the same edge.

Let γ^a, $a, b, c = 1, 2, 3, 4$, be four-dimensional Dirac matrices. The signature is assumed to be Euclidean. Consequently, all Dirac matrices are Hermitian. The Hermitian matrix

$$\gamma^5 = \gamma^1\gamma^2\gamma^3\gamma^4, \quad \mathrm{tr}\gamma^5\gamma^a\gamma^b\gamma^c\gamma^d = 4\varepsilon^{abcd}, \qquad (1)$$

divides the "right" component φ and the "left" component χ of a Dirac spinor ψ. At each vortex a_i of complex $\mathfrak{K}$, Dirac spinors ψ_i and $\bar{\psi}_i$ are defined, on which the Dirac matrices act from left and right, respectively. It should be recalled that, in the case of a Euclidean signature, fields ψ_i and $\bar{\psi}_i$ are assumed to be independent variables, which are transformed into each other in the case of Hermitian conjugation. We juxtapose each oriented edge a_ia_j to an element of group Spin(4):

$$\Omega_{ij} = \Omega_{ji}^{-1} = \exp\left(\frac{1}{2}\omega_{ij}^{ab}\sigma^{ab}\right), \quad \sigma^{ab} = \frac{1}{4}[\gamma^a, \gamma^b]. \quad (2)$$

Holonomy element Ω_{ij} of the gravitational field executes a parallel translation of spinor ψ_j from vertex a_j of edge a_ia_j to neighboring vertex a_i. We denote by V a linear space with basis γ^a. Let each oriented edge a_ia_j be put in correspondence with element $\hat{e}_{ij} \equiv e_{ij}^a\gamma^a \in V$, such that

$$\hat{e}_{ij} = -\Omega_{ij}\hat{e}_{ji}\Omega_{ij}^{-1}. \qquad (3)$$

Index A enumerates 4-simplices. The notation $\bar{\psi}_{Ai}$, ψ_{Ai}, $\hat{e}_{Aij}$, and Ω_{Aij} indicates that edge a_ia_j belongs to the 4-simplex with index A.

By hypothesis, complex $\Re$ has a disk topology. For such a complex, the concept of orientation can be introduced. We define the orientation of the complex by defining the orientation of each 4-simplex. In this case, if two 4-simplices have a common tetrahedron, the two orientations of the tetrahedron, which are defined by the orientations of these two 4-simplices, are opposite. In our case, the complex obviously has only two orientations.

Let a_{Ai}, a_{Aj}, a_{Ak}, a_{Al}, and a_{Am} be all five vertices of a 4-simplex with index A and $\varepsilon_{Aijklm} = \pm 1$ depending on whether the order of vertices $a_{Ai}a_{Aj}a_{Ak}a_{Al}a_{Am}$ defines the positive or negative orientation of this 4-simplex. In addition, $\varepsilon_{Aijklm} = 0$ if at least two indices coincide. We can now write the Euclidean action in the model in question:

$$I = \frac{1}{5\times 24}\sum_A \sum_{i,j,k,l,m} \varepsilon_{Aijklm}\mathrm{tr}\gamma^5 \times\left\{-\frac{1}{l_P^2}\Omega_{Ami}\Omega_{Aij}\Omega_{Ajm}\hat{e}_{Amk}\hat{e}_{Aml} + \frac{1}{24}\hat{\Theta}_{Ami}\hat{e}_{Amj}\hat{e}_{Amk}\hat{e}_{Aml}\right\}, \quad (4)$$

$$\hat{\Theta}_{Aij} = \frac{i}{2}\gamma^a(\bar{\psi}_{Ai}\gamma^a\Omega_{Aij}\psi_{Aj} - \bar{\psi}_{Aj}\Omega_{Aji}\gamma^a\psi_{Ai}) \equiv \Theta^a_{Aij}\gamma^a.$$

The volume of a 4-complex is given by

$$V_A = \frac{1}{4!}\times\frac{1}{5!} \times\sum_A\sum_{i,j,k,l,m}\varepsilon_{Aijklm}\varepsilon^{abcd}e^a_{Ami}e^b_{Amj}e^c_{Amk}e^d_{Aml}.$$

Here, factor 1/4! is required since the volume of a four-dimensional parallelepiped with generatrices e^a_{Ami}, e^b_{Amj}, e^c_{Amk}, and e^d_{Aml} is 4! times larger than the volume of a 4-simplex with the same generatrices, while factor 1/5! is due to the fact that all five vertices of each simplex are taken into account independently.

The dynamic variables are quantities Ω_{ij} and $\hat{e}_{ij}$, which describe the gravitational degrees of freedom, and fields $\bar{\psi}_i$ and ψ_i, which are material fermion fields (other material fields are not considered here).

Action (4) is of interest in connection with fermion doubling (or Wilson doubling) since its fermion part possesses the following properties.

1. Action (4) is local.

2. In the naive continuum limit, action (4) is transformed into the gravitational action in the Palatini form plus the action for Dirac fields that are connected to the minimal extent with the gravitational field.

3. The fermion part of action (4) is phase-invariant and also γ^5 invariant; i.e., it is invariant relative to the following transformations:

(a) $\psi \longrightarrow \exp(i\alpha)\psi$, $\bar{\psi} \longrightarrow \bar{\psi}\exp(-i\alpha)$,

(b) $\psi \longrightarrow \exp(i\beta\gamma^5)\psi$, $\bar{\psi} \longrightarrow \bar{\psi}\exp(-i\beta\gamma^5)$,

where α and β are real-valued continuous global parameters.

It is well known [5–8] that, on a hypercubic lattice, Wilson doubling takes place for any fermion action possessing properties 1–3. In addition, it is known [9] that Wilson doubling also takes place on periodic lattices on which the fermion action has the form

$$I = \sum_{\mathbf{x},\mathbf{y}}\bar{\psi}_{\mathbf{x}}\hat{H}(\mathbf{x}-\mathbf{y})\psi_{\mathbf{y}} \quad (5)$$

(**x** and **y** are the radius vectors of the lattice site) and possesses the three above-mentioned properties. However, it remains unclear whether a fermion action with properties 1–3 leads to Wilson doubling on any lattice. Here, we give examples of lattices (simplicial complexes) on which Wilson doubling does not take place for action (4). The absence of Wilson doubling for a γ^5-invariant action is equivalent to the possibility of introducing a single Weyl field on a lattice. For example, in order to introduce a "right" Weyl field, the following substitutions should be made in action (4):

$$\psi \longrightarrow (1/2)(1+\gamma^5)\psi, \quad \bar{\psi} \longrightarrow \bar{\psi}(1/2)(1-\gamma^5).$$

According to Hartle and Hawking [10], the main problem in quantum gravity is to calculate the fundamental functional integral/statistical sum—the domain of the fields is a D-dimensional disk, and the wave function of the generated Universe depends on the fields defined on the disk boundary (sphere S^{D-1}). In our case, instead of the integral over continuum, we have a finite-multiplicity integral (since complex $\Re$ is finite) over all fundamental fields with weight $\exp I$. Naturally, the fundamental statistical sum of the Universe must also include summation over the lattices themselves. Consequently, the numbers of simplices and the method for their combination into a complex (and, hence, the numbers of physical degrees of freedom and their constraints) are not fixed and should be determined statistically proceeding from the principle of statistical sum saturation. The dimension of the complex itself must be fixed in a similar way.

In order to solve the problem of Wilson doubling we are interested in, we must assume that the Universe has expanded to such an extent that we can disregard the

fluctuations of the gravitational field and study the eigenmodes of the discrete Dirac operator in relations (4). For solving this problem, we must idealize the situation in the indicated direction. Consequently, we will henceforth assume that

$$\Omega_{ij} = 1, \quad (e^a_{ij} + e^a_{jk} + \ldots + e^a_{li}) = 0. \tag{6}$$

Here, the sum in the parentheses is taken over any closed path formed by 1-simplices. Equations (6) indicate that the curvature and twist are equal to zero. Thus, the geometric realization of complex $\mathfrak{K}$ is in the D-dimensional Euclidean plane, e^a_{ij} being the components of the vector in a certain orthogonal basis in this plane, and the beginning and end of this vector being located at vertices a_i and a_j, respectively. It should be noted that, if relations (6) hold, we have

$$\Theta^a_{ij} = -\Theta^a_{ji}. \tag{7}$$

Let us demonstrate the problem of Wilson doubling using the simple example of the Dirac theory in a 2D Minkowski space. We denote by φ and χ the upper and lower components of complex Dirac field ψ and suppose that the fermion Hamiltonian in the continuum case has the form

$$\mathcal{H} = \int dx\left[\varphi^\dagger\left(-i\frac{\partial}{\partial x} - eA_x\right)\varphi + \chi^\dagger\left(i\frac{\partial}{\partial x} + eA_x\right)\chi\right]. \tag{8}$$

Here, A_x is the x component of the gauge field (for simplicity, we choose an Abelian field). In the free case, the equations for eigenmodes and their solutions have the form

$$-i\frac{\partial}{\partial x}\varphi_p = \epsilon^\varphi_p\varphi_p, \tag{9a}$$
$$\varphi_p = e^{ipx}, \quad -\infty < p < +\infty, \quad \epsilon^\varphi_p = p,$$

$$i\frac{\partial}{\partial x}\chi_p = \epsilon^\chi_p\chi_p, \tag{9b}$$
$$\chi_p = e^{ipx}, \quad -\infty < p < +\infty, \quad \epsilon^\chi_p = -p.$$

By definition, the spatial component of the vector current has the form

$$J_x \equiv \frac{\delta\mathcal{H}}{\delta A_x} = e(\varphi^\dagger\varphi - \chi^\dagger\chi). \tag{10}$$

Let us now consider the simplest generalization of Hamiltonian (8) to a lattice and confine the analysis to upper component φ. In our case, the lattice is one-dimensional. For simplicity, we assume that the lattice consists of N vertices on a circle, which are enumerated successively, their numbers being determined by modulus N. A complex quantity φ_n is defined at each vertex

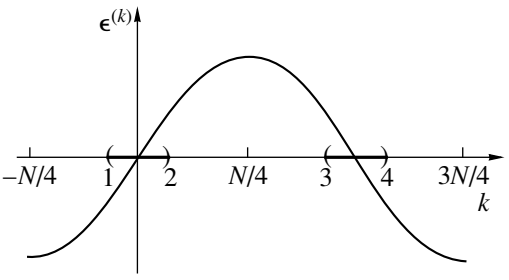

Fig. 1.

of the lattice with number n and a phase factor $\exp(ieA_n)$ is defined on each 1-simplex $a_n a_{n+1}$. The real-valued Hamiltonian, which is transformed into Hamiltonian (8) in the continuum limit, can be written in the form

$$\mathcal{H} = \frac{1}{2i} \times \sum_{n=1}^{N} \varphi^\dagger_n(\exp(-ieA_n)\varphi_{n+1} - \exp(ieA_{n-1})\varphi_{n-1}), \tag{11}$$

while an analog of the spatial component of vector current (10) has the form

$$J_n \equiv -\frac{\delta\mathcal{H}}{\delta A_n} = \frac{e}{2}[\varphi^\dagger_n\exp(-ieA_n)\varphi_{n+1} + \varphi^\dagger_{n+1}\exp(ieA_n)\varphi_n]. \tag{12}$$

In the free case ($A_n = 0$), we have the following equation for the eigenmodes of Hamiltonian (12) and its solution:

$$\frac{1}{2i}(\varphi_{n+1} - \varphi_{n-1}) = \epsilon\varphi_n,$$
$$\varphi^{(k)}_n = \frac{1}{\sqrt{N}}\exp\left(\frac{2\pi ikn}{N}\right), \quad \epsilon^{(k)} = \sin\frac{2\pi k}{N}, \tag{13}$$

$$k = -\frac{N}{4}, -\frac{N}{4} + 1, \ldots, 3\frac{N}{4} - 1. \tag{14}$$

We assume for simplicity that number N is a multiple of number 4. Integer k enumerates the modes; when we enumerate all independent modes, this number runs continuously through N values. Figure 1 shows the curve describing the dependence of $\epsilon^{(k)}$ on number k for large values of N. We introduce a quasi-continuous parameter $p = 2\pi k/N$ and consider the region $|p| \ll 1$. The range of values of k corresponding to this region p is contained between parentheses 1 and 2 in Fig. 1. In this region, which will be referred to as trivial, we have approximately

$$\epsilon_p = p, \quad |p| \ll 1. \tag{15}$$

Thus, as $N \longrightarrow \infty$, the spectrum of field φ on the lattice in the trivial region coincides with spectrum (8) and the lattice contribution to current (12) coincides with the contribution to current (10) from the upper component of the Dirac field. The zero and soft modes from the trivial region of the spectral parameter will be referred to as trivial.

Let us now carry out the substitution

$$k = \frac{N}{2} + \frac{N}{2\pi}p, \quad |p| \ll 1. \tag{16}$$

In this case, the range of spectral parameter k is confined between parentheses 3 and 4 in Fig. 1. This region of the spectral parameter and the corresponding modes will be referred to as nontrivial. Since the nontrivial region (as well as the trivial one) is relatively small, the dependence of the spectrum on the spectral parameter can be linearized:

$$\epsilon_p = -p, \quad |p| \ll 1. \tag{17}$$

In accordance with relations (14) and (16), for nontrivial modes we have

$$\varphi_{n+1}^{(p)} = -\varphi_n^{(p)} e^{ip} = (-1)^{n+1}\varphi_0 e^{ip(n+1)}, \quad |p| \ll 1. \tag{18}$$

It follows hence that the configurations of the "modes"

$$\chi_n^{(p)} = (-1)^n \varphi_n^{(p)}, \quad \chi_{n+1}^{(p)} = \chi_n^{(p)} e^{ip}, \quad |p| \ll 1, \tag{19}$$

have a continuum limit. If we express the contribution to vector current (12) from nontrivial modes in terms of modes $\chi^{(p)}$, we obtain formula (10). Together with formula (17) for the spectrum, this means that an aggregate of nontrivial modes can be treated in the continuum limit as the lower component of Dirac field χ; in this case, Hamiltonian (7) can be used for describing the entire system.

It is also worth noting that a nontrivial mode in the given case has two branches, each of which separately has a continuum limit. For example, in the case of even n, we mark by primes vertices a_n and the values φ_n of nontrivial modes corresponding to them, while in the case of odd n, we mark the corresponding quantities by two primes. Aggregates of quantities $\{\varphi_n'\}$ and $\{\varphi_n''\}$ form two branches of the nontrivial mode. The above formulas show that both branches have continuum limits, and Eq. (13) for the zero mode does not mix these two branches.

It should be noted, irrespective of the above example, that nontrivial modes that do not split into their branches, each of which has a continuum limit, can hardly be treated as physical modes in the continuum limit.

It can be seen from the example considered above that the phenomenon of Wilson doubling means that, if only one left (right) Weyl field is introduced explicitly on lattices, we inevitably have only Dirac fields in the continuum limit.

We will show in this work that the "no go" theorem, which was proved for an action of type (5) [9], is generally invalid for amorphous lattices. This statement is substantiated by considering examples of amorphous lattices in two, three, and four dimensions for action (4) or its analogs for which Wilson doubling is absent. In the final section, we will consider possible consequences of this result for the problem of axial anomaly. It is concluded that the version of discrete quantum gravity proposed in [1] is transformed in the continuum limit into the quantum theory of gravitation constructed with the help of the dynamic quantization method [2].

2. TWO-DIMENSIONAL LATTICES

We begin the analysis of Wilson doubling with the simplest case, when $\mathfrak{K}$ is a two-dimensional simplicial complex. We assume that the geometrical realization of complex $\mathfrak{K}$ is a two-dimensional surface with the disk topology, $\partial\mathfrak{K}$ being a one-dimensional simplicial complex with the topology of a circle. For definiteness, we assume that two-dimensional Dirac matrices γ^a, $a = 1, 2$, are $\gamma^1 = \sigma^1$ and $\gamma^2 = \sigma^2$, where σ^α, $\alpha = 1, 2, 3$, are the Pauli matrices. By definition, $\gamma^5 = i\gamma^1\gamma^2$. At each vertex a_i of complex $\mathfrak{K}$, Dirac spinors ψ_i and $\bar{\psi}_i$ are defined, which are two-dimensional column and row matrices, respectively. We juxtapose each oriented edge $a_i a_j$ to an element of group Spin(2) (which is Abelian in the two-dimensional case), denoted by $\Omega_{ij} = \Omega_{ij}^{-1}$. In two dimensions, index A enumerates triangles of complex $\mathfrak{K}$. Compound indices (Ai), (Aij), etc., indicate the fact that vertex a_{Ai}, edge $a_{Ai}a_{Aj}$, etc., belong to a triangle with index A. The complexes considered here permit the introduction of orientation. We define the orientation of the complex by defining the orientation of each triangle. If two triangles share an edge, the two orientations of the edge defined by the orientations of these two triangles are opposite. By definition, $\varepsilon_{Aijk} = \pm 1$ depending on whether the order of vertices $a_{Ai}a_{Aj}a_{Ak}$ defines the positive or negative orientation of the corresponding triangle.

We can now write the fermion part of the action (cf. formulas (4)):

$$I_\psi = \frac{1}{6}\sum_A \sum_{i,j,k} \varepsilon_{Aijk}\varepsilon_{ab}\Theta_{Aij}^a e_{Aik}^b. \tag{20}$$

Two-dimensional action (20) possesses all three of the properties listed after formula (4) for four-dimensional action (4). In accordance with the same considerations

as those given in the Introduction, equalities (6) and (7) also hold in the two-dimensional case.

We can now write the equations for the eigenmodes of a discrete Dirac operator. We fix two neighboring vertices a_i and a_j and single out the contribution to action (20), proportional to Θ^a_{ij}. Figure 2 shows a part of the complex, containing 1-simplex a_ia_j; indices i, j, k, l enumerate vertices, while index A enumerating triangles assumes two values (1 and 2) in this case. We have everywhere $s^a_{ij} = \varepsilon_{ab}e^b_{ij}$; i.e., vector s^a_{ij} can be obtained by rotating vector e^a_{ij} clockwise through angle $\pi/2$. The sought contribution to the action is given by

$$\Delta I_{\psi ij} = \frac{1}{3}\Theta^a_{ij}S^a_{ij}, \quad S^a_{ij} = s^a_{kj} + s^a_{jl}. \tag{21}$$

Vectors S^a_{ij} can be referred to as an "umbrella" of vertex a_i from the neighboring vertex a_j. Umbrella S^a_{ij} is determined unambiguously from two given neighboring vertices a_i and a_j; it can be seen from Fig. 2 and relations (6) that $S^a_{ij} = -S^a_{ji}$. We separate from the complex a subcomplex $\mathfrak{v}_i$ consisting wholly of 2-simplices containing vertex a_i and will refer to this complex as a neighborhood of vertex a_i. We enumerate the vertices on boundary $\partial\mathfrak{v}_i$ in such a way that vertex a_{j+1} follows vertex a_j during a continuous counterclockwise traversing of boundary $\partial\mathfrak{v}_i$ and assume that index j is defined by (mod n), where n is the number of vertices on $\partial\mathfrak{v}_i$. The fact that index j enumerates the vertices on $\partial\mathfrak{v}_i$ is denoted by $j(i)$. Using formula (21), we can easily separate from action (20) the contribution proportional to spinor $\bar\psi_i$:

$$\Delta I_{\bar\psi_i} = \frac{1}{3}\sum_{j(i)}\Theta^a_{ij}S^a_{ij}. \tag{22}$$

Using formulas (6), (20), and (22), we obtain the following equation for eigenmodes of the discrete Dirac operator at internal vertices a_i:

$$\frac{\delta\Delta I_{\bar\psi_i}}{\delta\bar\psi_i} = \frac{i}{6}\sum_{j(i)}\hat S_{ij}\psi_j = \epsilon\left(\frac{1}{3}v_i\right)\psi_i. \tag{23}$$

Here, v_i is the area of the neighborhood $\mathfrak{v}_i$. In accordance with the second equality in (6), we have the identity

$$\sum_{j(i)}\hat S_{ij} \equiv 0. \tag{24}$$

Indeed, each vector $s^a_{j(i),\,j(i)+1}$ is contained in two and only two umbrellas in the latter sum. It follows from identity (24) that Eq. (23) is of the difference nature; i.e., the left-hand side of this equation is a function of differences $(\psi_{j(i)} - \psi_{k(i)})$ only.

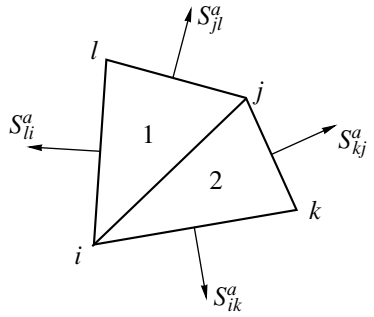

Fig. 2.

The system of equations for eigenmodes has a more elegant form in complex notation. Let us suppose that x^a_j represents the Cartesian coordinates of vertex a_j and $z_j = x^1_j + ix^2_j$ is its complex coordinate. We denote the upper and lower components of Dirac spinor ψ by φ and χ, respectively. In this case, Eq. (23) assumes the form

$$-\frac{1}{2}\sum_{j(i)}(\bar z_{j+1} - \bar z_{j-1})\chi_j = \epsilon v_i\varphi_i, \tag{23'a}$$

$$\frac{1}{2}\sum_{j(i)}(z_{j+1} - z_{j-1})\varphi_j = \epsilon v_i\chi_i; \tag{23'b}$$

for the zero mode ($\epsilon = 0$), we have

$$\sum_{j(i)}(z_{j+1} - z_{j-1})\varphi_j = 0$$
$$\longleftrightarrow \sum_{j(i)} z_j(\varphi_{j+1} - \varphi_{j-1}) = 0. \tag{25}$$

In the subsequent analysis, we will use the following notation for difference variables: $\psi_{i,j} \equiv \psi_i - \psi_j$.

We are interested in all zero modes of the discrete Dirac operator with zero boundary conditions for difference variables $\varphi_{k,k'}$. One of these solutions is obvious: $\varphi_i = \text{const}$. This solution will be referred to as a trivial zero mode.

In order to clarify the situation with zero modes, we consider a concrete example.

Suppose that $\partial\mathfrak{v}_i$ has an even number of vertices. In this case, vertex a_i is called even. Then the set of indices $j(i)$ can be divided into two groups containing equal numbers of elements. Indices of the first group will be marked by one prime and those of the second group, by

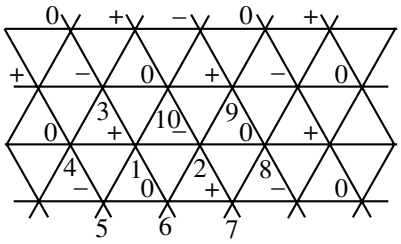

Fig. 3.

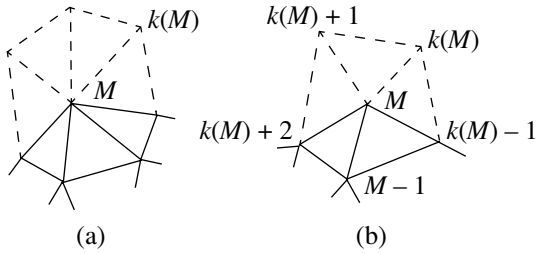

Fig. 4.

two primes; as we traverse continuously along $\partial\mathfrak{v}_i$, vertices with primed and double primed indices alternate. In the case under consideration, Eq. (25) can be rewritten in the form

$$\left[\sum_{j'(i)} z_{j'}(\varphi_{j'+1} - \varphi_{j'-1})\right] + \left[\sum_{j''(i)} z_{j''}(\varphi_{j''+1} - \varphi_{j''-1})\right] = 0. \quad (26)$$

We now assume that all internal vertices of the complex have an even number of neighboring vertices. In addition, we assume that the entire set of internal vertices splits into a finite number of subsets (in our case, we have three subsets $\{a_{i'}\}$, $\{a_{i''}\}$, and $\{a_{i'''}\}$) so that system of equations (25) for the zero mode contains only differences $(\psi_{j'_1} - \psi_{j'_2})$, $(\psi_{j''_1} - \psi_{j''_2})$, and $(\psi_{j'''_1} - \psi_{j'''_2})$. It is important to note that the coordinates of the vertices are in the common position. We will refer to fields $\psi_{j'}$, $\psi_{j''}$, and $\psi_{j'''}$ as the branches of the zero mode and of soft modes close to it. In this case, the phenomenon of Wilson doubling obviously takes place (the effect of the boundary for $\alpha_0 \longrightarrow \infty$ can be neglected). Figure 3 precisely illustrates this example, in which the vertices from three such subsets of vertices are marked by indices $0, \pm$. A nontrivial zero mode can be taken, for example, in the form

$$\varphi^0 = c \neq 0, \quad \varphi^{\pm} = [\exp(\pm 2\pi i/3)]c.$$

Here, φ^0 and $\varphi^{\pm}$ are the values of field φ at the vertices marked in Fig. 3 by indices 0 and $\pm$, respectively. This nontrivial mode is orthogonal to the trivial mode (in the natural measure $\sum_i \bar{\psi}_i \psi_i$ on a regular lattice) and is hence independent. In this example, we have three branches of the nontrivial zero mode.[1]

[1] In connection with the problem analyzed here, we can mention review [11], in which the difference Laplace operator factorized into first-order difference operators is considered on regular triangular lattices. The latter operators level out the values of variables at neighboring vertices and differ in this respect from the operator in formulas (23) and (25).

The following question arises: does a lattice, on which nontrivial zero modes are absent, exist? In order to answer this question, we must study some properties of system of equations (25), which requires additional constructions.

For the Wilson doubling to take place, the zero set of values of variables $\psi_{k,k-2}$ on boundary $\partial\mathfrak{N}$ must correspond to different solutions to system of equations (25). In other words, for zero values of variables $\psi_{k,k-2}$ on $\partial\mathfrak{N}$, nonzero solutions to system of equations (25) must exist for certain variables $\varphi_{i,j}$. Further, we prove that all internal difference variables $\psi_{j(i),j(i)-2}$ vanish on the so-called odd complexes for zero values of variables $\psi_{k,k'}$ on $\partial\mathfrak{N}$ in accordance with system of equations (25).

The fact that internal variables $\psi_{j(i),j(i)-2}$ are equal to zero does not always mean that Wilson doubling is absent. Indeed, for the lattice depicted in Fig. 3, we have

$$\psi_{j'_1} = \psi_{j'_2} = \ldots, \quad \psi_{j''_1} = \psi_{j''_2} = \ldots,$$
$$\psi_{j'''_1} = \psi_{j'''_2} = \ldots,$$

but

$$\psi_{j'} \neq \psi_{j''} \neq \psi_{j'''}.$$

Let us now consider a complex with properties opposite to those of the complex shown in Fig. 3 in a certain sense. This is a complex for which the boundary $\partial\mathfrak{v}_i$ of each neighborhood has an odd number of vertices. Such complexes will be referred to as odd.

The inductive procedure of constructing odd complexes can be described as follows. At the first stage, we can take any complex consisting of an odd number of triangles with one common vertex, which is the only internal vertex. Suppose that an odd complex with $(M-1)$ internal vertices has already been constructed. We take any vertex on its boundary, denote it by a_M, and make it an internal vertex by supplementing the complex with new elements.

Solid lines in Fig. 4 depict the old part of the complex with $(M-1)$ internal vertices, while dashed lines show the part of the new complex, which is added to the

old one. In Fig. 4a (4b), boundary vertex a_M first belonged to the boundaries of an even (odd) number of 1-simplices. Consequently, in the additional construction, an odd (even) number of new vertices and the required number of 1-simplices are added in the case depicted in Fig. 4a (4b). If the external boundary angle at vertex a_M in Fig. 4b is acute, the additional construction of the complex may consist in adding a single 1-simplex with boundary vertices $a_{k(M)-1}$ and $a_{k(M)+2}$; as a result, the number of boundary vertices is reduced by one.

In the subsequent analysis, we will consider only the complexes constructed by induction in accordance with the above procedure. The property of oddness is not necessary in this case.

We will refer to difference variable $\varphi_{k,i}$ as a regular internal variable if $a_k \in \partial\Re$ and $a_i \notin \partial\Re$. A set of regular internal variables $\{\varphi_{k_i,i}\}_M$, $i = 1, \ldots, M$ (where M in the number of internal vertices) is an independent regular set of internal variables if all internal vertices a_i are pairwise different. The adjective "regular" will be henceforth omitted since doing so will not lead to misunderstanding. For the remaining $(L-1)$ independent variables, we take independent difference variables $\{\varphi_{k_\alpha,k'_\alpha}\}$, a_{k_α}, $a_{k'_\alpha} \in \partial\Re$. Here, L is the number of vertices on $\partial\Re$.

Let us consider the system of M equations (25) for a complex with M internal and L boundary vertices for $M+L-1$ independent variables $\{\varphi_{k_i,i}\}_M$ and $\{\varphi_{k_\alpha,k'_\alpha}\}$:

$$\sum_{j=1}^{M} X_{i,j}\varphi_{k_j,j} + \sum_{\alpha=1}^{L-1} Y_{i,\alpha}\varphi_{k_\alpha,k'_\alpha} = 0, \quad i = 1, \ldots, M. \tag{27}$$

Here, coefficients $X_{i,j}$ and $Y_{i,\alpha}$ can be expressed linearly in terms of variables z_i.

Statement 1. An $M \times M$ matrix $\|X_{i,j}\|$ is nondegenerate if M is an even number.

Proof. Let us consider a complex with $M = 2$. For example, this can be a subcomplex of the complex depicted in Fig. 5 and consisting of triangles with numbers from 1 to 8.

We take differences $\{\varphi_{10,1}, \varphi_{7,2}\}$ as internal variables and differences $\{\varphi_{8,3}, \varphi_{9,7}, \varphi_{10,8}, \varphi_{11,9}, \varphi_{8,7}\}$ as boundary variables. Then the matrix in the system of two equations (27) at vertices a_1 and a_2 (we are using notation $z_{i,j} = z_i - z_j$) has the form

$$\|X_{i,j}\| = \begin{pmatrix} 0 & -z_{9,3} \\ z_{9,3} & 0 \end{pmatrix}, \quad \det X_{i,j} = z_{9,3}^2 \neq 0.$$

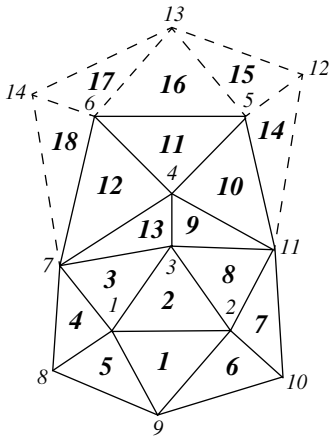

Fig. 5.

Let us now consider a complex with $M = 4$, which is a subcomplex of the complex shown in Fig. 5 and consists of triangles with numbers from 1 to 13. We choose the internal and boundary difference variables in the form $\{\varphi_{10,1}, \varphi_{7,2}, \varphi_{8,3}, \varphi_{7,4}\}$ and $\{\varphi_{8,6}, \varphi_{9,7}, \varphi_{10,8}, \varphi_{11,9}, \varphi_{10,5}, \varphi_{11,6}\}$, respectively. In this case, the matrix in the system of four equations (27) at vertices a_1, a_2, a_3, and a_4 has the form

$$\|X_{i,j}\| = \begin{pmatrix} 0 & -z_{9,3} & z_{7,2} & 0 \\ z_{9,3} & 0 & -z_{11,1} & 0 \\ -z_{7,2} & z_{11,1} & 0 & -z_{11,7} \\ 0 & 0 & z_{11,7} & 0 \end{pmatrix},$$

$$\det X_{i,j} = z_{9,3}^2 z_{11,7}^2 \neq 0.$$

We assume that the statement has been proved for an even number $M-2$ and establish its validity for complexes with M internal vertices.

Let us consider the following case of completion of the complex depicted in Fig. 5: vertices a_5 and a_6 are successively made internal by adding the triangles with numbers 14, 15, 16 and then 17 and 18 to the complex. We can assume that vertices a_5 and a_6 are identified with vertices a_{M-1} and a_M, respectively. Let the numbers of equations in the new system of M equations (27) correspond to the numbers of internal vertices. We express the old boundary variables of forms $\varphi_{k_1,M-1}$ and $\varphi_{k_2,M}$, which have become internal variables in the completed complex, in terms of new internal variables $\varphi_{k(M-1),M-1}$, $\varphi_{k(M),M}$ and new boundary variables. Thus, as we pass from the complex with $M-2$ internal vertices to the complex with M internal vertices, nonzero coefficients $X_{M-2,M-1}$ and $X_{M-2,M}$ appear only in the $(M-2)$th equation in the new system of equations (27). It is important that remaining coefficients

$X_{i,j}$, $i, j = 1, \ldots, M-2$, do not change in this case. On the other hand, in the $(M-1)$th and Mth equations (see Fig. 5), we have

$$X_{M-1,j} = 0, \quad X_{M,j} = 0, \quad j = 1, 2, \ldots, M-3.$$

Let us consider the linear combination of the last two rows of matrix $X_{i,j}$:

$$Y_j = c_{M-1}X_{M-1,j} + c_M X_{M,j}, \quad c_{M-1}^2 + c_M^2 > 0.$$

Since vertices a_{M-1} and a_M are adjacent, we have $Y_{M-1}^2 + Y_M^2 > 0$ (which can be verified directly). Suppose that the equality

$$Y_j = \sum_{i=1}^{M-2} c_i X_{i,j}, \quad 1 \le j \le M \tag{28}$$

holds, where c_i are certain numbers. In this case, we have two alternatives.

1. $Y_j = 0$, $1 \le j \le M-2$. Then, in view of inductive hypothesis and conservation of matrix $X_{i,j}$, $1 \le i, j \le M-2$, in transition $(M-2) \longrightarrow M$, equality (28) for $1 \le j \le M-2$ will be satisfied if $c_1 = \ldots = c_{M-2} = 0$. However, equality (28) is not satisfied in this case for $j = M-1, M$.

2. $Y_{M-2} \neq 0$. It can be seen from Fig. 5 that $c_{M-3} \neq 0$ in this case. The latter inequality implies in turn that $c_{M-4} \neq 0$. This means that equality (28) is contradictory. Indeed, $Y_j = 0$ for $j = 1, \ldots, M-4$, while the corresponding terms on the right-hand side of Eq. (28) can be equal to zero only if $c_1 = \ldots = c_{M-4} = 0$.

Thus, equality (28) cannot be satisfied, which means that the Statement is valid in the case depicted in Fig. 5.

The Statement 1 can be proved analogously in all remaining cases. It is only important that vertices a_{M-1} and a_M must be adjacent in the proof by induction. □

In the case of complexes with an odd number of internal vertices, the statement is incorrect in the variables used here. However, an insignificant modification of the variables makes it possible to formulate and prove an analogous statement. Indeed, a complex with an odd number of internal vertices can be obtained from a complex with an even number of vertices by the additional construction shown in Fig. 4. Suppose that a complex with an even number of internal vertices has $M-1$ internal vertices $a_1, \ldots, a_{M-1}$, while the completed complex has M internal vertices $a_1, \ldots, a_M$. We consider a set of M variables, consisting of $M-1$ internal variables $\{\varphi_{k,i}\}$, $i = 1, \ldots, M-1$, and a boundary variable $\varphi_{k(M),k(M)-1}$. We choose a set of the remaining independent variables so that it consists of all boundary variables of the initial complex with $M-1$ internal vertices as well as the missing boundary variables of the completed complex with M internal vertices. In these variables, system of M equations (25) for the completed complex has the form

$$\sum_{j=1}^{M-1} X_{i,j}\varphi_{k_j,j} + \ldots = 0, \quad i = 1, \ldots, M-1,$$

$$\sum_{j=1}^{M-1} X'_{M,j}\varphi_{k,j} + X'_{j+1}\varphi_{k(M),k(M)-1} + \ldots = 0,$$

$$X'_{j+1} \neq 0.$$

Dots denote the contributions from all other difference variables. In this system of equations, the determinant of the minor for variables $\{\varphi_{k,i}\}$, $i = 1, \ldots, M-1$ and $\varphi_{k(M),k(M)-1}$ is equal to $(x'_{j+1}\det X_{i,j}) \neq 0$ since, in accordance with Statement 1, $\det X_{i,j} \neq 0$.

At this stage, it is expedient to formulate a more exact criterion for the presence of Wilson doubling. Suppose that we can choose in system of equations (27) M independent internal variables and the required number of independent boundary variables in such a way that the total number of independent variables is smaller than a number $M + L - 1$. For this reason, vanishing of the boundary variables contained in system of equations (27) does not mean vanishing (in accordance with system (27)) of all difference variables $\varphi_{i,j}$ for internal vertices a_i and a_j. In this case, Wilson doubling takes place. This criterion of Wilson doubling can also be applied to a part of the complex. If, for example, we take the part of the complex depicted in Fig. 3 and bounded by two internal and eight boundary vertices (the vertices of this part of the complex are enumerated by indices from 1 to 10 in Fig. 3), then the system of two equations (27),

$$z_{10,6}\varphi_{7,1} = -z_{10,8}\varphi_{9,7} + z_{9,1}\varphi_{10,6} - z_{9,7}\varphi_{8,6},$$
$$z_{10,6}\varphi_{3,2} = z_{6,4}\varphi_{5,3} - z_{5,3}\varphi_{6,4} + z_{3,2}\varphi_{10,6},$$

contains only seven independent difference variables (two internal variables $\varphi_{7,1}$ and $\varphi_{3,2}$ and five boundary variables $\varphi_{5,3}$, $\varphi_{9,7}$, $\varphi_{6,4}$, $\varphi_{10,6}$, and $\varphi_{8,6}$. The total number of independent difference variables on this subcomplex is nine. In accordance with the above system of equations, vanishing of the five boundary variables indicated above leads to vanishing of internal variables $\varphi_{1,7}$ and $\varphi_{2,3}$, but not difference variable $\varphi_{1,2}$; according to our criterion, this indicates the presence of Wilson doubling. For an indefinitely large subcomplex of the complex shown in Fig. 3, the result is identical.

It can easily be seen that, in the case of an odd lattice, system of equations (27) inevitably contains all

$M + L - 1$ independent difference variables. This follows directly from the identity

$$\varphi_{j(i)+1,\,j(i)} \equiv \sum_{0 \le k \le (n-1)/2} \varphi_{j(i)+2k+2,\,j(i)+2k}, \qquad (29)$$

where n is an (odd) number of vertices on boundary $\partial\mathfrak{v}_i$. Equality (29) shows that any difference variables can be expressed in terms of the difference variables contained in system of equations (25). Consequently, vanishing of all boundary variables in system of equations (27) leads to vanishing of all variables $\varphi_{i,j}$.

This result can be reformulated as follows. We consider system of equations (25) for finite subcomplexes of an odd complex with M internal vertices. Let vertex a_i and at least one of vertices $a_{j(i)}$ and $a_{j(i)-2}$ be internal. On the odd complex, there are M independent variables of form $\varphi_{j(i),\,j(i)-2}$, which are contained in system of equations (25) and via which all difference variables of form $\varphi_{i,j}$ can be expressed, the minor of these variables differing from zero. Indeed, a transition from an independent system of regular internal variables to an independent system of variables of form $\varphi_{j(i),\,j(i)-2}$ on an odd complex can be reduced to a linear nondegenerate transformation of variables.[2] Consequently, for $M \longrightarrow \infty$, the minors for any independent sets of variables $\{\varphi_{j(i),\,j(i)-2}\}$, in terms of which all difference variables of form $\varphi_{i,j}$ can be expressed in a finite region of the complex, differ from zero. Consequently, we can formulate

Statement 2. On odd complexes, Wilson doubling is absent.

3. MULTIDIMENSIONAL LATTICES

We will now prove that among complexes with a dimension of $d > 2$, complexes exist that are analogous, in a certain sense, to two-dimensional odd complexes.

We consider the case when $\mathfrak{K}$ is a three-dimensional complex embedded in a three-dimensional Euclidean space. In this case, index A enumerates tetrahedra. We assume that γ matrices are four-dimensional. All other notations are the same as in the previous sections. The orientation of the complex is determined by (or determines) the orientation of each tetrahedron; if two tetrahedra have a common triangle, the two orientations of the triangle determined by the orientations of these two tetrahedra are opposite. Analogously to the two-dimensional case, $\varepsilon_{Aijlk} = \pm 1$ depending on whether the order of vertices $a_{Ai}a_{Aj}a_{Ak}a_{Al}$ defines the positive or negative orientation of the corresponding tetrahedron.

[2] Such a transformation of variables is impossible for the complex depicted in Fig. 3.

The fermion part of the action (cf. formulas (4) and (20)) can be written as

$$I_\psi = \frac{1}{2\times 4\times 6}\sum_A \sum_{i,j,k,l} \varepsilon_{Alijk}\varepsilon^{abc}\Theta^a_{Ali}e^b_{Alj}e^c_{Alk}. \qquad (30)$$

We assume that order $a_{Al}a_{Ai}a_{Aj}a_{Ak}$ defines the positive orientation. Factor 1/4 in relation (30) takes into account the fact that each tetrahedron is included in the sum four times, while factor 1/6 in Eq. (30) is required since the volume of the parallelepiped with generatrices e_{Ali}, a_{Alj}, a_{Alk} has a volume six times as large as the volume of the tetrahedron $a_{Al}a_{Ai}a_{Aj}a_{Ak}$.

We denote by $\mathfrak{v}^{(3)}_i$ a three-dimensional subcomplex consisting of all three simplices of the complex, which contain internal vertex a_i. We enumerate vertices on $\partial\mathfrak{v}^{(3)}_i$ by index $j(i)$. Let

$$\mathfrak{v}^{(2)}_{j(i)} \in \partial\mathfrak{v}^{(3)}_i$$

denote a two-dimensional subcomplex consisting of all 2-simplices of subcomplex $\partial\mathfrak{v}^{(3)}_i$, which contain vertex $a_{j(i)}$. We assume that boundary $\partial\mathfrak{v}^{(2)}_{j(i)}$ is traversed in the positive direction if the circumvention appears counterclockwise to an "observer" located at vertex $a_{j(i)}$ and looking at vertex a_i. We enumerate the vertices on boundary $\partial\mathfrak{v}^{(2)}_{j(i)}$ by index $j'(i,j)$, the value of this index increasing by unity upon a transition from one vertex to an adjacent one in the case of positive motion along $\partial\mathfrak{v}^{(2)}_{j(i)}$. We also assume that index $j'(i,j)$ is defined in (mod n), where n is the number of vertices on $\partial\mathfrak{v}^{(2)}_{j(i)}$. We denote by $s^a_{j(i),\,j'(i,j),\,j'(i,j)+1}$ a vector equal in magnitude to the area of triangle $a_{j(i)}a_{j'(i,j)}a_{j'(i,j)+1}$ and directed along the outward normal of this triangle relative to subcomplex $\mathfrak{v}^{(3)}_i$:

$$s^a_{j(i),\,j'(i,j),\,j'(i,j)+1} = \frac{1}{2}\varepsilon^{abc}e^b_{j(i),\,j'(i,j)}e^c_{j(i),\,j'(i,j)+1}. \qquad (31)$$

The umbrella of vertex a_i from vertex $j(i)$ is the vector

$$S^a_{ij} = \sum_{j'(i,j)=1}^{n} s^a_{j(i),\,j'(i,j),\,j'(i,j)+1}. \qquad (32)$$

Umbrella (32) can also be presented as the sum

$$S^a_{ij} = (2!)^{-2}\sum_{A(i,j)}\sum_{k,l}\varepsilon^{abc}\varepsilon_{A(i,j)ijkl}e^b_{A(i,j)ik}e^c_{A(i,j)il}, \qquad (33)$$

where index $A(i,j)$ enumerates all tetrahedra containing edge a_ia_j.

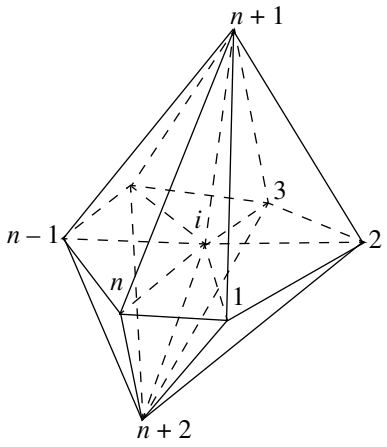

Fig. 6.

It can easily be seen that the contribution to action (30), proportional to $\bar{\psi}_i\gamma^a\psi_j$, is given by

$$\Delta I_{\bar{\psi}_i\psi_j} = \frac{i}{12}\bar{\psi}_i\gamma^a\psi_j S^a_{i,j}. \tag{34}$$

Since the volume of the complex can be represented in the form (cf. the four-dimensional case)

$$V = \frac{1}{3!}\frac{1}{4!}\sum_A \sum_{i,j,k,l} \varepsilon_{Aijk}\varepsilon^{abc}e^a_{li}e^b_{lj}e^c_{lk},$$

the equation for the eigenmodes of the Dirac operator (an analog of Eq. (23)) has the form

$$\frac{i}{3}\sum_{j(i)}\hat{S}_{ij}\psi_j = \epsilon v_i\psi_i, \tag{35}$$

where v_i is the volume of a neighborhood of $\mathfrak{v}^{(3)}_i$.

We will also write the formula that will be used in the subsequent analysis and that can be derived using Eq. (33):

$$\sum_{j(i)} S^a_{ij}e^b_{ij} = 3v_i\delta^{ab}. \tag{36}$$

Indeed, using Eq. (33), we obtain

$$\sum_{j(i)} S^a_{ij}e^b_{ij} = \frac{1}{4}\varepsilon^{abc}\sum_{j,k,l}\varepsilon_{Aijkl}e^b_{Aij}e^c_{Aik}e^d_{Ail}. \tag{36'}$$

The latter sum is equal to $6v_i\varepsilon^{bcd}$.

It should be noted that, if the summation in Eq. (36) were carried out not over all vertices $\{j(i)\}$, but over a certain subset of vertices $\{j'(i)\}$, this sum would not be proportional to $v_i\delta^{ab}$. This can be seen from the expression for the sum on the right-hand side of Eq. (36'), in which the summation over all $j(i)$ were changed to the summation over subset $j'(i)$. However, sum $\sum_{j'}\sum_{k,l}\varepsilon_{Aij'kl}e^b_{ij'}e^c_{ik}e^d_{il}$ is not proportional to quantity ε^{bcd}; it has a more complex structure and substantially depends on the positions of vertices.

Obviously, each vector $s^a_{j(i),j'(i,j),j'(i,j)+1}$ (31) appears in the three umbrellas of vertex a_i from vertices $a_{j(i)}$, $a_{j'(i,j)}$, and $a_{j'(i,j)+1}$: $S^a_{i,j}$, $S^a_{i,j'}$, and $S^a_{i,j'+1}$. Consequently, the following identity, analogous to identity (24), holds:

$$\sum_{j(i)} S^a_{i,j(i)} \equiv 0; \tag{37}$$

in this sum, each vector (31) appears three and only three times.

As in the two-dimensional case, Wilson doubling exists if set of vertices $\{a_{j(i)}\}$ on $\partial\mathfrak{v}^{(3)}_i$ can be split into two intersecting subsets $\{a_{j'(i)}\}$ and $\{a_{j''(i)}\}$ such that

$$\sum_{j'(i)} S^a_{i,j'(i)} \equiv 0, \quad \sum_{j''(i)} S^a_{i,j''(i)} \equiv 0. \tag{38}$$

It was shown above that a similar splitting takes place for a set of adjacent vertices of each even vertex in the case of two-dimensional complexes. Figure 6 shows part of a three-dimensional complex, consisting of neighborhood $\mathfrak{v}^{(3)}_i$ of vertex a_i. It can be seen from the figure that the set of vertices on $\partial\mathfrak{v}^{(3)}_i$ splits into two subsets $(a_1, \ldots, a_n)$ and (a_{n+1}, a_{n+2}), so that

$$\sum_{j=1}^{n} S^a_{i,j} \equiv 0, \quad \sum_{j=n+1}^{n+2} S^a_{i,j} \equiv 0. \tag{39}$$

If, however, n is an even number, the first identity in (39) splits into two more identities:

$$\sum_{j=1}^{n/2} S^a_{i,2j-1} \equiv 0, \quad \sum_{j=1}^{n/2} S^a_{i,2j} \equiv 0. \tag{40}$$

Obviously, each vector (31) on $\partial\mathfrak{v}^{(3)}_i$ appears exactly twice in the first sum in (39), once in the second sum of (39) and once in each sum from (40).

If identity (37) splits into two (as in the case of (39) or three (as in the case of even n in (40)) independent identities, we will say that the set of umbrellas of vertex a_i can be expanded into complete subsets. Complete subsets of umbrellas can be either simple or double. By definition of a complete simple (double) subset of

umbrellas, each vector $s^a_{j,j',j'+1}$ is contained in one (two) and only one (two) umbrella (umbrellas) of this subset. In accordance with this definition, the complete subsets umbrellas $\{S^a_{i,n+1}, S^a_{i,n+2}\}$ and $\{S^a_{i,j}, j = 1, \ldots, n\}$ in the example depicted in Fig. 6 are simple and double, respectively. It should be noted that, if the set of umbrellas of any vertex can be decomposed into two complete subsets, one of these subsets is simple and the other double is in accordance with identity (37).

Let us formulate the criterion for decomposability of the set of umbrellas of an arbitrary vertex a_i. For this purpose, it is convenient to visualize each vertex on $\partial\mathfrak{v}^{(3)}_i$ as a small ball either red or white in color.

Statement 3. The set of umbrellas of vertex a_i is decomposable if and only if all vertices on $\partial\mathfrak{v}^{(3)}_i$ can be painted red and white so that each 2-simplex on $\partial\mathfrak{v}^{(3)}_i$ has one red and two white vertices.

Proof. (1) We assume that the set of umbrellas of vertex a_i is decomposable and consider the complete simple subset of umbrellas $(S^a_{i,j'}, j' = 1, \ldots, \alpha\}$. We paint vertices $a_{j'}, j' = 1, \ldots, \alpha$ red and the remaining vertices $a_{j''}$ on $\partial\mathfrak{v}^{(3)}_i$ white. Then each 2-simplex on $\partial\mathfrak{v}^{(3)}_i$ has one red and two white vertices.

(2) If the coloring of the vertices on $\partial\mathfrak{v}^{(3)}_i$ mentioned in the statement exists, we take set of umbrellas $\{S^a_{i,j'}, j' = 1, \ldots, \alpha\}$, where $a_{j'}, j' = 1, \ldots, \alpha$ is the set of red vertices on $\partial\mathfrak{v}^{(3)}_i$. Obviously, this set of umbrellas is a complete simple subset of umbrellas. □

Remark. Since we have a natural one-to-one correspondence between umbrellas $S^a_{i,j(i)}$ and neighborhoods $\mathfrak{v}^{(2)}_{j(i)}$, as well as between vectors (31) and 2-simplices from $\partial\mathfrak{v}^{(3)}_i$, we can henceforth use, instead of the concepts "umbrella" and "vector (31)," the concepts "neighborhood" and "2-simplex," respectively.

It is expedient to introduce the following terminology. If a subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'}, j' = 1, \ldots, s\}$ exists on $\partial\mathfrak{v}^{(3)}_i$, a 2-simplex from $\partial\mathfrak{v}^{(3)}_i$, which is not contained in any neighborhood of this subset of neighborhoods, will be referred to as white; a 2-simplex contained in one of these neighborhoods will be referred to as yellow; the one contained in two neighborhoods will be called green, and that contained in three neighborhoods, blue. Thus, the coloring of 2-simplices is relative by nature and is determined unambiguously by the given subset of neighborhoods.

Let us describe the inductive process of constructing subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'}\}$ covering boundary $\partial\mathfrak{v}^{(3)}_i$ to the minimal extent. By definition, this means that all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ are colored identically (yellow, green, or blue) relative to this subset of neighborhoods. The process of construction begins with choosing neighborhood $\mathfrak{v}^{(2)}_{j'_1} \in \partial\mathfrak{v}^{(3)}_i$, where $a_{j'_1} \in \partial\mathfrak{v}^{(3)}_i$ is a certain vertex. If subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'}, j' = 1, \ldots, s\}$ (where s is smaller than or equal to the number of vertices on $\partial\mathfrak{v}^{(3)}_i$) has already been constructed, it may happen that all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ have the same color (yellow, green, or blue) relative to this subset. Otherwise, the process of constructing the subset of neighborhoods is continued, the following two conditions being satisfied on each step: (a) newly added neighborhood $\mathfrak{v}^{(2)}_{s+1} \in \partial\mathfrak{v}^{(3)}_i$ borders on at least one neighborhood from the already constructed subset of neighborhoods and (b) the number of colors of all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ relative to subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'}, j' = 1, \ldots, s\}$ is the minimal possible number (for example, the number of colors can be equal to two: white or yellow). The process of constructing the subset of neighborhoods from $\partial\mathfrak{v}^{(3)}_i$ described above is terminated when all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ acquire the same color (yellow, green, or blue); the subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'}\}$ resulting from such a construction will be referred to as a least covering subset.

The following two lemmas obviously hold.

Lemma 1. If all vertices on $\partial\mathfrak{v}^{(3)}_i$ can be painted red and white in the way indicated in Statement 3, and the above-described process of constructing the subset of neighborhoods starts from the neighborhood of any red vertex, it terminates when all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ become yellow.

Lemma 2. If the above-described process of constructing the subset of neighborhood under the conditions of Lemma 1 begins from any white vertex from $\partial\mathfrak{v}^{(3)}_i$, it ends by coloring all 2-simplices from $\partial\mathfrak{v}^{(3)}_i$ either yellow or green.

Indeed, an inductive analysis shows that boundary $\partial\mathfrak{v}^{(2)}_{j'_1}$ of any neighborhood from subset of neighborhoods $\{\mathfrak{v}^{(2)}_{j'_1}, j' = 1, \ldots, s\}$ contains a red vertex. Consequently, the added neighborhood $\mathfrak{v}^{(2)}_{s+1}$ can only be a neighborhood of a white vertex. □

These two lemmas lead to one more lemma.

Lemma 3. If, irrespective of the first and subsequent steps, the above-described procedure of constructing

the subset of neighborhoods terminates only when all 2-simplices from $\partial\mathfrak{v}_i^{(3)}$ are colored blue, the vertices on $\partial\mathfrak{v}_i^{(3)}$ cannot be painted red or white in the way described in Statement 3. □

In the example depicted in Fig. 6, vertices are painted red and white: vertices a_1, …, a_n are white, while vertices a_{n+1} and a_{n+2} are red. If number n is even, two more red and white colorings of vertices exist: vertices a_1, a_3, …, a_{n-1} are red, while the remaining vertices are white, or vice versa.

Let $\mathfrak{K}$ be a three-dimensional simplicial complex realized in a three-dimensional Euclidean space and let the set of tetrahedra of complex $\mathfrak{K}$ fill a compact region in a Euclidean space with the topology of a three-dimensional sphere. Suppose that neighborhood $\mathfrak{v}_i^{(3)}$ is a boundary neighborhood; i.e., boundaries $\partial\mathfrak{K}$ and $\partial\mathfrak{v}_i^{(3)}$ have at least one common 2-simplex, which will be denoted by $a_{j_1(i)}a_{j_2(i)}a_{j_3(i)}$. Suppose also that the above-mentioned red and white coloring of vertices exists on boundary $\partial\mathfrak{v}_i^{(3)}$. We assume that vertices $a_{j_1(i)}$ and $a_{j_2(i)}$ are white, while vertex $a_{j_3(i)}$ is red. We complete complex $\mathfrak{K}$ by supplementing it with one more boundary vertex $a_{k(i)}$ and three boundary edges $a_{j_1(i)}a_{k(i)}$, $a_{j_2(i)}a_{k(i)}$, and $a_{j_3(i)}a_{k(i)}$. The new complex contains, instead of tetrahedron $a_ia_{j_1(i)}a_{j_2(i)}a_{j_3(i)}$, three tetrahedra $a_ia_{j_1(i)}a_{j_2(i)}a_{k(i)}$, $a_ia_{j_2(i)}a_{j_3(i)}a_{k(i)}$, and $a_ia_{j_3(i)}a_{j_1(i)}a_{k(i)}$. It can easily be verified that the newly constructed complex has no red or white vertices on boundary $\partial\mathfrak{v}_i'^{(3)}$ (objects on the completed complex are primed).

Indeed, we start the above process of inductive construction of the subset of neighborhoods $\{\mathfrak{v}_{j'}'^{(2)}\}$ from vertex $a_{j_3(i)}$. Obviously, after n steps (n is the number of red vertices on $\partial\mathfrak{v}_i^{(3)}$), 2-simplex $a_{j_1(i)}a_{j_2(i)}a_{k(i)}$ on $\partial\mathfrak{v}_i'^{(3)}$ remains white, while all remaining 2-simplices become yellow. The next, $(n+1)$th, step must consist in adding one of three neighborhoods $\mathfrak{v}_{j_1}^{(2)}$, $\mathfrak{v}_{j_2}^{(2)}$, or $\mathfrak{v}_k^{(2)}$ to the constructed subset of neighborhoods. In any case, all 2-simplices on $\partial\mathfrak{v}_i'^{(3)}$ will ultimately be yellow (as 2-simplex $a_{j_1(i)}a_{j_2(i)}a_{k(i)}$) or green. For definiteness, we assume that neighborhood $\mathfrak{v}_{j_1}^{(2)}$ is added. As a result, all triangles from $\mathfrak{v}_{j_1}^{(2)}$, except for triangle $a_{j_1(i)}a_{j_2(i)}a_{k(i)}$, become green. For this reason, the inductive procedure must be continued until, after a certain step, triangle $a_{j_1(i)}a_{j_2(i)}a_{k(i)}$ becomes green. However, this is possible only by adding neighborhood $\mathfrak{v}_{j_2}^{(2)}$ or neighborhood $\mathfrak{v}_k^{(2)}$. In any case, blue triangles will appear in neighborhood $\partial\mathfrak{v}_i'^{(3)}$. For example, if we adjoin neighborhood $\mathfrak{v}_k^{(2)}$, triangle $a_{j_1}a_ka_{j_3}$ becomes blue. It follows hence that the inductive process of constructing the subset of neighborhoods $\{\mathfrak{v}_{j'}'^{(2)}\}$ on $\partial\mathfrak{v}_i'^{(3)}$ terminates when all 2-simplices on $\partial\mathfrak{v}_i'^{(3)}$ become blue in color. Consequently, it follows from Lemma 3 that neighborhood $\partial\mathfrak{v}_i'^{(3)}$ does not contain red or white vertices.

Let us now describe the inductive procedure of constructing a three-dimensional complex with an arbitrary number of vertices, whose inner vertices possess the property that a set of umbrellas for each of these vertices cannot be decomposed into complete subsets. This process begins with a complex containing only one internal vertex a_1. If neighborhood $\partial\mathfrak{v}_1^{(3)}$ contains no vertices with red or white colors, the process of constructing continues by adding one more internal vertex. If red and white coloring exists on $\partial\mathfrak{v}_1^{(3)}$, neighborhood $\mathfrak{v}_1^{(3)}$ is reconstructed in the way described above.

Suppose that a three-dimensional complex $\mathfrak{K}_{M-1}$ with $M-1$ internal vertices and the required properties has already been constructed in a three-dimensional space. We complete this complex to a complex with M internal vertices and the required properties, choose an arbitrary vertex on its boundary, and denote it by a_M. Let a_{k_1}, …, a_{k_s} be a set of boundary vertices of the complex which are nearest to vertex a_M, so that 1-simplices $a_{k_1}a_{k_2}$, $a_{k_2}a_{k_3}$, …, $a_{k_{s-1}}a_{k_s}$, $a_{k_s}a_{k_1}$ belong to the boundary of the complex and form a closed broken line l. In a Euclidean space, we construct a two-dimensional simplicial complex S_M with boundary l such that all its internal points do not belong to complex $\mathfrak{K}_{M-1}$ and we denote by $a_{k_1'}$, …, $a_{k_r'}$ the set of internal vertices of complex S_M. We bring complex $\mathfrak{K}_{M-1}$ to complex $\mathfrak{K}_M$ by adding to $\mathfrak{K}_{M-1}$ all simplices of complex S_M, as well as all 1-simplices $a_Ma_{k_1'}$, …, $a_Ma_{k_r'}$, all 2-simplices bounded by the old and new 1-simplices, and all 3-simplices bounded by the old and new 2-simplices. After this, vertex a_M becomes an internal vertex of complex $\mathfrak{K}_M$. In accordance with the above arguments, complex S_M can be chosen so that the set of umbrellas of vertex a_M cannot be decomposed into complete subsets.

In four dimensions, the formulas

$$I_\psi = \frac{1}{5\times 6\times 24} \times \sum_A \sum_{i,j,k,l,m} \varepsilon_{Aijklm}\varepsilon^{abcd}\Theta^a_{Aij}e^b_{Aik}e^c_{Ail}e^d_{Aim}, \tag{41}$$

$$S^a_{ij} = (3!)^{-2}\sum_{A(i,j)k,j,m}\sum \varepsilon^{acdf}\varepsilon_{A(i,j)ijklm}e^c_{A(i,j)ik} \times e^d_{A(i,j)il}e^f_{A(i,j)im}, \tag{42}$$

$$\frac{i}{4}\sum_{j(i)}\hat{S}_{ij}\psi_j = \epsilon v_i\psi_i, \tag{43}$$

$$\sum_{j(i)} S^a_{ij}e^b_{ij} = 4v_i\delta^{ab}, \tag{44}$$

are, respectively, analogs of (30), (33), (35), and (36). The notation in these formulas corresponds to the notation introduced for two and three dimensions. In particular, v_i is the 4-volume of a neighborhood of internal vertex a_i. If $\{j'(i)\}$ is a certain subset of indices of set $\{j(i)\}$, the remark following formula (36') concerning partial sum $\sum_{j'(i)} S^a_{ij'}e^b_{ij'}$ remains valid. In four dimensions, identity (37) also holds and the inductive procedure of constructing such 4-complexes, in which identity (37) at each internal vertex cannot be split into two independent identities (as in relations (38)), also exists. We do not prove these statements here since this can easily be done by a trivial generalization of the same statements formulated in two and three dimensions.

4. ABSENCE OF WILSON DOUBLING

The presence of Wilson fermion doubling on a certain lattice is interpreted here as the situation when Eqs. (35) or (43) have qualitatively different solutions for $\epsilon \longrightarrow 0$. In order to distinguish between these solutions, we introduce the subscripts in parentheses: $\psi_{(1)\epsilon}$, $\psi_{(2)\epsilon}$, … (the remaining indices of a mode are indicated when required). The following orthogonality conditions are observed:

$$\sum_i v_i\bar{\psi}_{(\alpha_1)\epsilon_1 i}\psi_{(\alpha_2)\epsilon_2 i} = 0, \quad \alpha_1 \neq \alpha_2. \tag{45}$$

It can easily be seen that, among infrared or low-energy modes ($\epsilon \longrightarrow 0$), modes with a continuum limit always exist. We denote these modes by $\psi_{(1)}$ or just by ψ and will call these modes trivial. Indeed, suppose that the values of field ψ_i at neighboring vertices almost coincide. Then, for adjacent vertices a_i and a_j, we have

$$\psi_j = \psi_i + e^a_{ij}\partial_a\psi_i + \ldots, \tag{46}$$

where dots indicate the contribution from higher order derivatives of field ψ_i. We assume that x^a_i are the Cartesian coordinates of vertex a_i and $e^a_{ij} = x^a_j - x^a_i$. Substituting relation (46) into Eq. (35) or (43) and using formulas (36) or (44), we can rewrite Eqs. (35) and (43) in the continuum form:

$$i\gamma^a\partial_a\psi(x) + \kappa^{ab}(x)\partial_a\partial_b\psi(x) + \ldots = \epsilon\psi(x). \tag{47}$$

It is extremely important that the term with the lowest order derivative on the left-hand side of Eq. (47) is universal by nature and does not depend on the detailed structure of a simplicial complex. This term coincides with a continuum Dirac field on which the continuum Dirac operator is acting. All remaining terms on the left-hand side of Eq. (47) contain, first, higher order derivatives and, second, nonuniversal coefficients that depend explicitly on the lattice structure. In particular, $\kappa^{ab}(x)$ in Eq. (47) denotes 4×4 matrix functions (or 2×2 functions in a two-dimensional space), which change substantially upon a change in argument x by $\Delta x \sim a$, where a is the lattice scale. Since variables $\{x^a_i\}$ are dynamic variables in the quantum theory of gravitation formulated in the Introduction, over which integration is performed, functions $\kappa^{ab}(x)$ should be treated as random quantities. This means that, for example, the density matrix for propagation of Dirac particles must be averaged over field $\kappa^{ab}(x)$. In this case, correlator $\langle\kappa^{ab}(x)\kappa^{cd}(x')\rangle$ behaves analogously to correlator (A.2) (see Appendix).

It will be shown in Appendix for a nonrelativistic particle that, if the Hamiltonian is the sum of the free Hamiltonian and a perturbation proportional to a higher power of the momentum and containing a random factor, then the averaged density matrix in the long-wave limit coincides with the density matrix for a free particle. The same conclusion is valid in the relativistic case also. This leads to the statement that, in the theory of gravitation considered here, trivial smooth fermion modes exist in the long-wave limit.

It is natural to assume that nontrivial fermion modes $\psi_{(\alpha)i}$ in the long-wave limit split into smooth branches $\psi'_{(\alpha)}(x)$, $\psi''_{(\alpha)}(x)$, … . Precisely such nontrivial fermion modes were studied in connection with the problem of Wilson doubling. Each branch can be obtained in the continuum limit from the values of field $\psi_{i'}$, $\psi_{i''}$, … on subsets of vertices $\{a_{i'}\}$, $\{a_{i''}\}$, …, respectively. Splitting of nontrivial modes into smooth branches in the long-wave limit distinguishes these modes from trivial fermion modes.

Let us prove that nontrivial fermion modes do not exist on "odd" lattices (by odd lattices, we mean the lattices on which identities (37) do not split into individual identities of type (38)).

We consider a nontrivial mode $\psi_{(\alpha)}$ in the limit $\varepsilon \longrightarrow 0$. Let the set of vertices $a_{j(i)}$ split into subsets of vertices $a_{j'(i)}$, $a_{j''(i)}$, …, on which the values of branches of mode $\psi'_{j'(i)}$, $\psi''_{j''(i)}$, … are defined. In accordance with the above assumption, the following expansions hold for the branches of a nontrivial mode:

$$\psi^{\xi}_{j^{\xi}(i)} = \psi^{\xi}(x_i) + e^a_{ij^{\xi}}\partial_a\psi^{\xi}(x_i) + \ldots, \quad \xi = ', '', \ldots. \tag{48}$$

It is natural to assume that

$$\left|\psi^{\xi_1}(x) - \psi^{\xi_2}(x)\right| \sim 1, \quad \xi_1 \neq \xi_2. \tag{49}$$

Since in the long-wave limit we have

$$\partial_a\psi^{\xi} \longrightarrow 0, \quad \xi = ', '', \ldots, \tag{50}$$

quantities $\psi^{\xi}(x_i)$ can be regarded as constant in compact regions of the complex, including a large number of vertices. Let us consider one of such regions and denote it by $\mathfrak{V}$. In accordance with the above arguments, we can assume that

$$\psi'(x_i)\big|_{\mathfrak{V}} = c'_{\mathfrak{V}}, \quad \psi''(x_i)\big|_{\mathfrak{V}} = c''_{\mathfrak{V}}, \ldots, \tag{51}$$

where $c'_{\mathfrak{V}}$, $c''_{\mathfrak{V}}$, … are certain numerical constants. In the long-wave limit, in the main approximation (in the ratio of the lattice scale to the wavelength), we can disregard all terms except the first in expansion (48). In the same approximation, we must neglect the right-hand sides of Eqs. (35) and (43). Thus, Eqs. (35) and (43) are reduced to the following system of equations:

$$\sum_{j'(i)} \hat{S}_{ij'}c'_{\mathfrak{V}} + \sum_{j''(i)} \hat{S}_{ij''}c''_{\mathfrak{V}} + \ldots, \quad a_i \in \mathfrak{V}. \tag{52}$$

Since only identities (37) are observed on the complex, in the case when $c'_{\mathfrak{V}} \neq c''_{\mathfrak{V}} \neq \ldots$, system of equalities (52) imposes constraints on independent variables e^a_{ij}, the number of the constraints being equal to the number of vertices in all regions $\mathfrak{V}$ minus the number of different constants $c'_{\mathfrak{V}}$, $c''_{\mathfrak{V}}$, … . (It should be noted that the latter number is on the order of unity, while the number of vertices in subcomplex $\mathfrak{V}$ may be indefinitely large.) Consequently, relation (49) is inadmissible. If, however, we assume that $c'_{\mathfrak{V}} = c''_{\mathfrak{V}} = \ldots$, Eq. (43) in the long-wave limit assumes the form

$$\frac{i}{4}\gamma^a\left[\sum_{j'(i)} S^a_{ij'}e^b_{ij'}\partial_b\psi'(x_i) + \sum_{j''(i)} S^a_{ij''}e^b_{ij''}\partial_b\psi''(x_i) + \ldots\right] = \epsilon v_i\psi(x_i). \tag{53}$$

Since quantities $\sum_{j'(i)} S^a_{ij'}e^b_{ij'}$, $\sum_{j''(i)} S^a_{ij''}e^b_{ij''}$, … have a complex structure that substantially depends on microscopic details of the amorphous lattice (see the remark following formula (36)), Eq. (53) differs qualitatively from Eq. (47) for trivial modes in the long-wave limit: in the case of nontrivial modes, the differential Dirac operator has no continuum limit even in the main approximation and is determined to a considerable extent by microscopic details of the lattice. Obviously, smooth branches of nontrivial modes cannot exist in this case. For this reason, the only remaining possibility is $\partial_a\psi' = \partial\psi'' = \ldots$. However, this means that only trivial long-wave modes exist in the theory of gravitation under investigation.

Let us now suppose that, in the theory of gravitation with action (4), in evaluating the statistical sum

$$Z = \sum e^{-I}, \tag{54}$$

summation must also be carried out over the types of simplicial complexes. In this case, the mean numbers of "even" and "odd" vertices of complexes are commensurate. Obviously, the above arguments concerning the absence of Wilson fermion doubling remain in force in such a theory.

5. ANOMALY-FREE QUANTIZATION OF GRAVITY

Let us introduce in the theory a gauge field: we juxtapose each edge a_ia_j to a real-valued quantity

$$\mathcal{A}_{ij} = -\mathcal{A}_{ji}, \tag{55}$$

appearing in the fermion part of the action in the following manner (cf. relation (4)):

$$\Theta^a_{Aij} = \frac{i}{2}(\bar{\psi}_{Ai}\gamma^a\Omega_{Aij}e^{-ie\mathcal{A}_{Aij}}\psi_{Aj} - \bar{\psi}_{Aj}\Omega_{Aji}\gamma^a e^{-ie\mathcal{A}_{Aji}}\psi_{Ai}). \tag{56}$$

As a result, the action is found to be invariant to local gauge transformations

$$\psi_i \longrightarrow e^{ie\alpha_i}\psi_i, \quad \bar{\psi}_i \longrightarrow e^{-ie\alpha_i}\bar{\psi}_i, \quad \mathcal{A}_{ij} \longrightarrow \mathcal{A}_{ij} + \alpha_j - \alpha_i. \tag{57}$$

Here, $\{a_i\}$ are arbitrary real numbers.

We introduce the Weyl field

$$\psi_\pm = \left(\frac{1 \pm \gamma^5}{2}\right)\psi, \quad \bar{\psi}_\pm = \bar{\psi}\left(\frac{1 \mp \gamma^5}{2}\right). \tag{58}$$

Fermion part (56) of action (4) obviously splits into the sum of actions of the "right" and "left" Weyl fields,

$$I_\psi = I_{\psi+} + I_{\psi-}, \tag{59}$$

so that action $I_{\psi+}$ (of $I_{\psi-}$) can be obtained from fermion part (56) of action (4) by inserting projector $(1 + \gamma^5)/2$ (or $(1 - \gamma^5)/2$ immediately to the left of field ψ.

We define the fermion measure as

$$D\bar{\psi}D\psi = \prod_i d\bar{\psi}_i d\psi_i, \tag{60}$$

where $d\psi_i$ is the product of the differentials of all independent components of spinor ψ_i and $d\bar{\psi}_i$ is the product of the differentials of all components of conjugate spinor $\bar{\psi}_i$. Obviously, we have

$$d\psi_i = d\psi_{i+}d\psi_{i-}, \quad d\bar{\psi}_i = d\bar{\psi}_{i+}d\bar{\psi}_{i-}, \tag{61}$$

where $d\psi_{i+}(d\psi_{i-})$ is the product of the differentials of all independent components of spinor $\psi_{i+}(\psi_{i-})$ (the same holds for $d\bar{\psi}_{i+}(d\bar{\psi}_{i-})$). By virtue of relations (61), functional measure (60) can be factorized:

$$\begin{aligned} D\bar{\psi}D\psi &= (D\bar{\psi}_+D\psi_+)(D\bar{\psi}_-D\psi_-), \\ D\bar{\psi}_+D\psi_+ &= \prod_i d\bar{\psi}_{i+}d\psi_{i+}, \\ D\bar{\psi}_-D\psi_- &= \prod_i d\bar{\psi}_{i-}d\psi_{i-}. \end{aligned} \tag{62}$$

Measures $(D\bar{\psi}_+D\psi_+)$ and $(D\bar{\psi}_-D\psi_-)$, as well as actions $I_{\psi+}$ and $I_{\psi-}$, are separately invariant to gauge transformations (57).

We introduce the following notation:

$$Z_\pm\{\Omega, \mathcal{A}\} \equiv \int D\bar{\psi}_\pm D\psi_\pm \exp(-I_{\psi\pm}). \tag{63}$$

Here, either upper or lower signs are taken. The total statistical sum or chiral statistical sums as functions of an arbitrary electromagnetic field can be written in the form

$$Z\{\mathcal{A}\} = \sum_{\{\Omega\}} e^{-I_\Omega} Z_+\{\Omega, \mathcal{A}\}Z_-\{\Omega, \mathcal{A}\}, \tag{64}$$

$$Z_\pm\{\mathcal{A}\} = \sum_{\{\Omega\}} e^{-I_\Omega} Z_\pm\{\Omega, \mathcal{A}\}, \tag{65}$$

where I_Ω is a part of action (4) independent of fermion fields.

In accordance with the conclusion drawn in the previous section, chiral theories (65) have a continuous low-energy limit. Since Wilson doubling is absent, each of these theories contains one right or left Weyl field in the continuum limit.

Let us prove that functional $Z_\pm\{\mathcal{A}\}$ is gauge invariant:

$$Z_\pm\{\mathcal{A} + \partial_\mu\alpha\} = Z_\pm\{\mathcal{A}\}. \tag{66}$$

This equality can be proved easily and exactly on a lattice even prior to summation over gravitational degrees of freedom for functional (63). Indeed, neither the actions nor the measures in relation (63) change upon substitution of variables (57); consequently, equalities (66) hold both on a lattice and in the continuum limit.

Let us now carry out in integrals (63) the substitution of only fermion variables of type (57) with an infinitesimal parameter

$$\alpha_i = \varepsilon\delta_{ij}, \quad \varepsilon \longrightarrow 0. \tag{67}$$

The electromagnetic field does not change in this case. We now take into account the fact that the substitution in the integration variables does not change the integrals or the measures in these integrals. This substitution of variables changes only the actions in the integrals. In the first order in ε, these changes in the continuum limit are given by

$$\varepsilon(x_j)\partial_\mu J^\mu_\pm(x_j), \quad J^\mu_\pm(x) = \frac{1}{2}\bar{\psi}\gamma^\mu(1 \pm \gamma^5)\psi(x). \tag{68}$$

Consequently, from the equality of integrals (63) before and after such a substitution of variables, we obtain the following conservation laws:

$$\partial_\mu J^\mu_\pm\{\mathcal{A}\}(x) \equiv \langle \partial_\mu J^\mu_\pm(x)\rangle_{F,\Omega} = 0. \tag{69}$$

Here, averaging is carried out only over fermion and gravitational degrees of freedom.

On the other hand, variation of functionals (65) relative to gauge vector field $\mathcal{A}$ has the form

$$\delta \ln Z_\pm = e\int dx \delta\mathcal{A}_\mu(x) J^\mu_\pm\{\mathcal{A}\}(x). \tag{70}$$

If $\delta\mathcal{A}_\mu = \partial_\mu\delta\alpha$, we again arrive at equalities (66) in accordance with relations (69) and (70).

Thus, in the discrete quantum gravity considered here, Wilson doubling of fermions and, hence, anomalous divergence of chiral currents are absent. Anomalous divergence of the axial current obviously does not exist either.

6. CONTINUUM LIMIT

In [1], arguments were given in favor of the fact that infrared divergence appearing in integration of tetrads e^a_{ij} with respect to field in the lattice gravitational statistical sum lead to degeneracy of discrete quantum gravity to a continuous quantum theory of gravitation.

The continuous quantum theory of gravitation obtained in this way must envisage, among other things, the possibility of anomaly-free inclusion of a chiral fermion field. In this connection, the problem of explicit description of such a continuum theory arises.

In this section, we give arguments in favor of the fact that the continuum theory of gravitation proposed by the author in a number of previous articles [2, 12, 13] and constructed on the basis of the dynamic quantization method is the continuum limit of the discrete quantum gravity we are interested in.

The ideology of the dynamic quantization method is described in detail in [2, 12]; an exactly solvable example (two-dimensional quantum gravity) demonstrating this method is given in [13]. For this reason, we will only describe here some required basic properties of the general covariant theory quantized with the help of the dynamic quantization method.

In the case of dynamic quantization, the theory is constructed in a space with a pseudo-Euclidean signature of the metric, and the general covariant theory is assumed to be regularized in the ultraviolet spectral region so that the following axioms hold.

Axiom 1. All physical states of the theory can be obtained from the ground state $|0\rangle$ with the help of operators $A_N^\dagger$ with $|N| < N_0$:

$$\begin{aligned} |N_1, \ldots, N_s\rangle &= A^\dagger_{N_1}, \ldots, A^\dagger_{N_s}|0\rangle, \\ A_N|0\rangle &= 0, \quad [A_{N_1}, A^\dagger_{N_2}] = \delta_{N_1N_2}. \end{aligned} \tag{71}$$

States (71) form an orthogonal basis of space F' of the physical states in the theory.

Here, N is a point of a certain countable set with a norm, so that condition $|N| < N_0$ singles out a finite subset of this set. As $N_0 \longrightarrow \infty$, the number of elements of this subset tends to infinity as a certain positive power of number N_0. Only nonzero commutators and anticommutators are written everywhere.

Axiom 2. Fundamental dynamic variables (fields) $\Phi(x)$ transform state (71) into a superposition of states of the same form, containing all states in which one of the occupation numbers modulo differs by unity from the occupation numbers of state (71), while the remaining occupation numbers coincide with those of state (71).

For the Dirac field we are interested in, Axiom 2 indicates that the following expansion holds:

$$\begin{aligned} \psi(x) &= \sum_{|N|<N_0} (a_N\psi_N^{(+)}(x) + b_N^\dagger\psi_N^{(-)}(x)) + \ldots, \\ \{a_{N_1}, a^\dagger_{N_2}\} &= \{b_{N_1}, b^\dagger_{N_2}\} = \delta_{N_1N_2}. \end{aligned} \tag{72}$$

Here, instead of generalized creation and annihilation operators $\{A_N, A_N^\dagger\}$, Fermi creation and annihilation operators $\{a_N, b_N, a_N^\dagger, b_N^\dagger\}$ are used. The set of modes $\{\psi_N^{(+)}(x), \psi_N^{(-)}(x)\}$ forms a complete set in which any spinor field $\psi(t, \mathbf{x})$ can be expanded at any instant t. We denote by $\mathbf{x}$ the set of spatial coordinates. The set of modes $\{\psi_N^{(+)}(x)\}$ is a positive-frequency set, while the set of modes $\{\psi_N^{(-)}(x)\}$ is a negative-frequency set. Such a division of modes takes place at a certain instant $t = t_0$, positive- and negative-frequency modes corresponding to the massless one-particle Dirac Hamiltonian at the same instant. At any instant, set of modes $\{\psi_N^{(+)}(x), \psi_N^{(-)}(x)\}$ can be determined by solving the massless Dirac equation. Dots in Eq. (72) indicate the nonlinear contribution relative to operators $\{A_N, A_N^\dagger\}$, among which both Fermi and Bose operators exist.

Axiom 3. Equations of motion and constraints for physical fields coincide, to within the transposition of operators, with corresponding classical equations and constraints.

Axiom 3 indicates that operators $\{A_N, A_N^\dagger\}$ commute (at least, in a weak sense) with all first-order constraints or with the Hamiltonian of the theory. Indeed, it is only in this case that the regularized equations of motion and constraints do not change (modulo the constraints existing in the theory) in the course of regularization (i.e., vanishing of pairs of operators $(A_N, A_N^\dagger)$ with $|N| > N_0$) and regularized first-order constraints preserve their order.

Suppose that gravitational fields (tetrad and connectivity) contain classical parts that are preserved after the formal exclusion of all operators $\{A_N, A_N^\dagger\}$. This assumption is necessary in respect to gravitational fields. In the problem under investigation, it is convenient to assume that the electromagnetic field is arbitrary, but not expandable in the creation and annihilation operators. Quantum fluctuations of electromagnetic field can then be taken into account by pairing electromagnetic field in accordance with Wick's theorem and by replacing pair correlators with the corresponding propagator. However, such computations are of no interest for our analysis.

It should be borne in mind that the expansion of fields and equations in nonlinearities relative to operators $\{A_N, A_N^\dagger\}$ is in fact an expansion in the coupling constants of the theory. Consequently, the exact massless Dirac equation

$$ie_a^\mu\gamma^a D_\mu\psi = 0 \tag{73}$$

can be reduced, using Eq. (72) in the lowest approxima-

tion, to the equation

$$ie_a^{(0)\mu}\gamma^a D_\mu^{(0)} \times\left[\sum_{|N|<N_0}(a_N\psi_N^{(+)}(x)+b_N^\dagger\psi_N^{(-)}(x))\right]=0. \quad (74)$$

In Eqs. (73) and (74), $e_a^\mu(x)$ is a tetrad and D_μ is a covariant derivative including the electromagnetic field. Subscript (0) indicates the fields and operators that can be obtained by formal exclusion of all operators $\{A_N, A_N^\dagger\}$.

It follows from Eq. (74) that

$$e_a^{(0)\mu}\partial_\mu(\bar\psi^{(1)}\gamma^a\psi^{(1)})=0, \quad (75)$$

$$e_a^{(0)\mu}\partial_\mu(\bar\psi^{(1)}\gamma^a\gamma^5\psi^{(1)})=0, \quad (76)$$

where

$$\psi^{(1)}(x)=\sum_{|N|<N_0}(a_N\psi_N^{(+)}(x)+b_N^\dagger\psi_N^{(-)}(x)). \quad (77)$$

It is extremely important that fermion field (77) is regularized. It is for this reason that Eq. (74) leads to equalities (75) and (76). Indeed, the bilinear forms constructed from Dirac (or Weyl) fields do not require regularization; differential operators can be applied to such field directly, without a preliminary regularization.

Equalities (75) and (76) were obtained in the one-loop approximation. It is well known that the inclusion of higher order loops makes zero contribution to an anomalous divergence of vector, axial-vector, and chiral currents. In our case, this means that the expansion into a power series in operators $\{A_N, A_N^\dagger\}$ does not change results (75) and (76).

Obviously, the conservation of the chiral fermion current in the chiral theory can be obtained in a similar way.

It should be stated that the method of dynamic quantization differs in principle from the Feynman quantization. This can be seen even from the obtained result (75), (76). It also follows from the fact that, in the dynamic quantization method, vacuum for $t \longrightarrow -\infty$ generally differs qualitatively from vacuum for $t \longrightarrow +\infty$. Indeed, the operator equations in the theory of gravitation with the dynamic quantization method in the lowest approximation coincide with the Einstein equations, while quantum corrections are taken into account against the background of classical solutions. However, even on the classical level, the evolution equations in the theory of gravitation lead from one singularity to another, and these singularities may be quantitatively different. On the contrary, in the Feynman quantization, it is assumed that the ground state evolves over an infinitely long time into a state differing from the initial state only in a phase factor. This hypothesis leads to the rule of circumvention of the poles in the propagators used in the Feynman diagram technique and, hence to the possibility of Wick rotation towards a Euclidean space. Consequently, the continuum quantum field theory formulated in a Euclidean space is automatically quantized in the Feynman sense. In the dynamic quantization method, Wick rotation to a Euclidean space is ruled out. This can be seen if only from evolution equation (74) which makes sense only for a pseudo-Euclidean signature.

Thus, we arrive at the following conclusion.

The continuum theory of gravitation obtained with the help of the dynamic quantization method is a continuum limit of the discrete quantum gravity formulated in [1]. This statement can be substantiated as follows.

1. Both theories contain a finite number of physical degrees of freedom.

2. In both theories, it is possible to introduce only one chiral Weyl field, the corresponding chiral current being conserved.

ACKNOWLEDGEMENTS

I am grateful to A. Ioselevich, M. Skvortsov, and I. Kolokolov for the clarification of some concepts concerning the problem of localization of quantum particles in the presence of random factors.

This study was financed by the Federal Program of Support for Leading Scientific Schools (project no. 2044.2003.2)

APPENDIX

Let us consider, in a three-dimensional space, the one-particle Hamiltonian

$$\mathcal{H}=\frac{\mathbf{p}^2}{2m}+\mathbf{p}^2\kappa(\mathbf{x})\mathbf{p}^2 \quad (A.1)$$

with a random function $\kappa(\mathbf{x})$. Suppose that the density matrix for system (A.1) must be averaged over function $\kappa(\mathbf{x})$ in accordance with the rule

$$\langle\kappa(\mathbf{x})\rangle=0, \quad \langle\kappa(\mathbf{x})\kappa(\mathbf{x}')\rangle=\frac{a^7}{m^2}\delta^{(3)}(\mathbf{x}-\mathbf{x}'), \quad (A.2)$$

while the mean values of higher powers of $\kappa(\mathbf{x})$ are calculated in accordance with the Wick theorem. We state that the density matrix of system (A.1) in the limit $p^2 \longrightarrow 0$ tends to the density matrix of a free particle with Hamiltonian $\mathcal{H}_0=\mathbf{p}^2/2m$; consequently, the localization effect is absent in the present case.

Indeed, let

$$K_0(\mathbf{k}', \mathbf{k}, t) = \exp\left(-i\frac{\mathbf{k}^2}{2m}t\right)(2\pi)^3\delta^{(3)}(\mathbf{k}' - \mathbf{k})\theta(t) \quad \text{(A.3)}$$

be the amplitude of transition for a free particle with Hamiltonian $\mathcal{H}_0$) in the momentum space. The exact amplitude of the transition satisfies the equation

$$\left(\frac{\partial}{\partial t} + i\mathcal{H}_0' - iV'\right)K(\mathbf{k}', \mathbf{k}, t) = \delta(t)(2\pi)^3\delta^{(3)}(\mathbf{k}' - \mathbf{k}), \quad \text{(A.4)}$$

$$V = -\mathrm{p}^2\kappa\mathbf{p}^2.$$

We expand K in operator V,

$$\begin{aligned}&\langle \mathbf{k}', t|K|\mathbf{k}, 0\rangle \\ &= \langle \mathbf{k}', t|\{K_0 + K_0(iV)K_0 + \ldots\}|\mathbf{k}, 0\rangle \\ &= K_0(\mathbf{k}', \mathbf{k}, t) + \mathbf{k}'^2\mathbf{k}^2 Q\{\mathbf{k}', \mathbf{k}, \kappa, t\};\end{aligned} \quad \text{(A.5)}$$

for $\mathbf{k}'^2\mathbf{k}^2 \longrightarrow 0$, operator Q in Eq. (A.5) has a finite limit. The density matrix can be expressed in terms of the transition amplitude:

$$\rho(\mathbf{k}_1', \mathbf{k}_2'; \mathbf{k}_1, \mathbf{k}_2; t) = K(\mathbf{k}_1', \mathbf{k}_1, t)K^*(\mathbf{k}_2', \mathbf{k}_2, t). \quad \text{(A.6)}$$

Density matrix (A.6) must be averaged in accordance with rules (A.2). Taking into account relation (A.5), we obtain

$$\begin{aligned}\langle \rho(\mathbf{k}_1', \mathbf{k}_2'; \mathbf{k}_1, \mathbf{k}_2; t)\rangle &= \exp\left(-i\frac{\mathbf{k}_1^2 - \mathbf{k}_2^2}{2m}t\right) \\ &\times (2\pi)^6\delta^{(3)}(\mathbf{k}_1' - \mathbf{k}_1)\delta^{(3)}(\mathbf{k}_2' - \mathbf{k}_2) \\ &+ (\mathbf{k}_1'^2)(\mathbf{k}_2'^2)R(\mathbf{k}_1', \mathbf{k}_2'; \mathbf{k}_1, \mathbf{k}_2; t).\end{aligned} \quad \text{(A.7)}$$

In this relation, operator $R(\mathbf{k}_1', \mathbf{k}_2'; \mathbf{k}_1, \mathbf{k}_2; t)$ is regular in the limit $\mathbf{k}_1'^2 \longrightarrow 0$, $\mathbf{k}_2'^2 \longrightarrow 0$. This leads to the statement formulated at the beginning of the Appendix.

REFERENCES

1. S. N. Vergeles, Zh. Éksp. Teor. Fiz. **120**, 1069 (2001) [JETP **93**, 926 (2001)].
2. S. N. Vergeles, Zh. Éksp. Teor. Fiz. **118**, 996 (2000) [JETP **91**, 859 (2000)].
3. L. S. Pontryagin, *Fundamentals of Combinatorial Topology* (Nauka, Moscow, 1976).
4. P. Hilton and S. Wylie, *Homology Theory: An Introduction to Algebraic Topology* (Cambridge Univ. Press, Cambridge, 1960; Mir, Moscow, 1966).
5. K. G. Wilson, Erice Lecture Notes (1975).
6. J. Kogut and L. Susskind, Phys. Rev. D **11**, 395 (1975).
7. L. Susskind, Phys. Rev. D **16**, 3031 (1977).
8. M. Luscher, hep-th/0102028.
9. H. B. Nielsen and M. Ninomiya, Nucl. Phys. B **185**, 20 (1981); Nucl. Phys. B **193**, 173 (1981).
10. J. B. Hartle and S. W. Hawking, Phys. Rev. D **28**, 2960 (1983).
11. S. P. Novikov and I. A. Dynnikov, Usp. Mat. Nauk **52**, 175 (1997).
12. S. N. Vergeles, Teor. Mat. Fiz. **112**, 132 (1997).
13. S. N. Vergeles, Zh. Éksp. Teor. Fiz. **117**, 5 (2000) [JETP **90**, 1 (2000)].

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1093–1101.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1222–1231.

NUCLEI, PARTICLES, AND THEIR INTERACTION

Extrapolation of Triplet Phases of Proton–Proton Scattering to Low Energies

V. V. Pupyshev
Joint Institute for Nuclear Research, Dubna, Moscow oblast, 141980 Russia
e-mail: pupyshev@thsun1.jinr.ru
Received June 2, 2003

Abstract—It is shown that, to correctly extrapolate the triplet phases of *pp* scattering to a range of energies below several megaelectronvolts, one should take into account, together with the Coulomb and nuclear interactions, the interactions of the magnetic moment of a proton with the Coulomb field and the magnetic moment of another proton. A simple method is proposed for such an extrapolation. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

This paper is dedicated to the 70th birthday of professor V.B. Belyeav and represents a continuation of the studies of low-energy expansions for systems of several quantum particles that were carried out by his students [1–8].

The knowledge of the energy dependence of the scattering characteristics (phases δ, amplitudes f, cross sections $d\sigma$, analyzing power A_y, etc.) in the limit of low collision energies ($E \longrightarrow 0$) enables one to solve two important problems: an applied problem of extrapolating these characteristics to low energies that are inaccessible for direct experimental investigation, and the inverse problem, which is aimed at the recovery of an interaction by the experimental data available. Therefore, one of the main problems in scattering theory is to investigate the low-energy behavior of the scattering characteristics and to derive explicit expressions for their low-energy expansions.

It is well known [9, 10] that, in the low-energy limit, the energy dependence of the scattering phases of two elementary or composite nuclear particles is substantially affected by the long-range power-law terms $V^d \sim r^{-d}$, $d \geq 3$ of the total effective interaction

$$V^{\text{eff}}(r) = V^s(r) + V^d(r),$$

where r is the distance between the centers of mass of particles and V^s is a rapidly decreasing ($V^s = o(V^d)$, $r \longrightarrow \infty$) effective interaction induced by nuclear forces. In the system (N, N) of two nucleons, the following interactions are among such terms of electromagnetic origin: the polarization interaction of protons [11, 12],

$$V^d(r) = V^p(r) = \alpha_e r^{-4},$$
$$\alpha_e = (1.07 \pm 0.11)10^{-3}\ \text{fm}^3, \quad d = 4, \tag{1}$$

the interaction of the magnetic moments of nucleons [13],

$$V^d(r) = V^{mt}(r) = b_t r^{-3} S_{12}, \quad d = 3, \tag{2}$$

and the interaction of the magnetic moment of a neutron n or a proton p with the Coulomb field of another proton [13],

$$V^d(r) = V^{mls}(r) = b_{ls} r^{-3}(\mathbf{l}\cdot\mathbf{s}), \quad d = 3. \tag{3}$$

In formulas (1)–(3), we used standard notation from the theory of *NN* interactions [14]: α_e is the electric polarizability of a proton; $\mathbf{l}$ and $\mathbf{s} = \mathbf{s}_1 + \mathbf{s}_2$ are the total angular momentum and the spin of a two-nucleon system, respectively, where $\mathbf{s}_1$ and $\mathbf{s}_2$ are spins of the nucleons; S_{12} is a known tensor operator; and b_{ls} and b_t are constants that are different for the systems (n, n), (n, p), and (p, p).

A comprehensive analysis of the role of the polarization potential (1) in the *pp* scattering and in the reaction $dd \longrightarrow e + \nu_e$ has been carried out in surveys [1, 2]. One of the results of this analysis is as follows: the contribution of the polarization interaction (1) to the elastic *pp* scattering and to the cross section of the reaction $dd \longrightarrow e + \nu_e$ is negligible because the Coulomb interaction V^c between protons is repulsive and the constant α_p of the polarization potential is small.

In the system (n, n), there are no interactions (1) or (3); however, the total interaction $V = V^s + V^{mt}$ contains the long-range tensor term (2). The role of this term in the triplet *nn* scattering ($s = 1$) was first mentioned in [3] and then investigated in [4, 5]. In these works, a neutron–neutron analog of the Ramsauer effect [15, 16] was first theoretically predicted, and it was shown that this analog results from the interference between the *nn* scattering by the V^s and V^{mt} interactions and should manifest itself as a deep minimum in the total cross sec-

tion of the triplet scattering of neutrons at an energy of $E \approx 20$ keV in the system of their centers of mass. As was shown in [3], this phenomenon is of interest for the experimental investigation of the cross sections of the *nn* scattering and the reaction $\pi^- d \longrightarrow \gamma nn$. The experimental verification of another feature of the triplet *nn* scattering, namely, a linear (in the scattering momentum) decrease in the ${}^3P_j$ phases, which, according to the phase analysis [5], is due to the interaction V^{mt}, is also of interest.

The necessity of taking into account the total magnetic interaction $V^m \equiv V^{mls} + V^{mt}$ for the correct theoretical interpolation of the experimental data of the *np* and *pp* scattering has been repeatedly pointed out. Although various approaches to solving this problem and analyzing its state of the art have been discussed in detail in surveys [6] and [17], it is worth mentioning once again the most interesting conclusions made in [6] and [18–20].

In [6], the present author first showed that, when the interaction V^{mls} is theoretically taken into account, the function $d\sigma A_{y,np}$ must decrease as $O(E^{1/2})$ for $E \longrightarrow 0$, whereas, when V^{mls} is not taken into account, this function decreases much faster, namely, as $O(E^{3/2})$. In [18], involving the same interaction in the theoretical analysis of *np* scattering, Hogan and Seyler explained the spikelike behavior of the analyzing power $A_{y,np}(\theta)$ for energies of $E_{\text{lab}} = 25$–210 MeV and angles of $\theta < 5°$.

The analyzing power $A_{y,pp}$ has been the main subject of numerous investigations (see [17]) into the role of the interaction V^{mls} in the elastic *pp* scattering. The common feature of all known methods that take into account this interaction is the application of the Born approximation. For example, in [19], the amplitude f^m (additional to the Coulomb–nuclear amplitude f^{cs}) induced by the interaction V^{mls} was calculated in the Born approximation for a plane wave ($f^m \approx f^m_B$); moreover, it was shown that this method of taking into account the interaction V^{mls} for energies $E_{\text{lab}} > 150$ MeV hardly improves the agreement between the theoretical and experimental values of the analyzing power $A_{y,pp}$. In [20], the Born approximation for a plane wave, distorted by the Coulomb interaction ($f^m \approx f^m_{BC}$), was applied to calculating f^m; it was shown that the absolute values of the amplitudes f^m_B and f^m_{BC} are approximately equal, but their phases differ substantially; therefore, the function $A_{y,pp}(\theta)$ is characterized by a spikelike behavior in the range of small angles θ.

In [17], it was pointed out that, for an energy of $E_{\text{lab}} = 9.75$ MeV, taking into account the interaction V^{mls} improves the agreement between calculated values of the function $A_{y,pp}(\theta)$ and the experimental results obtained in the range of small angles $\theta < 30°$; for an energy of $E_{\text{lab}} = 5.5$ MeV, a similar result was obtained in a broader range of angles, $\theta < 90°$. Hence, one can assume that a further decrease in energy will increase the contribution of the interaction V^{mls} to the observed characteristic $A_{y,pp}(\theta)$ at large angles as well. Since the expression for $A_{y,pp}$ in terms of the phases of *pp* scattering is well known [17, 21], the first stage in the study of this contribution in the low-energy limit consists in analyzing the specific features of the low-energy behavior of the phases of *pp* scattering that are associated with the interactions V^{mls} and V^{mt} and their sum V^m. The study of the features in the behavior of these phases that are induced by the mutual effect of the nuclear and magnetic interactions V^s and V^m is of no less interest. In spite of the fact that the role of the interaction V^m in *pp* scattering has been studied for a long time, the question of the theoretical existence of the above-mentioned features still remains open. The present work has been stimulated by the author's wish to answer this question. The paper is organized as follows. In Section 2, we formulate the model of the *pp* scattering that is used in this paper. In Section 3, we describe methods for the exact and approximate calculation of the phases of the *pp* scattering. The phases obtained by the numerical analysis are presented in Section 4 and summarized in the Conclusions.

2. A MODEL OF PROTON–PROTON SCATTERING

Suppose that a system (p, p) is described by the nonrelativistic Schrödinger equation [9]. In the system of the center of mass of protons, we rewrite this equation as

$$[\Delta_r + k^2 - V^{ca}(\mathbf{r})]\Psi(\mathbf{r}; \mathbf{k}) = 0, \quad k^2 = \frac{m_p}{\hbar^2}E,$$

where Ψ is the wavefunction of the protons; $\mathbf{k}$ and E are their relative momentum and energy, respectively; $\mathbf{r}$ is a vector directed from one proton to another; and m_p is the proton mass.

We assume that the total interaction $V^{ca} = V^c + V^a$ is a superposition in which the interaction V^a decreases as r increases faster than the central Coulomb potential,

$$V^c(r) = \frac{m_p e^2}{\hbar^2 r} = \frac{1}{Rr}, \quad R \equiv \frac{\hbar^2}{m_p e^2}, \tag{4}$$

where e is the electron charge and R is the Bohr radius of the *pp* system. Theoretically, three cases are possible: $a = s, m, ms$. In the first case ($a = s$), $V^a = V^s$ is a short-range nuclear interaction; in the second case ($a = m$), $V^a = V^m$ is a magnetic interaction; and, in the third, the most realistic, case ($a = ms$), $V^a = V^{ms} = V^m + V^s$ is a superposition of the magnetic and nuclear interactions.

Following [13], we assume that V^m is a superposition $V^m = V^{mt} + V^{mls}$ whose components are defined by the formulas

$$V^{mt} \equiv \frac{b_t S_{12}}{r^3}, \quad b_t \equiv -\frac{m_p}{\hbar^2}\mu_p^2\mu_0^2 = -\mu_p^2\frac{m_e}{m_p}r_e,$$
$$S_{12} \equiv \frac{3(\mathbf{s}_1\cdot\mathbf{r})(\mathbf{s}_2\cdot\mathbf{r}) - r^2(\mathbf{s}_1\cdot\mathbf{s}_2)}{4r^2}, \tag{5}$$

and

$$V^{mls} = \frac{b_{ls}(\mathbf{l}\cdot\mathbf{s})}{r^3},$$
$$b_{ls} \equiv -\frac{m}{\hbar^2}8\mu_0^2\left(\mu_p - \frac{1}{4}\right) = -2\left(\mu_p - \frac{1}{4}\right)\frac{m_e}{m_p}r_e. \tag{6}$$

Here, m_e is the electron mass, μ_p is the magnetic moment of a proton expressed in nuclear magnetons μ_0, r_e is the classical radius of electron,

$$\mu_0 \equiv \frac{e\hbar}{2m_p c}, \quad r_e \equiv \frac{e^2}{m_e c^2}.$$

In our calculations, we use the Reed interaction with a soft core [22] as the nuclear interaction V^s and the well-known constants [23]

$$m_p = 938.2796 \text{ MeV}, \quad \mu_p = 2.7927,$$
$$\hbar^2/m_p = 41.4969 \text{ MeV}\cdot\text{fm}^{-2},$$
$$m_e = 0.5110034 \text{ MeV}, \quad r_e = 2.817938 \text{ fm},$$
$$\text{Ry} = 13.605804 \text{ eV};$$

according to (4)–(6), for these constants, we obtain

$$R = 28.8064\ldots\text{fm}, \quad b_{ls} = -0.005371\ldots\text{fm},$$
$$b_t = -0.001534\ldots\text{fm}.$$

It is clear from physical considerations that, at distances that are an order of magnitude less than the nucleon size (≈1 fm), both magnetic interactions should be described by other formulas that are nonsingular for $r \longrightarrow 0$. Since such formulas are not presently available, we can set $V^{mt} \equiv 0$ and $V^{mls} \equiv 0$ for $r \le 1.0$ fm. There is another reason why one should neglect both these interactions in the range of distances $r \le r^s$, where r^s is the effective radius of nuclear interaction. Let us consider this reason and show that the choice of the Reed interaction with a soft core does not restrict the generality of our analysis.

It is well known from quantum mechanics [9] and the method of phase functions [10] that, at large collision energies, the scattering scenario of two particles mainly depends on the structure of their interaction at small distances, while the main features of the scattering at low energies depend on the behavior of the interaction at large distances, i.e., on the behavior of the interaction's "tail." All modern phase-equivalent *NN* interactions have identical, rather rapidly decreasing, Yukawa tails $V^s \sim \exp(-m_\pi r)/r$, where m_π = 134.9630 MeV is the mass of a π meson. This tail determines the behavior of the parameters of the Coulomb–nuclear *pp* scattering at low energies; therefore, these parameters weakly depend on the choice of nuclear interaction. Another physical reason for such a weak dependence is the complete screening of the nuclear interaction at small distances by repulsive Coulomb and centrifugal potentials $1/Rr$ and $l(l+1)/r^2$. Therefore, when considering the triplet phases of the *pp* scattering, one can restrict the analysis to the calculation of these phases for a certain one phase-equivalent nuclear interaction without loss of generality. In the present work, we use the Reed interaction with a soft core as such an interaction. This interaction well describes the available experimental data for energies $E > 10$ MeV and therefore contains information both on the nuclear interaction and on the magnetic interaction, which is effectively taken into account for finite distances. From the physical point of view, the upper boundary r^s of this internal domain is the effective radius [1] of the potential V^s for $E > 10$ MeV. One usually estimates this radius as $r^s \approx 4$ fm [14]. To avoid a repeated inclusion of the magnetic interaction in our calculations in the domain $r \le r^s$, we assume henceforth that $V^{mls} \equiv 0$ and $V^{mt} \equiv 0$ for $r \le 4$ fm.

3. METHOD

Like other realistic nuclear interactions [14], the Reed interaction contains, along with short-range central terms that are independent of $\mathbf{l}$, $\mathbf{s}_1$, and $\mathbf{s}_2$, short-range spin–orbit and tensor interactions:

$$V^{sls} = V^{sls}(\mathbf{r})(\mathbf{l}\cdot\mathbf{s}), \quad V^{st} = V^{st}(\mathbf{r})S_{12}. \tag{7}$$

The first of these interactions preserves the angular momentum l, spin s, total momentum $\mathbf{j} = \mathbf{l} + \mathbf{s}$, and total isospin $T = 1$ of the system (p, p), while the second term preserves s and j but, in general, does not preserve $l = j, j \pm 1$. Therefore, in the general case, the triplet *pp* state $|sj\rangle$ with certain total momentum j and spin $s = 1$ represents a superposition of the basis *pp* states $|slj\rangle$ with $l = j \pm 1$:

$$|sj\rangle = a|s, j-1, j\rangle + b|s, j+1, j\rangle, \quad a^2 + b^2 = 1. \tag{8}$$

In the case under consideration ($s = 1$ and $T = 1$), there is no mixing in the state ${}^3P_j$ with $j = 0, 1$ and in the states with $j = l > 1$. The states $|slj\rangle$ with definite l are called pure, while all the other states $|sj\rangle$ are called mixed. For example, the state ${}^3P_2$–${}^3F_2$ is mixed and is represented

by the superposition (8) of two basis states with $l = 1$ and $l = 3$.

Magnetic interactions (5) and (6) contain the same operators $\mathbf{l} \cdot \mathbf{s}$ and S_{12} as nuclear interactions (7) but decrease as $r \longrightarrow \infty$ much more slowly. Therefore, taking into account magnetic interactions does not change the classification of states of the system (p, p) but should change the energy dependence of the scattering parameters of the phases $\delta_{l,j}^{c,a}$ and the mixing parameters ε_j^a introduced by Stapp *et al.* [21]. By definition, $\delta_{l,j}^{c,a}(k)$ is the difference between the phase $\delta_{l,j}^{c,a}(k)$ of the scattering by the superposition $V^c + V^a$ and the Coulomb phase $\delta_l^c(k)$. In the case $a = s$, the phase $\delta_{l,j}^{c,s}(k)$ is usually called a Coulomb–nuclear phase [9]. Therefore, in the case $a = m$, it seems reasonable to call the phase $\delta_{l,j}^{c,m}(k)$ Coulomb–magnetic when $a = m$ and Coulomb–magnetic–nuclear when $a = ms$. The physical meaning of the phase $\delta_{l,j}^{c,a}(k)$ is more correctly conveyed by the longer term: the scattering phase induced by the interaction V^a in the Coulomb field V^c.

Among all known approaches to the qualitative and numerical analysis of the energy dependence of the functions $\delta_{l,j}^{c,a}(k)$ and $\varepsilon_j^{c,a}(k)$ and the contributions, to these functions, of the parameter of the interaction V^a, which is taken into account either everywhere or only in a chosen range of distances, the physically transparent method of phase functions [10] seems to be the most convenient one. In this method, the phases $\delta_{l,j}^{c,a}(k)$, $l = j$, and $j \pm 1$, and the mixing parameter $\varepsilon_j^{c,a}(k)$ induced by the interaction V^a in the Coulomb field V^c, are defined by the expressions

$$\delta_{l,j}^{c,a}(k) \equiv \lim_{r\to\infty} \delta_{l,j}^{c,s}(r;k) \quad \text{and} \quad \varepsilon_j^{c,a}(k) \equiv \lim_{r\to\infty} \varepsilon_j^{c,a}(r;k)$$

as the limits of the corresponding phase functions $\delta_{l,j}^{c,s}(r;k)$ and $\varepsilon_j^{c,a}(r;k)$ that vanish for $r = 0$ and, for any $r = b$, represent the phases and the mixing parameter induced by the same—but truncated at the point $r = b$—interaction $V^a(r)$. The phase functions satisfy the following, computationally rather simple equations [10]:

$$\begin{aligned}\partial_r \delta_{l,j}^{c,a} &= -k^{-1}\sec(2\varepsilon_j^{c,a})\{V_{l,l}^a(P_l^2\cos^4\varepsilon_j^{c,a} \\ &- Q_l^2\sin^4\varepsilon_j^{c,a}) - V_{n,n}^a\sin^2(2\varepsilon_j^{c,a})\frac{P_n^2 - Q_n^2}{4} \\ &- V_{l,n}^a\sin(2\varepsilon_j^{c,a})[P_lQ_n\cos^2\varepsilon_j^{c,a} - P_nQ_l\sin^2\varepsilon_j^{c,a}]\}, \quad (9)\\ \partial_r\varepsilon_j^{c,a} &= -k^{-1}\{V_{l,n}^a(P_lP_n\cos^2\varepsilon_j^{c,a} \\ &+ Q_lQ_n\sin^2\varepsilon_j^{c,a}) - \frac{1}{2}\sin(2\varepsilon_j^{c,a})\sum_{n=j\pm1} V_{n,n}^a P_nQ_n\}.\end{aligned}$$

Here, l, $n = j \pm 1$ and $l \neq n$ for mixed states, and $l = n = j$ and $\varepsilon_j^{c,a} \equiv 0$ for pure states;

$$P_l \equiv F_l\cos\delta_{l,j}^{c,a} + G_l\sin\delta_{l,j}^{c,a};$$

$$Q_l \equiv F_l\sin\delta_{l,j}^{c,a} - G_l\cos\delta_{l,j}^{c,a};$$

$F_l(\rho, \eta)$ and $G_l(\rho, \eta)$ are the Coulomb functions [24] of the dimensionless argument $\rho \equiv kr$ and the Sommerfeld parameter $\eta \equiv 1/kR$; and $V_{l,n}^a$ are the matrix elements of the interaction V^a in the basis of vector spherical functions. For example, we have

$$\begin{aligned}V_{l,l}^{mt}(r) &= 2b_tr^{-3}\left\{\delta_{l,j} - \frac{l\delta_{l,j-1} + (l+1)\delta_{l,j+1}}{2j+1}\right\}, \\ V_{l,n}^{mt}(r) &= b_tr^{-3}\frac{\sqrt{6j(j+1)}}{2j+1}, \quad l \neq n\end{aligned} \quad (10)$$

for interaction (5), and $V_{l,n} \equiv 0$ for $l \neq n$ and

$$\begin{aligned}V_{l,l}^{mls}(r) &= b_{ls}r^{-3}[j(j+1) - l(l+1) - s(s+1)], \\ j &= l, l \pm s, s = 1\end{aligned} \quad (11)$$

for interaction (6).

The method proposed for analysis of the effect of magnetic interactions on the energy dependence of the phases $\delta_{l,j}^{c,a}$, $a = m, ms$ is extremely simple and consists in comparing the graphs of the phases calculated for different energies in the three theoretically possible cases $a = s, m, ms$.

Before proceeding to the discussion of the numerical results, we will try to predict the main features of the behavior of the phases $\delta_{l,j}^{c,a}$. To this end, we consider the first iteration of Eqs. (9), which is implemented by the substitution of $\delta_{l,j}^{c,a} \equiv 0$ and $\varepsilon_j^{c,a} \equiv 0$ into the right-hand sides of these equations. This yields the following representation as a sum of the Born phases $\tilde{\delta}_{l,j}^{c,s}$ and $\tilde{\delta}_{l,j}^{c,m}$ for the expected approximations $\tilde{\delta}_{l,j}^{c,ms}$ of the phases $\delta_{l,j}^{c,ms}$:

$$\begin{aligned}\delta_{l,j}^{c,ms}(k) &\approx \tilde{\delta}_{l,j}^{c,ms}(k) \equiv \tilde{\delta}_{l,j}^{c,s}(k) + \tilde{\delta}_{l,j}^{c,m}(k), \\ \tilde{\delta}_{l,j}^{c,a}(k) &\equiv -k^{-1}\int_b^\infty dr V_{l,l}^a(r)F_l^2(\rho,\eta);\end{aligned} \quad (12)$$

here, $a = s$ or $a = m$ and $b = 0$. To calculate the integrals in these formulas, it is convenient to pass to the dimensionless variables $x \equiv r/R$ and $q \equiv kR$. It is well

known [25] that the Born Coulomb–nuclear phase decreases very rapidly for any l and $E \longrightarrow 0$:

$$\tilde{\delta}^{c,s}_{l,j}(k) \sim (kR)^{2l+1}\exp(-\pi\eta), \tag{13}$$

while the Born Coulomb–magnetic phase decreases much slower:

$$\tilde{\delta}^{c,m}_{l,j}(k) = -V^{m}_{l,l}(r)r^{3}\frac{2l+1-2\eta\chi_l(\eta)}{2l(l+1)(2l+1)}(1+o(1)),$$
$$\chi_l(\eta) \equiv \frac{\pi}{2} - \mathrm{Im}\psi(l+1+i\eta), \quad \psi \equiv \frac{\Gamma'}{\Gamma}. \tag{14}$$

Indeed,

$$\tilde{\delta}^{c,m}_{l,j}(k) = -\frac{k^3}{3R^2}V^{m}_{l,l}(r)r^3(1+o(1)), \tag{15}$$

if $\eta \ll l$, which is true for

$$E \ll \frac{1}{2l^2}\frac{m_e}{m_p}\mathrm{Ry} \approx 12.5l^{-2}\ \mathrm{keV}.$$

An approximation, more accurate than (14), can be obtained from perturbation theory [8].

Due to drastically different falloff rates of the Born phases (13)–(15), for sufficiently low energies, we have

$$\left|\tilde{\delta}^{c,s}_{l,j}(k)\right| \ll \left|\tilde{\delta}^{c,m}_{l,j}(k)\right|, \quad \tilde{\delta}^{c,ms}_{l,j}(k) \approx \tilde{\delta}^{c,m}_{l,j}(k),$$

$$E < E^{\mathrm{lower}}_{l,j}.$$

Therefore, at such energies, one can neglect the nuclear interaction, but the magnetic interaction should be taken into account. At sufficiently high energies, where $|V^m| \ll E$, the inverse relations

$$\left|\tilde{\delta}^{c,s}_{l,j}(k)\right| \gg \left|\tilde{\delta}^{c,m}_{l,j}(k)\right|, \quad \tilde{\delta}^{c,ms}_{l,j}(k) \approx \tilde{\delta}^{c,s}_{l,j}(k),$$

$$E > E^{\mathrm{upper}}_{l,j}$$

must hold; therefore, one can neglect the magnetic interaction but should take into consideration the nuclear interaction. In the intermediate region $E^{\mathrm{lower}}_{l,j} < E < E^{\mathrm{upper}}_{l,j}$, where the moduli of the phases $\tilde{\delta}^{c,s}_{l,j}$ and $\tilde{\delta}^{c,m}_{l,j}$ are of about the same order of magnitude, there is interference between the particles scattered by the nuclear and magnetic interactions; therefore, one should take into account both these interactions to describe this interference. If the phases $\tilde{\delta}^{c,s}_{l,j}$ and $\delta^{c,m}_{l,j}$ have different signs in this region, then their sum $\delta^{c,ms}_{l,j}$ vanishes at a certain value of energy.

Thus, if we assume that the approximations of the exact phases $\delta^{c,ms}_{l,j}$ by the phases $\tilde{\delta}^{c,ms}_{l,j}$ given by formulas (12) are acceptable, then we should expect the following two features in the behavior of the phases $\delta^{c,ms}_{l,j}$: a slow decrease ($\delta^{c,ms}_{l,j} \sim \delta^{c,m}_{l,j} \sim k^3$) as $E \longrightarrow 0$ for any l and j, and a sign change for a certain nonzero energy, which, however, occurs only when the phases $\tilde{\delta}^{c,s}_{l,j}$ and $\tilde{\delta}^{c,m}_{l,j}$ have different signs.

The next feature is due to the fact that the matrix elements (10) and (11) of the magnetic interactions (5) and (6) depend only on j; therefore, this feature should manifest itself in any approximate and exact calculations of the phases. The matrix elements $V^{mls}_{l,n}$ with $l \neq n$ increase with j, while the elements $V^{mls}_{l,l}$ and all the elements $V^{mt}_{l,n}$ remain bounded. Therefore, one should expect that, as j increases, the contribution of the interaction V^{mt} to the phases $\delta^{c,a}_{l,l\pm1}$, $a = ms$ will decrease as $1/j$, while the contributions of these interactions to the phases $\delta^{c,a}_{l,l}$, $a = m, ms$ remain of the same order of magnitude.

The method of phase functions allows one to qualitatively substantiate the approximation

$$\delta^{c,ms}_{l,j}(k) \approx \delta^{c,s}_{l,j}(k) + \tilde{\delta}^{c,m}_{l,j}(k), \tag{16}$$

which is physically more realistic than representation (12). To this end, we set $a = ms$. Integrating Eqs. (9) over the interval $r \leq r^s$, where $V^{ms} = V^s$, we obtain the values of Coulomb–nuclear phases $\delta^{c,s}_{l,j}(k) \approx \delta^{c,s}_{l,j}(r^s; k)$ as the values of the corresponding phase functions at point r^s. We use these values as the boundary values for analyzing Eqs. (9) in the domain $r \geq r^s$, where $V^{ms} \approx V^m$. The first iteration of these equations yields a representation in the form of the sum (16), while the subsequent iterations give rise to additional terms; each nth term ($n = 2$, 3, ...) decreases, as $E \longrightarrow 0$, faster than the preceding one, namely, as $(\tilde{\delta}^{c,m}_{l,j}(k))^n$. Therefore, at low energies, representation (16) is an approximation that contains, as a term, the exact Coulomb–nuclear phase. It remains to determine its asymptotic behavior as $E \longrightarrow 0$.

We begin with auxiliary formulas. First, we single out an entire function Θ_l with parameter q^2 from the Coulomb function G_l. To this end, we rewrite the Lambert formula (formula (3.25) in [26]) as

$$G_l(\rho,\eta) = \tilde{G}_l(\rho,\eta) + h^c(q)F_l(\rho,\eta),$$
$$\tilde{G}_l(\rho,\eta) \equiv \frac{\Theta_l(x,q)}{C_l(q)}, \tag{17}$$

where $C_l(q)$ and $h^c(q)$ are expressed in terms of the known functions [24] $C_l(\eta)$ and $h(\eta)$:

$$C_l(q) \equiv G(2l+2)q^l C_l(\eta)$$
$$= (2q)^l \exp\left(-\frac{\pi\eta}{2}\right)|\Gamma(l+1+i\eta)|,$$
$$h^c(q) \equiv \frac{h(\eta)}{qC_0^2(q)}, \quad h(\eta) \equiv \mathrm{Re}\psi(i\eta) - \ln\eta.$$

Now, we modify the well-known Bessel–Clifford expansions (see formulas (14.4.1)–(14.4.4) in [24]) that contain polynomials $b_n(\eta)$ with parameter k^2 and the modified Bessel functions $I_n(z)$ and $K_n(z)$ of the variable $z \equiv 2x^{1/2}$. Collecting the terms with equal powers of k^2 in these expansions, we obtain the required representation

$$F_l(\rho,\eta) = qC_l(q)\sum_{n=0}^{\infty} q^{2n} f_{ln}(x),$$
$$\tilde{G}_l(\rho,\eta) = C_l^{-1}(q)\sum_{n=0}^{\infty} q^{2n} g_{ln}(x). \tag{18}$$

Here,

$$\left\{\begin{array}{c} 2f_{ln}(x) \\ (2l+1)g_{ln}(x) \end{array}\right\}$$
$$\equiv 2^{-2n}\sum_{m=2n}^{3n} a_{nm} z^{m+1} \left\{\begin{array}{c} I_{2l+m+1}(z) \\ (-1)^{-m}K_{2l+m+1}(z) \end{array}\right\},$$

and the energy-independent coefficients a_{nm} satisfy the recurrent chains ($m = 2n, ..., 3n$ for each $n = 1, 2, ...$) of equations

$$2ma_{nm} + 2(2l+m)a_{n-1,m-2} + a_{n-1,m-3} = 0,$$

here, $a_{00} \equiv 1$, and $a_{nm} \equiv 0$ if $n > 0$ and $m < 2n$ or $m > 3n$.

Next, we pass to the tangents of the phase functions in Eqs. (9). Then, we replace the tangents by the required series,

$$\tan\delta_{l,j}^{c,s}(r;k) = -qC_l^2(q)\sum_{n=0}^{\infty} q^{2n} A_{l,j,n}(x;h^c),$$
$$\tan\varepsilon_j^{c,s}(r;k) = -qC_{j-1}(q)C_{j+1}(q)\sum_{n=0}^{\infty} q^{2n} B_{j,n}(x;h^c), \tag{19}$$
$$A_{l,j,0}(x;h^c) = A_{l,j}(x)[1 + h^c q C_l^2 A_{l,j}(x)]^{-1},$$
$$B_{j,0}(x;h^c) = B_j(x)[1 + h^c q C_{j-1}(q)C_{j+1}(q)B_j(x)]^{-1},$$

and, using formulas (17) and (18), represent the Coulomb functions as series in which the argument x is separated from the parameter q. Finally, letting $q \longrightarrow 0$, we obtain the energy-independent equations

$$\partial_x A_{l,j} = R^2\{V_{l,l}^s[f_l - A_{l,j}q_l]^2 + V_{n,n}^s B_j^2 g_n^2$$
$$- 2B_j V_{l,n}^s[f_l - A_{l,j}g_l]g_n\},$$
$$\partial_x B_j = R^2 V_{l,n}^s[(f_l - A_{l,j}g_l)(f_n - A_{n,j}g_n)$$
$$+ B_j^2 g_l g_n] - R^2 B_j \sum_{n=j\pm1} V_{n,n}^2(f_n - A_{n,j}g_n)g_n, \tag{20}$$

where, just as in the original equations (9), l, $n = j \pm 1$, and $l \neq n$ for mixed states and $l = n = j$, $B_j \equiv 0$, for pure states. By virtue of (19), the required solutions to Eqs. (20) vanish at $x = 0$ and, due to the exponential fall-off of the nuclear interaction, are everywhere bounded. Therefore, we can pass to the limit as $r \longrightarrow \infty$ in (19) and obtain the required asymptotics:

$$\delta_{l,j}^{c,s}(k) \approx -\arctan\frac{qC_l^2(q)A_{l,j}^{c,s}}{1 + h^c(q)qC_l^2(q)A_{l,j}^{c,s}},$$
$$A_{l,j}^{c,s} \equiv \lim_{x\to\infty} A_{l,j}(x). \tag{21}$$

Since the leading terms of these asymptotics differ from the asymptotics of the corresponding Born phases (13) only by numerical factors, approximation (16) leads to the same features in the behavior of the phases $\delta_{l,j}^{c,ms}$ as those obtained from the supposed Born approximation (12).

We suggest using formula (16) for extrapolating the phases $\delta_{l,j}^{c,ms}$ to low energies. This formula is sufficiently simple: its second term $\tilde{\delta}_{l,j}^{c,m}$ is expressed in terms of known functions via equality (14), and the first term $\delta_{l,j}^{c,s}$ can be approximated by the asymptotics (21) with coefficient the $A_{l,j}^{c,s}$, which can easily be calculated as the limit, for $x \longrightarrow \infty$, of the function $A_{l,j}(x)$ subject to Eqs. (20). Now, to verify that the extrapolation formula proposed is sufficiently exact, we discuss the results of numerical analysis of the phases.

4. RESULTS OF CALCULATIONS

The results discussed in this section have been obtained by the numerical integration of the differential equations derived from Eqs. (9) by changing the dimensional variables r and k to dimensionless ones, $x \equiv r/R$ and $q \equiv kR$. As the phases, we used appropriate values of the phase functions that reproduce the phases accurate up to five decimal places for a sufficiently large distance of $x = 10^2$ in the case of $a = s$ and for $x = 10^6$ in

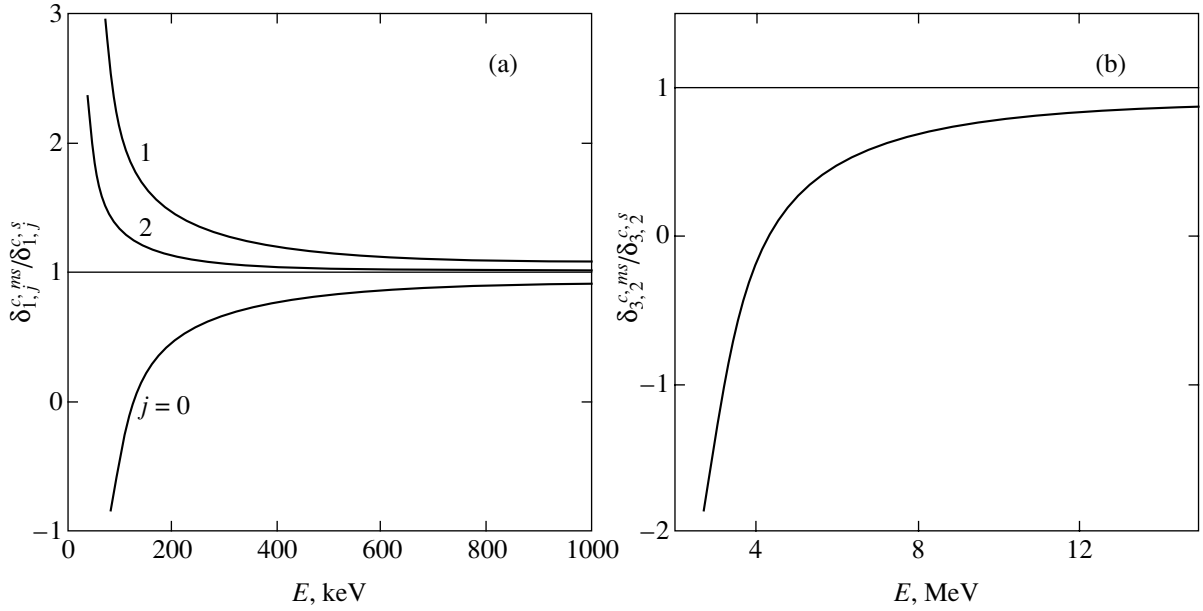

Fig. 1. Phase ratios $\delta_{l,j}^{c,ms}/\delta_{l,j}^{c,s}$; (a) $l = 1$ and $j = 0, 1, 2$, and (b) $l = 3$ and $j = 2$.

the cases of $a = m$ and $a = ms$. As the coefficients $A_{l,j}^{c,s}$ of asymptotics (21), we took the solutions $A_{l,j}^{c,s}(x)$ of Eqs. (20) at the point $x = 10^2$, which also guaranteed an accuracy of five decimal places.

The exact phases thus calculated were compared with the approximate phases determined by formulas (14), (21), or (16) for the cases $a = m$, $a = s$, or $a = ms$, respectively. We have established that, for energies below 15 MeV, the relative accuracy of all these approximations for $j = 0, 1, 2$ is no worse than 0.001.

We present the calculated values of the coefficients $A_{l,j}^{c,s}$ as the products

$$A_{1,0}^{c,s} \approx -26.743 d_1, \quad A_{1,1}^{c,s} \approx 15.116 d_1,$$

$$A_{1,2}^{c,s} \approx -8.739 d_1, \quad A_{3,2}^{c,s} \approx -39\,205 d_3$$

with the cofactors $d_1 \equiv (3!)^{-2}R^{-3}$ and $d_3 \equiv (7!)^{-2}R^{-7}$.

Note that the replacement of the phases $\delta_{l,j}^{c,s}$ by the corresponding Born integrals (12) is unacceptable, because the value of the ratio $|\delta_{l,j}^{c,s}/\tilde{\delta}_{l,j}^{c,s}|$ ranges from 0.2 to 1.5 when energy varies from 0 to 15 MeV. Therefore, representation (12) is not an approximation, although it describes all the qualitative features in the behavior of the phases $\delta_{l,j}^{c,ms}$.

The graphs shown in Figs. 1 and 2 have been obtained by the numerical integration of Eqs. (9) and do not differ from those obtained by the approximate formulas (14), (16), or (21) to within graphical resolution.

Figure 1 represents the graphs of the ratios

$$\frac{\delta_{l,j}^{c,ms}(k)}{\delta_{l,j}^{c,s}(k)}, \quad l = 1, \quad j = 0, 1, 2; \quad l = 3; \quad j = 2.$$

Figure 1a shows that, for $l = 1$, these ratios are appreciably different from 1 in the domain of sufficiently small energies ($E < E_{1,j}^{\text{upper}} \approx 1$ MeV). Hence, to correctly describe the ${}^3P_j$ phases with $j = 0, 1$ and the ${}^3P_2$–${}^3F_2$ phases with $l = 1$ in this domain, one should take into account the magnetic interaction V^m, whereas, in the domain of high energies ($E > E_{1,j}^{\text{upper}}$), one can neglect V^m compared with the nuclear interaction V^s. According to Fig. 1b, to correctly describe the phases $\delta_{1,2}^{c,ms}$, $l = 3$, one should take into account the magnetic interaction V^m in the range of energies from zero to a value of $E = E_{3,2}^{\text{upper}} \approx 15$ MeV, which is an order of magnitude greater than that in the previous case $l = 1$.

In Fig. 2, the solid curves represent the phases $\delta_{l,j}^{c,a}$, $a = s, m, ms$, and the dashed curves represent the phases $\bar{\delta}_{l,j}^{c,a}$, $a = m, ms$, calculated when interaction (5) is switched on ($b_t = 0$) but interaction (6) is switched off. Figures 2a–2c show that

$$\delta_{1,j}^{c,ms}(k) \approx \delta_{1,j}^{c,m}(k) \gg \delta_{1,j}^{c,s}(k), \quad j = 0, 1, 2,$$
$$E < E_{1,j}^{\text{lower}} \approx 20 \text{ keV}.$$

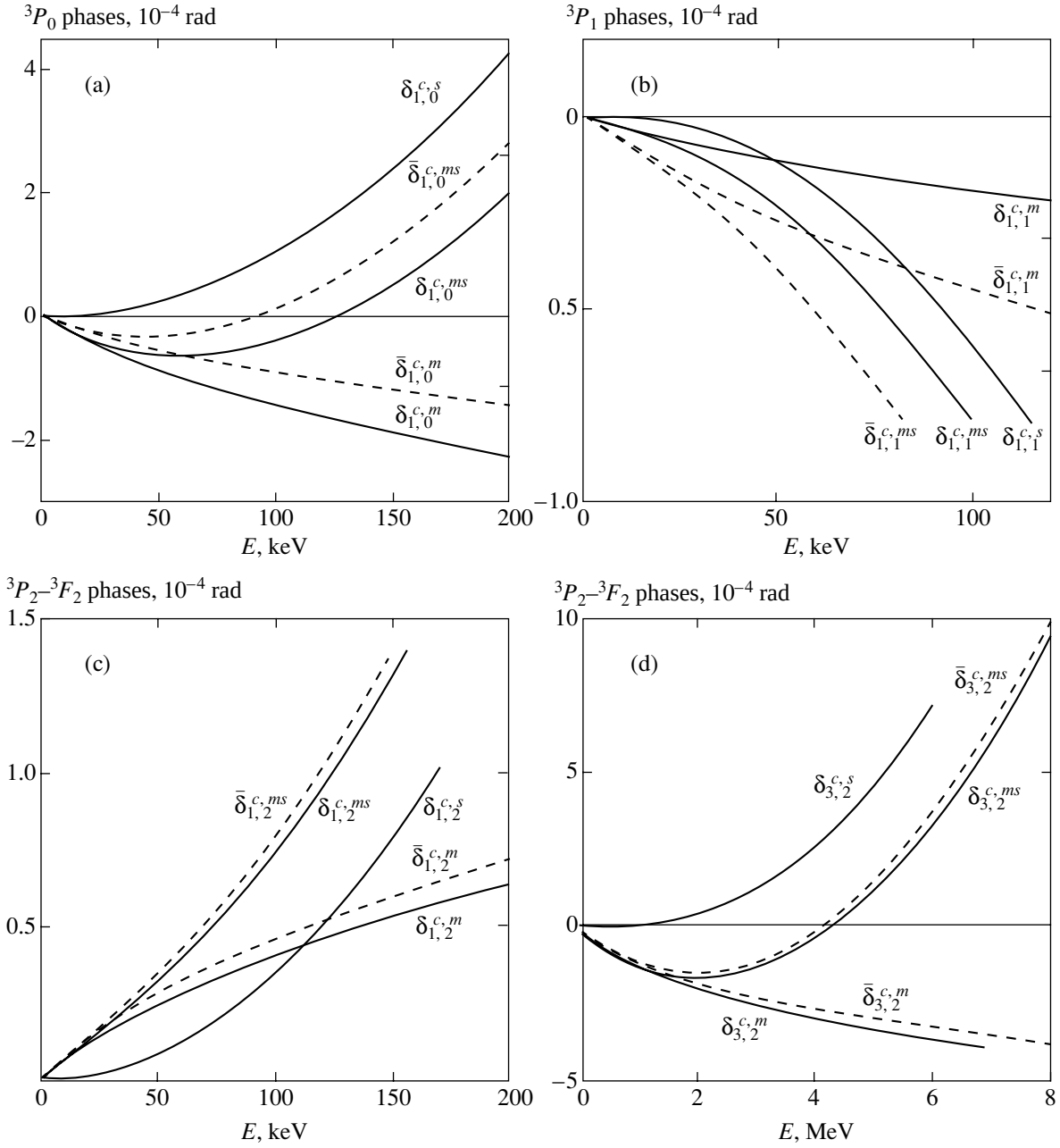

Fig. 2. Phases $\delta_{l,j}^{c,a}$, $a = s, m, ms$ (solid curves) and $\bar{\delta}_{l,j}^{c,a}$, $a = m, ms$ (dashed curves).

Hence, in the domain of sufficiently low energies ($E <$ 20 keV), the contributions of the nuclear interaction V^s to the phases $\delta_{1,j}^{c,ms}$, $j = 0, 1, 2$ are negligible. According to Fig. 2d, the contribution of V^s to the phase $\delta_{3,2}^{c,ms}$ remains negligible up to an energy of $E = E_{3,2}^{\text{lower}} \approx$ 2 MeV.

Next, the phases $\delta_{1,0}^{c,s}$ and $\delta_{1,0}^{c,m}$ depicted in Fig. 2a have different signs; as a result of interference between the particles scattered by a nuclear interaction and by the sum of two magnetic interactions (5) and (6), the phase $\delta_{1,0}^{c,ms}$ changes its sign at $E \approx 120$ keV. According to Fig. 2a, for $E < 200$ keV, the phases $\bar{\delta}_{1,0}^{c,a}$, $a = m, ms$, appreciably differ from the corresponding phases $\delta_{1,0}^{c,a}$. Hence, both magnetic interactions (5) and (6) produce a comparable (in order of magnitude) effect on the formation of the Coulomb–magnetic–nuclear $^3P_0$ phase $\delta_{1,0}^{c,ms}$; therefore, none of these interactions can be neglected compared with certain other ones. According

to Fig. 2b, a similar result holds for the $^3P_1$ phase $\delta_{1,j}^{c,ms}$; however, as follows from Figs. 2c and 2d, one may neglect the tensor magnetic interaction (5) when calculating the $^3P_2$–$^3F_2$ phases $\delta_{l,2}^{c,ms}$ with $l = 2 \pm 1$. Finally, as is shown in Fig. 2d, the phase $\delta_{3,2}^{c,ms}$ has a zero at $E \approx 4$ MeV.

We complete this section with the following conclusions: formula (16) allows one to extrapolate the phases $\delta_{l,j}^{c,ms}$ with $j = 0, 1, 2$ to energies of $E < 15$ MeV with a relative accuracy of 0.001; all the features of the energy dependence of phases that were predicted analytically in Section 3 have been confirmed numerically.

5. CONCLUSIONS

Let us summarize the main results of the analysis of triplet phases of the *pp* scattering. The interactions between the magnetic moment of a proton and the Coulomb field and the magnetic moment of another proton have a substantial effect on the behavior of the triplet phases at energies below several megaelectronvolts. Owing to these interactions, in the limit of zero collision energy, all triplet phases should be proportional to the cube of the collision momentum and the $^3P_0$ phase and the $^3P_2$–$^3F_2$ phase with $l = 3$ should change their signs at energies of $E \approx 120$ keV and $E \approx 4$ MeV, respectively. All the features of the energy dependence pointed out above are described to a good accuracy by the simple extrapolation formula (16), which is independent of the choice of a model of nuclear interaction among all phase-equivalent interactions. The Coulomb–magnetic and Coulomb–nuclear terms of this formula can readily be determined by formulas (14) and (21) to a good accuracy for $E < 15$ MeV. To calculate, to a high accuracy, the coefficients $A_{l,j}^{c,s}$ and $B_j^{c,s}$ of higher order terms in the low-energy representations of the Coulomb–nuclear phases and the mixing parameters, we suggest applying the energy-independent equations (20). A full analysis of these equations seems to be important for extending perturbation theory [8] and the method of phase functions [10] to the case of the superposition of the Coulomb interaction and the short-range central, spin–orbit, and tensor interactions.

In conclusion, we note once again that, because a direct experimental investigation of the triplet *NN* scattering in the range of energies below several megaelectronvolts is impossible at the present state of the art, a theoretical study of the role of electromagnetic corrections to the nuclear *NN* interaction in this energy domain remains an interesting and topical problem.

REFERENCES

1. V. V. Pupyshev and O. P. Solovtsova, Int. J. Mod. Phys. A **7**, 2713 (1992).
2. V. V. Pupyshev and O. P. Solovtsova, Fiz. Élem. Chastits At. Yadra **27**, 859 (1996) [Phys. Part. Nucl. **27**, 353 (1996)].
3. V. V. Pupyshev and O. P. Solovtsova, in *Proc. of the International Conference on Mesons and Nuclei at Intermediate Energies*, Ed. by M. Kh. Khankhasaev and Zh. B. Kurmanov (JINR, Dubna, 1994), p. 84.
4. V. V. Pupyshev and O. P. Solovtsova, Phys. Lett. B **354**, 1 (1995).
5. V. V. Pupyshev and O. P. Solovtsova, Yad. Fiz. **59**, 1807 (1996) [Phys. At. Nucl. **59**, 1745 (1996)].
6. V. V. Pupyshev, Fiz. Élem. Chastits At. Yadra **28**, 1457 (1997) [Phys. Part. Nucl. **28**, 586 (1997)].
7. V. V. Pupyshev and S. A. Rakityansky, Z. Phys. A **348**, 227 (1994).
8. V. V. Pupyshev, J. Phys. A: Math. Gen. **28**, 3305 (1995).
9. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 3: *Quantum Mechanics: Non-Relativistic Theory*, 3rd ed. (Nauka, Moscow, 1974; Pergamon, New York, 1977).
10. V. V. Babikov, *The Method of Phase Functions in Quantum Mechanics* (Nauka, Moscow, 1976).
11. V. I. Gol'danskiĭ, O. A. Karpukhin, A. V. Kutsenko, and V. V. Pavlovskaya, Zh. Éksp. Teor. Fiz. **38**, 1695 (1960) [Sov. Phys. JETP **11**, 1223 (1960)].
12. V. A. Petrun'kin, Fiz. Élem. Chastits At. Yadra **12**, 692 (1981) [Sov. J. Part. Nucl. **12**, 278 (1981)].
13. W. A. Barker and F. N. Glover, Phys. Rev. **99**, 317 (1955).
14. G. E. Brown and A. D. Jackson, *The Nucleon–Nucleon Interaction* (North-Holland, Amsterdam, 1976; Atomizdat, Moscow, 1979).
15. C. Ramsauer and R. Kollath, Ann. Phys. (Leipzig) **3**, 54 (1929).
16. J. Z. Holtsmark, Z. Physik **66**, 49 (1930).
17. V. G. J. Stoks and J. J. de Swart, Phys. Rev. C **42**, 1235 (1990).
18. W. S. Hogan and R. G. Seyler, Phys. Rev. C **1**, 17 (1970).
19. G. Breit and H. M. Ruppel, Phys. Rev. **127**, 2123 (1962).
20. L. D. Knutson and D. Chiang, Phys. Rev. C **18**, 1958 (1978).
21. H. P. Stapp, T. J. Ypsilantis, and M. Metropolis, Phys. Rev. **105**, 302 (1957).
22. Jr. R. V. Reid, Ann. Phys. (N.Y.) **50**, 411 (1968).
23. M. Aguilar-Benitez, R. L. Grawford, R. Frosch, *et al.*, Phys. Lett. B **111B**, 1 (1982).
24. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and I. A. Stegun, 2nd ed. (Dover, New York, 1971; Nauka, Moscow, 1979).
25. R. O. Berger and L. Spruch, Phys. Rev. **138**, B1106 (1965).
26. E. Lambert, Helv. Phys. Acta **42**, 667 (1969).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1102–1111.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1232–1243.

ATOMS, SPECTRA, RADIATION

Spectroscopic Manifestations of Saturation of Optical Transitions by Spontaneous Emission

É. G. Saprykin[a], S. N. Seleznev[b], and V. A. Sorokin[b]

[a]*Novosibirsk State University, Novosibirsk, 630090 Russia*

[b]*Institute of Automation and Electrometry, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

e-mail: saprykin@gorodok.net, sorokin@iae.nsk.su

Received November 30, 2002; in final form, May 13, 2003

Abstract—Stimulated absorption of spontaneous emission of a gas-discharge plasma can cause a significant increase in the population of the emitting states, compared to their population determined by inelastic collisions and spontaneous decay from upper levels. By imposing a magnetic field, this self-saturation is reduced and its contribution to the level population can be identified. The magnetic-field-dependent changes in Doppler profiles due to radiative transitions caused by saturating spontaneous emission are analyzed in the case of absorption of a weak monochromatic wave. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The role played by stimulated radiative transitions in the kinetics of quantum states has been known since the time of Einstein. However, extensive studies of their diverse consequences began only after the invention of lasers, because stimulated transitions manifest themselves when their probabilities (depending on the intensity of the stimulating radiation) are comparable to those of spontaneous transitions. The required intensities are easily attained using laser light. As a result, the populations of the states involved in an optical transition become comparable, the level lifetimes are reduced, and other manifestations of stimulated radiative transitions are observed (e.g., see [1]). In the absence of saturating laser light, the rates of stimulated radiative transitions induced by the spontaneous emission of a low-pressure gas-discharge plasma are generally negligible as compared to those of spontaneous decay. For this reason, they are difficult to identify among other processes taking place in the plasma (as inelastic collisions with electrons or atoms and spontaneous decay from upper levels), and their contribution has been neglected. The impact of stimulated transitions on the characteristics of gas-discharge plasmas was recognized only recently. In particular, it was shown in [2–4] that stimulated transitions can play a significant role in typical low-temperature gas-discharge plasmas under certain conditions.

These observations apply primarily to metastable states (which cannot decay spontaneously), and stimulated absorption can be the mechanism responsible for their decay [2]. As estimated in [2], the frequency ν of absorption-induced transitions per metastable atom per 1 cm^3 is 10^6 s^{-1} under typical conditions of a neon glow-discharge plasma. When the concentration of the excited atoms whose emission induces stimulated transitions is proportional to the electron concentration, the rates of radiative de-excitation of metastables are higher than the rates of electron-impact quenching by an order of magnitude.

Moreover, stimulated transitions can manifest themselves in exchange of magnetic coherence between degenerate levels of comparable width. Under these conditions, even if the frequency of stimulated exchange of magnetic coherence between levels is low as compared to the spontaneous relaxation rate A_{mn} ($\nu_{nm} \sim 10^6$–10^7 $s^{-1} \sim 0.1A_{mn}$), the magnetic resonance widths can change significantly (decrease or increase) [3]. Stimulated transfer and exchange of magnetic coherence are of interest because breakdown of the coherence of Zeeman sublevels in a weak magnetic field suggests that they can be observed directly in experiment. Observations of stimulated transfer and exchange of magnetic coherence (alignment) between levels induced by spontaneous emission of a discharge were reported in [4]. It should be noted that, even though stimulated transitions have low probabilities, the amplitudes of the magneto-optical alignment resonances transferred from lower levels turn out to be smaller than, but comparable to, the amplitudes of the natural alignment resonances for the level under study. The reason is that, unlike relaxation times (determined by the frequency of stimulated transitions per atom), the amplitudes depend on the total rate of transition of aligned atoms to the upper level, which is proportional to the frequency of stimulated excitation from the lower state multiplied by the concentration of aligned atoms on the lower level. Accordingly, a high concentration of

aligned atoms on lower levels compensates for a low transfer probability.

However, phenomena associated with alignment (rank-two polarization moment) must also manifest themselves in level population (rank-zero polarization moment). Again, the low probability of "upward" transitions induced by spontaneous emission can be compensated for by a high concentration of atoms on lower levels. Whereas the change in the difference of level populations induced by spontaneous emission of a discharge (normally used as a saturation criterion) remains small, the relative increase in the population of a weakly populated level can be large enough to be detected spectroscopically. Under such conditions, spontaneous emission becomes a factor that additionally increases the population of a decaying state. However, in contrast to the case of magnetic coherence, no mechanism has so far been found for identifying the contribution of radiative transitions induced by spontaneous emission to the population.

The present study was conducted to single out the contribution of transitions induced by spontaneous emission to level populations (self-saturation of transitions). It was motivated by the results reported in [5], where the line profile associated with absorption of monochromatic emission by the $3s_2$–$2p_4$ neon transition was found to be asymmetric as a function of longitudinal magnetic field when the laser frequency was detuned from the transition line center. According to the analysis presented in [5], only magnetic-field-dependent transfer of population from lower levels to the level under study could be responsible for the asymmetry. However, the transfer mechanism has remained somewhat unclear.

The magnetic-field dependence of radiation-induced population transfer that underlies the method proposed here is qualitatively explained as follows. Consider an excited two-level gas with a triply degenerate lower or upper level in a magnetic field that splits the spontaneous emission line. Suppose that the splitting interval is wider than the line width. Three linearly polarized spectral components will then be emitted in the direction orthogonal to the magnetic field. The wave polarized parallel to the magnetic field (π component) is not shifted by the magnetic field and does not contribute to the effect in question. The remaining two components have similar polarizations orthogonal to the magnetic field (σ components), but their respective frequency shifts induced by the magnetic field have opposite signs. When the Zeeman splitting interval is sufficiently large, each of these spontaneous-emission components induces radiative transitions (saturates the transition) between the corresponding pair of sublevels. However, as the magnetic field strength is reduced, these spectral components overlap and each transition is stimulated by the sum of "intrinsic" and "extrinsic" emission intensities. This increases the probability of stimulated transitions, the extent of self-saturation, and

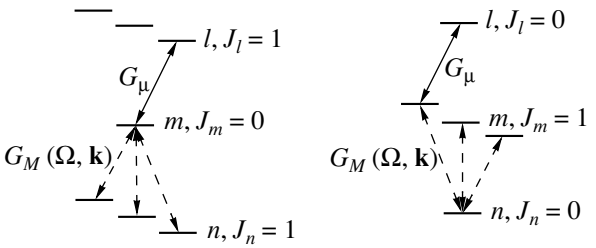

Fig. 1. Optical transition schemes.

the population of the upper level. The resulting additional population is a function centered at the zero of the magnetic field. It can be detected by various methods, including measurement of spontaneous emission and probing by laser light. Thus, we now have an experimental approach that can be used to identify self-saturation among the processes contributing to level population.

The present study was conducted to determine the line shapes recorded by laser-based spectroscopy for levels with different angular momenta and Landé factors in the presence of isotropic saturating spontaneous radiation emitted by atoms in thermal motion in a discharge under different observation conditions.

2. SATURATION IN LONGITUDINAL MAGNETIC FIELD

We analyze absorption of a monochromatic probe wave by the three-level system schematized in Fig. 1 in the presence of isotropic broadband radiation. Suppose that the level populations are such that $N_n \gg N_m \gg N_l$. For example, the difference in population between the lower energy levels can be as large as four orders of magnitude [6]. Circularly polarized laser light is resonant with the upper pair of levels. The corresponding absorption spectrum is measured by magnetic scanning. Following [5], we consider a magnetic field parallel to the wave vector $\mathbf{k}_\mu$ of the probe wave propagating along the z axis. We assume that the saturating isotropic broadband radiation is generated by spontaneous emission from level m and can be represented as a superposition of plane waves propagating in all directions.

2.1. A Cascade System with a Nondegenerate Intermediate Level

When level m is nondegenerate, any change in the magneto-optical spectra of absorption in transitions from this level must be attributed to effects depending on its population. Accordingly, both upper and lower levels (l and n) must be triply degenerate with respect to the magnetic quantum number (their magnetic moments must be $J_l = J_n = 1$), while the Landé factor g_n of level n may differ from g_l. To calculate the work

done by the probe field under these conditions, we use the expansion of expression (8.79) in [1, p. 137] in powers of the saturating radiation intensity (see also (13.4)) obtained by modeling in terms of relaxation constants:

$$P_\mu = P_\mu^{(0)} + P_\mu^{(1)},$$

$$P_\mu^{(0)} = -2\hbar\omega_\mu \left\langle \frac{\Gamma_{lm} N_m W(\mathbf{v}) |G_\mu|^2}{\Gamma_{lm}^2 + (\Omega_\mu + \Delta_\mu - \mathbf{k}_\mu \cdot \mathbf{v})^2} \right\rangle_{\mathbf{v}},$$

$$P_\mu^{(1)} = -2\hbar\omega_\mu \left\langle \frac{\Gamma_{lm} N_n W(\mathbf{v}) |G_\mu|^2}{\Gamma_{lm}^2 + (\Omega_\mu + \Delta_\mu - \mathbf{k}_\mu \cdot \mathbf{v})^2} \right. \tag{1}$$

$$\left. \times \sum_{M=-1}^{1} \frac{2\Gamma_{mn}}{\Gamma_{mm}} \frac{|G_M(\Omega, \mathbf{k})|^2}{\Gamma_{mn}^2 + (\Omega + \Delta M - \mathbf{k} \cdot \mathbf{v})^2} \right\rangle_{\mathbf{v}, \mathbf{k}, \Omega}.$$

Here, Γ_{lm}, Γ_{mn}, and Γ_{mm} are the relaxation constants in the system of terms *l*, *m*, and *n*; $W(\mathbf{v})$ is the Maxwellian velocity distribution; G_μ is the Rabi frequency of the probe field; $\Omega_\mu = \omega_\mu - \omega_{lm}$ and $\Omega = \omega - \omega_{mn}$ are the respective mismatches between the probe and saturating-radiation frequencies (ω_μ and ω) and the transition line centers ω_{lm} and ω_{mn}; $\Delta = \mu_B g_n H$ and $\Delta_\mu = \mu_B g_l H$ determine the Zeeman frequency shifts of the magnetic sublevels of levels *n* and *l*, respectively; the index *M* runs over the magnetic sublevels of level *n*; and $G_M(\Omega, \mathbf{k})$ is the Rabi frequency for spontaneous *m*–*n* emission in the direction defined by the vector $\mathbf{k}$, which depends on the magnetic quantum number *M* of level *n*. Summing over the magnetic sublevels of level *n* is equivalent to summing over the polarizations of spontaneous emission represented in a spherical basis with unit vectors $\mathbf{e}_{+1}$, $\mathbf{e}_z$, and $\mathbf{e}_{-1}$ (the *z* axis is parallel to the magnetic field). The angle brackets in (1) denote averaging over the variables written as subscripts outside the angle brackets: the atom velocity $\mathbf{v}$, the direction of the wave vector $\mathbf{k}$, and the mismatch Ω of the saturating radiation. The sum over the magnetic sublevels in (1) can be treated as proportional to the first nonlinear correction to the population of level *m*.

In the linear approximation, only the averaging over the longitudinal atom velocity v_z is required in (1). In the first nonlinear approximation, averaging over all velocity components must be performed. With a view to integrating over Ω, only the terms associated with the change in level *m* population induced by the saturating spontaneous radiation are retained in (1). Since the contributions due to nonlinear interference will vanish, there is no need to include them in the starting expressions. We also neglect the terms that represent field-induced splitting in (1), because they are small when $N_n \gg N_m$ (see [1, p. 210]).

The saturating radiation intensity $|G_M(\Omega, \mathbf{k})|^2$ is represented in (1) by three components: $|G_{+1}(\Omega, \mathbf{k})|^2$, $|G_0(\Omega, \mathbf{k})|^2$, and $|G_{-1}(\Omega, \mathbf{k})|^2$. The distribution $|G_M(\Omega, \mathbf{k})|^2$ is isotropic with respect to azimuthal orientation φ. Its dependence on the polar angle Θ is determined by the projections of atomic dipole oscillations on the plane perpendicular to the propagation direction (defined by vector $\mathbf{k}$). It is well known that electric dipole oscillation represented in a spherical basis with mutually perpendicular unit vectors $\mathbf{e}_{+1}$, $\mathbf{e}_z$, and $\mathbf{e}_{-1}$ satisfies the magnetic quantum-number selection rules for dipole radiation. Dipole oscillation along the quantization axis *z* (magnetic field) gives rise to a radiation component with zero frequency shift and polarization parallel to the magnetic field (π component). Oscillation perpendicular to the *z* axis gives rise to Zeeman-shifted right- and left-hand polarized σ components of radiation. Projection of dipole oscillation on the plane perpendicular to the vector $\mathbf{k}$ is performed by using the Wigner matrix $D^1(0\beta0) = d^1(\beta)$ with $\beta = \Theta$ to rotate the coordinate system about the *y* axis so that the laboratory frame ($z \parallel \mathbf{H}$) is transformed into a coordinate system with $z \parallel \mathbf{k}$ [7]. The Euler angles α and γ in $D^1(\alpha\beta\gamma)$ can be set to zero. They only shift the phases of the circularly polarized radiation components without changing the radiation intensity, because they are incoherent in the case of spontaneous emission. Next, the polarization vectors of the saturating radiation must again be represented in the laboratory coordinate frame:

$$\mathbf{E}_1 = d^1(-\Theta) P_E d^1(\Theta) \mathbf{E},$$

$$\mathbf{E} = \begin{bmatrix} E_{+1} \\ E_0 \\ E_{-1} \end{bmatrix}, \quad P_E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

Here, E_{+1}, E_0, and E_{-1} are the plane waves associated with atomic dipole oscillation and P_E is the projection matrix constructed by taking into account the transverse nature of the electromagnetic wave. Since E_{+1}, E_0, and E_{-1} have random phases in the case of spontaneous emission, the interference between E_{+1}, E_0, and E_{-1} can be neglected. Finally, by virtue of the Maxwellian velocity distribution for the emitting atoms, $|G_M(\Omega, \mathbf{k})|^2$ is expressed as

$$|G_{+1}(\Omega, \mathbf{k})|^2 = I_0 d^2 \left\{ \exp\left[-\left(\frac{\Omega + \Delta}{k V_T} \right)^2 \right] \left(\frac{1 + \cos^2\Theta}{2} \right)^2 \right.$$

$$\left. + \exp\left[-\left(\frac{\Omega}{k V_T} \right)^2 \right] \frac{(\cos\Theta \sin\Theta)^2}{2} \right.$$

$$+ \exp\left[-\left(\frac{\Omega-\Delta}{kV_T}\right)^2\right]\left(\frac{1-\cos^2\Theta}{2}\right)^2\Bigg\},$$

$$|G_0(\Omega, \mathbf{k})|^2 = I_0 d^2\Bigg\{\exp\left[-\left(\frac{\Omega+\Delta}{kV_T}\right)^2\right]\frac{(\cos\Theta\sin\Theta)^2}{2}$$

$$+ \exp\left[-\left(\frac{\Omega}{kV_T}\right)^2\right]\sin^4\Theta \qquad (3)$$

$$+ \exp\left[-\left(\frac{\Omega-\Delta}{kV_T}\right)^2\right]\frac{(\cos\Theta\sin\Theta)^2}{2}\Bigg\},$$

$$|G_{-1}(\Omega, \mathbf{k})|^2 = I_0 d^2\Bigg\{\exp\left[-\left(\frac{\Omega+\Delta}{kV_T}\right)^2\right]\left(\frac{1-\cos^2\Theta}{2}\right)^2$$

$$+ \exp\left[-\left(\frac{\Omega}{kV_T}\right)^2\right]\frac{(\cos\Theta\sin\Theta)^2}{2}$$

$$+ \exp\left[-\left(\frac{\Omega-\Delta}{kV_T}\right)^2\right]\left(\frac{1+\cos^2\Theta}{2}\right)^2\Bigg\}.$$

Here, I_0 is proportional to the integral intensity of spontaneous radiation (determined by the level population N_m and by the corresponding spontaneous emission coefficient) and d is the reduced dipole moment for the m–n transition. Since the distribution $|G_M(\Omega, \mathbf{k})|^2$ is isotropic with respect to azimuthal orientation φ, this expression is independent of φ; $\cos\Theta = \mathbf{k}\cdot\mathbf{H}/kH$. The expressions in (3) are not rigorous with regard to electromagnetic field normalization, but they are sufficiently accurate for quasiclassical description of spatial characteristics, polarization, and spectra of spontaneous emission of an ensemble of atoms in magnetic field. It can readily be shown that the total radiation intensity, $|G_+|^2 + |G_-|^2 + |G_0|^2$, is independent of Θ when $\Delta = 0$, as should be expected in the case of isotropic spontaneous decay. Moreover, the components $|G_+|^2$, $|G_-|^2$, and $|G_0|^2$ do not give rise to any components of nonzero rank in the polarization tensor integrated over the solid angle (see [1, p. 157, Eq. (10.33)]).

The resonance conditions that follow from (3) are independent of the magnetic field magnitude only for a wave propagating along the z axis (in which case $\Theta = 0$). A photon of this kind emitted in a transition involving a change in the magnetic quantum number (with $M_m - M_n = +1$ or -1) is absorbed in a similar transition. The photons emitted at a nonzero angle Θ relative to the z axis in transitions of a certain type (e.g., with $M_m - M_n = +1$) are absorbed in transitions of all types. In a nonzero magnetic field, the radiation components absorbed in "extrinsic" transitions do not satisfy the resonance conditions. This is obvious when the propagation direction is perpendicular to the z axis. In this case, both Zeeman-shifted σ components have similar linear polarizations perpendicular to the magnetic field. These components interact separately with both transitions for which $M_m - M_n = \pm 1$, and only one-half of the radiation satisfies the resonance conditions. The photons polarized parallel to the magnetic field (π component) satisfy the resonance conditions either when $\Theta = 0$ (no radiation) or when $\Theta = \pi/2$ (radiation of the highest intensity). The highest intensity of nonresonant radiation is attained for the π component when $\Theta \approx 51°$.

The expression for the linear part of the work done by the probe field in (1) has been analyzed in the context of various problems (e.g., see [1, p. 252]). It is of interest here only as compared to the nonlinear part. By performing the standard averaging over the longitudinal atom velocity V_z and over Ω (for a large Doppler broadening, i.e., kV_T, $k_\mu V_T \gg \Gamma_{lm}$, Γ_{mn}, and for Ω_μ that are not very large as compared to $k_\mu V_T$), the nonlinear part of the work done by the probe field can be represented by three components:

$$P_\mu^{(1)} = \frac{2\hbar\omega_{lm}|G_\mu|^2 I_0 d^2 N_n\sqrt{\pi}}{(k_\mu V_T)(kV_T)}\exp\left[-\left(\frac{\Omega_\mu+\Delta_\mu}{k_\mu V_T}\right)^2\right]$$

$$\times\Bigg\langle\left(\frac{3}{2}+\frac{3\cos^4\Theta}{2}-\cos^2\Theta\right)\exp(-Z^2)$$

$$+ \cos^2\Theta\sin^2\Theta\Bigg\{\exp\left[-\left(Z-\frac{\Delta}{kV_T}\right)^2\right] \qquad (4)$$

$$+ \exp\left[-\left(Z+\frac{\Delta}{kV_T}\right)^2\right]\Bigg\} + \frac{\sin^4\Theta}{4}\Bigg\{\exp\left[-\left(Z-\frac{2\Delta}{kV_T}\right)^2\right]$$

$$+ \exp\left[-\left(Z+\frac{2\Delta}{kV_T}\right)^2\right]\Bigg\}\Bigg\rangle_{\Theta,\varphi,V_\rho/V_T},$$

where

$$Z = \frac{\Omega_\mu+\Delta_\mu}{k_\mu V_T}\cos\Theta + \frac{V_\rho}{V_T}\sin\Theta\cos\varphi.$$

The first term in angle brackets represents interaction between spontaneous emission and the corresponding transition and is independent of magnetic field. The second and third terms correspond to interactions with the Zeeman-shifted "extrinsic" transitions.

Performing the averaging over the Maxwellian distribution of the transverse velocity components V_ρ (in explicit form), one obtains integrals of the form

$$F(A, B) = \frac{1}{\pi}\int_0^\infty \exp[-(A + Bx)^2 - x^2]x\,dx$$

$$= \frac{1}{2\pi(1 + B^2)}\left\{\exp(-A^2) - \sqrt{\frac{\pi}{1 + B^2}}\right. \tag{5}$$

$$\left.\times AB\exp\left(-\frac{A^2}{1 + B^2}\right)\left[1 - \mathrm{erf}\left(\frac{AB}{\sqrt{1 + B^2}}\right)\right]\right\},$$

where

$$B = \sin\Theta\cos\varphi, \quad A = \frac{\Omega_\mu + \Delta_\mu}{k_\mu V_T}\cos\Theta \pm \frac{N\Delta}{kV_T},$$

$$x = \frac{V_\rho}{V_T}.$$

The factor N can be 0, 1, or 2. For the first (resonant) term in the angle brackets in (4), $N = 0$. For the second and third terms (which represent nonresonant absorption of spontaneous radiation), $N = 1$ and 2, respectively. The averaged over φ nonlinear part of the work done by the field can be expressed in terms of integrals of the form

$$F_\varphi(A, s) = \int_0^{2\pi} F(A, s, \cos\varphi)d\varphi, \quad s = \sin\Theta. \tag{6}$$

The integral in (6) cannot be expressed in terms of elementary functions. However, it can be calculated exactly for $s = 0$ and $s = 1$:

$$F_\varphi(A, 0) = \exp(-A^2), \quad F_\varphi(A, 1) = \frac{\exp(-A^2/2)}{\sqrt{2}}.$$

Since $|s|$ varies from 0 to 1 in (6), good accuracy (within 0.1%) is achieved at intermediate values when (6) is represented by the interpolation formula

$$F_\varphi(A, s) = \exp(-A^2)(1 - 1.506s^2 + 0.506s^4) + \frac{\exp(-A^2/2)}{\sqrt{2}}(1.551s^2 - 0.551s^4). \tag{7}$$

The coefficients in (7) are found by the least-squares method for $|A| < 3$. A final expression for the work done by the probe field is obtained by calculating the integral over the polar angle Θ in (4):

$$P_\mu = -2\hbar\omega_\mu|G_\mu|^2\frac{\sqrt{\pi}}{k_\mu V_T}\exp(-x^2)$$

$$\times\left\{N_m + \frac{N_n I_0 d^2}{kV_T}[F_0(x) + F_1(x, y) + F_2(x, y)]\right\},$$

$$F_0(x) = \int_0^\pi F_\varphi(x\cos\Theta, \sin\Theta)$$

$$\times\left(\frac{3}{2} + \frac{3\cos^4\Theta}{2} - \cos^2\Theta\right)\sin\Theta d\Theta,$$

$$F_1(x, y) = \int_0^\pi [F_\varphi(x\cos\Theta + y, \sin\Theta)$$

$$+ F_\varphi(x\cos\Theta - y, i\sin\Theta)]\cos^2\Theta\sin^3\Theta d\Theta, \tag{8}$$

$$F_2(x, y) = \int_0^\pi [F_\varphi(x\cos\Theta + 2y, \sin\Theta)$$

$$+ F_\varphi(x\cos\Theta - 2y, \sin\Theta)]\frac{\sin^4\Theta}{4}\sin\Theta d\Theta,$$

$$x = \frac{\Omega_\mu + \Delta_\mu}{k_\mu V_T}, \quad y = \frac{\Delta_\mu g_n}{kV_T g_l}.$$

Figure 2 shows the functions $F_0(x)$, $F_1(x, y)$, and $F_2(x, y)$ describing the change in the population of level m due to absorption of spontaneous emission. One can easily find approximate expressions for these functions.

The quantity in braces in Eq. (8) is the total level m population. Since the integral radiation intensity I_0 is proportional to N_m, we can factor N_m out of the braces. Now, we see that the nonlinear correction to the Doppler profile is proportional to the lower level population N_n multiplied by the probability of spontaneous decay from the upper level. Since N_n can be higher than N_m by several orders of magnitude, even spontaneous emission having a low integral intensity can cause a substantial change in the level m population as compared to the initial N_m. In a nonzero magnetic field, the additional population is partially reduced by self-saturation. This effect manifests itself, in particular, by changing the symmetry properties of the magneto-optical profile.

Figure 3a shows the derivatives of magneto-optical spectra calculated in the case when half the population

N_m[1] is created by absorption of spontaneous emission, $k_\mu = k$, and $g_n/g_l = 1$. We see that the asymmetry of magneto-optical profiles increases with Ω_μ. The asymmetry manifests itself by the difference between the low- and high-frequency "tails" of the profiles when

$$0 < \left| \frac{\Omega_\mu}{k_\mu V_T} \right| < 0.5.$$

A numerical analysis shows that the asymmetry is mainly determined by the function $F_2(x, y)$, which is narrower on the scale of Zeeman splitting. The contribution due to $F_1(x, y)$ is less pronounced when $\Omega_\mu \neq 0$. By "switching off" the nonresonant processes associated with $F_1(x, y)$ and $F_2(x, y)$, the asymmetry of the profiles is eliminated and only the barely visible distortion of the Doppler profile due to the function $F_0(x)$ is retained. The nonresonant processes are "switched off" when $g_n/g_l \ll 1$ or $|\Omega_\mu/k_\mu V_T| \gg 1$ (the probe is detuned too far from resonance). When $g_n/g_l > 1$, the symmetry is even more pronounced. It is obvious that the asymmetry can be observed only in the case of significant population transfer by magnetic scanning of the absorption line. Figure 3b demonstrates that the asymmetry cannot be observed if the probe-absorption profile is measured by frequency scanning.

When the polarization of the probe wave is changed from right- to left-hand circular in (8), the sign before Δ_μ must be reversed. This is equivalent to mirror reflection of the graphs shown in Fig. 3a with respect to the vertical axis passing through the point $\Delta_\mu = 0$. When the probe wave is linearly polarized, the magneto-optical absorption profile can be represented as the sum of contributions due to the right- and left-hand polarized components of the probe field. The resulting absorption profile is symmetric, as in the absence of population transfer, but its shape is more complicated. When the probe is detuned too far from resonance, the resulting line has a bimodal profile. Since population transfer modifies the absorption profile without violating its symmetry, the resulting changes are difficult to notice visually and can be found only by numerical analysis.

2.2. A Cascade System with a Degenerate Intermediate Level

Analogous calculations can readily be performed for a system of levels with $J_l = J_n = 0$ and $J_m = 1$. The sum over magnetic sublevels in (1) reduces to a single term associated with the magnetic sublevel of level m that is common to both probe and saturating fields. The corresponding magneto-optical absorption profile is similar to that described by (8), but the functions $F_0(x)$, $F_1(x, y)$, and $F_2(x, y)$ are different:

$$F_{0(010)}(x) = \int_0^\pi F_\varphi(x\cos\Theta, \sin\Theta)\left(\frac{1+\cos^2\Theta}{2}\right)\sin\Theta d\Theta,$$

$$F_{1(010)}(x, y) = \int_0^\pi [F_\varphi(x\cos\Theta + y, \sin\Theta)] \times \frac{\cos^2\Theta\sin^3\Theta}{2} d\Theta,$$

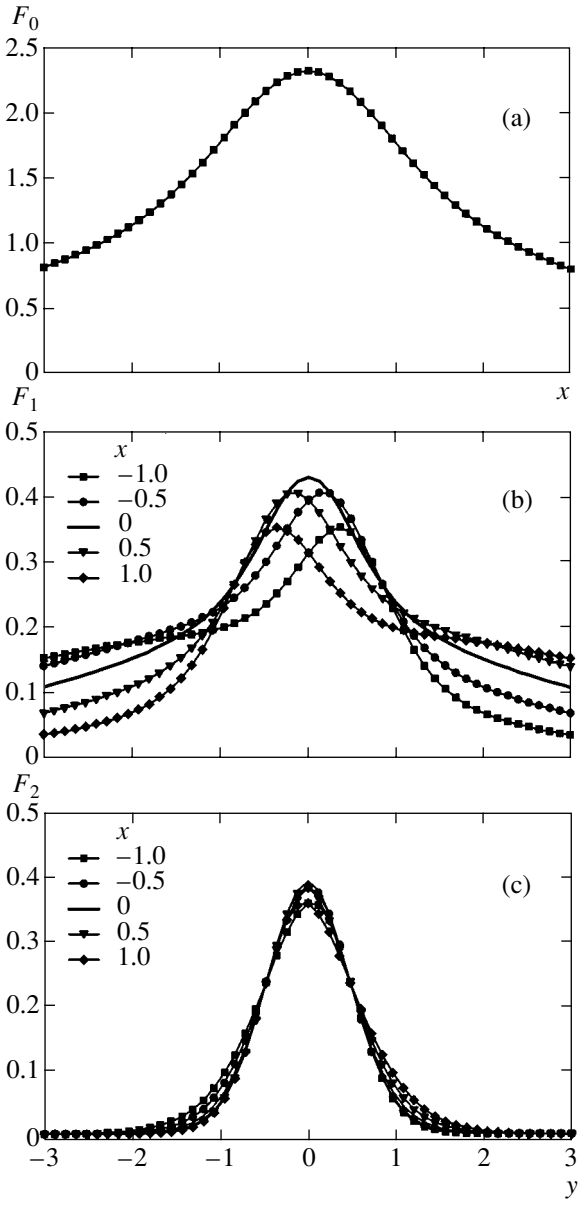

Fig. 2. Approximating functions: (a) $F_0(x)$; (b) $F_1(x, y)$; (c) $F_2(x, y)$.

[1] This value of the additional population due to self-saturation is used to ensure qualitative agreement of the calculated asymmetry with that reported in [5].

$$F_{2(010)}(x, y) = \int_0^{\pi} [F_{\varphi}(x\cos\Theta + 2y, \sin\Theta)] \times \frac{\sin^4\Theta}{4}\sin\Theta d\Theta.$$

The graphs of these functions are qualitatively similar to those presented in Fig. 2 (parenthesized subscripts correspond to the $J = 0 \longrightarrow 1 \longrightarrow 0$ three-level system), except for the relative contribution of the nonresonant term $F_{2(010)}(x, y)$ (as compared to that of $F_{0(010)}(x)$), which is larger (by more than two times). The resulting asymmetry corresponding to an equal population transfer is more pronounced than in the case of a $J = 1 \longrightarrow 0 \longrightarrow 1$ three-level system. Moreover, population transfer is also more pronounced in the $J = 0 \longrightarrow 1 \longrightarrow 0$ system, because both statistical weight and population of the term responsible for spontaneous emission are three times higher. In the case of linear polarization, population transfer does not cause any line asymmetry.

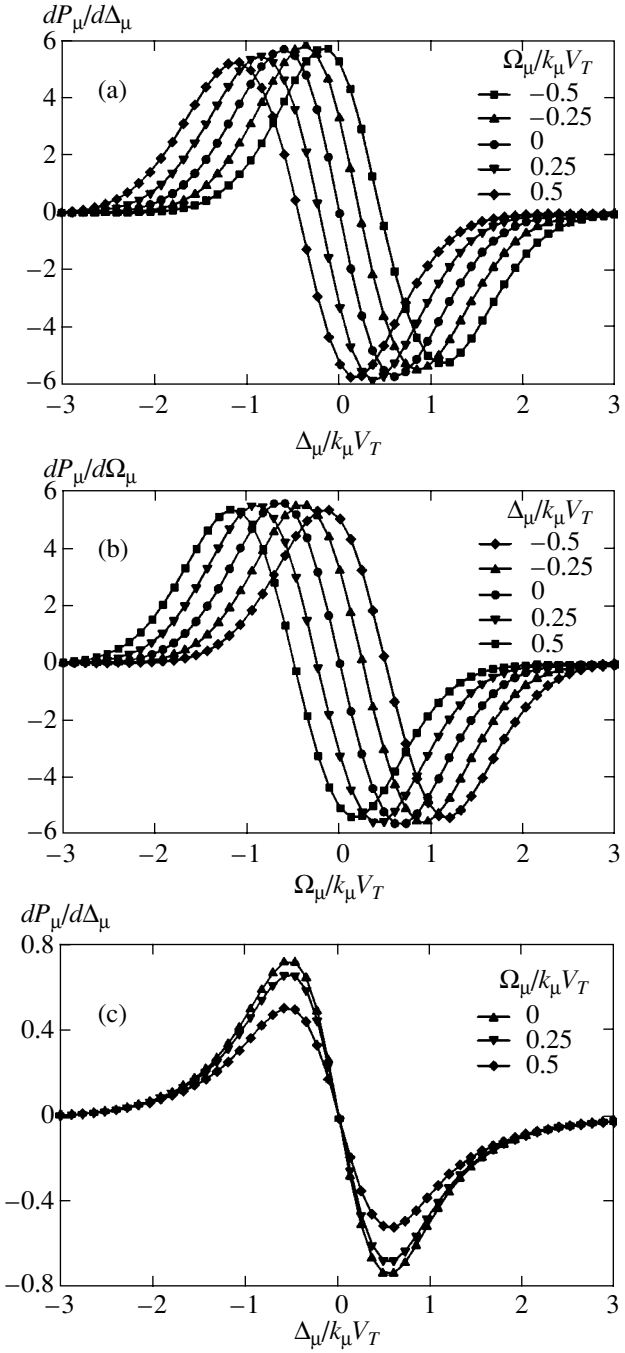

Fig. 3. Derivatives of absorption profiles for probe field: (a) longitudinal magnetic field scanning with various Ω_μ; (b) frequency scanning with various Δ_μ; (c) π-polarized probe wave and transverse orientation of scanned magnetic field.

When the angular moments of the levels are arbitrary and their g-factors are equal, irreducible spherical tensor operators should be employed. The corresponding expressions in (1) would have a similar structure, whereas summing over magnetic sublevels should be replaced by summing over polarizations represented in terms of polarization moments, and a qualitatively similar final result would be obtained. In the case of optically induced population transfer, asymmetric $P_\mu(\Delta)$ would be obtained by scanning the longitudinal magnetic field.

When both angular moments and g-factors of the levels are arbitrary, the situation is more complicated. When g_l differs from g_m (while $J_l, J_m \geq 1$), an asymmetric linear-absorption profile qualitatively similar to (8) will be observed without any population transfer. The linear-absorption profile is a sum over transitions with $M_m - M_l = +1$:

$$P_\mu^{(0)} \propto \sum_M \left|\langle J_l M J_m - (M-1)|11\rangle\right|^2 \times \exp\left[-\left(\frac{\Omega_\mu + \mu_B H[g_l M - g_m(M-1)]}{k_\mu V_T}\right)^2\right] = \sum_M \left|\langle J_l M J_m - (M-1)|11\rangle\right|^2 \times \exp\left[-\left(\frac{\Omega_\mu + \Delta + M\delta\Delta}{k_\mu V_T}\right)^2\right]. \quad (9)$$

Here, $\langle\ldots|\ldots\rangle$ is the Wigner 3j-symbol determining the intensities of transitions for different magnetic quantum numbers,

$$\Delta = \mu_B g_m H, \quad \delta = \frac{g_l - g_m}{g_m}.$$

As a function of magnetic field strength, the sum in (9) consists of several components with different intensi-

ties, which are centered at different values of

$$H_0 = \frac{\Omega_\mu}{\mu_B(g_l M - g_m M + g_m)},$$

and have different widths $k_\mu V_T(1 + \delta M)$. The overall magneto-optical profile is asymmetric when $\Omega_\mu \neq 0$, as in the case of upward population transfer by stimulated transitions, and particular effects can be identified only by analyzing an experimental profile numerically, as in the case of linear polarization.

When $g_l = g_m \neq g_n$, there is no asymmetry in the linear-absorption profile and population transfer from level n, again, breaks the symmetry of the magneto-optical profile. When the Landé factors of levels m and n are different (J_n, $J_m \geq 1$), the asymmetry includes a contribution due to axially propagating radiation, because the work done by the probe field is given by an expression analogous to (8), which holds for waves propagating along the z axis. When $\Theta = 0$, the work done by the field contains terms of the form

$$\exp\left[-\left(\frac{\Omega_\mu + \Delta}{k_\mu V_T}\cos\Theta + \frac{V_\rho}{V_T}\sin\Theta\cos\varphi \pm \frac{M\delta\Delta}{kV_T}\right)^2\right],$$

which are analogous to the terms associated with nonaxial radiation in (4). The resulting level m population depends on the magnetic field strength, and P_μ is asymmetric when $\Omega_\mu \neq 0$. The contribution of the asymmetric distortion in the case of an axially propagating wave obviously increases with the difference between the g-factors of the levels involved in the transition. Moreover, the result depends on the angular momenta of the levels. In particular, the asymmetry is less pronounced when $J_l = 1$, $J_m = 2$, and $J_n = 1$ as compared to the case of $J_l = 1$, $J_m = 2$, and $J_n = 2$. This is explained by the difference between the transition probabilities corresponding to different M_m. When an axially propagating wave is absorbed in an "extrinsic" transition (in the case of emission with a certain M_m and off-resonance absorption by transition with a different M_m), the effect of "extrinsic" radiation on the system of levels with $J_l = 1$, $J_m = 2$, and $J_n = 2$ is stronger than that observed in the case of $J_l = 1$, $J_m = 2$, and $J_n = 1$.

The magneto-optical profile for the $J = 1/2 \longrightarrow 1/2 \longrightarrow 1/2$ three-level system must be weakly asymmetric. In this case, the expression for P_μ does not contain any function analogous to $F_2(x, y)$ (which is mainly responsible for the asymmetry).

3. SATURATION IN TRANSVERSE MAGNETIC FIELD

The results of calculations can easily be extended to the case of transverse magnetic field with respect to $\mathbf{k}_\mu$. By virtue of the assumed isotropy of the saturating spontaneous emission, the absolute direction of magnetic field is irrelevant with regard to the additional population due to self-saturation. However, as the magnetic field orientation is varied relative to the wave vector of the probe field,[2] the additional population manifests itself against the varying linear-absorption profiles observed in the absence of population transfer. This implies a greater diversity in choosing the polarization of the probe field. When the electric field of the probe wave is parallel to the magnetic field (in the case of π polarization), the probe field interacts with a transition whose frequency is independent of the magnetic field strength. Absorption of the probe wave depends on magnetic field only in the case of magnetic-field-dependent population transfer to level m, and the resulting magneto-optical profile of probe-wave absorption is completely determined by stimulated transfer processes. The final expression for the work P_μ done by the probe field is equivalent to (9) with a zero g-factor of level l. The variables x and y in the corresponding functions $F_0(x)$, $F_1(x, y)$, and $F_2(x, y)$ are

$$x = \frac{\Omega_\mu}{k_\mu V_T}, \quad y = \frac{\Delta}{kV_T}.$$

Figure 3c shows magneto-optical profiles for the π component of the probe field. Variation of Ω_μ affects only the amplitudes of magneto-optical profiles, but does not lead to asymmetric distortion. The case when the electric field of the probe wave is orthogonal to the magnetic field (σ-polarized) is analogous to that of longitudinal magnetic field and a linearly polarized probe wave. The resulting profile P_μ of probe-wave absorption is symmetric irrespective of population transfer from level n and bimodal when the mismatch is large. When the probe wave is circularly polarized, a symmetric overall profile will also be observed. In the last two cases, stimulated population transfer cannot be identified qualitatively, and a detailed numerical analysis of the profiles is required. Note also that difference between the g-factors of levels does not lead to profile asymmetry in transverse magnetic field, but it affects the line width.

Measurement of the π polarization of the probe wave in transverse magnetic field would seem to be the most effective method for detecting stimulated population transfer against zero background. However, special care should be taken to ensure both the purity of polarization and the absence of any contribution to the magneto-optical profile due to the probe component polarized perpendicular to the magnetic field. When the probe wave propagates along the axis of an oblong cell, the use of transverse magnetic field may be advantageous at low pressures, when spontaneous emission is anisotropic and stimulated transitions can be induced by its longitudinal component.

4. MULTILEVEL SYSTEMS

Now, we discuss some spectroscopic manifestations of self-saturation of optical transitions in real multilevel

[2] Normally, its direction is parallel to the axis of the discharge tube.

systems. The case of probing of absorption in transitions from resonant levels provides the closest analogy to three-level systems of the kind considered in this study. In the case of neon, the levels in question are $1s_2$ and $1s_4$ in Paschen's notation. (Note that profile asymmetry due to this effect was originally observed in neon [5], which has always been a benchmark medium in nonlinear spectroscopy.) The mean free path of the photons emitted in spontaneous decay of these levels is short, and the spontaneous emission is isotropic even in capillary discharge tubes at actual neon pressures. When processing magneto-optical spectra to obtain quantitative data, one should bear in mind that these photons are strongly reabsorbed and the distorted profile due to the additional population created by stimulated transitions is non-Gaussian, whereas Doppler profiles with widths determined by the discharge temperature can be used to describe "normal" absorption. One should also allow for the modification of linear absorption due to the influence of magnetic field on electron temperature and concentration [8]. However, this effect is characterized by a different dependence on the magnetic field strength, which can be taken into account when the observed profiles are analyzed.[3] Probing of absorption in transitions from the metastable level $1s_5$ can be suggested as a benchmark test to distinguish between these two effects. Since this experiment does not involve spontaneous decay, upward population transfer induced by spontaneous emission is impossible.

In the case of absorption in transitions from higher $2p$ levels, the additional population due to nonlinear effects will manifest itself in a more complicated manner. One reason is that each of these levels (except for $2p_9$) decays to two or three ones. For example, the level $2p_4$ involved in the neon $3s_2$–$2p_4$ transition, with a virtually normal Zeeman splitting ($g_l = 1.302$, $g_m \approx 1.298 \approx g_l$), can decay to the lower states $1s_2$, $1s_4$, and $1s_5$. The populations of these states exceed that of level $2p_4$ by several orders of magnitude [6]. This implies that expression (9) must involve several N_n such that $N_n \gg N_m$, but the corresponding profiles have different widths on the scale of Zeeman splitting because of difference in the Landé factors of $1s$ levels and the Doppler widths of transitions. It is clear that the recorded signal will also reflect the magnetic-field-dependent additional population induced on resonant levels $1s_2$ and $1s_4$ by stimulated transitions and transferred upwards by inelastic collisions with electrons. This effect complicates the profile shape, at the same time increasing the magnetic-field-dependent additional population on these levels.

When discussing the influence of magnetic field on level populations in gas-discharge plasmas, one cannot ignore the change in level populations due to latent alignment of levels in a coordinate system tied to a moving atom (the Kallass–Chaika effect) [9]. This phenomenon, as well as the one considered above, should also be attributed to the σ-polarized field components. However, it manifests itself on the scale of Zeeman splitting corresponding to level widths, because it results from interference in coherent interactions between the field components and Zeeman sublevels. Latent alignment is characterized by opposite signs of alignment for slow- and fast-moving atoms. The Doppler line widths associated with the latently aligned levels change accordingly. A weak magnetic field that splits the levels and breaks the alignment restores the natural line width, and the corresponding change in reabsorption coefficients results in magnetic-field dependence of level populations (including those of degenerate levels). However, this effect can easily be distinguished from the one discussed in this paper because of their disparity on the scale of Zeeman splitting (in terms of level width compared to the Doppler line width).

5. CONCLUSIONS

It should be reiterated that the influence of self-saturation of transitions on the absorption line profile can be detected only when the line is split by magnetic field. Profile asymmetry, as well as other manifestations of stimulated population transfer in the work done by the field, can be observed only on the scale of Zeeman splitting. Frequency scanning does not create conditions under which stimulated transitions can be observed, and the spectral profile retains its shape. When the spectrum is scanned on a frequency scale, self-saturation manifests itself only by a change in the absorption profile amplitude. Spectral frequency scanning in a nonzero magnetic field will obviously change the absorption line profile. However, this change will be similar to those normally caused by magnetic field, whereas self-saturation will manifest itself only in the changed profile amplitude. Therefore, no information about self-saturation can be obtained by conducting an experiment of this kind only.

Since asymmetry of the magneto-optical profile is a qualitative effect, even a small extent of population transfer can be detected by special methods designed to identify the asymmetric part of an absorption coefficient. When self-saturation of transitions is ignored, both spectral line profiles and cross sections of excitation by electron impact may be determined incorrectly and the dependence of discharge characteristics on current may be misinterpreted. One example of the essential role played by the effect in question can be found in [5]. In [10], a new radiative process was predicted: (spontaneous or stimulated) transfer of optical coherence (dipole moment) from one atomic transition to another. It manifests itself in asymmetry of Doppler line profiles. One would naturally try to identify the resulting change in a line profile (several tenths of per-

[3] This dependence is characterized by a profile wider than the Doppler one and by opposite signs corresponding to magnetic fields parallel and perpendicular to the discharge axis.

cent in the case of spontaneous transfer) by a magnetic scanning method. The experiments described in [5] were conducted to validate this approach by applying it to a transition unaffected by transfer of optical coherence. However, an asymmetry (about 5%) due to self-saturation was revealed, which thwarted all hopes for detecting transfer of optical coherence by magnetic scanning.

Note that the influence of self-saturation on the results of magnetic scanning is only part of the overall effect. An analysis of the two-dimensional model outlined in the Introduction shows that the additional population detected by a magnetic scanning method is only one-fourth of the total increase in population due to stimulated transitions. The overall effect does not include the contribution of the π component, which is responsible for one-half of line intensity, while the effect due to the σ components manifests itself by the difference in the saturation caused by their total intensity and half-intensity. In practice, the influence of magnetic field on level populations may be further reduced by effects associated with thermodynamic equilibrium.

It should be expected that self-saturation of optical transitions would manifest itself in the absence of magnetic field. To observe the most clear-cut evidence of self-saturation, one should look for a difference in population between resonant and metastable levels. However, in addition to excitation by electron impact and de-excitation by spontaneous emission, imprisonment of resonant radiation must also play an important role in increasing resonant-level populations (e.g., see [6]). This mechanism of optical pumping described in the literature differs from the self-saturation of optical transitions discussed in this paper. It is associated with increase in the effective level lifetimes under conditions of radiation imprisonment. The effective lifetime of a state with respect to a resonant transition depends on transition probability, ground-state population, and cell length. These parameters are also important for the effect discussed here. In particular, cell length determines the characteristic radiation-imprisonment time, i.e., its intensity [2]. Of primary importance is the population in the excited state, but this parameter is never mentioned in the available explanations of the increase in level lifetime due to radiation imprisonment. As the population in an emitting state increases with electron concentration (discharge current), this state must be additionally populated by increasing self-saturation. However, radiation imprisonment can only decrease the population through the reduction of the lifetimes of excited states due to increase in frequency of inelastic collisions with electron concentration.[4]

Thus, self-saturation must manifest itself in experiments with variable discharge current. Indeed, some evidence of this effect was found in data concerning the cross sections of excitation of resonant neon states by electron impact (see [6, Table 3]), whereas no such effect was observed in analogous data for metastable states, for which self-saturation is impossible [6, Table 4]. The dependence on discharge current revealed in [6] for resonant levels (increase in cross sections, except for the measurements performed at a minimal current) was attributed to incorrect probe measurements of electron concentration and temperature, while its absence in the case of metastable states was interpreted as accidental. However, we believe that the author should not have questioned the measurement accuracy. The error (if any) would be significant only at the minimal discharge current. We are certain that self-saturation of resonant transitions manifested itself in the experiments reported in [6].

ACKNOWLEDGMENTS

We thank S.G. Rautian and A.M. Shalagin for helpful discussions. This work was supported under the State Program "Universities of Russia," project no. UR.01.01.054, and by the Russian Foundation for Basic Research, project no. 02-02-17923.

REFERENCES

1. S. G. Rautian, G. I. Smirnov, and A. M. Shalagin, *Nonlinear Resonances in Atomic and Molecular Spectra* (Nauka, Novosibirsk, 1979).
2. L. A. Vaĭnshteĭn, V. R. Mironenko, S. G. Rautian, and É. G. Saprykin, Opt. Spektrosk. **87**, 372 (1999) [Opt. Spectrosc. **87**, 341 (1999)].
3. S. G. Rautian and É. G. Saprykin, Opt. Spektrosk. **92**, 385 (2002) [Opt. Spectrosc. **92**, 342 (2002)].
4. É. G. Saprykin, S. N. Seleznev, and V. A. Sorokin, Pis'ma Zh. Éksp. Teor. Fiz. **76**, 322 (2002) [JETP Lett. **76**, 264 (2002)].
5. I. V. Barybin, V. A. Sorokin, and A. E. Churin, Opt. Spektrosk. **95** (6), 933 (2003) [Opt. Spectrosc. **95**, 873 (2003)].
6. S. É. Frish, in *Spectroscopy of the Gas-Discharge Plasma* (Nauka, Leningrad, 1970), p. 244.
7. L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Mechanics* (Addison-Wesley, Reading, Mass., 1981; Mir, Moscow, 1984).
8. V. L. Granovskiĭ, *Electric Current in a Gas* (Nauka, Moscow, 1971), p. 462.
9. M. P. Chaĭka, *Interference of Degenerate Atomic States* (Leningr. Gos. Univ., Leningrad, 1975).
10. S. G. Rautian, Pis'ma Zh. Éksp. Teor. Fiz. **61**, 461 (1995) [JETP Lett. **61**, 473 (1995)]; Zh. Éksp. Teor. Fiz. **108**, 1186 (1995) [JETP **81**, 651 (1995)].

Translated by A. Betev

[4] Note that radiation imprisonment can be increased to the same extent by increasing either the concentration of absorbing atoms or the reactor size. In the former case, the mean free path of spontaneously emitted photons decreases, and so does the characteristic radiation-imprisonment time. This reduces the effect of self-saturation when the population of the emitting state remains constant. In the latter case, the converse effect may be observed.

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1112–1122.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1244–1254.

ATOMS, SPECTRA, RADIATION

Radiofrequency Muon Depolarization in the Muonium + Nuclear Spin System

S. A. Moiseev and V. G. Nikiforov

Zavoĭskiĭ Physicotechnical Institute, Kazan Scientific Center, Russian Academy of Sciences, Sibirskiĭ trakt 10/7, Kazan, 420029 Tatarstan, Russia

e-mail: moiseev@kfti.knc.ru

Received March 26, 2003

Abstract—The effect of a considerable strengthening of muon depolarization in ALC resonance experiments was predicted for the muonium + nuclear spin system in the presence of a radiofrequency field. A mathematical approach was developed for obtaining analytic solutions that described the muon spin dynamics in ALC experiments, including a particular exact solution that contained much information about the system studied in fairly low magnetic radiofrequency fields. An analysis of these solutions and numerical calculations allowed us to comprehensively analyze muon depolarization patterns in a radiofrequency field. The results reveal the potential of muon depolarization strengthening for considerably increasing the sensitivity of experimental studies of muonium interactions with neighboring nuclear spins and for obtaining new spectroscopic information. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

When implanted into a substance, the positively charged muon (μ^+) often captures an electron to produce hydrogen-like paramagnetic muonium ($\mu^+ - e^-$). The properties of muonium (Mu) attract much interest in relation to the fundamental problems of quantum electrodynamics and weak interactions and to testing of the standard model (e.g., see [1, 2]). Review [1] contains a consistent analysis of electromagnetic and weak interactions in muonium that determine the hyperfine splitting of Mu levels in vacuum and the effective magnetic moments of the electron and muon, which were measured most accurately by magnetic resonance techniques (also see [3]). Studies of hyperfine interactions and the spin dynamics of Mu are also of importance for experimentally determining the molecular structure and physicochemical properties of various substances [4–6]. The formation of Mu is to a great extent determined by Coulomb interactions between the electron and muon and does not strongly influence the initial polarization of the muon and electron. Thanks to this circumstance, we can study the spin dynamics of the muon using the muon spin rotation (μSR) technique to obtain information about the spectroscopic parameters of Mu. Of considerable interest are hyperfine interactions of Mu with neighboring nuclei, which provide important information about atoms and chemical bonds [4–6]. The close similarity of the physicochemical properties of the muonium and hydrogen atoms allow muonium to be used as a model of impurity hydrogen atoms in various compounds [7].

One of the frequently used methods for studying the spin interactions between Mu and surrounding atoms (Mu + Nu) is based on using cross-relaxation processes with polarization transfer from the muon to nuclear spins via hyperfine interactions with electron spins [8]. Abragam was the first to note that such a transfer of muon polarization was possible [9]; soon afterward, this phenomenon was observed experimentally [10]. Muon polarization transfer to a nucleus occurs under the conditions of quasi-crossing of energy levels of the total quantum system [avoided level crossing (ALC) resonance] including the muon, electron, and nuclear spins. These conditions are attained by additionally tuning the static magnetic field H_z. Because of the fairly short muon lifetime, the ALC resonance can only be successfully observed when a muon is rapidly depolarized; that is, when hyperfine interactions responsible for polarization transfer are strong. Interpreting the results of such experiments requires the nuclei that interact with the muonium be accurately identified. Currently, the unknown spectral parameters of nuclear spins are determined from other experiments. In spite of considerable progress in the μSR techniques of ALC resonance measurements, they are likely incapable of providing information comparable to the amount of that obtained by spectroscopy with the use of stationary and, especially, pulsed magnetic resonance techniques (e.g., see [11, 12]). To obtain more spectroscopic information about the muonium + nuclear spin system, we here suggest using additional irradiation of substances by a radiofrequency field in ALC experiments.

It is unclear a priori how a radiofrequency field can affect the muon spin dynamics and ALC signal parameters, because the Mu + Nu quantum system acquires fairly diverse properties under these conditions. The Mu + Nu system is characterized by the presence of a

large number of quantum levels, spin quantum coherence at the instant when the Mu + Nu system forms, and, lastly, the appearance of entangled quantum states, which play a key role in polarization transfer by the Abragam mechanism. Briefly announcing the results of this work, note that the action of a radiofrequency field on the Mn + Nu system causes several transitions between four quantum levels and, together with cross-relaxation processes, strengthens the depolarization of the muon subsystem. The influence of a radiofrequency field on the spin dynamics at the ALC resonance point cannot be described by simple magnetic resonance models based on applying the two-level approach to spectroscopic transitions in an external radiofrequency field. In the problem under consideration, in which an electromagnetic field strongly couples all four Mu + Nu system levels, we found an analytic solution that fairly accurately describes the dynamics of the behavior of this system and the shape of the ALC resonance line. This analytic solution, in particular, exactly describes the system at the center of the ALC resonance. The numerical analysis that we performed showed that the exact solution contained much information about the magnitude of the effect. In particular, a radiofrequency field can considerably increase the amplitude of the ALC signal, which opens up possibilities of observing ALC spectra in substances with weak hyperfine interactions, which are exceedingly difficult or virtually impossible to study by the traditional techniques of ALC experiments. This result can be of great practical significance, for it offers a means of increasing the amplitude of ALC signals and, accordingly, the sensitivity of ALC measurements and of obtaining additional information about the spectroscopic parameters of the muonium and nuclear spins interacting with it.

2. PHYSICAL MODEL

Our study of the influence of a radiofrequency field on the ALC resonance signal is based on the simplest quantum model of the interaction of muonium with one nuclear spin (Mu + Nu($S = 1/2$)). Nevertheless note that the selected model can serve as a basis for describing the process under consideration in more complex Mu + Nu(1/2) systems, in particular, those with a larger number of particles and particles with larger spins [8]. The Hamiltonian of the selected model will be written as

$$H = H_0 + V(t), \tag{1}$$

where H_0 is the energy of the muonium + nuclear spin system,

$$\begin{aligned} H_0 = {} & \hbar\boldsymbol{\sigma}\cdot\tilde{A}_{\mu e}\cdot\boldsymbol{\tau} - \hbar\omega_\mu\sigma_z + \hbar\omega_e\tau_z \\ & - \hbar\omega_n S_n + \hbar\mathbf{S}\cdot\tilde{A}_{ne}\cdot\boldsymbol{\tau}, \end{aligned} \tag{2}$$

and $V(t)$ is the interaction energy of three spins with the magnetic radiofrequency radiation field H_1,

$$\begin{aligned} V(t) &= -\hbar\gamma_\Sigma S^\Sigma H_1\cos(\omega t), \\ \gamma_\Sigma S^\Sigma &\equiv \gamma_\mu\sigma_x + \gamma_n S_x - \gamma_e\tau_x. \end{aligned} \tag{3}$$

Here, $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, and $\mathbf{S}$ are the spin operators of the muon, electron, and nucleus, respectively; $\tilde{A}_{\mu e}$ and $\tilde{A}_{ne}$ are the tensors of the hyperfine interaction between the muon spin and the electron and nuclear spins; $\omega_\mu = \gamma_\mu H_z$, $\omega_e = \gamma_e H_z$, and $\omega_n = \gamma_n H_z$ are the Zeeman frequencies of the muon, electron, and nucleus, respectively, in the external constant magnetic field H_z oriented along the z axis; H_1 is the amplitude of the radiofrequency magnetic field; and γ_μ, γ_n, and γ_e are the gyromagnetic ratios of the muon, nucleus, and electron. Below, we consider the fairly frequently encountered situation of axially symmetrical hyperfine interactions, for which $A_{zz} = A^{\uparrow\uparrow}$ and $A_{xx} = A_{yy} = A^\perp$. We then have

$$\hbar\boldsymbol{\sigma}\cdot\tilde{A}_{\mu e}\cdot\boldsymbol{\tau} = \hbar\left\{A^{\uparrow\uparrow}_{\mu e}\sigma_z\tau_z + \frac{1}{2}A^\perp_{\mu e}(\sigma^+\tau^- + \sigma^-\tau^+)\right\}, \tag{4}$$

$$\hbar\mathbf{S}\cdot\tilde{A}_{ne}\cdot\boldsymbol{\tau} = \hbar\left\{A^{\uparrow\uparrow}_{ne}S_z\tau_z + \frac{1}{2}A^\perp_{ne}(S^+\tau^- + S^-\tau^+)\right\}. \tag{5}$$

Isotropic hyperfine interactions will be described using the notation

$$A^{\uparrow\uparrow}_{\mu e} = A^\perp_{\mu e} = \omega_0, \quad A^{\uparrow\uparrow}_{ne} = A^\perp_{ne} = \Omega.$$

We have $\omega_0 \gg \Omega$ in order of magnitude; for instance, for silicon, $\omega_0 = 2\pi \times 2006$ MHz, and for iron, $\Omega(^{19}\text{F spin } 1/2) \approx 100$ MHz [8].

Following [8], we will first describe the most important properties of the Mu + Nu(1/2) system in the external constant magnetic field H_z. Tuning to the ALC resonance requires using high H_z fields, when the Zeeman electron energy in the magnetic field ($\hbar\omega_e$) becomes predominant in the series of energies (1). For this reason, far from level crossing, the cross-relaxation operators of hyperfine interactions from (4) and (5),

$$\begin{aligned} V^{(+,-)}_{\mu,e} &= \frac{1}{2}\hbar A^\perp_{\mu e}(\sigma^+\tau^- + \sigma^-\tau^+), \\ V^{(+,-)}_{n,e} &= \frac{1}{2}\hbar A^\perp_{ne}(S^+\tau^- + S^-\tau^+), \end{aligned}$$

have a weak influence on the spin dynamics and can be included as perturbations. Let us determine energy levels using the spin state functions $|m_\mu, m_e, m_n\rangle$, where m_μ, m_e, and m_n are the magnetic quantum numbers of the muon, electron, and nucleus, as basis functions. At

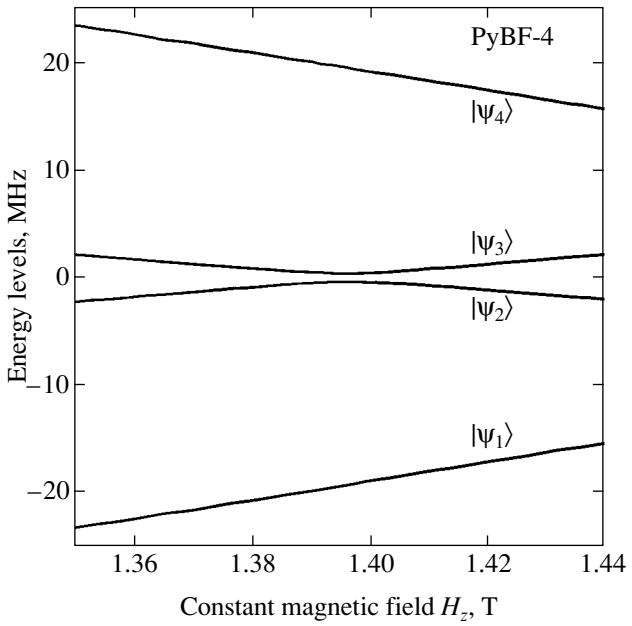

Fig. 1. Energy levels (see Table 2) in the vicinity of the ALC resonance for the example of PyBF-4, $H_{\text{res}} = 1.396$ T.

$\omega_0 \gg \Omega$, we must primarily take into account the hyperfine electron–muon spin interaction described by the $V_{\mu,e}^{(+,-)}$ term, which entangles states with opposite muon and electron spin orientations ($|\uparrow_e\downarrow_\mu\rangle$, $|\downarrow_e\uparrow_\mu\rangle$). For a small $\omega_0/\omega_e < 1$, the energies of eight possible states (E_{1+}, … , E_{4-}) far from level crossing are largely determined by the terms of zeroth order in the interaction $V_{n,e}^{(+,-)}$ (see Table 1). The four energy levels E_{3+}, …, E_{4-} with the opposite predominant electron spin orientation $|\downarrow_e\rangle$ are separated by a considerable energy interval (around $\hbar\omega_e > \hbar\omega_0$). For this quartet of levels, the alternating field frequency does not coincide with the resonance transition frequencies of Mu + Nu that cause muon spin flip. For this reason, the spin dynamics for this quartet of levels will not be considered.

Close to the ALC resonance at the constant magnetic field value

$$H_{\text{res}} = \frac{1}{2}\frac{A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}}{\gamma_\mu - \gamma_n},$$

the energies of the $|\varphi_{1-}\rangle$ and $|\varphi_{2+}\rangle$ levels coincide. In the vicinity of $H_z \approx H_{\text{res}}$, the weak hyperfine electron–nucleus interaction $V_{n,e}^{(+,-)}$ begins to play an important role; it entangles states with oppositely oriented electron and nucleus spins ($|\uparrow_e\downarrow_n\rangle$, $|\downarrow_e\uparrow_n\rangle$) and thereby entangles the states $|\varphi_{1-}\rangle$ and $|\varphi_{2+}\rangle$. The new system of wave functions and the corresponding energy eigenval-

Table 1. Wave functions and their energies far from the ALC resonance

Wave function	Energy level in $\hbar$ units
$\|\varphi_{1+}\rangle = \|\uparrow_\mu\uparrow_e\uparrow_n\rangle$	$E_{1+} = \frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} + A_{ne}^{\uparrow\uparrow}) + \omega_- - \frac{1}{2}\omega_n$
$\|\varphi_{1-}\rangle = \|\uparrow_\mu\uparrow_e\downarrow_n\rangle$	$E_{1-} = \frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}) + \omega_- + \frac{1}{2}\omega_n$
$\|\varphi_{2+}\rangle = \sin\xi\|\uparrow_\mu\downarrow_e\uparrow_n\rangle + \cos\xi\|\downarrow_\mu\uparrow_e\uparrow_n\rangle$	$E_{2+} \approx -\frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}) + \omega_+ - \frac{1}{2}\omega_n$
$\|\varphi_{2-}\rangle = \sin\xi\|\uparrow_\mu\downarrow_e\downarrow_n\rangle + \cos\xi\|\downarrow_\mu\uparrow_e\downarrow_n\rangle$	$E_{2-} \approx -\frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} + A_{ne}^{\uparrow\uparrow}) + \omega_+ + \frac{1}{2}\omega_n$
$\|\varphi_{3+}\rangle = \|\downarrow_\mu\downarrow_e\uparrow_n\rangle$	$E_{3+} = \frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}) - \omega_- - \frac{1}{2}\omega_n$
$\|\varphi_{3-}\rangle = \|\downarrow_\mu\downarrow_e\downarrow_n\rangle$	$E_{3-} = \frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} + A_{ne}^{\uparrow\uparrow}) - \omega_- + \frac{1}{2}\omega_n$
$\|\varphi_{4+}\rangle = \cos\xi\|\uparrow_\mu\downarrow_e\uparrow_n\rangle - \sin\xi\|\downarrow_\mu\uparrow_e\uparrow_n\rangle$	$E_{4+} \approx -\frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} + A_{ne}^{\uparrow\uparrow}) - \omega_+ - \frac{1}{2}\omega_n$
$\|\varphi_{4-}\rangle = \cos\xi\|\uparrow_\mu\downarrow_e\downarrow_n\rangle - \sin\xi\|\downarrow_\mu\uparrow_e\downarrow_n\rangle$	$E_{4-} \approx -\frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}) - \omega_+ + \frac{1}{2}\omega_n$

Note: $\cot 2\xi = (\omega_\mu + \omega_e)/A_{\mu e}^{\perp}$.

ues [8] are listed in Table 2, where the following notation is used:

$$\omega' = \frac{1}{4}(A_{\mu e}^{\uparrow\uparrow} + A_{ne}^{\uparrow\uparrow}) - \frac{1}{2}\omega_\mu - \frac{1}{2}\omega_n, \quad (6)$$

$$\omega_x = \omega_\mu - \omega_n - \frac{1}{2}(A_{\mu e}^{\uparrow\uparrow} - A_{ne}^{\uparrow\uparrow}), \quad (7)$$

$$\cot 2\alpha = \frac{\omega_x}{\omega_G}, \quad (8)$$

$$\omega_G = \frac{2}{\hbar}\langle \varphi_{1-}|H|\varphi_{2+}\rangle = A_{ne}^{\perp}\sin\xi, \quad (9)$$

$$\omega_{GX} = \sqrt{\omega_x^2 + \omega_G^2}. \quad (10)$$

In the particular case of maximum entangling of the electron and nuclear spin states, which takes place when the magnetic field is tuned to the condition $E_{1-} = E_{2+}$ ($\omega_x = 0$, $\alpha = \pi/4$), the $|\psi_3\rangle$ and $|\psi_2\rangle$ wave functions take the simpler form

$$|\Psi_3\rangle = (1/2)^{1/2}\{|\varphi_{1-}\rangle + |\varphi_{2+}\rangle\}, \quad (11)$$

$$|\Psi_2\rangle = (1/2)^{1/2}\{|\varphi_{1-}\rangle - |\varphi_{2+}\rangle\}, \quad (12)$$

the difference of the energy levels of the $|\psi_3\rangle$ and $|\psi_2\rangle$ states becomes

$$E_{3-} - E_2 = \hbar\omega_G$$

(see Fig. 1), and the ALC resonance signal reaches a maximum.

We assume that the experiment is conducted at room temperature, when, in paramagnetic systems, the initial polarization of the electronic and nuclear systems that form a chemical compound with the muon is virtually absent. Under these conditions, the spin density matrix of the ensemble of quantum Mu + Nu(1/2) systems at the initial time of muonium formation becomes

$$\rho(0) = \frac{1}{4}|\uparrow_\mu\rangle\langle\uparrow_\mu| \otimes \{|\downarrow_e\rangle\langle\downarrow_e| + |\uparrow_e\rangle\langle\uparrow_e|\} \otimes \{|\uparrow_n\rangle\langle\uparrow_n| + |\downarrow_n\rangle\langle\downarrow_n|\}. \quad (13)$$

Muonium atoms that arise with the opposite initial orientation of the electron spin $|\downarrow_e\rangle$ determine the nonoscillating background ($P_{z,\downarrow e}^{\mu}(t)$) of muon polarization, because at the given initial electron spin state, the transfer of muon polarization, like the influence of the radiofrequency field (see comments to Table 1), is virtually suppressed,

$$P_{z,\downarrow e}^{\mu}(t) = \frac{1}{2}\exp\left(-\frac{t}{\tau_\mu}\right) \quad (14)$$

Table 2. Wave functions and their energies close to the ALC resonance

Wave function	Energy level in $\hbar$ units
$\|\Psi_4\rangle = \|\varphi_{1+}\rangle$	$E_4 = \omega' + \frac{1}{2}\omega_e$
$\|\Psi_3\rangle = \sin\alpha\|\varphi_{1-}\rangle + \cos\alpha\|\varphi_{2+}\rangle$	$E_3 = \frac{1}{2}\omega_{GX} + \frac{1}{2}\omega_e$
$\|\Psi_2\rangle = \cos\alpha\|\varphi_{1-}\rangle - \sin\alpha\|\varphi_{2+}\rangle$	$E_2 = -\frac{1}{2}\omega_{GX} + \frac{1}{2}\omega_e$
$\|\Psi_1\rangle = \|\varphi_{2-}\rangle$	$E_1 = -\omega' + \frac{1}{2}\omega_e$

($\tau_\mu \approx 2.19703(4)$ μs is the muon lifetime [13]). Let us rewrite (13) only leaving the density matrix related to the initial electron spin state $|\uparrow_e\rangle$, which is responsible for the appearance of the ALC signal,

$$\rho(0) = \frac{1}{4}|\uparrow_\mu\rangle\langle\uparrow_\mu| \otimes |\uparrow_e\rangle\langle\uparrow_e| \otimes \{|\uparrow_n\rangle\langle\uparrow_n| + |\downarrow_n\rangle\langle\downarrow_n|\}. \quad (15)$$

In what follows, initial condition (15) is used to consider the influence of a radiofrequency field on spin polarization near the ALC resonance.

3. THE INFLUENCE OF A RADIOFREQUENCY FIELD ON MUON POLARIZATION: ANALYTIC AND NUMERICAL SOLUTIONS

The wave functions given in Tables 1 and 2 can be used to find the matrix elements of the generalized magnetic moment $(\gamma_\Sigma S_x^\Sigma)_{nk} = \langle\Psi_n|\gamma_\Sigma S_x^\Sigma|\Psi_k\rangle$, which determines the strength of transitions in an external alternating magnetic field of radiofrequency radiation,

$$(\gamma_\Sigma S_x^\Sigma)_{mn} = (\gamma_\Sigma S_x^\Sigma)_{nm}, \quad (16.1)$$

$$(\gamma_\Sigma S_x^\Sigma)_{12} = -\frac{1}{2}\sin\alpha\gamma_n + \frac{1}{2}\cos\alpha\gamma_{\mu e}(\xi), \quad (16.2)$$

$$(\gamma_\Sigma S_x^\Sigma)_{13} = \frac{1}{2}\cos\alpha\gamma_n + \frac{1}{2}\sin\alpha\gamma_{\mu e}(\xi), \quad (16.3)$$

$$(\gamma_\Sigma S_x^\Sigma)_{14} = (\gamma_\Sigma S_x^\Sigma)_{23} = 0, \quad (16.4)$$

$$(\gamma_\Sigma S_x^\Sigma)_{24} = \frac{1}{2}\cos\alpha\gamma_n - \frac{1}{2}\sin\alpha\gamma_{\mu e}(\xi), \quad (16.5)$$

$$(\gamma_\Sigma S_x^\Sigma)_{34} = \frac{1}{2}\sin\alpha\gamma_n + \frac{1}{2}\cos\alpha\gamma_{\mu e}(\xi), \quad (16.6)$$

where

$$\gamma_{\mu e}(\xi) = \gamma_\mu \cos\xi - \gamma_e \sin\xi.$$

Equations (16) show that four transitions between all four levels are allowed,

$$|\psi_1\rangle \longleftrightarrow |\psi_2\rangle, \quad |\psi_1\rangle \longleftrightarrow |\psi_3\rangle,$$
$$|\psi_2\rangle \longleftrightarrow |\psi_4\rangle, \quad |\psi_3\rangle \longleftrightarrow |\psi_4\rangle,$$

and, no matter what the constant magnetic field H_z value, the transitions

$$|\psi_2\rangle \longleftrightarrow |\psi_3\rangle, \quad |\psi_1\rangle \longleftrightarrow |\psi_4\rangle$$

remain forbidden. The spin dynamics of muonium under external periodic perturbations can often be described by quantum transitions within some pair of levels (e.g., see [14]). As follows from (16), the radiofrequency field approximately equally couples all four levels of Mu + Nu, and we cannot specify a single pair of quantum states. In this situation, a general exact analytic solution cannot be obtained. It is therefore of interest to find solutions that allow the most important properties of the system under consideration to be described. Below, we use the mathematical approach developed in this work to obtain an analytic solution that, according to numerical analysis data, contains the most important information on the influence of a radiofrequency field on the amplitude of the ALC signal. Note that this solution is exact at the center of the ALC resonance. Numerical muon polarization calculations were performed for the PyBF-4 compound, which is of interest for experiments.

3.1. Partition of the Hamiltonian

Let us use (16) to rewrite the Hamiltonian in the rotating wave approximation for the radiofrequency field. The H_0 and $V(t)$ energies can then conveniently be written in terms of the $\hat{P}_{ij} = |\Psi_i\rangle\langle\Psi_j|$ operators,

$$V_{nm} = V_{mn} = -\frac{H_1}{2}(\gamma_\Sigma S_x^\Sigma)_{mn}, \tag{17}$$

$$H_0 = \hbar\left(\omega'(\hat{P}_{44} - \hat{P}_{11}) + \frac{\omega_{GX}}{2}(\hat{P}_{33} - \hat{P}_{22})\right), \tag{18}$$

$$V = \hbar(\hat{P}_{12}V_{12} + \hat{P}_{13}V_{13} + \hat{P}_{24}V_{24} + \hat{P}_{34}V_{34}) \times \exp(i\omega t) + \text{H.c.} \tag{19}$$

Further, we use a unitary transformation of $U_0(t)$ to obtain the new representation

$$|\Psi(t)\rangle = U_0(t)|\phi(t)\rangle,$$
$$U_0(t) = \exp(-i\omega t(\hat{P}_{44} - \hat{P}_{11})), \tag{20}$$

where the behavior of the $|\phi(t)\rangle$ wave function is determined by the time-independent Hamiltonian

$$\tilde{H} + U_0^+(t)(H_0 + V(t))U_0(t) - \omega(\hat{P}_{44} - \hat{P}_{11}), \tag{21}$$

$$\tilde{H} = \tilde{H}_0 + \tilde{V},$$

$$\tilde{H}_0 = \hbar\left(\Delta(\hat{P}_{44} - \hat{P}_{11}) + \frac{\omega_{GX}}{2}(\hat{P}_{33} - \hat{P}_{22})\right),$$

$$\tilde{V} = \hbar(V_{12}(\hat{P}_{12} + \hat{P}_{21}) + V_{13}(\hat{P}_{13} + \hat{P}_{31}) + V_{24}(\hat{P}_{24} + \hat{P}_{42}) + V_{34}(\hat{P}_{34} + \hat{P}_{43})). \tag{22}$$

Here, $\Delta = \omega' - \omega$. The Hamiltonian $\tilde{H}$ can conveniently be rewritten in the basis of the quantum states

$$|\chi_1\rangle = \cos\alpha|\Psi_1\rangle - \sin\alpha|\Psi_4\rangle, \quad |\chi_2\rangle = |\Psi_2\rangle,$$
$$|\chi_3\rangle = |\Psi_3\rangle, \quad |\chi_4\rangle = \sin\alpha|\Psi_1\rangle + \cos\alpha|\Psi_4\rangle,$$

using the $P_{nm}^\chi = |\chi_n\rangle\langle\chi_m|$ operators. The Hamiltonian then takes the form

$$H^\chi = \tilde{H}_0^\chi(\alpha, \Delta) + \tilde{V}^\chi(\alpha, \Delta), \tag{23}$$

$$\tilde{H}_0^\chi(\alpha, \Delta) = \tilde{H}_{0,1}^\chi(\alpha, \Delta) + \tilde{H}_{0,2}^\chi(\alpha, \Delta),$$
$$[\tilde{H}_{0,1}^\chi(\alpha, \Delta), \tilde{H}_{0,2}^\chi(\alpha, \Delta)] = 0, \tag{24}$$

$$\tilde{H}_{0,1}^\chi(\alpha, \Delta) = -\cos2\alpha \cdot \hbar\Delta P_{11}^\chi - \frac{1}{2}\hbar\omega_G P_{22}^\chi - \frac{1}{4}\hbar(\gamma_{\mu e}(\xi) - \sin2\alpha \cdot \gamma_n)H_1(P_{12}^\chi + P_{21}^\chi), \tag{25}$$

$$\tilde{H}_{0,2}^\chi(\alpha, \Delta) = \cos2\alpha \cdot \hbar\Delta P_{44}^\chi + \frac{1}{2}\hbar\omega_G P_{33}^\chi - \frac{1}{4}\hbar(\gamma_{\mu e}(\xi) + \sin2\alpha \cdot \gamma_n)H_1(P_{34}^\chi + P_{43}^\chi), \tag{26}$$

$$\tilde{V}^\chi(\alpha, \Delta) = -\sin2\alpha \cdot \hbar\Delta(P_{41}^\chi + P_{14}^\chi) - \frac{1}{4}\cos2\alpha \cdot (\hbar\gamma_n H_1)[(P_{24}^\chi + P_{42}^\chi) + (P_{13}^\chi + P_{31}^\chi)]. \tag{27}$$

The dependence on two parameters α and Δ introduced in the Hamiltonian $\tilde{H}(\alpha, \Delta)$ characterizes the degree of detuning from the exact ALC resonance; the resonance becomes exact at $\alpha = \pi/4$ and $\Delta = \omega' - \omega = 0$. The first condition corresponds to the usual tuning of the constant magnetic field H_z to the maximum of the ALC signal. According to the second condition, the alternating field frequency ω should coincide with the two-photon resonance frequency between the quantum states $|\Psi_1\rangle$ and $|\Psi_4\rangle$. All four muonium levels $|\psi_{1-4}\rangle$ are then coupled by the radiofrequency field. The interaction $\tilde{V}^\chi$ exactly equals zero at $\alpha = \pi/4$ and $\Delta = 0$, when the

Hamiltonian $\tilde{H}$ represents the sum of two commuting simpler terms,

$$\tilde{H}_0^\chi = \tilde{H}_{0,1} + \tilde{H}_{0,2}.$$

Such a partition of the Hamiltonian allows an analytic solution to be obtained and the influence of $\tilde{V}^\chi$ to be included by perturbation theory methods in wide ranges of α and Δ parameter variations. Of primary interest is the solution in the zeroth order of perturbation theory.

3.2. Zeroth-Order Perturbation Theory

In the zeroth order in $\tilde{V}^\chi$ the density matrix is obtained in the form

$$\rho_{\chi,0}(t|\alpha,\Delta) = U_0^\chi(t)\rho(0)U_0^{\chi+}(t), \quad (28)$$

where

$$U_0^\chi(t) = \exp\{-(i/\hbar)\tilde{H}_0^\chi(t)\} = U_{0,1}^\chi(t)U_{0,2}^\chi(t), \quad (29)$$

$$U_{0,1}^\chi(t) = \exp\left\{-\frac{i}{\hbar}H_{0,1}^\chi t\right\}$$

$$= \left\{1_\chi - (P_{11}^\chi + P_{22}^\chi)\left(1 - \cos\left[\sqrt{\alpha_1^2 + \delta^2/4}\,t\right]\right)\right. \quad (30)$$

$$\left. + i\frac{A_{0,1}}{\sqrt{\alpha_1^2 + \delta^2/4}}\sin\left[\sqrt{\alpha_1^2 + \delta^2/4}\,t\right]\right\}$$

$$\times \exp(i\bar{E}_1(P_{11}^\chi + P_{22}^\chi)t),$$

$$U_{0,2}^\chi(t) = \exp\left\{-\frac{i}{\hbar}H_{0,2}^\chi t\right\}$$

$$= \left\{1_\chi - (P_{33}^\chi + P_{44}^\chi)(1 - \cos[\sqrt{\alpha_2^2 + \delta^2/4}\,t])\right. \quad (31)$$

$$\left. + i\frac{A_{0,2}}{\sqrt{\alpha_2^2 + \delta^2/4}}\sin[\sqrt{\alpha_2^2 + \delta^2/4}\,t]\right\}$$

$$\times \exp(i\bar{E}_2(P_{33}^\chi + P_{44}^\chi)t),$$

$$\delta = \Delta\cos\alpha - \frac{1}{2}\omega_{GX}, \quad \alpha_1 = \frac{1}{4}H_1(\gamma_{\mu e}(\xi) - \gamma_n\sin 2\alpha),$$

$$\bar{E}_1 = \frac{1}{2}\left(\frac{1}{2}\omega_{GX} + \Delta\cos 2\alpha\right),$$

$$\alpha_2 = \frac{1}{4}H_1(\gamma_{\mu e}(\xi) - \gamma_n\sin 2\alpha),$$

$$\bar{E}_2 = -\frac{1}{2}\left(\frac{1}{2}\omega_{GX} + \Delta\cos 2\alpha\right), \quad 1_\chi = \sum_{n=1}^{4} P_{nn}^\chi,$$

$$A_{0,1} = \alpha_1(P_{12} + P_{21}) - \frac{1}{2}\delta(P_{22} - P_{11}),$$

$$A_{0,2} = \alpha_2(P_{34} + P_{43}) - \frac{1}{2}\delta(P_{44} - P_{33}).$$

Taking the initial state (15) of the $\rho(0)$ density matrix and the finite muon lifetime τ_μ into account in (28), we can write the solution for the z component of muon polarization observed in ALC experiments,

$$P_z^\mu(t|\alpha,\Delta)_0 = 2\langle\sigma_z(t|\alpha,\Delta)\rangle_0$$
$$= 2\mathrm{Sp}\{U_0^+(t)\sigma_z U_0(t)\rho_{\chi,0}(t|\alpha,\Delta)\}\exp\{-t/\tau_\mu\}. \quad (32)$$

Here, the index "0" denotes the zeroth order of perturbation theory. Simple but cumbersome calculations in (32) yield

$$P_z^\mu(t|\alpha,\Delta)_0 = \frac{1}{2}$$

$$\times\left\{\cos^2 2\alpha + \sin^2 2\alpha\cos\left[\left(\Delta\cos 2\alpha + \frac{1}{2}\omega_{GX}\right)t\right]\right. \quad (33)$$

$$\left.\times\left[\cos(W_1 t)\cos(W_2 t) - \sin(W_1 t)\sin(W_2 t)\frac{T}{W_1 W_2}\right]\right\}$$

$$\times\exp\{-t/\tau_\mu\},$$

where

$$W_1 = \left(\left(\frac{H_1}{4}\right)^2[\gamma_{\mu e} - \sin 2\alpha\cdot\gamma_n]^2\right.$$
$$\left. + \frac{1}{4}\left(\Delta\cos 2\alpha - \frac{1}{2}\omega_{GX}\right)^2\right)^{1/2},$$

$$W_2 = \left(\left(\frac{H_1}{4}\right)^2[\gamma_{\mu e} + \sin 2\alpha\cdot\gamma_n]^2\right. \quad (34)$$
$$\left. + \frac{1}{4}\left(\Delta\cos 2\alpha - \frac{1}{2}\omega_{GX}\right)^2\right)^{1/2},$$

$$T = \left(\frac{H_1}{4}\right)^2[\gamma_{\mu e}^2 - (\sin^2 2\alpha\cdot\gamma_n^2)]$$
$$+ \frac{1}{4}\left(\Delta\cos 2\alpha - \frac{1}{2}\omega_{GX}\right)^2.$$

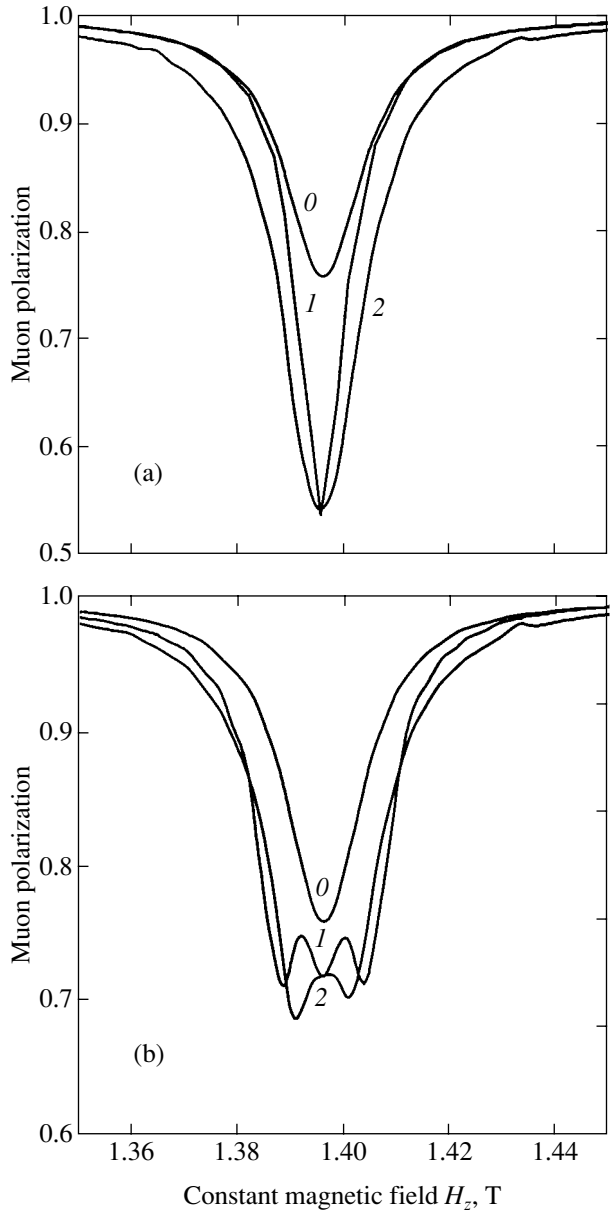

Fig. 2. Dependences of the ALC signal amplitude on the constant magnetic field H_z at radiofrequency field amplitudes H_1 = (a) 0.02 and (b) 0.06 T. The carrier field frequency ω is tuned in resonance to the two-quantum transition frequency $\omega' = (E_4 - E_1)/2\hbar$ (19.4 MHz) at the ALC resonance point: (*2*) muon polarization calculated by zeroth-order equation (33), (*1*) numerical calculations, and (*0*) muon polarization in the absence of a radiofrequency field.

For simplicity, it is taken into account in (33) that hyperfine interaction is much weaker than the Zeeman electron energy,

$$\sin^2\xi \approx (\omega_0/\omega_e)^2 \ll 1.$$

Let us qualitatively estimate the conditions of the applicability of zeroth-order perturbation theory and compare the solution $P_z^\mu(t|\alpha, \Delta)_0$ with the numerical solutions at various constant H_z and radiofrequency H_1 magnetic fields. The alternating field frequency will be fixed at the ALC resonance point.

3.3. Analysis of the Applicability of Zeroth-Order Perturbation Theory

As there exists an exact solution in the center of the ALC signal (details are given below), it is of the greatest importance to estimate the validity of the approximate solution on ALC resonance "wings." A comparison of $\tilde{H}_0$ and $\tilde{V}^\chi(\alpha, \Delta)$ shows that the influence of $\tilde{V}^\chi(\alpha, \Delta)$ can be ignored if the condition

$$\left|\frac{\omega_x}{\omega_G}\right| = \left|\frac{(\gamma_\mu - \gamma_n)(H_z - H_{\text{res}})}{\omega_G}\right| \gg 1 \qquad (35)$$

holds. If $|H_z - H_{\text{res}}|$ is insufficiently large to satisfy (35), the approximate solution describes the signal well if the inequality

$$\left|\frac{(\gamma_\mu + \gamma_n)(H_z - H_{\text{res}})}{\gamma_n H_1}\right| \gg \frac{1}{2} \qquad (36)$$

is satisfied. This inequality is valid if $\gamma_{\mu e} \sim \gamma_n$, as is characteristic of the ALC resonance region. The presence of two conditions (35) and (36) considerably broadens the scope of the applicability of zeroth-order perturbation theory, especially at the ALC resonance wings.

The integral value

$$\langle P_z^\mu \rangle = \frac{1}{\tau_\mu}\int_0^\infty P_z^\mu(t)dt$$

is measured experimentally.

The approximate zeroth-order solution $\langle P_z^\mu(\alpha, \Delta)_0\rangle$ is compared in Fig. 2 with numerical calculations for the example of the substance PyBF-4. The figure shows that the approximate solution well describes the shape and width of the ALC signal. Significant differences only arise close to the line center. If H_1 is small, the shape of the ALC signal described by the $\langle P_z^\mu(\alpha, \Delta)_0\rangle$ solution is broadened compared with the exact solution. The $\langle P_z^\mu(\alpha, \Delta)_0\rangle$ solution tends to the exact solution as H_1 increases, but there remain differences close to $H_z = H_{\text{res}}$. The $\langle P_z^\mu(\alpha, \Delta)_0\rangle$ solution

can be refined using higher orders of perturbation theory,

$$\rho_\chi(t) = U_0^\chi(t)U_1^\chi(t)\rho(0)U_1^\chi(t)^+U_0^\chi(t)^+, \quad (37)$$

$$\begin{aligned} U_1^\chi(t) &= 1 + \int_0^t dt_1\left(-\frac{i}{\hbar}\tilde{V}^\chi(t_1)\right) \\ &+ \int_0^t dt_1\left(-\frac{i}{\hbar}\tilde{V}^\chi(t_1)\right)\int_0^{t_1} dt_2\left(-\frac{i}{\hbar}\tilde{V}^\chi(t_2)\right) + \ldots \\ &= T\exp\left\{-\frac{i}{\hbar}\int_0^t dt\tilde{V}^\chi(t)\right\}, \end{aligned} \quad (38)$$

where

$$\tilde{V}^\chi(t) = U_0^\chi(t)^+\tilde{V}^\chi U_0^\chi(t)$$

and T is the Dyson chronological ordering operator. Equation (37) requires cumbersome calculations, which appear to be unnecessary in this work because the $P_z^\mu(t|\alpha, \Delta)_0$ approximate solution already contains the most important information about the shape and amplitude of the ALC signal and can be used to theoretically analyze the main spectral parameters of the system under study. Of great interest is the exact solution, which can be obtained at the ALC resonance point ($H_z = H_{res}$).

3.4. Exact Solution

At the ALC resonance point ($\alpha = \pi/4$, $\Delta = \omega' - \omega = 0$), the solution for muon polarization (33) takes the form

$$\begin{aligned} P_z^\mu(t|\pi/4, 0)_0 &= P_z^\mu(t|\pi/4, 0) \\ &= 2\langle\sigma_z(t|\alpha = \pi/4, \Delta = 0)\rangle \\ &= \frac{1}{2}\left\{1 + \cos\left(\frac{\omega_G}{2}t\right)\left[\left(1 - \frac{1}{2}e(\omega_G, \Omega_1, \Omega_2)\right)\right.\right. \\ &\times \cos\{(\tilde{\Omega}_1 + \tilde{\Omega}_2)t\} + \frac{1}{2}e(\omega_G, \Omega_1, \Omega_2) \\ &\left.\left.\times \cos\{(\tilde{\Omega}_1 - \tilde{\Omega}_2)t\}\right]\right\}\exp\{-t/\tau_\mu\}. \end{aligned} \quad (39)$$

Equation (39) contains the following values:

$$\begin{aligned} W_{1,2}(\alpha = \pi/4; \Delta = 0) &= \tilde{\Omega}_{1,2} = \sqrt{\Omega_{1,2}^2 + (\omega_G/4)^2}, \\ \Omega_1 &= \frac{H_1}{4}(\gamma_{\mu e}(\xi) - \gamma_n), \\ \Omega_2 &= \frac{H_1}{4}(\gamma_{\mu e}(\xi) + \gamma_n), \\ \varepsilon(\omega_G, \Omega_1, \Omega_2) &= 1 - \frac{1}{(\tilde{\Omega}_1\tilde{\Omega}_2)}(\Omega_1\Omega_2 + (\omega_G/4)^2). \end{aligned} \quad (40)$$

Note that, in the absence of a radiofrequency field or at high radiofrequency fields when $\Omega_{1,2} \gg \omega_G$, the $e(\omega_G, \Omega_1, \Omega_2)$ function in (40) tends to zero. Using (39), we find the measured integral value

$$\begin{aligned} \langle P_z^\mu\rangle &= \frac{1}{4}\left\{2 + \left[1 - \frac{1}{2}e(\omega_G, \tilde{\Omega}_1, \tilde{\Omega}_2)\right]\right. \\ &\times\left[\frac{1}{1 + (\tilde{\Omega}_1 + \tilde{\Omega}_2 + \omega_G/2)^2\tau_\mu^2}\right. \\ &\left.+ \frac{1}{1 + (\tilde{\Omega}_1 + \tilde{\Omega}_2 - \omega_G/2)^2\tau_\mu^2}\right] \\ &+ \frac{1}{2}e(\omega_G, \tilde{\Omega}_1, \tilde{\Omega}_2)\left[\frac{1}{1 + (\tilde{\Omega}_1 - \tilde{\Omega}_2 + \omega_G/2)^2\tau_\mu^2}\right. \\ &\left.\left.+ \frac{1}{1 + (\tilde{\Omega}_1 + (-\tilde{\Omega}_2) - \omega_G/2)^2\tau_\mu^2}\right]\right\}. \end{aligned} \quad (41)$$

The dependence of polarization $\langle P_z^\mu\rangle$ on the radiofrequency field amplitude H_1 is shown in Fig. 3 for the example of PyBF-4. It is noteworthy that $\langle P_z^\mu\rangle$ has a deep intermediate minimum at $H_1 \approx (1/2)H_{1,\max}$ ($H_1 \approx 0.027$ T for PyBF-4), where $H_{1,\max}$ is found by the equation

$$\tilde{\Omega}_1(H_{1,\max}) - \tilde{\Omega}_2(H_{1,\max}) = \omega_G/2.$$

The presence of the minimum substantially weakens the condition imposed on the radiofrequency field amplitude by the requirement of substantial muon depolarization. Even magnetic radiofrequency fields on the order of 100 G substantially increase the ALC signal amplitude. It should be added that radiofrequency fields with an amplitude of $H_1 \le 300$ G increase the amplitude of the observed signal almost twofold without substantially changing the shape of the ALC signal. The important conclusion can be drawn that the influence of comparatively low radiofrequency fields on the

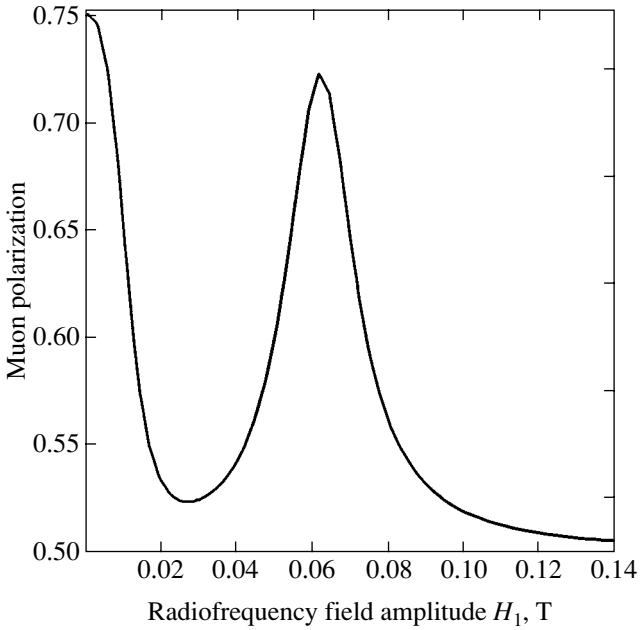

Fig. 3. Dependence of muon polarization on the radiofrequency field amplitude H_1 for PyBF-4 at the constant magnetic field $H_z = H_{res}$ and the alternating field frequency $\omega = \omega'$.

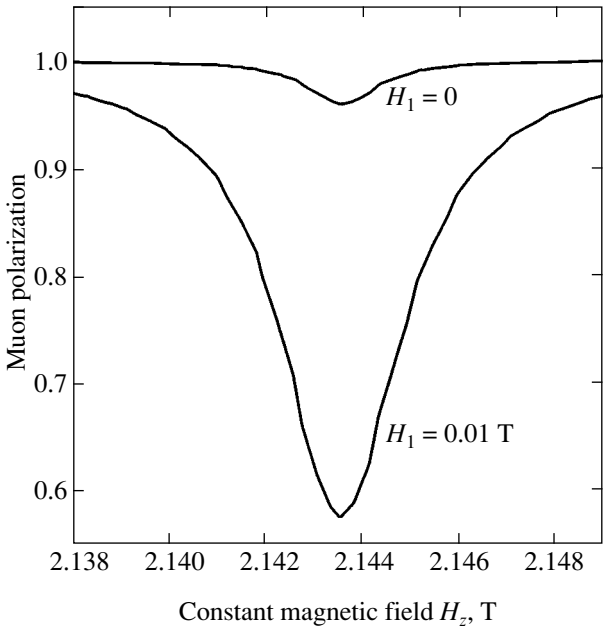

Fig. 4. Dependences of muon polarization on constant magnetic field H_z. The influence of a radiofrequency field (H_1 = 0.01 T) on the amplitude of the signal in a substance with weak electron–nucleus hyperfine interaction ($\omega_G\tau_\mu < 1$, ω_G = 0.03 MHz).

behavior of the ALC signal can be analyzed in terms of the exact fairly simple solution (39), (41), which describes the behavior of the center of the curve shown in Fig. 2.

Solution (28)–(31) for the density matrix allows us to easily find the equation for the polarization of nuclei under the action of a radiofrequency field at the ALC resonance point,

$$P_z^n(t) = \frac{1}{2}\sin(\omega_G t/2) \times \left\{ \frac{\omega_G}{4\tilde{\Omega}_1}\sin(\tilde{\Omega}_1 t)\cos(\tilde{\Omega}_2 t) + \frac{\omega_G}{4\tilde{\Omega}_2}\sin(\tilde{\Omega}_2 t)\cos(\tilde{\Omega}_1 t) \right\} \exp(-t/\tau_\mu). \quad (42)$$

In the absence of a radiofrequency field ($H_1 = 0$, $\Omega_{1,2} = 0$), (42) yields

$$P_z^n(t) = \frac{1}{4}(1 - \cos(\omega_G t))\exp(-t/\tau_\mu).$$

At high radiofrequency fields ($\Omega_1 \approx \Omega_2 = \Omega \gg \omega_G$), the oscillations of nuclear polarization take the form

$$P_z^n(t) = \frac{\omega_G}{8\Omega}\sin(\omega_G t/2)\sin(2\Omega t)\exp(-t/\tau_\mu) = \frac{1}{2}(\omega_G/\gamma_{\mu,e}H_1)\sin(\omega_G t/2)\sin\left(\frac{1}{2}\gamma_{\mu,e}H_1 t\right)\exp(-t/\tau_\mu). \quad (43)$$

According to (43), an increase in the radiofrequency field amplitude suppresses polarization transfer from the muon to the spin of the nucleus. The conclusion can be drawn that an increase in muon depolarization in a radiofrequency field is attained as a result of an increase in muon spin rotation rather than additional polarization of the nuclear subsystem.

4. DISCUSSION AND CONCLUSIONS

Currently, the development of high-resolution μSR spectroscopy again becomes related to the use of radiofrequency fields (see [6]); this in full measure refers to studying muonium in radiofrequency fields [15, 16]. The results obtained in this work lead us to conclude that the use of radiofrequency radiation for detecting ALC signals enables us to also extract information about transition frequencies in the muonium + nuclear spin systems and the parameters that characterize hyperfine interactions in these systems. It is exceedingly important that applying radiofrequency fields substantially increases the degree of muon depolarization and, accordingly, the amplitude of the observed ALC resonance signal. Equations (33), (39), and (41) are the most important results of this work. These equations can be used to study muon depolarization stimulation in ALC experiments.

Table 3. Parameters used in calculations of the ALC signal for the PyBF-4 substance

	ω_0, MHz	Ω, MHz	γ_μ, MHz/T	γ_e, MHz/T	γ_n, MHz/T	H_{res}, T	ω', MHz
PyBF-4	417.2	152	$2\pi135.53$	$2\pi28024.21$	$2\pi40.55$	1.396	19.4

As concerns the physical properties of the solutions, of interest are the special features of the behavior of muon polarization when the traditional ALC resonance technique is used. Muon polarization behavior is then described by the well-known equation

$$P^{\mu}_{z,1}(t) = \frac{1}{4}\{3 + \cos(\omega_G t)\}\exp(-t/\tau_\mu)$$

[this equation also follows from (39) at $H_1 = 0$], which shows that a maximum decrease in the mean degree of muon polarization amounts to 1/4 of the initial polarization ($\omega_G\tau_\mu \gg 1$). As follows from solution (39), (41) (e.g., see Fig. 4), the action of a radiofrequency field can decrease the initial muon polarization even by half the initial polarization value. Remarkably, this opens up possibilities of attaining considerable relative strengthening of muon depolarization at the ALC resonance point at small ω_G values ($\omega_G\tau_\mu < 1$), when the usual ALC signal is weak or inaccessible to observation,

$$\bar{P}^{\mu}_{z,1}\Big|_{\omega_G\tau_\mu \ll 1} \longrightarrow 1,$$

which happens if the electron–nucleus hyperfine interaction is weak. The results of the corresponding calculations are plotted in Fig. 4, which shows that the action of a radiofrequency field increases the signal more than tenfold (in the absence of a radiofrequency field, the minimum polarization is $P(H_{res}) = 0.97$, whereas at $H_1 = 100$ G, the minimum polarization decreases to $P(H_{res}) = 0.57$). We stress that performing such experiments requires varying the static magnetic field H_z and the carrier field frequency ω in fairly wide ranges to scan the magnetic field near $H_z = H_{res}$ and tune the radiofrequency field to the nuclear frequencies of the ALC resonance transition.

It can be concluded by analogy that, provided the hyperfine interaction between electrons and nuclei is fairly weak, the number of levels covered by a radiofrequency field will be 6 and 8 for nuclear spins 1 and 3/2, respectively. This noticeably complicates obtaining the corresponding analytic solutions. Based on the results of this work, we, however, expect that the use of radiofrequency fields should also strengthen muon depolarization in such systems, although to a lesser extent than in systems with nuclear spin 1/2. It is at present difficult to estimate the corresponding effects.

To summarize, the use of radiofrequency fields in ALC experiments can increase the sensitivity of the μSR technique and the amount of information about hyperfine muonium interactions obtained in such experiments. The mathematical approach developed in this work can be applied to describe muon echo, which inspires hopes for advances in μSR experiments [17–21], because the use of the spin echo technique allows the quantum dynamics of the system under study to be controlled at long times. The muon echo in ALC experiments will be formed in the four-level quantum system (for nuclear spin 1/2), whose nonclassical physical properties are determined by the entangled states of the muon, electron, and nuclear spins. We should therefore expect the properties of this echo to be substantially different from those of the spin echo in two-level systems. Also note that the analytic solution obtained in this work for the behavior of the four-level system in a quasi-stationary field is a rare case of an exact solution in spectroscopy. Generalizing this solution to other quantum systems may be of independent interest.

ACKNOWLEDGMENTS

One of us (S.A.M.) thanks his colleagues N.M. Suleimanov and R.G. Mustafin for valuable discussions, which stimulated the undertaking of this study, and K.M. Salikhov for valuable discussion of the results. This work was financially supported by the Russian Foundation for Basic Research (project nos. 00-02-16192 and 00-15-97410), the CRDF, and the NIOKR Foundation of Tatarstan (project no. 14-79).

REFERENCES

1. V. W. Hughes and G. zu Putlitz, in *Advanced Series on Direction in High Energy Physics*, Ed. by T. Kinoshita (1990), Vol. 7, p. 822.
2. Y. Kuno and Y. Okada, Rev. Mod. Phys. **73**, 151 (2001).
3. V. W. Hughes and T. Kinoshita, Rev. Mod. Phys. **71**, S133 (1999).
4. A. Schenck, *Muon Spin Rotation Spectroscopy: Principles and Applications in Solid State Physics* (Hilger, Bristol, 1985).
5. V. P. Smilga and Yu. M. Belousov, *The Muon Method in Science* (Nauka, Moscow, 1991; Nova Sci., New York, 1994).
6. *Proceedings of the Eighth International Conference on Muon Spin Rotation, Relaxation and Resonance μSR'99*, Physica B (Amsterdam) **289–290** (2000).
7. S. F. J. Cox, P. J. C. King, W. G. Williams, *et al.*, Physica B (Amsterdam) **289–290**, 538 (2000).
8. B. D. Patterson, Rev. Mod. Phys. **60**, 69 (1988).
9. A. Abragam, C. R. Acad. Sci., Ser. II **299**, 95 (1984).

10. R. F. Keifl, S. Kreitzman, M. Celio, *et al.*, Phys. Rev. A **34**, 681 (1986).
11. R. R. Ernst, G. Bodenhausen, and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Clarendon Press, Oxford, 1987; Mir, Moscow, 1990).
12. A. Schweiger and G. Jeschke, *Principles of Pulse Electron Paramagnetic Resonance* (Oxford Univ. Press, Oxford, 2001).
13. Review of Particle Properties, Phys. Lett. B **204**, 1 (1988).
14. Yu. M. Belousov and V. P. Smilga, Zh. Éksp. Teor. Fiz. **102**, 211 (1992) [Sov. Phys. JETP **75**, 112 (1992)].
15. O. Kormann, J. Major, I. D. Reid, *et al.*, Physica B (Amsterdam) **289**–**290**, 530 (2000).
16. R. Scheuermann, H. Dilger, E. Roduner, *et al.*, Physica B (Amsterdam) **289**–**290**, 698 (2000).
17. S. R. Kreitzman, D. L. Williams, N. Kaplan, *et al.*, Phys. Rev. Lett. **61**, 2890 (1988).
18. S. A. Moiseev and N. M. Suleĭmanov, Pis'ma Zh. Éksp. Teor. Fiz. **64**, 500 (1996) [JETP Lett. **64**, 544 (1996)].
19. S. P. Cottrel, S. F. J. Cox, J. S. Lord, *et al.*, Appl. Magn. Reson. **15**, 469 (1998).
20. S. A. Moiseev, R. G. Mustafin, V. G. Nikiforov, *et al.*, Phys. Rev. B **61**, 5891 (2000).
21. N. M. Suleimanov, S. A. Moiseev, M. A. Clark-Gayther, *et al.*, Physica B (Amsterdam) **289**–**290**, 676 (2000).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1123–1130.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1255–1263.

ATOMS, SPECTRA, RADIATION

Photoluminescence of Er^{3+} Ions in Layers of Quasi-Ordered Silicon Nanocrystals in a Silicon Dioxide Matrix

P. K. Kashkarov[a], M. G. Lisachenko[a], O. A. Shalygina[a], V. Yu. Timoshenko[a], B. V. Kamenev[a,b], M. Schmidt[c], J. Heitmann[c], and M. Zacharias[c]

[a]*Faculty of Physics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia*
[b]*Department of Electrical and Computer Engineering, New Jersey Institute of Technology University Heights Newark, New Jersey 07102-1982, USA*
[c]*Max-Planck-Institut für Mikrostrukturphysik, 06120 Halle, Germany*
e-mail: pavel@vega.phys.msu.su
Received April 24, 2003

Abstract—The spectra and kinetics of photoluminescence from multilayered structures of quasi-ordered silicon nanocrystals in a silica matrix were studied for undoped samples and samples doped with erbium. It was shown that the optical excitation energy of silicon nanocrystals could be effectively transferred to Er^{3+} ions, which was followed by luminescence at a wavelength of 1.5 μm. The effectiveness of energy transfer increased as the size of silicon nanocrystals decreased and the energy of exciting light quanta increased. The excitation of erbium luminescence in the structures was explained based on dipole–dipole interaction (the Förster mechanism) between excitons in silicon nanocrystals and Er^{3+} ions in silica surrounding them. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Much attention has been given to erbium ion Er^{3+} luminescence in crystalline and amorphous silicon in recent years (e.g., see collected papers [1, 2]). This is explained by the demand for silicon devices that effectively emit at a wavelength of 1.5 μm (the $^4I_{13/2} \longrightarrow {}^4I_{15/2}$ transitions in the inner 4*f* shell of Er^{3+}), which corresponds to maximum transmittance of fiber communication lines. Quite a number of unsolved problems, however, prevent creating the desired optoelectronic device. For instance, when crystalline silicon (c-Si) is used as a matrix for Er^{3+}, strong temperature quenching of erbium luminescence is observed as a result of nonradiative deexcitation of Er^{3+} ions via the back transfer of the energy to the matrix [3]. As a consequence, the quantum yield of luminescence from c-Si:Er samples is exceedingly low at room temperature. Temperature quenching of luminescence at 1.5 μm is much weaker for amorphous hydrogenated silicon (a-Si:H) doped with erbium [4]. An analysis of time dependences (kinetics) of photoluminescence from Er^{3+} ions in a-Si:H showed that the energy of electron–hole pairs was transferred to the ions in fairly short (submicrosecond) times, which provided a high effectiveness of their excitation [5–7]. Nevertheless, because of the presence of various nonradiative loss channels, the intensity of erbium luminescence in a-Si:H(Er) is still insufficient for using this material in light-emitting devices.

An attractive approach to overcoming the difficulties mentioned above is the use of layers of erbium-doped silicon nanocrystals (nc-Si) embedded into a dielectric matrix [8–11]. Note that, although the wavelength of erbium luminescence is nearly independent of the nature of the matrix because of screening of the "working" 4*f* shell of Er^{3+} by the outer electron shells, the effectiveness of the excitation of ions can be controlled by changing the properties of the matrix, such as its forbidden band width and/or the density of defect electronic states and impurities [1, 3]. This is easily achieved with nc-Si structures because the forbidden band width of the nanocrystals depends on their size [12, 13]. In addition, Si nanocrystals can simultaneously ensure high charge carrier localization in small spatial regions close to the Er^{3+} ions and fairly long (hundreds of microseconds) electronic excitation lifetimes [12, 13]. The energy released in the recombination of a photoexcited electron–hole pair can then effectively be transferred to an Er^{3+} ion. Indeed, intense and stable photoluminescence of Er^{3+} ions is observed for erbium-doped nc-Si layers in a SiO_2 matrix even at room temperature [9, 10]. The effectiveness and lifetimes of photoluminescence then strongly depend on the technology used to prepare nc-Si/SiO_2 structures and the size of the nanocrystals [9]. Layers of quasi-ordered silicon nanocrystals in multilayered nc-Si/SiO_2 structures therefore show promise for applications

because the size of and the distance between the nanocrystals in them can be effectively controlled [11].

This work presents the results of a comparative study of the spectra and kinetics of photoluminescence from multilayered nc-Si/SiO_2 structures both doped with and free of erbium. We were able to quantitatively estimate the effectiveness of the transfer of the electronic excitation energy from silicon nanocrystals of various sizes to Er^{3+} ions in surrounding silica.

2. SAMPLES AND EXPERIMENTAL DETAILS

The samples were prepared based on superlattices of amorphous SiO/SiO_2 layers formed by successively depositing SiO and SiO_2 on a c-Si substrate by reactive sputtering [10, 11]. The thickness of SiO and SiO_2 layers was varied from 2 to 6 nm and from 2 to 4 nm, respectively. The structures comprised 30–50 pairs of layers, whose total thickness was 200–300 nm. The samples were annealed at 1100°C in nitrogen for 60 min. As a result, layers of closely spaced quasi-ordered Si nanocrystals separated by SiO_2 layers were formed [11]. According to the electron microscopy (see inset in Fig. 1) and X-ray diffraction data, the mean size d of the nanocrystals was close to the thickness of the initial SiO layers. The variance of nanocrystal sizes δd was about 0.5 nm. Part of the structures were used for the implantation of Er^{3+} ions with a 300 keV energy in doses of 5×10^{14} and 2×10^{15} cm^{-2}. Similar doses of ions were also implanted into homogeneous amorphous SiO_2 layers 250 nm thick. These amorphous layers were used as a reference in studying nanocrystalline structures. After implantation, all samples were additionally annealed at 950°C from 5 min to 1 h to remove radiation-induced defects. The mean concentration of Er atoms in the samples N_{Er} was 10^{19} and 4×10^{19} cm^{-3} for the smaller and larger implantation doses, respectively. These values were obtained taking into account the mean projective range R_p of Er ions in the SiO_2 matrix ($R_p \approx 120$ nm for ions with an energy of 300 keV), a spread of the mean projective range $\Delta R_p \approx 40$ nm, and the experimental observation that there was no substantial blurring of implanted particle profiles after annealing [14]. The concentration of Si nanocrystals in the nc-Si/SiO_2 structures was on the order of 10^{19} cm^{-3} according to the transmission electron microscopy data [11].

Photoluminescence was excited by a pulsed N_2 laser (quantum energy $\hbar\omega_1 = 3.7$ eV, pulse width $\tau \sim 10$ ns, pulse energy $E \leq 1$ μJ, pulse repetition frequency $\nu \sim 100$ Hz), a pulsed copper vapor laser ($\hbar\omega_2 = 2.4$ eV, $\hbar\omega_3 = 2.1$ eV, $\tau \sim 20$ ns, $E \leq 10$ μJ, $\nu \sim 12$ kHz), and a continuous He–Ne laser ($\hbar\omega_4 = 1.96$ eV, radiation power up to 10 mW). Laser radiation was focused on the samples into a spot 1.5 mm in diameter.

The photoluminescence spectra were recorded on an automated spectrometer equipped with an InGaAs photodiode. The spectra were corrected for the spectral response of the system. The photoluminescence spectra were measured with a resolution of about 2 nm in the forward current mode without using phase-sensitive accessories. The kinetics of photoluminescence in the visible range was recorded using a photomultiplier with

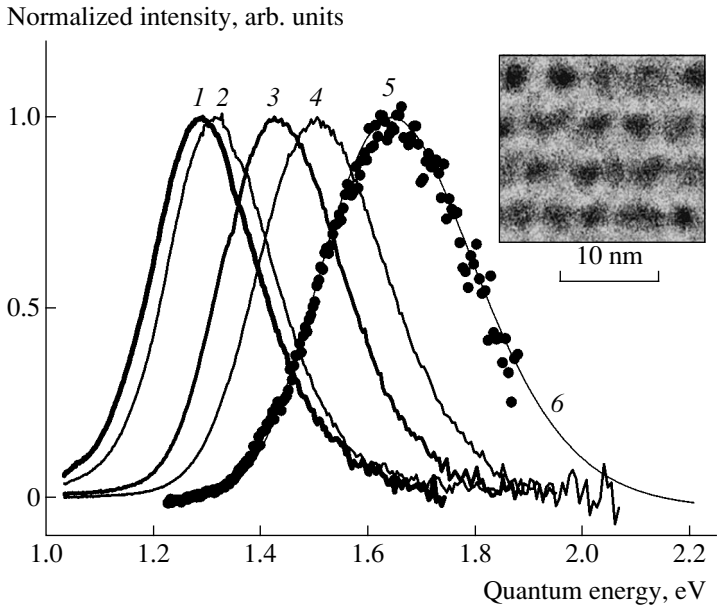

Fig. 1. Photoluminescence spectra of samples with mean nanocrystal sizes d of (*1*) 6, (*2*) 5, (*3*) 4, (*4*) 3, and (*5*) 2 nm excited by light with $\hbar\omega_1 = 3.7$ eV at $T = 300$ K. Curve *6* approximates spectrum *5* by a Gauss function. Shown in the inset is an electron microscopic image of the structure of nc-Si/SiO_2 with $d = 3.5$ nm.

a time constant of about 30 ns. In the infrared range, an InGaAs photodiode with a time constant of about 1 μs was used. Because of an insufficient sensitivity of the photodiode, it was only used to record the initial kinetics region and the integral photoluminescence intensity was then measured in the wavelength range 1.1–1.6 μm. Long-term photoluminescence relaxation components were studied with a more sensitive InGaAs photodiode (time constant 0.5 ms). The spectral resolution in kinetics experiments was 2 nm.

Most experiments in which photoluminescence spectra and kinetics were recorded were performed in air at 300 K. Several photoluminescence spectra were also measured in vacuum in the temperature range 6–450 K with the use of a DE-204N (Advanced Research Systems) closed-cycle helium cryostat.

3. RESULTS AND DISCUSSION

3.1. Photoluminescence Spectra at Room Temperature

The undoped nc-Si/SiO_2 structures excited by light with a quantum energy of $\hbar\omega_1$ gave fairly intense photoluminescence with an external quantum yield of 0.1 to 1% at $T = 300$ K. The normalized photoluminescence spectra of samples with different silicon nanocrystal sizes d are shown in Fig. 1. According to this figure, the photoluminescence band maximum shifts to higher quantum energies as d decreases. This shift is usually explained by an increase in the forbidden band width in nanocrystals caused by the quantum size effect, and the band itself is assigned to radiative recombination of excitons in nc-Si [11–13]. The photoluminescence band has a fairly large width, which increases from 0.23 to 0.34 eV at half-height as the mean size of nanocrystals decreases from 6 to 2 nm. The broadening of the photoluminescence spectrum at smaller d is likely to be related to strengthening forbidden band width fluctuations in nanocrystals as the $\delta d/d$ parameter increases. In our view, an additional reason for the excitonic photoluminescence band broadening in nc-Si can be the interaction of excitons with phonons of silicon and surrounding SiO_2. Indeed, photoluminescence bands 0.12–0.15 eV wide are observed even for isolated silicon quantum dots in a SiO_2 matrix [14].

The implantation of erbium ions caused substantial (~100-fold) suppression of excitonic photoluminescence and the appearance of an intense band at 0.81 eV (Fig. 2). This band is characteristic of the $^4I_{13/2} \longrightarrow {}^4I_{15/2}$ intracenter transitions in Er^{3+} ions implanted into a solid matrix [1, 2]. The Er^{3+} ions are formed by the transfer of erbium electrons into a bound state in SiO_2, as is typical of all lanthanides in dielectric matrices. Erbium donates its electrons to neighboring oxygen atoms or defects with the formation of an ionic bond.

The quenching of excitonic luminescence and the appearance of the erbium band were observed for all

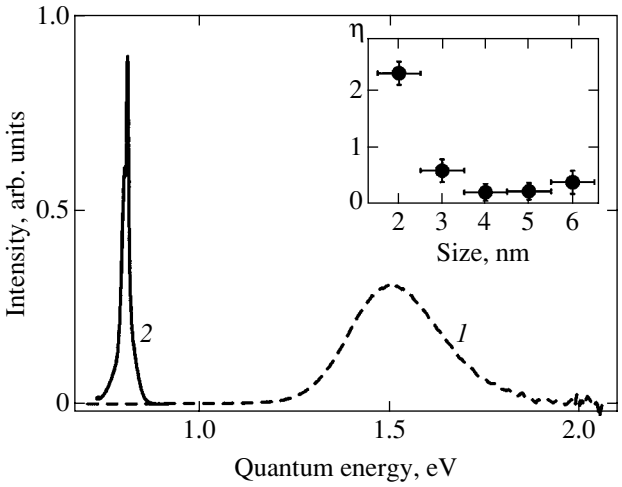

Fig. 2. Photoluminescence spectra of (*1*) undoped and (*2*) erbium-doped samples ($d = 3$ nm) excited by light with $\hbar\omega_1 = 3.7$ eV. Shown in the inset is the dependence of transfer factor η obtained by integrating the photoluminescence spectra on the size of nanocrystals; $T = 300$ K.

structures that we studied. At the same time, homogeneous a-SiO_2:Er^{3+} layers gave extremely weak photoluminescence at about 0.8 eV. This was evidence that the excitation of Er^{3+} occurred as a result of energy exchange with the matrix that absorbed a photon rather than direct light quantum absorption.

The ratio between the photoluminescence intensities of erbium-doped and undoped structures leads us to conclude that the larger part of the energy absorbed by the nanocrystals is transferred to the optically active Er^{3+} ions. The effectiveness of energy transfer can conveniently be quantitatively characterized by the ratio (called "transfer factor" in what follows) $\eta = \tilde{I}_{\mathrm{Er}}/\tilde{I}_{nc}$, where

$$\tilde{I}_{\mathrm{Er}} = \int \frac{I_{\mathrm{Er}}(\nu)}{\nu} d\nu, \quad \tilde{I}_{nc} = \int \frac{I_{nc}(\nu)}{\nu} d\nu. \qquad (1)$$

Here, $I_{\mathrm{Er}}(\nu)$ and $I_{nc}(\nu)$ are the photoluminescence spectra of the samples with and without erbium, respectively. The integration is performed over the spectral ranges of the erbium (0.75–0.85 eV) and excitonic (1.1–2.0 eV) photoluminescence bands.

The transfer factors η for structures containing nanocrystals of different sizes with the mean concentration of ions fixed at $N_{\mathrm{Er}} \approx 4 \times 10^{19}$ cm^{-3} are shown in the inset to Fig. 2. According to this figure, η values lie in the range 0.3–0.4 for structures with $d = 4$–6 nm and substantially increase for samples containing smaller nanocrystals. The η value for structures with $d = 2$ nm exceeds 2. The number of photoluminescence quanta emitted by the nc-Si/SiO_2:Er structure is therefore two times larger than that emitted by the undoped sample at the same optical excitation level. This is evidence that

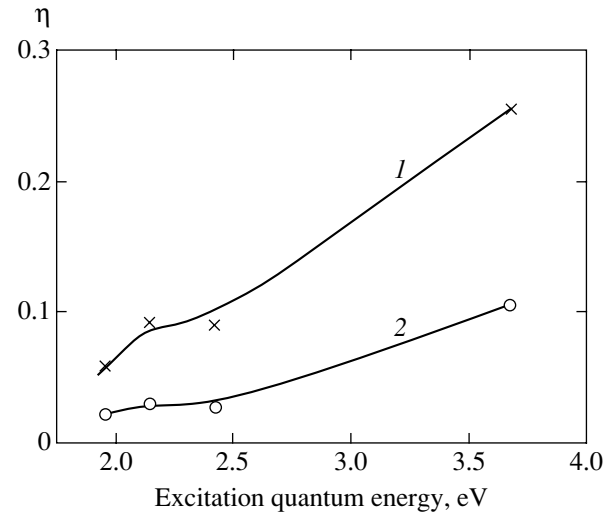

Fig. 3. Dependences of transfer factor η on the energy of excitation quanta for structures with d = (*1*) 2.5 and (*2*) 3.5 nm at T = 300 K.

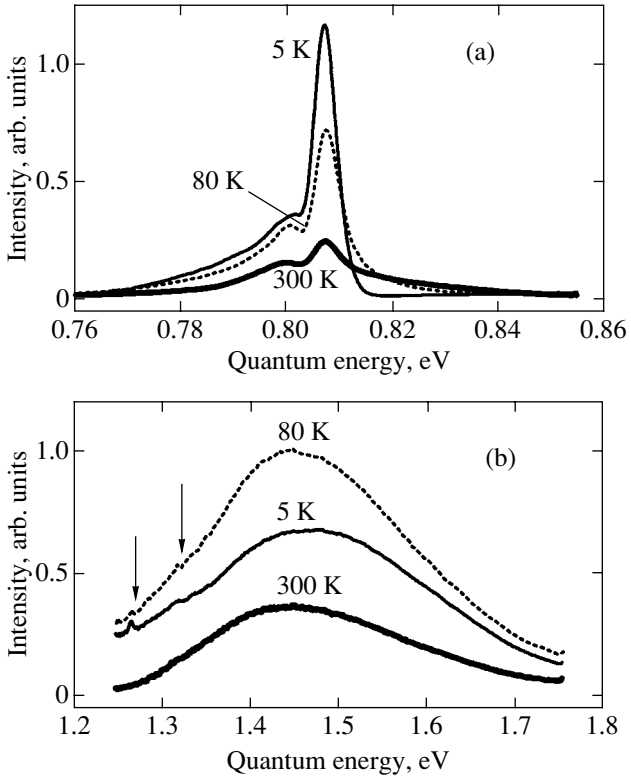

Fig. 4. Photoluminescence spectra of erbium-doped sample with d = 3.5 nm in the regions of (a) Er^{3+} and (b) nanocrystal luminescence at different temperatures.

the introduction of Er^{3+} ions creates an additional effective channel of radiative (at a wavelength of 1.5 μm) relaxation of optical excitation energy, which competes with nonradiative relaxation in the structures under consideration.

Our experiments showed that the η parameter increased as the energy of light quanta used to excite photoluminescence grew larger (Fig. 3). This cannot be explained by an increase in the absorption coefficient of nc-Si, because such an increase would equally influence doped and undoped samples and, therefore, should not contribute to η variations. The increase in η observed when high-energy pumping quanta are used can be explained by a contribution of high-energy exciton states to energy transfer to the Er^{3+} ions. Note also that the absolute η value is larger for structures with smaller nanocrystals, and its increase with the energy of pumping quanta is more substantial in such structures (Fig. 3, dependence *1*). This is also evidence that high-energy exciton states are involved in the excitation of Er^{3+} ions.

3.2. Temperature Dependence of Photoluminescence Spectra

The photoluminescence spectra of erbium-doped structures recorded at various temperatures are shown in Fig. 4. The intensity at the band maximum at a photon energy of 0.81 increases as temperature lowers (Fig. 4a). Simultaneously, the short-wave edge of the spectrum becomes suppressed because of a decrease in the population of the upper states in the fine structure of the Er^{3+} ion energy levels. The total width of the line therefore decreases. To within the accuracy of our measurements, we did not observe changes in the positions of the spectral band maxima.

The excitonic photoluminescence spectra of the erbium-doped structures are shown in Fig. 4b. As mentioned, the intensity of this photoluminescence is almost two orders of magnitude lower than that of the corresponding band in the spectra of the undoped samples. The intensity of excitonic luminescence exhibited a nonmonotonic behavior as the temperature varied, and its spectrum changed. The most substantial changes were observed at helium temperatures, at which a narrow line at a 1.26 eV energy and, simultaneously, regions of partial suppression of a broad photoluminescence band in the energy range 1.26–1.33 eV appeared. The 1.26 eV energy is known to correspond to the ${}^4I_{11/2} \longrightarrow {}^4I_{15/2}$ transition in the Er^{3+} ion. The presence of this line in indirectly excited erbium photoluminescence spectra is evidence of excitation transfer to still higher ion levels.

Such a feature as the suppression of low-temperature photoluminescence in the energy range 1.26–1.33 eV can be attributed to energy transfer from excitons in nc-Si to the second excited state of Er^{3+} ions. This transfer can be accompanied by the emission of silicon phonons, whose maximum energy is known to be $E_{\Gamma(0)} \approx 64$ meV. The regions of the strongest suppression of excitonic photoluminescence (Fig. 4b, arrows) are situated at precisely this distance from the 1.26 eV energy. Processes with the emission of phonons corre-

sponding to vibrational excitation of the O–Si–O bond (about 140 meV) can additionally contribute to energy transfer from the nanocrystals to Er^{3+} ions. On the whole, the fine structure of the regions of fluorescence quenching is not clearly defined, which can be explained by the superposition of processes with the emission of phonons of various types and energies and by phonon spectrum changes in small nanocrystals. Note in addition that the contribution of phonon-related features to the total level of exciton fluorescence quenching is less than 0.1%. This is evidence of the presence of a much stronger mechanism of energy transfer from excitons to Er^{3+} ions. A similar conclusion of the presence of an effective phononless interaction mechanism between excitons and Er^{3+} can be drawn based on the data reported in [9], where dips multiple to phonon frequencies in the low-temperature photoluminescence spectra of erbium-doped structures with nc-Si were observed for the first time.

Consider the temperature dependences of the effectiveness of photoluminescence in the samples. The $\tilde{I}_{Er}$ and $\tilde{I}_{nc}$ integral values are shown in Fig. 5 as functions of the inverse temperature. A decrease in the temperature from 300 to 60 K increases the yield of photoluminescence two- to threefold for both undoped and erbium-doped samples. Interestingly, both photoluminescence bands behave similarly in this temperature range. It is likely that decreasing the temperature causes the suppression of the nonradiative channel of the recombination of electron–hole pairs on defects (such as broken silicon bonds). This increases the concentration of excitons and, therefore, the yield of excitonic luminescence from undoped samples, on the one hand, and the rate of Er^{3+} excitation in the interaction with excitons in erbium-doped structures, on the other. Note that the photoluminescence of the doped structures depends more strongly on the temperature than that of the undoped samples (Fig. 5, dependences *2* and *1*, respectively). This result is easy to explain if it is taken into account that not all radiation-induced defects could be removed during postimplantation annealing.

The $\tilde{I}_{Er}$ value remains nearly constant in the temperature range 10–60 K (Fig. 5, dependence *3*) and even grows when the temperature decreases further. This shows that, first, the majority of the ions are well isolated with respect to back energy transfer to the matrix. Secondly, ions weakly bound with nc-Si or, conversely, situated inside the Si nanocrystals can contribute to erbium photoluminescence at helium temperatures. The probability of back energy transfer from the ions inside nanocrystals to the matrix decreases as the temperature lowers, as has, for instance, been observed for c-Si:Er [1, 2].

The $\tilde{I}_{nc}$ value (Fig. 5, dependences *1* and *2*) decreases as the temperature decreases. This can be explained by the transition of excitons to the triplet

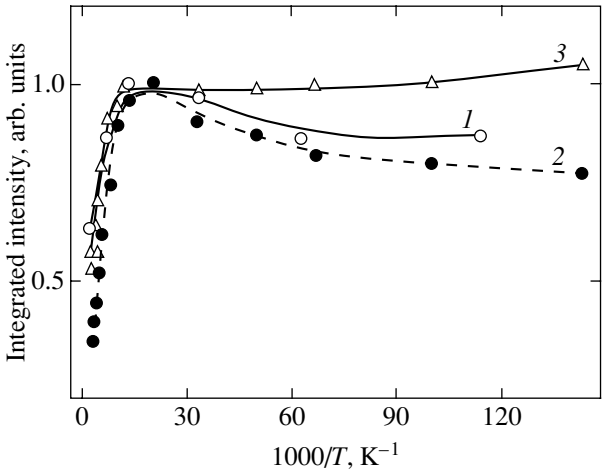

Fig. 5. Dependences of intensities integrated over the spectrum of photoluminescence of silicon nanocrystals with d = 3.5 nm in (*1*) nc-Si/SiO_2 and (*2*) nc-Si/SiO_2:Er and (*3*) of the erbium band in nc-Si/SiO_2:Er on inverse temperature.

state, which is more favorable energetically and is characterized by a much longer radiative recombination time [13]. As a consequence, the yield of photoluminescence decreases at a fixed rate of nonradiative recombination. The more noticeable decrease in $\tilde{I}_{nc}$ for samples with erbium at $T < 60$ K is easy to understand, the Er^{3+} ions being excitonic photoluminescence-quenching centers. Only a small number of nanocrystals (about 1%) that interact comparatively weakly with ions contribute to excitonic photoluminescence in erbium-containing structures. The conclusion can therefore be drawn that this interaction becomes stronger as temperature decreases, clearly because of an increase in the lifetimes of excitons, which become triplet at helium temperatures.

3.3. The Kinetics of Photoluminescence

Several typical kinetics of the relaxation of the intensity of excitonic luminescence in undoped and erbium-doped nc-Si/SiO_2 structures after the action of a laser pulse are shown in Fig. 6. These kinetics are not monoexponential, but can be well approximated by the so-called "stretched" exponential function

$$I_{PL}(t) = I_0 \exp[-(t/\tau_0)^{\beta}], \qquad (2)$$

where τ_0 is the mean time and β is a numerical parameter. Note that photoluminescence with a kinetics described by (2) is usually observed for disordered solid systems characterized by a variance of recombination times, for instance, for a-Si:H [5–7] and porous silicon [13].

An analysis of the kinetics of photoluminescence from undoped structures shows that the τ_0 parameter increases from several to tens of microseconds as the

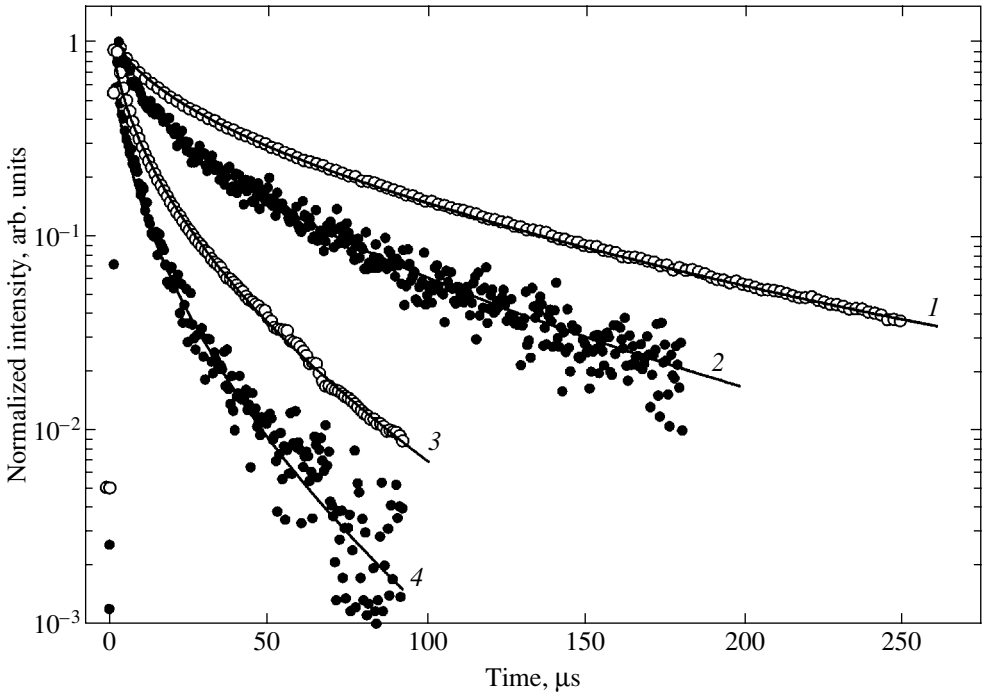

Fig. 6. Photoluminescence kinetics of (*1*, *3*) nc-Si/SiO_2 and (*2*, *4*) nc-Si/SiO_2:Er measured for quantum energies of (*1*, *2*) 1.49 and (*3*, *4*) 1.97 eV. The experimental values are given by symbols, and the lines are the approximations by (2).

energy of photoluminescence quanta changes from 2 to 1.5 eV. As concerns the β parameter, its value of about 0.5 remains almost unchanged. For erbium-containing structures, time τ_0 decreases approximately 2–2.5 times, whereas the β value is nearly constant. The constant β value is evidence of a weak influence of the electronic states of Er^{3+} and related defects on the variance of recombination parameters.

The implantation of Er^{3+} ions reduces the intensity of excitonic photoluminescence by two orders of magnitude, whereas the mean photoluminescence lifetimes are reduced only by half. This observation lends support to the suggestion that the major part of Si nanocrystals in doped samples barely contribute to luminescence in the energy range 1.2–1.8 eV because of the complete energy transfer from these nanocrystals to the ions followed by luminescence in the region of 0.81 eV. At the same time, the less than 1% of nanocrystals that remain are characterized by photoluminescence times shortened by interaction with Er^{3+} ions. These times can also be shortened by nonradiative recombination processes on defects caused by the introduction of Er^{3+}. The absence of strong temperature quenching of photoluminescence from the samples under consideration is, however, evidence of a low concentration of such defects.

The kinetics of the relaxation of photoluminescence of Er^{3+} ions measured for samples with nanocrystals of different sizes are shown in Fig. 7. Erbium photoluminescence is seen to be characterized by a nearly exponential kinetics. Similar dependences were obtained for all samples. The mean relaxation time τ_0 for photoluminescence of erbium ions determined by using exponential functions to approximate the experimental curves decreased from 3.4 to 2.2 ms as the size of nanocrystals increased from 2 to 6 nm. Such long relaxation times are characteristic of the intrinsic radiative lifetime of Er^{3+} ions. For instance, in c-Si:Er, such lifetimes are

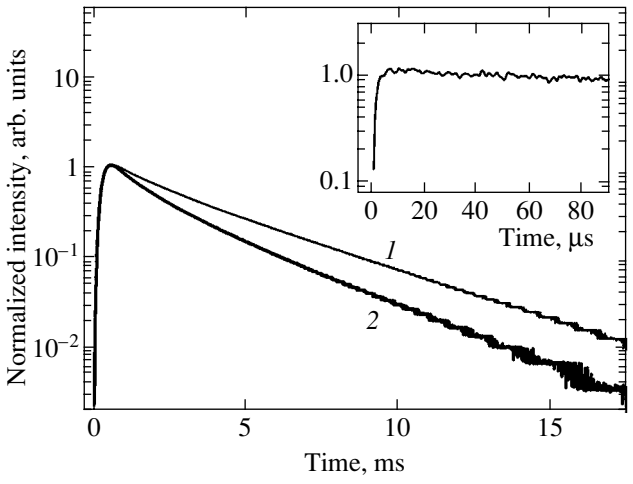

Fig. 7. Photoluminescence kinetics of Er^{3+} ions in nc-Si/SiO_2:Er samples with d = (*1*) 2 and (*2*) 6 nm (time resolution 0.5 ms). Shown in the inset is the initial kinetics region measured with a time resolution of 1 μs. Excitation by light with $\hbar\omega_1$ = 3.7 eV, T = 300 K.

only observed at liquid helium temperatures, at which deexcitation processes are suppressed [1, 2]. Some shortening of the lifetime of erbium photoluminescence observed in our experiments when the size of silicon nanocrystals increases can be explained by a stronger influence of local electric fields. Additional electric fields in dielectrically nonuniform matrices such as nc-Si/SiO_2 structures may arise from image charges induced at the boundaries between medium regions with different permittivities. The larger the size of nonuniformities, the stronger the field that acts on erbium. As a result, the electronic *f* orbitals of Er^{3+} experience additional distortion, which increases the matrix dipole moment of the transition between the first excited and ground Er^{3+} states. This should shorten photoluminescence lifetimes.

The initial region of the kinetics of erbium photoluminescence measured with a microsecond time resolution is shown in the inset to Fig. 7. The Er^{3+} photoluminescence rise times do not exceed 5 μs, which is noticeably shorter than the relaxation times of the nc-Si photoluminescence band. This lends support to the above suggestion of a high effectiveness of energy transfer from nanocrystals to Er^{3+} ions.

3.4. The Mechanism of Erbium Photoluminescence Excitation

The most probable reason for the excitation of Er^{3+} ions in the structures under consideration is electronic excitation (exciton) energy transfer in nc-Si to the ions by the mechanism of resonance dipole–dipole interaction (the Ferster mechanism) [16]. This results in the excitation of high-lying Er^{3+} energy states, whose levels can be substantially broadened because of fluctuations of electric fields in the given solid matrix (Fig. 8). In the structures under consideration, the nanocrystals are closely spaced in the oxide matrix, their density being no less than 10^{19} cm^{-3}, and are therefore separated by barriers as thin as 1–3 nm. For this reason, such a mechanism of energy transfer from nanocrystals to the ions present in the matrix is quite probable. This transfer is still more probable when an Er^{3+} ion is situated directly within a nanocrystal or on its surface.

The effectiveness of exciting erbium photoluminescence when energy is transferred from excitons substantially increases in structures containing nanocrystals with d = 2–3 nm (see Fig. 2). The excitonic photoluminescence spectrum of such structures (Fig. 1) is situated in the regions of transitions from the fourth ($^4F_{9/2}$), third ($^4I_{9/2}$), and second ($^4I_{11/2}$) excited states to the ground Er^{3+} level ($4I_{15/2}$) (Fig. 8). This increases the overlap integral between the emission spectrum of silicon nanocrystals (energy donors) and the absorption spectrum of ions (energy acceptors) and thereby increases the probability of energy transfer by the Ferster mechanism [16].

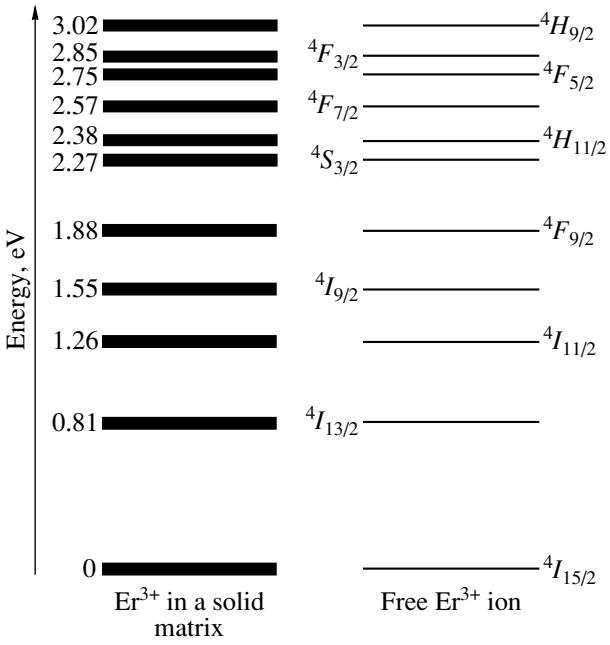

Fig. 8. Scheme of the electronic states of the Er^{3+} ion, free and implanted into a solid matrix.

4. CONCLUSIONS

We studied the luminescent properties of multilayered nc-Si/SiO_2 structures. It was shown that the energy absorbed by Si nanocrystals could be transferred with a high effectiveness to Er^{3+} ions in the surrounding oxide and further luminesced in the region of 1.5 μm. The effectiveness of energy transfer increased as the energy of pumping quanta grew and the size of the nanocrystals became smaller. These dependences were explained by effective dipole–dipole resonance interaction between excitons in silicon nanocrystals and excited states of Er^{3+} ions in SiO_2 surrounding the nanocrystals. The intensity of erbium photoluminescence additionally increased at helium temperatures because of the contribution of nonresonance processes with the participation of phonons and, possibly, the temperature-dependent contribution of erbium centers inside Si nanocrystals. It was found that the coupling between excitons and Er^{3+} ions in structures with nanocrystals 2 nm in diameter could be sufficiently strong for increasing the total yield of radiative recombination compared with undoped samples even at room temperature. The high effectiveness of the excitation of erbium photoluminescence attainable at room temperatures can, in our view, be of interest for applications and for developing optical amplifiers and light-emitting devices operating in the region of 1.5 μm.

ACKNOWLEDGMENTS

This work was financially supported by CRDF (grant no. RE2-2369), the Russian Foundation for Basic Research (project nos. 02-02-17259 and 03-02-16647), and programs of the Ministry of Science and Technologies of the Russian Federation.

REFERENCES

1. G. S. Pomrenke, P. B. Klein, and D. W. Langer, Mater. Res. Soc. Symp. Proc. **301** (1993).
2. K. Iga and S. Kinoshita, *Progress Technology for Semiconductors Lasers* (Springer, Berlin, 1996), Springer Ser. Mater. Sci., Vol. 30.
3. F. Priolo, G. Franzo, S. Coffa, *et al.*, J. Appl. Phys. **78**, 3874 (1995).
4. W. Fuhs, I. Ulber, G. Weiser, *et al.*, Phys. Rev. B **56**, 9545 (1997).
5. E. A. Konstantinova, B. V. Kamenev, P. K. Kashkarov, *et al.*, J. Non-Cryst. Solids **282**, 321 (2001).
6. B. V. Kamenev, V. Yu. Timoshenko, E. A. Konstantinova, *et al.*, J. Non-Cryst. Solids **299**, 668 (2002).
7. B. V. Kamenev, V. I. Emel'yanov, E. A. Konstantinova, *et al.*, Appl. Phys. B **74** (2), 151 (2002).
8. A. J. Kenyon, C. E. Chryssou, C. W. Pitt, *et al.*, J. Appl. Phys. **91**, 367 (2002).
9. K. Watanabe, M. Fujii, and S. Hayashi, J. Appl. Phys. **90**, 4761 (2001).
10. M. Schmidt, M. Zacharias, S. Richter, *et al.*, Thin Solid Films **397**, 211 (2001).
11. M. Zacharias, J. Heitmann, R. Shcholz, *et al.*, Appl. Phys. Lett. **80**, 661 (2002).
12. D. J. Lokwood, Z. H. Liu, and J. M. Baribeau, Phys. Rev. Lett. **76**, 539 (1996).
13. A. G. Cullis, L. T. Canham, and P. D. J. Calcott, J. Appl. Phys. **82**, 909 (1997).
14. A. Polman, J. Appl. Phys. **82**, 1 (1997).
15. J. Valenta, R. Juhasz, and J. Linnros, Appl. Phys. Lett. **80**, 1070 (2002).
16. V. M. Agranovich and M. D. Galanin, *Electronic Excitation Energy Transfer in Condensed Matter* (Nauka, Moscow, 1978; North-Holland, Amsterdam, 1982).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1131–1136.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1264–1270.

ATOMS, SPECTRA, RADIATION

Modified Jaynes–Cummings Systems and a Quantum Algorithm for the Knapsack Problem

A. Ya. Kazakov

St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, 190000 Russia
e-mail: a_kazak@mail.ru

Received April 28, 2003

Abstract—Dynamics of a system of two-level atoms interacting simultaneously with classical and quantized modes are analyzed. Both atom and cavity are assumed to interact with classical fields. The possibility of using this system as a quantum computer that solves the knapsack problem is discussed. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The possibility of solving complex computational problems by means of quantum computers and quantum algorithms has been widely discussed over the past 20 years [1–3]. In most theoretical studies, these problems were analyzed in terms of manipulations with sets of qubits, whereas physical implementations of particular computational algorithms escaped analysis. This paper deals with a certain physical system whose intrinsic properties make it a promising candidate for realizing Feynman's idea of analog quantum computing [4]. Following Feynman's suggestion, one may consider a physical system whose evolution, in a sense, solves some mathematical problems. Appropriate measurement of their characteristics is equivalent to retrieval of computed results. As an example, a doubly driven two-photon Jaynes–Cummings system is considered, and its relation to the (NP-complete) knapsack problem is established [5].

Earlier studies were focused on two- or three-level atoms interacting simultaneously with classical and quantized radiation modes [6–11]. These systems are usually described by appropriately modified Jaynes–Cummings models [12]. The aforementioned studies provided descriptions of the nonclassical properties of the quantized radiation mode and the collective effects due to interaction of N identical atoms with fields. In particular, it was found that two-photon interaction of a quantized field with atoms can result in exponential superradiance. According to [10], this effect is due to an "exponential resource" associated with exponential growth of the state-space dimension with N. It is well known that quantum computing is based on this particular resource.

The physical system analyzed in the present study is more complicated than that considered in [10]. As in [10], an atom (or N-atom system) is assumed to interact simultaneously with a classical quasi-resonant field and with a quantized field (in two-photon resonance). Furthermore, the cavity containing the quantized field interacts with another classical field. It is shown that this physical system can be used to solve the knapsack problem.

The paper is organized as follows. In the next section, a physical model of a cavity–atom system interacting with classical and quantized fields is considered and an analytical description of its dynamics is obtained, including the occupation of the quantized mode. These results are used in Section 3 to analyze the case of an N-atom system. In Section 4, it is shown how this system can be employed as a quantum computer solving the knapsack problem.

2. SINGLE ATOM

2.1. Physical Model

Consider a two-level atom that interacts simultaneously with a classical field (quasi-resonant with the atomic transition) and a quantized cavity field (in two-photon resonance). Suppose that the perfect cavity interacts with another classical mode. In the model discussed in [10, 11], only the atom was assumed to interact with an external classical field. In the rotating-wave approximation, the Hamiltonian of a doubly driven two-photon Jaynes–Cummings system is written as

$$\begin{aligned}\mathbf{H} &= \omega\mathbf{a}^{+}\mathbf{a} + \mathrm{diag}\{E_2, E_1\} + \zeta[(\mathbf{a}^{+})^2\mathbf{J}_{-} + \mathbf{a}^2\mathbf{J}_{+}] \\ &\quad + \mu[\exp(i\Omega t)\mathbf{J}_{-} + \exp(-i\Omega t)\mathbf{J}_{+}] \\ &\quad + iA[\mathbf{a}^{+}\exp(-i\omega_c t) - \mathbf{a}\exp(i\omega_c t)].\end{aligned} \tag{1}$$

Here,

$$\mathbf{J}_{-} = \mathbf{J}_{+}^{T} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

E_k are the atomic state energies (k = 1, 2); ζ is the coupling constant for the two-photon atomic transition and

the cavity field; μ and Ω are the normalized amplitude and frequency of the classical field that interacts with the atom, respectively; A and ω_c are the normalized amplitude and frequency of the classical field that interacts with the cavity, respectively; and $\mathbf{a}$ and $\mathbf{a}^+$ are the annihilation and creation operators for the quantized mode. Without loss of generality, one can assume that $E_2 = -E_1 = \kappa$. The Hamiltonian written out above acts on the space $M = F \otimes C^2$, where the Fock space F corresponds to the cavity field and C^2 is the state space of the two-level atom. It differs from the Hamiltonian used in [9–11] by the term that represents the interaction of the quantized mode with a coherent external pump [13].

To solve the Schrödinger equation

$$i\frac{\partial\Psi(t)}{\partial t} = \mathbf{H}\Psi(t)$$

for the wave function $\Psi(t)$ taking values in M, define $\mathbf{J}_0 = \mathrm{diag}(1, -1)$ and substitute

$$\Psi(t) = \exp[-it\omega(\mathbf{a}^+\mathbf{a} + \mathbf{J}_0)]\Phi(t) \quad (2)$$

to eliminate the optical frequency. The result is

$$i\frac{\partial\Phi(t)}{\partial t} = \tilde{\mathbf{H}}\Phi(t) \quad (3)$$

with

$$\tilde{\mathbf{H}} = \zeta[(\mathbf{a}^+)^2\mathbf{J}_- + \mathbf{a}^2\mathbf{J}_+] + \begin{pmatrix} \kappa-\omega & \mu\exp(2i\nu t) \\ \mu\exp(-2i\nu t) & \omega-\kappa \end{pmatrix} + iA[\mathbf{a}^+\exp(i\delta t) - \mathbf{a}\exp(-i\delta t)],$$

where $\nu = \omega - \Omega/2$ and $\delta = \omega - \omega_c$.

The solution to the problem of interaction between the atom and a classical field,

$$i\frac{\partial}{\partial t}\Xi(t) = \begin{pmatrix} \kappa-\omega & \mu\exp(2i\nu t) \\ \mu\exp(-2i\nu t) & \omega-\kappa \end{pmatrix}\Xi(t),$$

$$\Xi(0) = I,$$

where I is the identity matrix, is found in explicit form:

$$\Xi(t) = \exp(i\nu t\mathbf{J}_0)U\exp\left(-\frac{iRt\mathbf{J}_0}{2}\right)U^{-1},$$

$$U = \begin{pmatrix} \mu & \Delta - R \\ R - \Delta & \mu \end{pmatrix},$$

$$R = \sqrt{\mu^2 + \Delta^2}, \quad \Delta = \kappa - \frac{\Omega}{2}.$$

2.2. Averaging Procedure and the Averaged Hamiltonian

Following [9–11], assume that

$$R_q \ll R_{cl} \ll \Omega, \quad (4)$$

where Ω is an optical frequency, R_{cl} is the Rabi frequency of the classical field, and R_q is the effective Rabi frequency of the quantized field. Under these assumptions, the quantum system can be decomposed into fast and slow components whose dynamics correspond to the interaction with the classical fields and the interaction of the atom field-dressed states with the quantized mode, respectively.

To decouple the fast and slow components of the physical system considered here, substitute

$$\Phi(t) = \Xi(t)\varphi(t) \quad (5)$$

into Eq. (3). The wave function $\varphi(t)$ obeys the equation

$$i\frac{\partial}{\partial t}\varphi(t) = \Lambda_1(t)\varphi(t), \quad (6)$$

where

$$\Lambda_1(t) = \zeta[\Xi(t)]^{-1}[(\mathbf{a}^+)^2\mathbf{J}_- + \mathbf{a}^2\mathbf{J}_+]\Xi(t) + iA[\mathbf{a}^+\exp(i\delta t) - \mathbf{a}\exp(-i\delta t)].$$

Hereinafter, the parameters A and ζ are assumed to be similar in order of magnitude.

To calculate the dynamics up to the leading order in the small parameter $(A, R_q)/R$, Eq. (6) should be averaged. Suppose that the condition $|A|, |\nu|, |\delta| \ll R$ holds throughout the analysis that follows. Alternative methods for decoupling the slow and fast components of Jaynes–Cummings systems interacting with classical fields have been discussed in the context of different physical problems [14–19].

The dynamics of the slow component are governed by an averaged Hamiltonian:

$$i\frac{\partial}{\partial t}\varphi(t) = \mathbf{H}_{av}\varphi(t),$$

$$\mathbf{H}_{av} = \langle\langle[\Xi(t)]^{-1}\{\zeta[(\mathbf{a}^+)^2\mathbf{J}_- + \mathbf{a}^2\mathbf{J}_+]\}\Xi(t)\rangle\rangle + iA[\mathbf{a}^+\exp(i\delta t) - \mathbf{a}\exp(-i\delta t)],$$

where $\langle\langle\ldots\rangle\rangle$ means deletion of fast harmonics. Under the present assumptions (see [9–11] for details),

$$\mathbf{H}_{av} = [\rho((\mathbf{a}^+)^2\exp(2i\nu t) + \mathbf{a}^2\exp(-2i\nu t))] \times U\mathbf{J}_0U^{-1} + iA[\mathbf{a}^+\exp(i\delta t) - \mathbf{a}\exp(-i\delta t)], \quad (7)$$

$$\rho = \frac{\zeta\mu}{2\sqrt{\mu^2 + \Delta^2}}.$$

It is important that the first term in (7) is the product of the matrix $U\mathbf{J}_0U^{-1}$ with a Fock operator, whereas the second term is a Fock operator. The vectors

$$\mathbf{e}_1 = \frac{1}{\sqrt{D}}\begin{pmatrix}\mu\\ R-\Delta\end{pmatrix}, \quad \mathbf{e}_2 = \frac{1}{\sqrt{D}}\begin{pmatrix}\Delta-R\\ \mu\end{pmatrix},$$

where $D = \mu^2 + (R - \Delta)^2$, are the mutually orthogonal normalized eigenvectors of $U\mathbf{J}_0U^{-1}$. The corresponding eigenvalues are $\vartheta_k = (-1)^{k+1}$ ($k = 1, 2$). By decomposition in this basis in C^2, Hamiltonian (7) splits into a pair of one-dimensional Fock operators that are identical up to sign. Therefore, the analysis can be restricted to the initial value problem for the equation

$$i\frac{\partial}{\partial t}\varphi(t) = \mathbf{H}_{\text{av}}^{(1)}\varphi(t) \tag{8}$$

with

$$\mathbf{H}_{\text{av}}^{(1)} = \rho[(\mathbf{a}^+)^2\exp(2i\nu t) + \mathbf{a}^2\exp(-2i\nu t)] + iA[\mathbf{a}^+\exp(i\delta t) - \mathbf{a}\exp(-i\delta t)].$$

2.3. Solution of the Initial Value Problem

Consider Schrödinger equation (8) written in the Bargmann–Fock representation. Representing the corresponding wave function as an analytic function $\eta(z, t)$ ($\mathbf{a}^+ \longrightarrow z$, $\mathbf{a} \longrightarrow \partial/\partial z$), write the Schrödinger equation

$$i\eta_t = \rho[\exp(2i\nu t)z^2\eta + \exp(-2i\nu t)\eta_{zz}] + iA[z\exp(i\delta t)\eta - \exp(-i\delta t)\eta_z]. \tag{9}$$

The solution to an initial value problem for this equation can be written in explicit form [10, 11]. To avoid cumbersome calculations, consider the particular case of a vacuum initial state: $\eta(0) = 1$. In this case, the solution to Eq. (9) is

$$\eta(z, t) = \exp[\lambda(t) + \xi(t)z + \gamma(t)z^2],$$

where

$$\gamma(t) = -\rho\exp(2i\nu t)\frac{\chi(t)}{\sigma(t)},$$

$$\sigma(t) = (\nu + \sqrt{\nu^2 - 4\rho^2})\exp(it\sqrt{\nu^2 - 4\rho^2}) - (\nu - \sqrt{\nu^2 - 4\rho^2})\exp(-it\sqrt{\nu^2 - 4\rho^2}),$$

$$\chi(t) = \exp(it\sqrt{\nu^2 - 4\rho^2}) - \exp(-it\sqrt{\nu^2 - 4\rho^2}),$$

$$\xi(t) = \frac{2\exp(i\nu t)p(t)}{\sigma(t)},$$

$$p(t) = iA\left\{\frac{\rho\exp[i(\nu - \delta - \sqrt{\nu^2 - 4\rho^2})t]}{\nu - \delta - \sqrt{\nu^2 - 4\rho^2}}\right.$$

$$+\frac{(\nu - \sqrt{\nu^2 - 4\rho^2})\exp[i(\delta - \nu - \sqrt{\nu^2 - 4\rho^2})t]}{2(\delta - \nu - \sqrt{\nu^2 - 4\rho^2})}$$

$$-\frac{\rho\exp[i(\nu - \delta + \sqrt{\nu^2 - 4\rho^2})t]}{\nu - \delta + \sqrt{\nu^2 - 4\rho^2}}$$

$$-\frac{(\nu + \sqrt{\nu^2 - 4\rho^2})\exp[i(\delta - \nu + \sqrt{\nu^2 - 4\rho^2})t]}{2(\delta - \nu + \sqrt{\nu^2 - 4\rho^2})}$$

$$\left.-\sqrt{\nu^2 - 4\rho^2}\frac{(2\nu + \rho) - \delta}{(\nu - \delta)^2 - (\nu^2 - 4\rho^2)}\right\}.$$

The unwieldy expression for $\lambda(t)$ is omitted here. These relations can be used to calculate any quantum-statistical characteristic of the quantized field. The occupation dynamics of the quantized field required for further analysis is given by

$$n(t) = \frac{4\rho^2\sin^2(t\sqrt{\nu^2 - 4\rho^2})}{\nu^2 - 4\rho^2} + \frac{(1 + 4|\gamma|^2)|\xi|^2 + 4|\gamma|^2\text{Re}(\xi^2\gamma^{-1})}{(1 - 4|\gamma|^2)^2}. \tag{10}$$

The following observations should be made here. When $A = 0$ (the classical field interacting with the cavity vanishes), this is equivalent to the result obtained in [9]. When $|\nu| < 2|\rho|$, the occupation of the quantized mode grows exponentially. In an N-atom system, this leads to exponential superradiance. Suppose that the reverse inequality, $|\nu| > 2|\rho|$, is true. (Recall that $\nu = \omega - \Omega/2$ and the classical-mode frequency Ω is an easily controllable parameter.) In the general case, the quantized-mode occupation given by (10) is a complicated function, but it is easy to see that it is a bounded oscillating function of time. However, if

$$\delta = \nu \pm \sqrt{\nu^2 - 4\rho^2},$$

then the dynamics described by $n(t)$ is qualitatively different, since it is a quadratic function of time. For

example, if $\delta = \nu + \sqrt{\nu^2 - 4\rho^2}$, then

$$\xi(t) = A\frac{(2\rho + \sqrt{\nu^2 - 4\rho^2} - \nu)\exp(i\nu t)}{\sigma(t)}t + \epsilon(t),$$

$$\epsilon(t) = -i\frac{2A\exp(i\nu t)}{\sigma(t)}$$
$$\times\left\{\frac{1}{4\sqrt{\nu^2 - 4\rho^2}}[2\rho\exp(-2it\sqrt{\nu^2 - 4\rho^2})\right.$$
$$+ (\nu + \sqrt{\nu^2 - 4\rho^2})\exp(2it\sqrt{\nu^2 - 4\rho^2})]$$
$$\left. - \frac{\nu + 2\rho + \sqrt{\nu^2 - 4\rho^2}}{4\sqrt{\nu^2 - 4\rho^2}}\right\},$$

where $\epsilon(t)$ is a bounded oscillating function of time. Accordingly, one can retain only the terms with the fastest growing amplitude to obtain

$$n(t) \simeq t^2\frac{A^2(2\rho + \sqrt{\nu^2 - 4\rho^2} - \nu)^2}{4(\nu^2 - 4\rho^2)^2}$$
$$\times[\nu^2 - 4\rho^2\cos(2t\sqrt{\nu^2 - 4\rho^2})$$
$$+ 4\rho\nu\sin^2(t\sqrt{\nu^2 - 4\rho^2})]. \quad (11)$$

Recall that $\delta = \omega - \omega_c$, where ω_c is the frequency of the classical mode that interacts with the cavity. This parameter is also easy to control in an experiment. Relation (11) plays a key role in the analysis presented below.

3. INTERACTION OF A MANY-ATOM SYSTEM WITH CLASSICAL AND QUANTIZED FIELDS

Now, consider a system of N two-level atoms interacting with classical fields and a quantized field in a cavity, which, in turn, interacts with a classical mode. Suppose that each atom interacts with a respective classical field whose parameters can be controlled. Let us determine the occupation dynamics of the quantized mode. The wave function of this physical system takes values in the space $M_N = F \otimes (C^2)^N$, where F is again the Fock space of states for quantized mode and each of the N copies of C^2 is an atomic state space. Define $\mathbf{J}_0^{(m)}$, $\mathbf{J}_\pm^{(m)}$, $U^{(m)}$, and $\Xi^{(m)}(t)$ as operators acting on the mth component of the wave function similar to $\mathbf{J}_0$, $\mathbf{J}_\pm$, U, and $\Xi(t)$, respectively. The dynamics of this system is determined by a natural extension of Hamiltonian (1). In the rotating-wave approximation, it is written as

$$\mathbf{H}_N = \omega\mathbf{a}^+\mathbf{a}$$
$$+ \sum_{m=1}^{N}\{\kappa\mathbf{J}_0^{(m)} + \zeta[(\mathbf{a}^+)^2\mathbf{J}_-^{(m)} + \mathbf{a}^2\mathbf{J}_+^{(m)}]\}$$
$$+ \sum_{m=1}^{N}[\mu_m(\exp(i\Omega t)\mathbf{J}_-^{(m)} + \exp(-i\Omega t)\mathbf{J}_+^{(m)})]$$
$$+ iA[\mathbf{a}^+\exp(-i\omega_c t) - \mathbf{a}\exp(i\omega_c t)].$$

Here, μ_m and Ω are the normalized amplitude and frequency of the classical field that interacts with the mth atom. Let us solve the initial value problem for the corresponding Schrödinger equation.

Using an analog of (2),

$$\Psi(t) = \exp\left[-it\omega\left(\mathbf{a}^+\mathbf{a} + \sum_{m=1}^{N}\mathbf{J}_0^{(m)}\right)\right]\Phi(t),$$

to eliminate the optical frequency, one obtains

$$i\frac{\partial\Phi(t)}{\partial t} = \tilde{\mathbf{H}}_N\Phi(t),$$

$$\tilde{\mathbf{H}}_N = \sum_{m=1}^{N}\{(\kappa - \omega)\mathbf{J}_0^{(m)} + \zeta[(\mathbf{a}^+)^2\mathbf{J}_-^{(m)} + \mathbf{a}^2\mathbf{J}_+^{(m)}]$$
$$+ \mu_m\exp(-2i\nu t)\mathbf{J}_-^{(m)} + \mu_m\exp(2i\nu t)\mathbf{J}_+^{(m)}\}$$
$$+ iA[\mathbf{a}^+\exp(-i\delta t) - \mathbf{a}\exp(i\delta t)],$$

where δ and ν are the quantities defined above.

The matrix

$$\Xi_N(t) = \prod_{m=1}^{N}\Xi^{(m)}(t)$$

solves the equation

$$i\frac{\partial}{\partial t}\Xi_N(t) = \sum_{m=1}^{N}\{(\kappa - \omega)\mathbf{J}_0^{(m)}$$
$$+ \mu_m\exp(-2i\nu t)\mathbf{J}_-^{(m)} + \mu_m\exp(2i\nu t)\mathbf{J}_+^{(m)}\}\Xi_N(t)$$

and equals the identity matrix at the initial moment. Substituting $\Phi(t) = \Xi_N(t)\varphi(t)$ leads to an equation

analogous to (6):

$$i\frac{\partial}{\partial t}\varphi(t) = \Lambda_N(t)\varphi(t),$$

$$\Lambda_N(t) = [\Xi_N(t)]^{-1}$$
$$\times \sum_{m=1}^{N} \zeta[(\mathbf{a}^+)^2\mathbf{J}_-^{(m)} + \mathbf{a}^2\mathbf{J}_+^{(m)}]\Xi_N(t)$$
$$+ iA[\mathbf{a}^+\exp(-i\delta t) - \mathbf{a}\exp(i\delta t)].$$

Again, this equation is averaged to eliminate fast harmonics under the assumption that $|\nu|$, $|\zeta|$, and $|A| \ll R$. The result is

$$i\frac{\partial}{\partial t}\varphi(t) = \mathbf{H}_{N,\,\mathrm{av}}\varphi(t),$$

$$\mathbf{H}_{N,\,\mathrm{av}} = \langle\langle\Lambda_N(t)\rangle\rangle$$
$$= [(\mathbf{a}^+)^2\exp(2i\nu t) + \mathbf{a}^2\exp(-2i\nu t)]$$
$$\times \sum_{m=1}^{N} \rho_m U^{(m)}\mathbf{J}_0^{(m)}(U^{(m)})^{-1} \qquad (12)$$
$$+ iA[\mathbf{a}^+\exp(-i\delta t) - \mathbf{a}\exp(i\delta t)],$$

where ρ_m is the analog of ρ for the mth atom. Two remarks should be made here. First, each ρ_m is related to the corresponding μ_m and can be controlled (within certain limits) in experiment. Second, the averaged Hamiltonian is again represented as the sum of a Fock operator and the product of a Fock operator with a matrix operator.

The eigenvectors can easily be found and represented as $|\mathbf{e}_{k_1}, \mathbf{e}_{k_2}, \ldots, \mathbf{e}_{k_N}\rangle$ ($k_m = 1, 2$; $m = 1, 2, \ldots, N$). The decomposition of the wave function in this basis has the form

$$\varphi(t) = \sum_{\sigma}\eta_\sigma(t)|\mathbf{e}_{k_1}, \mathbf{e}_{k_2}, \ldots, \mathbf{e}_{k_N}\rangle,$$

where the functions $\eta_\sigma(t)$ take values in the Fock space and σ denotes a set of N numbers $k_1, k_2, \ldots, k_N$ equal to 1 or 2. The above sum runs over all such sets. Recall that the eigenvalues of each matrix component in (12) are $\vartheta_1 = 1$ and $\vartheta_2 = -1$. Accordingly, each $\eta_\sigma(t)$ obeys the equation

$$i\frac{\partial}{\partial t}\eta_\sigma(t)$$
$$= S_\sigma[(\mathbf{a}^+)^2\exp(2i\nu t) + \mathbf{a}^2\exp(-2i\nu t)]\eta_\sigma(t) \qquad (13)$$
$$+ iA[\mathbf{a}^+\exp(-i\delta t) - \mathbf{a}\exp(i\delta t)],$$

with

$$S_\sigma = \sum_{m=1}^{N}\vartheta_{k_m}\rho_m, \qquad (14)$$

where the numbers k_m in the sum constitute the set σ. The solutions to the initial value problems for Eqs. (13) and (8) are identical (under the change $\rho \longrightarrow S_\sigma$).

Again, consider the quantized field evolving from the vacuum initial state. The occupation of the quantized mode is found by using previous results:

$$n(t) = \sum_{(\sigma)}|c_\sigma|^2 n_\sigma(t),$$

where c_σ is the projection of the initial state of the many-atom system on the eigenvector corresponding to a set σ and the sum runs over all sets σ. When the value of μ_m is known for each m, the initial atomic states can be prepared so as to ensure that $|c_\sigma| = 2^{-N/2}$ for all sets. Note that the corresponding sum consists of 2^N summands.

4. KNAPSACK PROBLEM

Consider a symmetric formulation of the knapsack problem that is equivalent to the standard one [5]. Take N numbers $b_1, b_2, \ldots, b_N$. Denote a set $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ of numbers ± 1 by σ. For every σ, calculate the sum

$$T_\sigma = \sum_{m=1}^{N}\varepsilon_m b_m. \qquad (15)$$

Now, take a number $B = T_{\sigma_B}$ corresponding to some (unique) set σ_B. In the knapsack problem, σ_B is sought, i.e., the value of every ε_m ($m = 1, 2, \ldots, N$) in this set. When the problem is solved by direct search, one must consider 2^{N-1} cases (by virtue of symmetry), and the time complexity is proportional to 2^N.

Now, consider relationship between the results obtained above and the knapsack problem. As a starting point, note the formal similarity of the expressions for S_σ and T_σ given by (14) and (15), respectively. To be specific, suppose that the initial atomic states are prepared so that $|c_\sigma| = 2^{-N/2}$ for all sets. As shown above, the classical field amplitudes μ_m can be used to vary the values of ρ_m within certain limits. By means of an appropriate scaling, the values of μ_m can be adjusted to obtain ρ_m whose ratios are equal to the respective ratios of the numbers b_m. By virtue of the scaling, the values of S_σ are equal to T_σ if ϑ_k are identical to ε_k ($k = 1, 2$). Furthermore, choose a value of ν such that $\nu^2 > 4S_\sigma^2$ for each σ and a value of δ such that

$$\delta = \nu + \sqrt{\nu^2 - 4S_{\sigma_B}^2}$$

for σ_B. Then, the sum in

$$n(t) = 2^{-N}\sum_{(\sigma)} n_\sigma(t)$$

contains exactly two summands growing as quadratic functions of time while the remaining ones are bounded oscillating functions of time. One of these summands corresponds to σ_B; the other, to the set with opposite signs. After a time interval of order $2^{N/2}$ has elapsed, the existence of quadratic summands can be detected by measuring the occupation of the quantized mode.

When a single set σ_B makes up the sum in question, the values of ε_m can be determined as follows. Note that S_{σ_B} is a monotone increasing or decreasing function of μ_m, depending on the value of ε_m. Suppose that $S_{\sigma_B} > 0$. Increase μ_m to increase ρ_m for a particular m and change the value of δ accordingly, assuming that $\varepsilon_m = 1$. If this results in a quadratic increase in $n(t)$, then $\varepsilon_m = 1$; otherwise, $\varepsilon_m = -1$. The set σ_B is determined by examining all atoms (in N experiments). Thus, the time complexity of the determination of σ_B is proportional to $N2^{N/2}$.

5. CONCLUSIONS

The interaction of N atoms with classical fields (specific for each atom) and a quantized mode interacting with yet another classical one is analyzed. It is shown that this system can be considered as a model of an analog quantum computer that solves the knapsack problem. According to the present analysis, this possibility relies on an exponential resource: the dimension of the phase space of the physical system in question is 2^N. The corresponding time complexity of the knapsack problem is proportional to $N2^{N/2}$.

It is understood that the physical model considered here is highly simplified. However, the study has been undertaken to explore the parallel between the knapsack problem and a (more or less) realistic quantum system in the spirit of Feynman's suggestion [4]. Factors that may impede practical implementation of this quantum computer have been left outside the scope of this study.

REFERENCES

1. P. W. Shor, SIAM J. Comput. **26**, 1484 (1997).
2. A. Steane, quant-ph/9708022.
3. J. Preskill, quant-ph/9712048.
4. R. P. Feynman, J. Theor. Phys. **21**, 467 (1982).
5. N. Koblits, *Course of the Theory of Numbers and Cryptography* (TVP, Moscow, 2001).
6. A. Ya. Kazakov, Quantum Semiclassic. Opt. **10**, 753 (1998).
7. R. A. Ismailov and A. Ya. Kazakov, Zh. Éksp. Teor. Fiz. **116**, 858 (1999) [JETP **89**, 454 (1999)].
8. R. A. Ismailov and A. Ya. Kazakov, Zh. Éksp. Teor. Fiz. **118**, 798 (2000) [JETP **91**, 691 (2000)].
9. A. Ya. Kazakov, J. Opt. B: Quantum Semiclassic. Opt. **3**, 97 (2001).
10. R. A. Ismailov and A. Ya. Kazakov, Zh. Éksp. Teor. Fiz. **120**, 1172 (2001) [JETP **93**, 1017 (2001)].
11. A. Ya. Kazakov, Int. J. Theor. Phys., Groups Theor. Nonlinear Opt. **8**, 75 (2002).
12. E. T. Jaynes and F. W. Cummings, Proc. IEEE **51**, 89 (1963).
13. J. P. Clemens and P. R. Rice, Phys. Rev. A **61**, 063810 (2000).
14. C. K. Law and J. H. Eberly, Phys. Rev. A **43**, 6337 (1991).
15. P. Alsing, D.-S. Guo, and H. J. Carmichael, Phys. Rev. A **45**, 5135 (1992).
16. I. V. Jyotsna and G. S. Agarwal, Opt. Commun. **99**, 344 (1993).
17. Y.-T. Chough and H. J. Carmichael, Phys. Rev. A **54**, 1709 (1996).
18. F.-L. Li and S.-Y. Gao, Phys. Rev. A **62**, 043809 (2000).
19. A. Joshi, Phys. Rev. A **62**, 043812 (2000).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1137–1145.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1271–1280.

PLASMA, GASES

Efficiency of Hydrodynamic Energy Transfer to an Arbitrarily Thick Flat Layer of Material during Pulsed Ablation

S. Yu. Gus'kov

Lebedev Physical Institute, Russian Academy of Sciences, Leninskii pr. 53, Moscow, 119991 Russia

e-mail: guskov@sci.lebedev.ru

Received April 3, 2003

Abstract—We obtained a general analytical solution of the problem of hydrodynamic energy transfer to a flat layer of material with an arbitrary initial thickness when ablation—the evaporation of material and the formation of a pressure gradient under the action of an external pulsed energy source—takes place at one of its surfaces. The solution was obtained in the form of a dependence of the fraction of the source energy transferred to the nonevaporated part of the layer on the intensity and duration of the energy source as well as on the initial layer thickness and density. The solution includes, as limiting cases, the previously obtained solutions for the hydrodynamic transfer coefficient during the ablation acceleration of a thin layer, through which the travel time of a shock or acoustic wave is much shorter than the duration of the energy source, and for the ablation loading efficiency when a shock wave propagates through a semiinfinite layer. *© 2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

An intense flux of radiation in a wide range of parameters acts on a solid material via ablation. The ablation process consists in the evaporation of material and the formation of pressure in the outer part of the target and, as a result, in the excitation of hydrodynamic motion in the nonevaporated part of the target. By the hydrodynamic energy transfer efficiency under such an action, we mean the fraction of the radiation energy that is transferred to the nonevaporated part of the target. The energy transfer to a material via ablation is of fundamental importance for a wide range of problems related to the action of intense radiation fluxes, such as laser and X-ray radiation, beams of heavy ions, etc., on materials. These problems primarily include inertial thermonuclear fusion and material processing.

Previously, solutions for the hydrodynamic energy transfer efficiency during ablation have been obtained for two limiting cases of the problem: a thin layer, through which the travel time of the hydrodynamic wave (a shock or acoustic wave) is much shorter than the laser pulse duration, [1] and a semiinfinite layer [2].

In [1], the hydrodynamic energy transfer efficiency (or, following the terminology of the authors of this paper, the hydrodynamic transfer coefficient) was determined by solving the problem of the acceleration of a thin layer under the pressure of the material evaporated at the layer boundary during its heating by an external energy source. The relations between the parameters of the materials from the evaporated and nonevaporated parts of the layer and the evaporation wave velocity were derived from the continuity conditions at the hydrodynamic discontinuity used to simulate the evaporation surface. The velocity of the nonevaporated part of the layer was determined by solving the equation of motion for a flat layer of variable mass under the pressure of the evaporated material. This statement of the problem is valid for times later than the time at which the shock or acoustic wave emerges at the back surface of the layer. Therefore, the solution from [1] is valid only for a sufficiently thin layer or a sufficiently long pulse of the energy source such that the pulse duration exceeds the travel time of the shock or acoustic wave through the layer. As a result, the solution is a function of only the fraction of the evaporated mass of the layer. This statement of the problem corresponds to the physical conditions for the ablation acceleration of a material when inertial thermonuclear fusion targets are compressed. Therefore, the solution obtained in [1] accurately describes the entire data set of the numerous experiments on the ablation acceleration of thin foils under the action of a laser pulse, including the data pertaining to the hydrodynamic energy transfer efficiency as a whole and the separate measurements of the evaporated mass and the final target velocity (see, e.g., [3]). In turn, the solution of a similar problem for a thin spherical layer [4, 5] agrees well with the experimental data on the acceleration toward the center of a spherical shell target under the action of a laser pulse [4, 6].

The problem of energy transformation from a pulsed energy source into shock energy during ablation on the surface of an infinitely thick flat layer of material was solved in [2]. The fraction of the energy of the

1063-7761/03/9706-1137$24.00

external source transformed into the energy of the shock wave propagating deep into the layer or, according to the term introduced in [2], the ablation loading efficiency of the material is a function of the density ratio of the evaporated and nonevaporated parts of the target. The solution from [2] is consistent with the results of the experiments on the action of a laser pulse on the surfaces of thick samples of various materials (in particular, metals) carried out in a wide range of radiation intensities, from 10^8 to 10^{14} W cm^{-2}. This solution gives an accurate value, for example, for the material destruction depth under the action of a laser pulse (see, e.g., [7]).

In this paper, based on an analysis of the dynamics of shock and acoustic waves that propagate via ablation in a finite-thickness layer of material, we obtain a general solution of the problem of the efficiency of hydrodynamic energy transfer from an external source to the layer of material (below, we use the term "hydrodynamic efficiency") at arbitrary values of the layer thickness and the duration of the energy source. It includes, as limiting cases, the solution for the hydrodynamic transfer coefficient during the ablation acceleration of a thin layer and the solution for the ablation loading efficiency when a hydrodynamic wave propagates through a semiinfinite layer.

2. STATEMENT OF THE PROBLEM

The external energy pulse is assumed to have a constant intensity. The hydrodynamic wave in the target produced by the external energy pulse can be either a shock or acoustic wave, depending on the intensity of the energy pulse from the external source, because the latter unambiguously determines the ablation pressure and, hence, the hydrodynamic perturbation pressure amplitude. We will consider the excess of the hydrodynamic perturbation travel velocity above the speed of sound in the unperturbed material of the layer to be the shock formation criterion; otherwise, we will assume that an acoustic wave is excited.

Next, we assume that the duration of the external energy source can be arbitrarily related to the travel time of the hydrodynamic wave through the layer. If the duration of the energy source is shorter than the travel time of the hydrodynamic wave through the layer, then energy is transferred as the first shock or acoustic wave propagates through the layer. In this case, energy is transferred not to the entire layer but only to the part of it that is bounded on the one side by the evaporation wave front and on the other side by the hydrodynamic wave front. If the duration of the energy source is longer than the travel time of the first hydrodynamic wave through the entire layer, energy is transferred by a sequence of waves whose period at constant intensity of the energy source decreases with decreasing thickness of the layer via evaporation. At constant source intensity, either all hydrodynamic waves or all but the first wave can be acoustic.

Naturally, the acceleration of a finite-thickness layer because of the propagation of a shock or acoustic wave through its entire thickness is described by the Newton integral equation that operates with the concepts of mass and velocity of the entire object. Indeed, in the travel time of the hydrodynamic wave through the entire thickness of the layer, $\delta t = \Delta_0/D_h$ (Δ_0 is the layer thickness, and D_h is the wave velocity), the velocity of the entire layer increases by a value equal to the velocity of the material behind the hydrodynamic wave front, $\delta u = V_h$. Hence,

$$\frac{\delta u}{\delta t} = \frac{V_h D_h}{\Delta_0}. \tag{1}$$

Next, we use the standard formulas for the velocities D_h and V_h of an acoustic wave,

$$D_h = c_0, \quad V_h = \frac{P_c}{\rho_0 c_0}, \tag{2}$$

and a shock wave (for simplicity, a strong shock wave),

$$D_h = \left(\frac{\gamma+1}{2}\frac{P_c}{\rho_0}\right)^{1/2}, \quad V_h = \left(\frac{2}{\gamma+1}\frac{P_c}{\rho_0}\right)^{1/2}. \tag{3}$$

Here, P_c, γ, and ρ_0 are, respectively, the pressure behind the wave fronts, the adiabatic index, and the density of the unperturbed material, and

$$c_0 = \left(\gamma\frac{P_c}{\rho_0}\right)^{1/2} \tag{4}$$

is the speed of sound in it. Substituting expression (2) or (3) into (1) yields for both cases

$$\frac{du}{dt} = \frac{P_c}{\Delta_0\rho_0}. \tag{5}$$

The longer the duration of the energy source compared to the travel time of the hydrodynamic wave through the layer, the higher the accuracy of describing the acceleration of the layer by the Newton integral equation.

Given the above circumstances, a model of the energy transfer to an arbitrarily thick flat layer can be constructed as follows. If the source duration exceeds the travel time of the first hydrodynamic wave through the layer, then at the initial stage of the process whose duration is limited by the time at which the hydrodynamic wave emerges at the back surface of the layer, t_b, the energy transfer is described as the result of hydrodynamic wave propagation through the layer. Once the hydrodynamic wave has emerged at the back surface of the layer, starting from a time $t \geq t_b$ when the entire nonevaporated part of the layer is in motion, we may assume that energy is transferred to the entire nonevap-

orated mass of the layer via its acceleration as a whole under the pressure of the evaporated target material. This process can be described by the Newton equation for the velocity and mass of the entire nonevaporated part of the layer. If the duration of the source is shorter than the travel time of the first hydrodynamic wave through the layer, then the energy transfer is described only as the energy transfer from the hydrodynamic wave. The following statements of the problem for two types of hydrodynamic waves correspond to the above model.

(1) The duration of the energy source is shorter than the travel time of the hydrodynamic wave through the layer.

(a) Shock wave. In this case, there are two hydrodynamic discontinuities: the evaporation boundary and the shock front. We assume that the spatial density, pressure, and velocity distributions of the material in all parts of the layer are uniform. The energy flux from an external source with intensity I falls on the evaporation boundary and transforms into the kinetic and thermal energy fluxes of the nonevaporated, I_c, and evaporated, I_a, parts of the layer. $I = I_c + I_a$. The relations between the parameters of the material on both sides of the evaporation boundary in the evaporated part of the target and the nonevaporated part of the target behind the shock front can be derived from the continuity conditions for the mass, momentum, and energy fluxes at discontinuity:

$$\begin{aligned} \rho_c D_{\text{ev}} &= \rho_a(v + u + D_{\text{ev}}), \\ P_c &= P_a + \rho_c D_{\text{ev}}(v + u), \\ I &= \rho_c D_{\text{ev}}\left[\varepsilon_a + \Omega + \frac{1}{2}(v + u)^2\right] + P_a(v + u). \end{aligned} \tag{6}$$

Here, $\rho_{a(c)}$ and $P_{a(c)}$ are, respectively, the density and pressure of the evaporated material (the material of the nonevaporated part of the target behind the shock front);

$$\varepsilon_a = \frac{P_a}{(\gamma_a - 1)\rho_a}$$

is the internal energy of the evaporated material; Ω is the binding energy of the material in the initial state; v, u, and D_{ev} are, respectively, the velocity of the evaporated material, the velocity of the material in the nonevaporated part of the layer behind the shock front, and the evaporation wave velocity, which are related by the Jouguet condition

$$v + u + D_{\text{ev}} = c_a \equiv \left(\gamma_a \frac{P_a}{\rho_a}\right)^{1/2}, \tag{7}$$

where c_a is the adiabatic speed of sound. The relations between the parameters of the unperturbed material and the material behind the shock front can be derived from the continuity conditions for the mass, momentum, and energy fluxes at the shock front. For a strong shock [8],

$$\begin{aligned} \rho_c &= \frac{\gamma_c + 1}{\gamma_c - 1}\rho_0, \quad P_c = \frac{2}{\gamma_c + 1}\rho_0 D_w^2, \\ u &= \frac{2}{\gamma_c + 1} D_w, \end{aligned} \tag{8}$$

where ρ_0 is the density of the layer material in a normal state, D_w is the shock front velocity, and γ_c is the adiabatic index for the layer material.

Thus, as the shock wave propagates, the specific kinetic and thermal energies transferred to the mass of the nonevaporated material of the layer behind the shock front,

$$M_c(t) = (D_w \rho_0 - D_{\text{ev}} \rho_c)t, \tag{9}$$

are

$$\varepsilon_k = \frac{u^2}{2}, \quad \varepsilon_t = \frac{1}{\gamma_c - 1}\frac{P_c}{\rho_c}. \tag{10}$$

(b) Acoustic wave. The parameters of the evaporated material and the nonevaporated material behind the shock front are related by the continuity equations (6) at the evaporation boundary with the Jouguet relation (7), while the parameters of the nonevaporated material behind the shock front and the unperturbed material are related by the acoustic relations

$$\rho_c \approx \rho_0, \quad u \approx \frac{P_c}{\rho_0 c_0}. \tag{11}$$

The formulas for the energies ε_k and ε_t and the mass of the nonevaporated part of the layer to which energy is transferred clearly differ from (9) and (10):

$$\begin{aligned} \varepsilon_k &= \frac{u^2}{2}, \quad \varepsilon_t = \frac{1}{\gamma_c - 1}\frac{P_c}{\rho_0}, \\ M_c(t) &= (c_0 - D_{\text{ev}})\rho_0 t. \end{aligned} \tag{12}$$

(2) The duration of the energy source is longer than the travel time of the hydrodynamic wave through the layer.

Initially, until the wave emerges at the back surface of the layer, the statement of the problem is identical to the previous case. Once the wave has emerged at the back surface, when the entire mass of the layer is in motion, the problem is described by the equation of motion for a layer of variable mass that decreases via evaporation

$$M\frac{du}{dt} = P_c, \quad \frac{dM}{dt} = -D_{\text{ev}}. \tag{13}$$

The pressure P_c and the evaporation wave velocity D_{ev} can be expressed in terms of the parameters of the external energy source and the initial parameters of the layer by solving the continuity equations at the evaporation boundary (6) using (7). The initial conditions are the velocity of the material behind the hydrodynamic wave front and the mass of the layer at the time the wave emerges at the back surface of the layer:

$$u_0 = u|_{t=t_b}, \quad M_0 = M|_{t=t_b}. \tag{14}$$

The latter can be determined from relations (6)–(10) or (6), (7), (11), (12) and by solving the problem at the first stage, respectively, for the shock and acoustic waves.

3. A GENERAL SOLUTION FOR THE HYDRODYNAMIC EFFICIENCY

In this section, we present the solutions of the problem in the approximation of a low density of the evaporated target material compared to the density of the layer in the unperturbed state, $\rho_a \ll \rho_c$, and in the approximation of a low binding energy of the material compared to the internal energy density of the evaporated material, $\Omega \ll \varepsilon_a$.

By simultaneously solving the continuity equations at the evaporation boundary (6) with the Jouguet condition (7), we obtain the following universal (for both types of initial hydrodynamic waves) expressions that relate the evaporation wave velocity and the pressure of the material behind the hydrodynamic wave front, respectively, to the speed of sound and the pressure of the material in the evaporated part of the target and, eventually, to the intensity of the energy source and the densities of the evaporated material and the nonevaporated material behind the hydrodynamic wave front:

$$D_{ev} = \frac{\rho_a}{\rho_c} c_a, \quad P_c = (\gamma_a + 1) P_a, \tag{15}$$

where

$$P_a = \frac{1}{\gamma_a}\rho_a c_a^2, \quad c_a = \left[\frac{2(\gamma_a - 1)}{\gamma_a + 1}\frac{I}{\rho_a}\right]^{1/3}. \tag{16}$$

Let us determine the conditions for the excitation of a particular type of initial hydrodynamic wave. When relation (15) between the pressure P_c in the nonevaporated part of the material and the ablation pressure P_a is substituted into (11), the shock excitation condition, $u_c > c_0$, can be written as

$$\frac{\rho_a c_a^2}{\rho_0 c_0^2} > \frac{\gamma_a}{\gamma_a + 1}. \tag{17}$$

Substituting expression (16) for the speed c_a into (17), we obtain the threshold intensity of the energy source that corresponds to the excitation of a shock wave:

$$I > \frac{1}{2}\left(\frac{\gamma_a}{\gamma_a + 1}\right)^{3/2}\frac{\gamma_a + 1}{\gamma_a - 1}\left(\frac{\rho_a}{\rho_0}\right)^{-1/2}\rho_0 c_0^3. \tag{18}$$

The larger the ratio of the density of the evaporated material to the initial density of the layer material, the lower the threshold intensity. The opposite signs in expressions (17) and (18) correspond to the excitation of an acoustic wave. The ratio

$$\beta = \frac{\rho_a c_a^2}{\rho_0 c_0^2}$$

will be called the adiabaticity parameter of the ablation or simply the adiabaticity parameter.

In the case of an initial shock wave, the evaporation wave velocity, the shock velocity, and the velocity of the material behind the shock front can be determined from relations (8), (15), and (16):

$$\begin{aligned} D_{ev} &= \frac{\gamma_c - 1}{\gamma_c + 1}\frac{\rho_a}{\rho_0} c_a, \\ D_w \equiv \frac{\gamma_c + 1}{2} u_c &= \left[\frac{(\gamma_c + 1)(\gamma_a + 1)}{2\gamma_a}\right]^{1/2}\left(\frac{\rho_a}{\rho_0}\right)^{1/2} c_a. \end{aligned} \tag{19}$$

In the case of an initial acoustic wave, its velocity, i.e., the speed of sound in the unperturbed material, c_0, is a fixed parameter of the problem, while the evaporation wave velocity and the velocity of the material behind the acoustic wave front can be determined from relations (11), (15), and (16):

$$D_{ev} = \frac{\rho_a}{\rho_0} c_a, \quad u_c = \frac{\gamma_a + 1}{\gamma_a}\beta c_0. \tag{20}$$

The problem contains three time variables: the duration of the energy source, t_p; the travel time of the hydrodynamic wave through the layer, t_b ($t_b = \Delta_0/c_0$ for an acoustic wave and $t_b = \Delta_0/D_w$ for a shock wave); and the total layer evaporation time, t_{ev} ($t_{ev} = \Delta_0/D_{ev}$ for an acoustic wave and

$$t_{ev} = \frac{\gamma_c - 1}{\gamma_c + 1}\frac{\Delta_0}{D_{ev}}$$

for a shock wave). The times t_b and t_{ev} can be deter-

mined from relations (19),

$$t_b = \left[\frac{2\gamma_a}{(\gamma_a+1)(\gamma_c+1)}\right]^{1/2}\left(\frac{\rho_a}{\rho_0}\right)^{-1/2}\frac{\Delta_0}{c_a}, \qquad t_{\mathrm{ev}} = \left(\frac{\rho_a}{\rho_0}\right)^{-1}\frac{\Delta_0}{c_a} \tag{21}$$

for a shock wave, and from relations (20),

$$t_b = \frac{\Delta_0}{c_0}, \quad t_{\mathrm{ev}} = \left(\frac{\rho_a}{\rho_0}\right)^{-1}\frac{\Delta_0}{c_a} \tag{22}$$

for an acoustic wave.

The solution for the hydrodynamic efficiency in the same general form for both types of initial waves depends on two dimensionless parameters, the ratios of the time parameters of the problem: $\tau_b = t_p/t_b$ and $\tau_{\mathrm{ev}} = t_p/t_{\mathrm{ev}}$. In various τ_b ranges, the solution has the following form.

(1) $0 \le \tau_b \le 1$, the duration of the energy source is shorter than the travel time of the hydrodynamic wave through the layer. Using relations (9), (10), (15), (16), and (19) for the shock problem or (12), (15), (16), and (20) for the acoustic problem, we obtain

$$\varepsilon_k = \varepsilon_T = \frac{1}{2}\left(\frac{\gamma_a+1}{\gamma_a}\right)^2\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2 c_a^2, \qquad M_c = \left(1-\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)\tau_b\rho_0\Delta_0. \tag{23}$$

Hence, the total hydrodynamic efficiency,

$$\eta = \frac{(\varepsilon_k+\varepsilon_T)M_c}{It_p},$$

and the hydrodynamic efficiencies in kinetic energy,

$$\eta_k = \frac{\varepsilon_k M_c}{It_p},$$

and thermal energy,

$$\eta_T = \frac{\varepsilon_T M_c}{It_p}$$

($\eta = \eta_k + \eta_T$), are

$$\eta = 2\eta_k = 2\eta_T = 2\frac{\gamma_a^2-1}{\gamma_a^2}\frac{\tau_{\mathrm{ev}}}{\tau_b}\left(1-\frac{\tau_{\mathrm{ev}}}{\tau_b}\right). \tag{24}$$

(2) $\tau_b \ge 1$, the duration of the source is longer than the travel time of the hydrodynamic wave through the layer. The solution of Eqs. (13) with the initial conditions (14) yields

$$\varepsilon_k = \frac{1}{2}\left(\frac{\gamma_a+1}{\gamma_a}\right)^2\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2 c_a^2\left(1+\frac{\tau_b}{\tau_{\mathrm{ev}}}\ln\frac{1-\tau_{\mathrm{ev}}/\tau_b}{1-\tau_{\mathrm{ev}}}\right)^2, \qquad M_c = (1-\tau_{\mathrm{ev}})\rho_0\Delta_0, \quad \varepsilon_T = \frac{1}{2}\left(\frac{\gamma_a+1}{\gamma_a}\right)^2\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2 c_a^2. \tag{25}$$

Next, using expression (25) and taking into account the thermal energy that was transferred to the layer by the wave (see (10) and (12)), we obtain for the hydrodynamic efficiency

$$\eta = \frac{\gamma_a^2-1}{\gamma_a^2}\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2\frac{1-\tau_{\mathrm{ev}}}{\tau_{\mathrm{ev}}} \times\left[1+\left(1+\frac{\tau_b}{\tau_{\mathrm{ev}}}\ln\frac{1-\tau_{\mathrm{ev}}/\tau_b}{1-\tau_{\mathrm{ev}}}\right)^2\right],$$
$$\eta_k = \frac{\gamma_a^2-1}{\gamma_a^2}\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2\frac{1-\tau_{\mathrm{ev}}}{\tau_{\mathrm{ev}}}\left(1+\frac{\tau_b}{\tau_{\mathrm{ev}}}\ln\frac{1-\tau_{\mathrm{ev}}/\tau_b}{1-\tau_{\mathrm{ev}}}\right)^2, \qquad \eta_T = \frac{\gamma_a^2-1}{\gamma_a^2}\left(\frac{\tau_{\mathrm{ev}}}{\tau_b}\right)^2\frac{1-\tau_{\mathrm{ev}}}{\tau_{\mathrm{ev}}}. \tag{26}$$

The ratio of τ_{ev} and τ_b is the ratio of the velocities of the evaporation and hydrodynamic waves: $\tau_{\mathrm{ev}}/\tau_b = D_{\mathrm{ev}}/D_h$. For shock and acoustic waves, this ratio is, respectively,

$$\frac{\tau_{\mathrm{ev}}}{\tau_b} \equiv \frac{D_{\mathrm{ev}}}{D_w} = \left[\frac{2\gamma_a}{(\gamma_a+1)(\gamma_c+1)}\right]^{1/2}\left(\frac{\rho_a}{\rho_0}\right)^{1/2}, \qquad \frac{\tau_{\mathrm{ev}}}{\tau_b} \equiv \frac{D_{\mathrm{ev}}}{c_0} = \beta^{1/2}\left(\frac{\rho_a}{\rho_0}\right)^{1/2}. \tag{27}$$

To within a factor that depends on the adiabatic constant, the ratio $\tau_{\mathrm{ev}}/\tau_b$ is a function of only the density ratio of the evaporated part of the target and the unperturbed material, ρ_a/ρ_0, for a shock wave and of the density ratio ρ_a/ρ_0 and the adiabaticity parameter β for an acoustic wave.

The solution obtained shows the following main peculiarities of the hydrodynamic efficiency for pulsed ablation of an arbitrarily thick layer of material. If the duration of the energy source is shorter than the travel time of the initial hydrodynamic wave through the entire thickness of the layer, then the hydrodynamic efficiency does not depend on the source duration and is the sum of the energy transfers in equal parts to the thermal and hydrodynamic components. Energy is transferred to the layer only from the initial hydrodynamic wave and corresponds to the approximate solution obtained for a semiinfinite layer in [2]. The main difference between the exact solution (24), (26) and the solution from [2] is that the former takes into account the

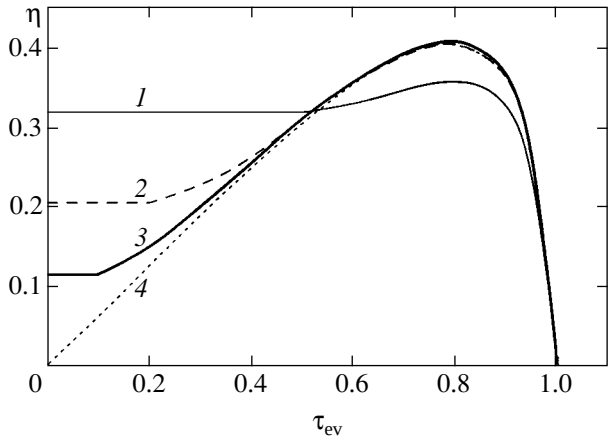

Fig. 1. Hydrodynamic efficiency versus τ_{ev}, the ratio of the laser pulse duration to the total layer evaporation time; $\tau_{ev}/\tau_b = 0.5$ (*1*), 0.2 (*2*), 0.1 (*3*), and 0 (*4*).

decrease in transferred energy due to the mass evaporation of the part of the layer through which the hydrodynamic wave propagates (the last factor in formula (24)). The hydrodynamic efficiency at $\tau_b \le 1$ depends only on the ratio of the travel time of the initial wave through the layer and the total layer evaporation time, τ_{ev}/τ_b. In this case, the hydrodynamic efficiency has a maximum at $\tau_{ev}/\tau_b = 1/2$. This value corresponds to the following density ratios ρ_a/ρ_0 for shock and acoustic waves, respectively:

$$\frac{\rho_a}{\rho_0} = \frac{1}{4}\frac{(\gamma_a+1)(\gamma_c+1)}{2\gamma_a}, \quad \frac{\rho_a}{\rho_0} = \frac{1}{4\beta} \equiv \frac{1}{4}\frac{\rho_0 c_0^2}{\rho_a c_a^2}.$$

The presence of a maximum stems from the fact that, on the one hand, the specific transferred energy increases with parameter τ_{ev}/τ_b (or density ratio ρ_a/ρ_0), and, on the other hand, the rate of decrease in the layer mass to which energy is transferred via the increase in evaporation wave velocity increases. However, the density ratios ρ_a/ρ_0 to which the maximum of the hydrodynamic efficiency correspond are close to unity. This ratio at $\gamma_a = \gamma_c = 5/3$ is 8/15 for the shock problem and within the range 0.25–1 for the acoustic problem. Ablation at such high densities in the evaporated part of the target can proceed only via the action of a high-energy source of radiation or particles. Under typical ablation conditions, for example, under laser radiation with a wavelength $\lambda \le 1$ μm, when the ratio $\rho_a/\rho_0 \ll 1$, the energy transfer efficiency increases with density ratio as $\eta \propto (\rho_a/\rho_0)^{1/2}$.

If the duration of the energy source is longer than the time it takes for the initial wave to emerge at the back surface of the layer, $\tau_b \ge 1$ (a thin layer), then the hydrodynamic efficiency ceases to be independent of the pulse duration. As long as the source duration is short compared to the total layer evaporation time, the hydrodynamic efficiency increases with pulse duration. At small ratios of the source duration to the total evaporation time, the increase in the velocity of the layer as a whole as its mass decreases via evaporation mainly contributes to the hydrodynamic efficiency. If the duration of the energy source is comparable to the total layer evaporation time, then the dependence on pulse duration can reach a maximum and then decreases to zero at the time of total evaporation of the layer material ($\tau_{ev} = 1$). The presence of a maximum is the well-known effect of material evaporation from the layer that leads to competition between two factors: the decrease in the mass of the layer and the increase in its velocity due to the same decrease in its mass. This effect is responsible for the maximum in the dependence of the hydrodynamic transfer coefficient on the fraction of the evaporated mass derived in (1) for a thin layer of material. The solution for an infinitely thin layer [1] can be obtained from the general solution (24), (26) by passing to the limit $\tau_b \longrightarrow \infty$:

$$\eta_k = \frac{\gamma_a^2 - 1}{\gamma_a^2}\left(\frac{1-\tau_{ev}}{\tau_{ev}}\right)\ln^2\left(\frac{1}{1-\tau_{ev}}\right). \qquad (28)$$

In Fig. 1, hydrodynamic efficiency is plotted against τ_{ev} at various ratios τ_{ev}/τ_b. The parameter τ_{ev} ranges from 0 to 1. The upper boundary of this range corresponds to the total evaporation of the entire layer mass. The curve for $\tau_{ev}/\tau_b = 0$ corresponds to an infinitely thin layer. As the ratio τ_{ev}/τ_b increases, i.e., as the ratio of the velocity of the initial hydrodynamic wave to the evaporation wave velocity decreases, for example, due to an increase in the ratio of the density in the region of the evaporated part of the target to the density of the unperturbed material (see (27)), the degree of influence of the propagation of the initial hydrodynamic wave on the hydrodynamic efficiency increases. The τ_{ev} range that corresponds to the energy transfer to the layer only from the initial wave and for which the hydrodynamic efficiency does not depend on τ_{ev}, i.e., the duration of the energy source, is $0 \le \tau_{ev} \le \tau_{ev}/\tau_b$. Therefore, the relative extent of the τ_{ev} range that corresponds to the energy transfer for a thick layer increases with ratio τ_{ev}/τ_b. In addition, the relative contribution of the energy transfer from the initial hydrodynamic wave to the hydrodynamic efficiency increases with τ_{ev}/τ_b. For the examples shown in the figure, the τ_{ev} range that corresponds to the energy transfer for a thick layer is from 0 for $\tau_{ev}/\tau_b = 0$ to 0.5 for $\tau_{ev}/\tau_b = 0.5$. At $\tau_{ev}/\tau_b = 0.2$, the hydrodynamic efficiency of the energy transfer from the initial wave is 0.2. This value accounts for 70% of the total hydrodynamic efficiency for $\tau_{ev} = 0.5$ and 50% for $\tau_{ev} = 0.8$, which corresponds to the maximum hydrodynamic efficiency (≈0.41). At $\tau_{ev}/\tau_b = 0.5$, the hydrodynamic efficiency of the energy transfer from the initial wave is 0.32. This value accounts for 75% of the total hydrodynamic efficiency for $\tau_{ev} = 0.8$.

4. THE HYDRODYNAMIC EFFICIENCY FOR LASER ABLATION OF A FLAT LAYER

When a source of monochromatic radiation with wavelength λ and intensity I in the range $10^{10} < I\lambda^2 < 10^{14}$ W cm^{-2}, which corresponds to ablation for the inverse bremsstrahlung absorption mechanism in the plasma of the evaporated material, acts on a material, the characteristic density of the absorption region is the critical plasma density

$$\rho_{cr} = 1.83 \times 10^{-3} \frac{A}{Z\lambda^2} \frac{\text{g}}{\text{cm}^3}. \quad (29)$$

Here, A is the atomic weight of the ions, Z is the degree of plasma ionization, and λ is measured in μm.

Neodymium glass, iodine, excimer, and CO_2 lasers are the most intense pulsed lasers that can produce a powerful action on materials. The wavelengths of Nd and CO_2 lasers at the fundamental frequency and the first two harmonics are, respectively, $0.33 < \lambda \leq 1.06$ μm and $0.2 < \lambda \leq 0.4$ μm. The wavelengths of excimer and iodine are 1.315 and 10.6 μm, respectively. The critical density of completely ionized plasma when acted upon by relatively short-wavelength Nd, iodine, and excimer lasers lies within the range $2 \times 10^{-3} < \rho_{cr} < 9 \times 10^{-2}$ g cm^{-3}. Thus, it is much lower than the Nd-laser density. The critical density for the radiation of long-wavelength CO_2 lasers is even lower, $\rho_{cr} \approx 3.2 \times 10^{-5}$ g cm^{-3}.

In this section, we discuss our solutions using the action of a short-wavelength laser pulse on a layer of a solid material of light elements as an example. We chose this problem for the following reasons. The hydrodynamic efficiency increases with increasing density of the evaporated material, i.e., with decreasing wavelength of the acting radiation. We chose a light material for the layer, because we can disregard the ionization state of the plasma in the evaporated part of the target by assuming it to be completely ionized and the radiative energy losses. Let us analyze the solutions for the hydrodynamic efficiency under the action of a laser pulse using the excitation of a shock wave as an example.

Using (29), we can write the adiabaticity parameter under the action of a laser pulse as

$$\beta \approx 8.9 \times 10^3 \left(\frac{\gamma_a - 1}{\gamma_a + 1}\right)^{2/3} \left(\frac{A}{Z}\right)^{1/3} \frac{I^{2/3}}{\lambda^{2/3}\rho_0 c_0^2}.$$

Here, I is measured in W cm^{-2}, ρ_0 is in g cm^{-3}, c_0 is in cm s^{-1}, and λ is in μm. Using criterion (18) and assuming that $\gamma_a = 5/3$ and $A/Z = 2$, we find the a shock wave is excited at a laser pulse intensity above

$$I \approx 3.4 \times 10^{-6} \lambda (\rho_0 c_0^2)^{3/2} \text{ W cm}^{-2}. \quad (30)$$

According to (30), the threshold intensity for the fundamental harmonic of a Nd laser is 4.7×10^{10} W cm^{-2} when it acts on polystyrene ($\rho_0 = 1$ g cm^{-3}, longitudinal speed of sound $c_0 = 2.35 \times 10^5$ cm s^{-1}), 3.9×10^{12} W cm^{-2} when it acts on aluminum ($\rho_0 = 2.7$ g cm^{-3}, $c_0 = 6.26 \times 10^5$ cm s^{-1}), and 1.7×10^{13} W cm^{-2} when it acts on beryllium ($\rho_0 = 1.85$ g cm^{-3}, $c_0 = 12.55 \times 10^5$ cm s^{-1}). The threshold intensity linearly decreases with decreasing radiation wavelength, because the ablation pressure (see (16)) increases with decreasing wavelength (increasing critical density).

Let us write solution (24), (26) for the excitation of a shock wave in the form of explicit dependences on the parameters of the laser ablation problem: radiation wavelength λ, pulse duration t_p, intensity I, and target density ρ_0. Substituting expressions (21) for τ_{ev} and τ_b into the solution using expression (16) for the speed of sound in the evaporated material and expression (29) for the critical density and assuming, as in the previous calculations of this section, that $\gamma_a = 5/3$ and $A/Z = 2$ and, in addition, $\gamma_c = 2$, we obtain the following results.

The condition for a thick layer, i.e., the condition that the layer is so thick that the shock wave does not traverse the entire layer in the laser pulse action time, is

$$\Delta_0 \geq 10.3 \frac{I^{1/3} t_p}{\lambda^{1/3} \rho_0^{1/2}}. \quad (31)$$

The limiting layer thickness at which the layer is thick and energy is transferred to the layer from the shock wave increases with decreasing laser wavelength and material density as well as with increasing laser pulse intensity and, naturally, duration. When the condition (31) is satisfied, the hydrodynamic efficiency depends neither on laser pulse duration nor on intensity, being

$$\eta = 5 \times 10^{-2} \frac{1}{\lambda \rho_0^{1/2}} \left(1 - 4 \times 10^{-2} \frac{1}{\lambda \rho_0^{1/2}}\right). \quad (32)$$

The hydrodynamic efficiency of the energy transfer to a thick layer increases with decreasing laser wavelength and density of the layer material.

The second range of parameters for the problem corresponds, on the one hand, to the condition that the pulse duration exceeds the travel time of the shock wave through the entire thickness of the layer and, on the other hand, to the condition that no evaporation of the entire layer material takes place in the laser pulse time:

$$0.41 \frac{I^{1/3} t_p}{\lambda^{4/3} \rho_0} \leq \Delta_0 \leq 10.3 \frac{I^{1/3} t_p}{\lambda^{1/3} \rho_0^{1/2}}. \quad (33)$$

In this range of parameters, the hydrodynamic

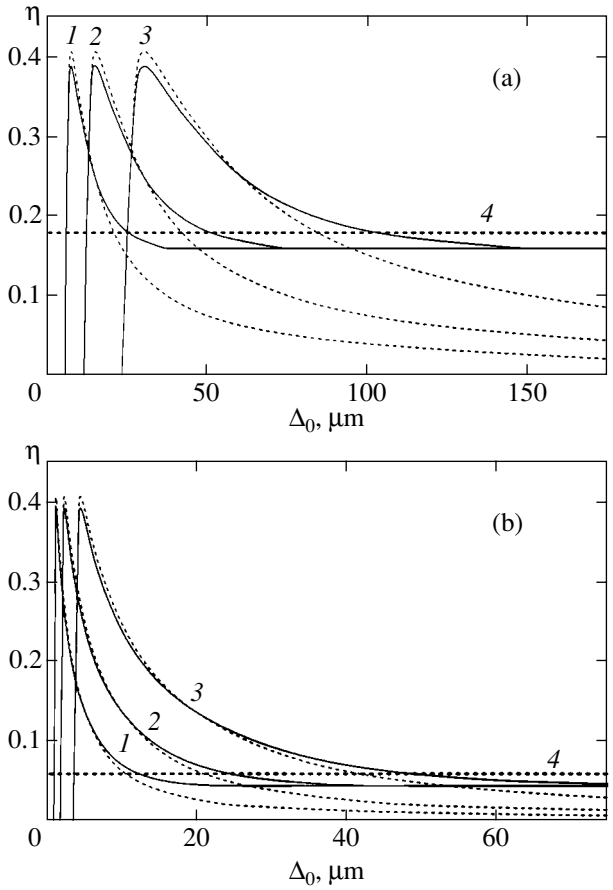

Fig. 2. Hydrodynamic efficiency versus polystyrene layer thickness for laser wavelengths $\lambda = 0.25$ μm (a) and 1.06 μm (b). Curves *1*, *2*, and *3* correspond to laser pulse durations of 0.5, 1, and 2 ns; the solid and dotted lines correspond to our solution and the solution for a thin layer [1], respectively. The dotted lines *4* represent the hydrodynamic efficiencies from the solution for a semiinfinite layer [2].

efficiency is

$$\eta = \frac{2.4\times10^{-3}\Delta_0}{\lambda^{2/3}I^{1/3}t_p}\left(1-\frac{4.1\times10^{-1}t_pI^{1/3}}{\Delta_0\lambda^{4/3}\rho_0}\right)$$

$$\times\left\{1+\left[1+25\lambda\rho_0^{1/2}\ln\left(\left(1-\frac{4.2\times10^{-2}}{\lambda\rho_0^{1/2}}\right)\right.\right.\right. \qquad (34)$$

$$\left.\left.\left.\times\left(1-\frac{4.1\times10^{-1}t_pI^{1/3}}{\Delta_0\lambda^{4/3}\rho_0}\right)^{-1}\right)\right]^2\right\}.$$

The quantities in expressions (30)–(34) are measured in the following units: I in 10^{12} W cm^{-2}, Δ_0 in μm, ρ_0 in g cm^{-3}, λ in μm, and t_p in ns.

In Fig. 2, the hydrodynamic efficiency of the energy transfer to a flat polystyrene layer ($\rho_0 = 1$ g cm^{-3}) is plotted against its thickness under laser radiation with wavelengths $\lambda = 0.25$ and 1.06 μm for laser pulse durations $t_p = 2$, 1, and 0.5 ns. The radiation intensity was chosen to be $I = 10^{14}$ W cm^{-2}. The figure also shows the solutions from [1] for a thin layer and from [2] for a semiinfinite layer. The presented data first illustrate the increase in the hydrodynamic efficiency of the ablation process with decreasing laser wavelength. Thus, for example, the hydrodynamic efficiency for thick layers under radiation with $\lambda = 0.25$ μm ($\eta = 0.16$) is higher by a factor of approximately 4 than that under radiation with $\lambda = 1.06$ μm ($\eta = 0.045$). Let us discuss the peculiarities of the energy transfer to a flat layer of material by using the data presented in Fig. 2a for $\lambda = 0.25$ μm. For a laser pulse of duration $t_p = 2$ ns, energy transfer to the nonevaporated part of the layer takes place only for a layer whose initial thickness exceeds $\Delta_0 = 23.6$ μm. Thinner layers are completely evaporated by the end of the pulse. The general solution is close to the solution from [1] for layers with thickness in the range $23.6 \le \Delta_0 \le 60$ μm. The maximum hydrodynamic efficiency, $\eta_{max} = 0.39$, corresponds to the initial layer thickness $\Delta_0 = 32$ μm and is reached when 76% of the layer mass has evaporated. According to [1], $\eta_{max} = 0.41$ corresponds to $\Delta_0 = 320$ μm and is reach when 81% of the layer mass has evaporated. Beginning with $\Delta_0 = 60$ μm, the solution from [1] for a thin layer shows a faster decrease in hydrodynamic efficiency than the general solution does. For a layer of thickness $\Delta_0 = 149.5$ μm equal to the distance to which the shock wave travels during the laser pulse, the general solution yields $\eta = 0.16$, a value that significantly exceeds the result of the solution from [1] for a thin layer, $\eta = 0.1$. For thick layers of $\Delta_0 > 149.5$ μm, the general solution yields a constant value of $\eta = 0.16$ close the ablation loading efficiency obtained in [2], while the solution for a thin layer from [1] yields a physically incorrect result—a decrease in hydrodynamic transfer efficiency and its approach to zero when $\Delta_0 \longrightarrow \infty$. For a laser pulse $t_p = 0.5$ ns in duration, the maximum thickness of the completely evaporating layer is 5.9 μm. In this case, the general solution differs from the solution for a thin layer [1] even beginning with layer thicknesses (20 μm) much smaller than those for $t_p = 2$ ns. A thick layer to which energy is transferred only by the shock wave corresponds to $\Delta_0 \ge 37.4$ μm; the minimum thickness is also much smaller than that for $t_p = 2$ ns.

As the laser wavelength increases, other things being equal, the minimum thickness of the layer that is not evaporated in the pulse time and the minimum thickness of the layer that may be considered thick decrease. At $t_p = 2$ ns for $\lambda = 1.06$ μm, these thicknesses are, respectively, 3.4 and 92.8 μm, while for $\lambda = 0.25$ μm, they are 23.6 and 149.6 μm.

5. CONCLUSIONS

The general solution for the hydrodynamic energy transfer efficiency under the ablation action of a pulsed energy source on a flat layer of material allows us to determine the energy transferred to the nonevaporated part of the target and the distribution of this energy between the kinetic and thermal components at arbitrary thicknesses of the layer of material and duration of the energy source. The most relevant area of applications of the solution is the investigation of the energy action of radiation pulses on solid materials in problems of acceleration and heating of inertial thermonuclear fusion targets and technological material processing.

ACKNOWLEDGMENTS

I am grateful to V.B. Rozanov, A. Caruso, and C. Strangio for helpful discussions of the results.

REFERENCES

1. Yu. V. Afanas'ev, E. G. Gamaliĭ, O. N. Krokhin, and V. B. Rozanov, Prikl. Mat. Mekh. **39**, 451 (1975).
2. K. S. Gus'kov and S. Yu. Gus'kov, Kvantovaya Élektron. (Moscow) **31**, 305 (2001).
3. N. G. Basov, Yu. A. Zakharenkov, A. A. Rupasov, *et al.*, in *Diagnostics of Dense Plasma*, Ed. by N. G. Basov (Nauka, Moscow, 1989).
4. N. G. Basov, P. P. Volosevich, E. G. Gamaliĭ, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **28**, 135 (1978) [JETP Lett. **28**, 125 (1978)].
5. Yu. V. Afanas'ev and S. Yu. Gus'kov, in *Nuclear Fusion by Inertial Confinement*, Ed. by G. Velarde *et al.* (CRC Press, Ann Arbor, 1993), p. 99.
6. N. G. Basov, P. P. Volosevich, E. G. Gamaliĭ, *et al.*, Zh. Éksp. Teor. Fiz. **78**, 420 (1980) [Sov. Phys. JETP **51**, 212 (1980)].
7. S. I. Anisimov, Yu. A. Imas, G. S. Romanov, and Yu. A. Khodyko, *Powerful Radiation Impact on Metals* (Fizmatgiz, Moscow, 1970).
8. Ya. B. Zel'dovich and Yu. P. Raĭzer, *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena* (Fizmatgiz, Moscow, 1963; Academic, New York, 1966 and 1967), Vols. 1 and 2.

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1146–1153.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1281–1290.

PLASMA, GASES

Electron Acceleration during the Breaking of an Intense Plasma Wave in an Inhomogeneous Plasma

V. I. Arkhipenko[a], V. N. Budnikov[†b], E. Z. Gusakov[b], A. K. Kapanik[a], V. A. Pisarev[a], and L. V. Simonchik[a]

[a] *Institute of Molecular and Atomic Physics, National Academy of Sciences of Belarus, Minsk, 220072 Belarus*
[b] *Ioffe Physicotechnical Institute, Russian Academy of Sciences, St. Petersburg, 194021 Russia*
e-mail: simon@imaph.bas-net.by
Received April 21, 2003

Abstract—We investigate the dynamics of the electron acceleration when an intense plasma wave breaks near resonance at the plasma frequency (focus) in an inhomogeneous magnetized plasma. The breaking threshold has been determined. We compare our experimental dependences of the current and energy of fast electrons on the intensity of the incident wave at various times with theoretical estimates. We show that when the breaking threshold is significantly exceeded, up to 50% of the electrons at plasma resonance are captured and accelerated by the wave. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The development of new highly efficient plasma methods for charged particle acceleration has attracted the attention of researchers since the late 1950s [1–7]. One of the schemes for the formation of a strong accelerating electric field involves the charge separation in plasma by the ponderomotive force exerted on electrons from an intense high-frequency electromagnetic wave. In this case, both the resonant, due to the decay instability of Raman scattering, and nonresonant, in the wake, excitation of plasma waves takes place. The growth of these waves is limited at a very high level by the breaking accompanied in modern experiments by electron acceleration to energies of 100 MeV [8].

In this paper, we present the results of our model experiment, in which we investigate in detail the electron acceleration that accompanies the breaking of an intense plasma wave in an inhomogeneous plasma. The generation of fast electrons is traced near the breaking threshold and when the latter is exceeded by four orders of magnitude in intensity. We compare the dependence of the current of accelerated electrons and their limiting energy on pumping power with predictions of the plasma-wave-breaking model.

This paper is structured as follows. After the description of our experiment and available diagnostic tools, we give theoretical estimates for the breaking threshold and discuss the expected dependences of the current of accelerated electrons and their limiting energy on pumping power. Subsequently, we present and discuss our experimental results and compare them with theoretical scalings. In the Conclusions, we discuss prospects for further experiments.

† Deceased.

2. DESCRIPTION OF THE EXPERIMENT

We carried out the experiments at the Granit linear plasma facility [9]. Plasma was produced via electron cyclotron breakdown in a quartz cylinder *1* (Fig. 1a) with an inner diameter of $2r_0 = 1.8$ cm and a length of about 1 m filled with argon at a pressure of about 2 Pa and placed in a magnetic field with a strength of ~3 kG. A monotonically falling (along the magnetic field) distribution of plasma electron density n_e was established in the middle part of the cylinder. This distribution can be fitted by

$$n_e(r, z) \sim \exp\left(-\frac{z}{l}\right)\left(1 - \frac{r^2}{r_0^2}\right)^{\beta}, \qquad (1)$$

where $l = 5$ cm is the plasma inhomogeneity scale length along the magnetic field and $\beta = 4$.

A microwave was applied to the plasma from one side via a 7.2×3.4 cm^2 waveguide *2* (Fig. 1a), with the electric field of the wave being parallel to the external magnetic field. The typical plasma parameters at the entrance were $n_e < 10^{12}$ cm^{-3} and $T_e = 2$ eV. When the density on the cylinder axis appreciably exceeded the critical density for frequency $f_0 = \omega_0/2\pi$, an oblique Langmuir wave was excited in the plasma predominantly in the form of a radial Trievelpiece–Gould mode. The dispersion relation for this mode in an inhomogeneous plasma is

$$k_\perp^2 = \left(\frac{\omega_{pe}^2(r, z)}{\omega_0^2} - 1\right)k_\parallel^2,$$

where $k_\parallel$ and $k_\perp$ are the parallel and perpendicular (relative to the magnetic field) wave vector components.

The transparency region for this wave is a dense plasma with a density above its critical value, $n_e > n_c$, where $n_c = \pi m f_0^2/e^2$. The near-axis plasma region (Fig. 1a) is a plasma waveguide for it with weak axial inhomogeneity; propagating through this waveguide toward lower densities, the wave decelerates. At the point at which the external magnetic field lines are perpendicular to the critical density surface, $n_e = n_c$ (focal point), the wave linearly transforms into a "warm" plasma wave. In this case, its field reaches the largest strengths given by the relation

$$E_0 = \left(\frac{2P_0'}{\omega_0}\right)^{1/2} \frac{k_0^{3/2}}{(3r_D^2 b k_0^3)^{1/2}} \times \exp\left[i\int_{-\infty}^{z}(k_0 + ik_0'')dz' - \frac{k_0}{2b}r^2 - i\omega_0 t\right] + \text{c.c.}, \quad (2)$$

where r_D is the Debye radius and $P_0' = \kappa P_0$ is the fraction of the power P_0 applied to the plasma that goes into the excitation of the fundamental radial Trievelpiece–Gould mode (according to [9], $\kappa \approx 0.2$). The real, $k_0 = k_0(z)$, and imaginary, $k_0'' = k_0''(z)$, parts of the component of the wave vector **k** along the external magnetic field can be determined near the focal point from the equation

$$3r_D^2(k_0 + ik_0'')^2 - \frac{z}{a} - \frac{2}{(k_0 + ik_0'')b} + i\eta'' = 0, \quad (3)$$

where $a \approx l = 5$ cm and $b \approx r_0\beta^{-0.5} = 0.4$ cm are the experimentally determined parameters of the plasma density distribution near the focal point. Thus, the longitudinal plasma permittivity is

$$\eta = 1 - \frac{\omega_{pe}^2(r,z)}{\omega_0^2}(1 + 3r_D^2k_0^2) + i\eta'',$$

$$\eta'' = \frac{\nu_{ea}}{\omega_0} - \pi\frac{\omega_{pe}^2}{k_0^2}\frac{\partial f_e}{\partial w}\bigg|_{w=\omega_0/k_0}.$$

Here, ν_{ea} is the electron–atom collision frequency, and $f_e(w)$ is the electron distribution in longitudinal velocities w normalized so that

$$\int_{-\infty}^{\infty} f_e(w)dw = 1.$$

The calculated factor that describes the damping of an oblique Langmuir wave is

$$\ln b_l = -\int_{-\infty}^{z} k''dz' = -\frac{\nu_{ea}}{\omega_0}k_0 a - \pi a\omega_0 f_e\left(\frac{\omega_0}{k_0}\right).$$

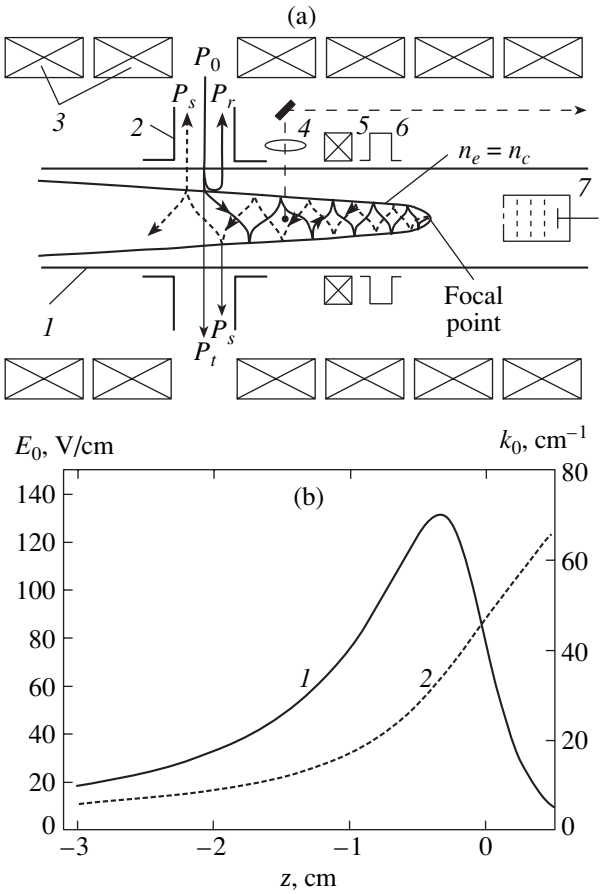

Fig. 1. (a) Experimental setup: P_0, P_r, P_s, P_t—incident, reflected, scattered, transmitted waves; n_c—critical density; *1*—quartz cylinder; *2*—waveguide; *3*—magnet coil; *4*—optical system; *5*—Rogowski coil; *6*—cavity; *7*—analyzer. (b) The distributions of electric field strength (1) and wave vector (2); $z = 0$ is the position of the focus.

The behavior of the electric field of an oblique Langmuir wave and its wave vector near a hybrid resonance calculated for the experimental parameters (T_e = 1.8 eV, $\nu_{ea} = 4.5 \times 10^7$ s^{-1}, $P_0 = 0.02$ W) is shown in Fig. 1b. The increase in electric field strength near the focus is so large that, even at a power ~10 mW applied to the plasma, nonlinear properties of the plasma manifest themselves and the parametric decay instability of stimulated backscattering is excited [10]. In the same range of microwave pumping powers, its Landau damping leads to a significant change of the electron velocity distribution function at suprathermal energies $W >$

10 eV. Thus, it can be fitted by the bi-Maxwellian function

$$n(\omega) = n_c\left[\left(\frac{m_e}{2\pi T_e}\right)^{1/2}\exp\left(-\frac{m_e w^2}{2T_e}\right) + \delta\left(\frac{m_e}{2\pi T_h}\right)^{1/2}\exp\left(-\frac{m_e w^2}{2T_h}\right)\right], \quad (4)$$

where δ and T_h are, respectively, the fraction of the accelerated electrons in the total density and their effective temperature.

One might expect the appearance of nonresonant nonlinear processes at high powers, $P \gg 10$ mW. Thus, because of the action of the Miller ponderomotive force, which excites decay instability at a power of 10 mW, one might expect the expulsion of electrons from the localization region of a strong microwave field when much higher powers are rapidly switched on. This expulsion should cause an increase in the quasi-static potential near the focus,

$$\varphi = \frac{E^2}{4\pi e n_c},$$

and, eventually, an acceleration of ions. At the pumping power $P_0 = 10$ kW attainable in our experiment, according to formula (2), this potential near the maximum of the microwave field, at first glance, can reach megavolts, $\varphi \approx 4 \times 10^5$ V. It should be noted, however, that expression (2) for the electric field distribution at a focus-type hybrid resonance was derived in the linear approximation. In particular, in its derivation, we used expressions from the linear theory for the Landau wave damping constant that includes the quasi-linear rearrangement of the electron distribution function. The damping of the plasma wave increases with its field, becoming nonlinear, and reaches the largest values during the wave breaking, when it captures electrons moving with velocities lower than the thermal velocity. The breaking condition is

$$v_b = \frac{\omega}{k} = \sqrt{\frac{2eE}{m_e k}}.$$

Given (2), the expression for the wave phase velocity at which the breaking occurs for sufficiently high pumping powers takes the form

$$\frac{\omega}{k_{wb}} = \left(\frac{e}{m_e}\right)^{2/5}(8\kappa P_0)^{1/5}. \quad (5)$$

It should be noted that, according to this estimate, the breaking of the plasma wave must be observed near the maximum of its electric field beginning from a pumping power $P_0 \approx 1$ W. As the pumping power increases, the breaking of the wave occurs at a smaller wave deceleration. The maximum energy of the electrons accelerated through the wave breaking can be estimated as

$$W_{wb} \approx 2m_e\left(\frac{\omega}{k_{wb}}\right)^2 = 4(2e^4 m_e \kappa^2 P_0^2)^{1/5}. \quad (6)$$

At a pumping power of 10 kW, this energy reaches $\varphi \approx 5 \times 10^3$ eV. The maximum plasma potential produced by the Miller force is related to this energy by $\varphi \approx W_{wb}/8e$. Because of the loss of accelerated electrons, this potential can numerically increase while retaining the same order of magnitude: $\varphi \approx W_{wb}/e$. The number of wave-accelerated electrons n_h can be estimated from the energy balance conditions as

$$\frac{n_h}{n_c} \approx \frac{1}{bk_{wb}} = \left(\frac{e}{m_e}\right)^{2/5}\frac{(8\kappa P_0)^{1/5}}{\omega_0 b}. \quad (7)$$

The rough estimate for the pulse of the current transferred by these electrons is

$$I \approx en_h\frac{2\omega}{k_{wb}}\frac{\pi b}{k_{wb}} = 2\left(\frac{e}{2m_e}\right)^{1/5}(\kappa P_0)^{3/5}. \quad (8)$$

Apart from the fast electron nonlinearities considered above, slower, in particular, ionization nonlinearities can show up in an experiment. Using (2), we write the expression for the electron vibrational energy at the focus as

$$W = \frac{P_0'}{\pi\omega_0 n_e}\frac{k_0^3\exp(-2ak_0\nu_{ea}/\omega_0)}{3r_D^2 bk_0^3 + 1}.$$

For $k_0 \sim 40$ cm^{-1} (this is a typical value for the wave number at the focus), the relation $W_\sim$[eV] $\approx 3.2P_0'$[W] can be obtained. For $P_0 = 25$ W, the electron vibrational energy is $W_\sim = 16$ eV, a value that is higher than the ionization energy for argon atoms, $E_i = 15.76$ eV. In this case, one might expect very fast ionization in the region of a strong microwave field and, as a result, a displacement of the point of hybrid resonance from the input region. This burning of the plasma waveguide channel must eventually lead to displacement of the point of hybrid resonance to the plasma boundary and suppression of associated nonlinear processes.

In this paper, in which we begin to investigate the propagation of an intense plasma wave through an inhomogeneous magnetized plasma, we present the results concerning the fast electron nonlinearities, in particular, the electron acceleration. We used a set of various diagnostics to study the processes in plasma during the propagation of a pulsed plasma wave. Figure 1a schematically shows the arrangement of elements of the diagnostic equipment. Thus, the plasma

density distribution was controlled by the cavity method; the recording of the spatial plasma emission distribution allowed the microwave power absorption region to be determined; the electron distribution function both in an unperturbed plasma and during the action of a pumping wave was controlled with a multigrid charged particle analyzer; the current of accelerated electrons was recorded with the Rogowski coil; and information about the wave processes in the plasma was extracted directly from the waveguide duct by analyzing the scattered signal.

The experiments were carried out for the following pumping parameters: frequency f_0 = 2840 MHz, pulse power $P_0 \le 5$ kW, pulse duration $t \sim 0.2$–1 μs, pulse front duration $t_f \sim 40$ ns, and pulse repetition frequency 300 Hz.

3. EXPERIMENTAL RESULTS

The shape of the microwave pulse, about 0.4 μs in duration, from the waveguide duct in the absence of plasma is shown in Fig. 2a. In the presence of plasma, the shape of the microwave pulse in the discharge cylinder at a pumping power lower than 50 W changes only slightly. However, at a power higher than 50 W, low-frequency oscillations with a frequency of 20–30 MHz appear at the end of the pulse. As the power increases, the time at which these oscillations appear is shifted to the beginning of the pulse. The multigrid analyzer located behind the focus at a distance of about 25 cm from the side of low densities records the electron current (Fig. 2b). The current increases in less than 0.1 μs and rapidly decreases when the microwave pulse ends.

Figure 2c shows an oscillogram for the signal from the Rogowski coil. The positive peak in this oscillogram corresponds to the initial increase in the electron current recorded by the analyzer (Fig. 2b), while the negative signal corresponds to the slower decrease in the current during the pulse.

Figure 2d shows an oscillogram for the current signal of the photomultiplier. The visible plasma emission is seen to increase in intensity near the focus almost immediately after the application of a microwave pulse. It should be noted that the light intensity slowly decreases during several microseconds after the completion of the microwave pulse.

The current pulse of the charged particle analyzer depends on the power of the applied microwave pulse and the retarding potential. In our experiment, the analyzer was placed at a distance of about 25 cm from the focal point from the side of low densities (Fig. 1a). Figure 3a shows oscillograms for the current pulses of the analyzer at a retarding potential U_a = –50 V for various powers of the applied microwave pulses. We see that the shape of the current pulses changes in a complicated way with microwave pulse power. As the power increases, the current peak is shifted to the beginning of the pulse.

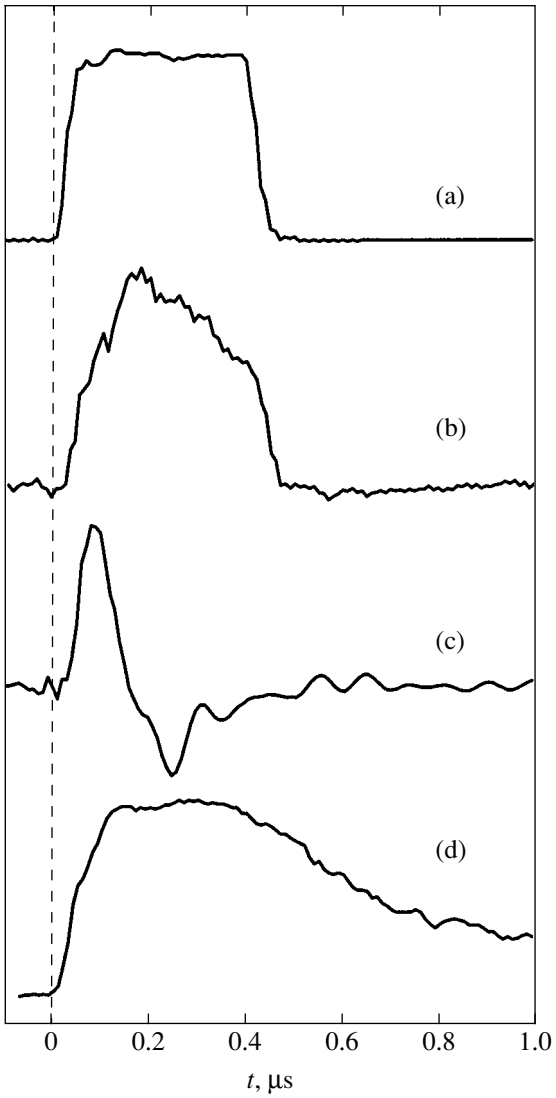

Fig. 2. Oscillograms for the pulses of the incident microwave emission (a), the current of the multigrid analyzer at a retarding potential of –50 V (b), the signal of the Rogowski coil (c), and the light intensity at the focus (d).

The current pulse also changes at various retarding potentials for a fixed power, as shown in Fig. 3b for a pulse of power P = 50 W. The presence of an electron current at retarding potentials of ~1000 V suggests that electrons with energies much higher than the initial plasma electron energy T_e are produced by the interaction of a microwave pulse with the plasma.

The current–voltage characteristics of the charge particle analyzer on a semilogarithmic scale (Fig. 4) have nearly linear segments at energies $W \gg T_e$, at which the electron energy can be characterized by the effective temperature T_h. The effective temperature depends on time and pulse power. Figure 4a shows the current–voltage characteristics at various times from the beginning of the pulse at power P = 50 W. We see that the characteristic approaches a straight line (solid

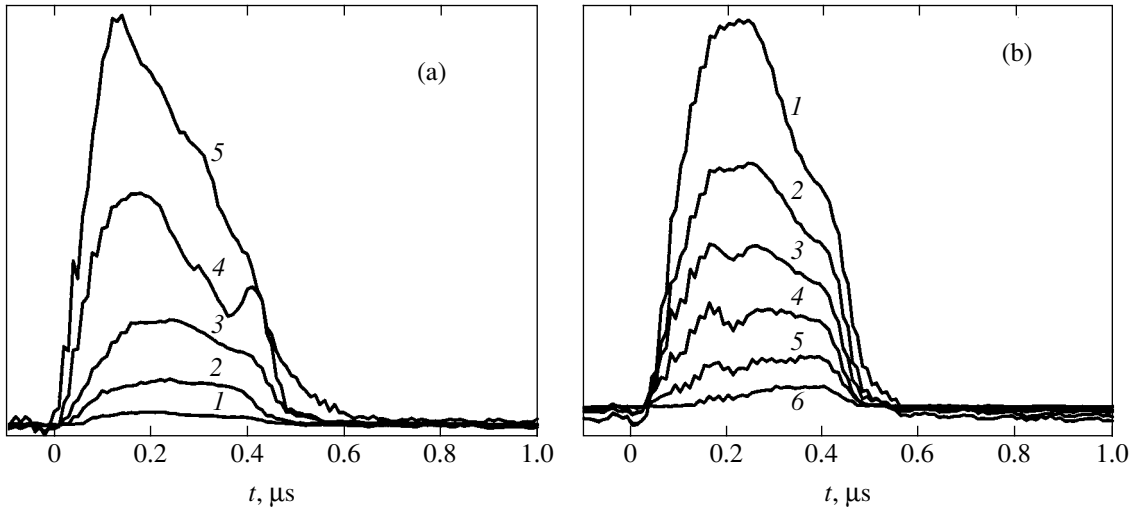

Fig. 3. Oscillograms for the current pulses of the charged particle analyzer (a) at the pumping powers P = 5 (*1*), 16 (*2*), 50 (*3*), 160 (*4*), 500 W (*5*), U_a = –50 V and (b) at the retarding potentials U_a = 0 (*1*), –50 (*2*), –200 (*3*), –500 (*4*), –700 (*5*), and –1200 V (*6*), P = 50 W.

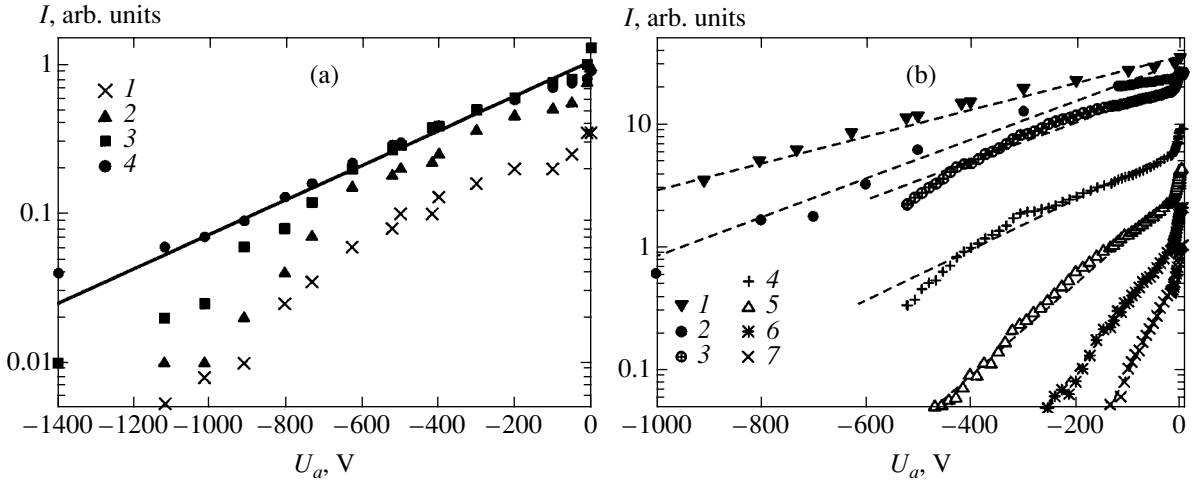

Fig. 4. Current–voltage characteristics of the charge particle analyzer (a) at the times during the pulse 0.05 (*1*), 0.1 (*2*), 0.15 (*3*), and 0.35 μs (*4*), and (b) at the pulse powers P = 50 (1), 25 (*2*), 16 (*3*), 5 (*4*), 2.5 (*5*), 1 (*6*), and 0.5 W (*7*).

line) by the end of the pulse (t > 0.2 μs). The effective temperature T_h corresponding to this straight line is 385 eV. At the same time, in the first half of the pulse, the characteristics deviate from the straight line, because there is a deficit of electrons with energies above 500–600 eV.

Figure 4b shows the changes of the current–voltage characteristics with power. All these characteristics were constructed for the same time from the beginning of the pulse, about 0.2 μs. The dashed lines represent the exponential dependences (4) for the corresponding effective temperatures T_h. We can see that the generation of electrons with energies above 400 eV is limited at powers of 5–25 W.

The current of accelerated electrons at retarding potentials U_a = 0 and –400 V is plotted against microwave pulse power in Fig. 5a. The solid line represents the dependence $P^{0.6}$. A deviation from this dependence is observed at high and low powers. The current at a power P ~ 5 kW decreases, because the ionization processes begin early and the plasma wave propagation conditions change already in 200 ns. The decrease in the current at P < 10 W for both retarding potentials probably stems from the fact that the breaking threshold is approached. This effect is more pronounced at U_a = –400 V. This can be explained by the small gain of energy during the breaking of a low-intensity wave that leads to the cutoff of the current–voltage characteristics (Fig. 4b) at energies above 500–600 eV. No high-energy electron current is recorded at a power on the order of several milliwatts.

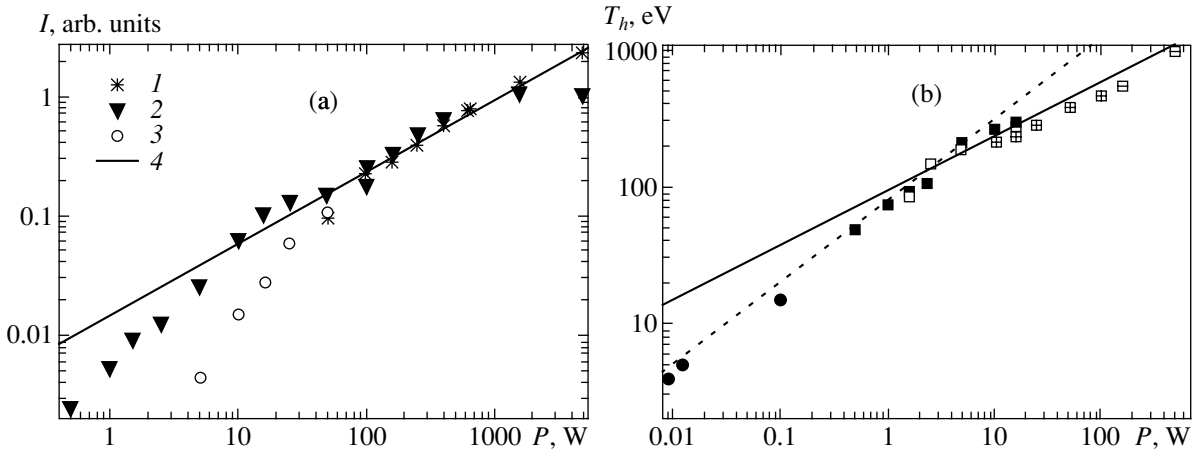

Fig. 5. (a) Current of the accelerated electrons versus microwave pulse power: measurements with the Rogowski coil (*1*), measurements with the particle analyzer at $U_a = 0$ (*2*) and –400 V (*3*), the dependence $I \propto P^{0.6}$ (*4*). (b) Effective temperature of the accelerated electrons versus microwave pulse power: the symbols represent the experimental data; the solid and dotted lines represent $T_h \propto P^{0.4}$ and $T_h \propto P^{0.6}$, respectively.

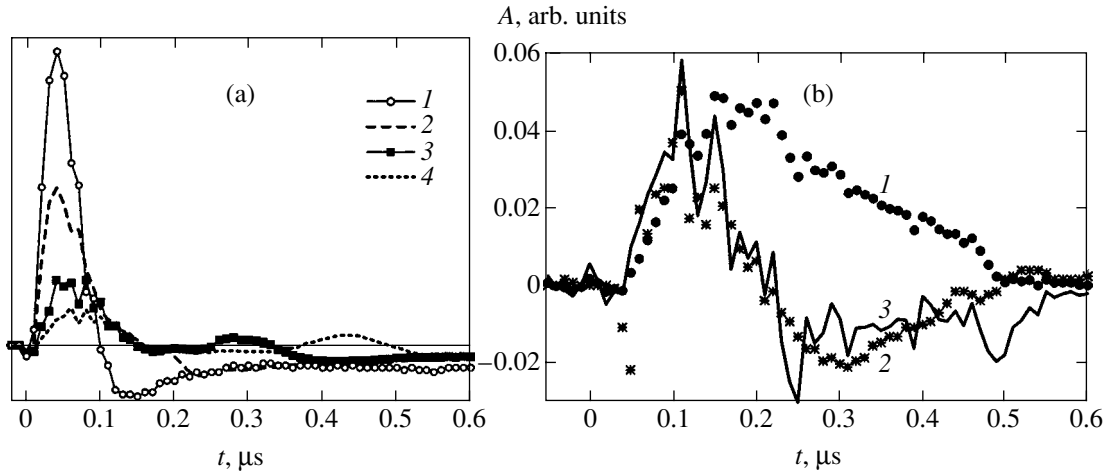

Fig. 6. (a) Oscillograms for the signal from the Rogowski coil at various microwave pulse powers: $P = 5000$ (*1*), 1600 (*2*), 400 (*3*), and 150 (*4*) W. (b) Oscillograms for the signal current from the Rogowski coil (1) and the particle analyzer (2) at $P = 50$ W, *3*—calculated signal.

The temperature of the accelerated electrons T_h is plotted against pumping wave power in Fig. 5b. Up to a power $P \sim 1$–2 W, the temperature of the accelerated electrons T_h increases as $P^{0.6}$ (the dotted line in Fig. 5b). Then, the rate of increase in temperature decreases. In this case, the dependence of the effective temperature on pumping power is close to $P^{0.4}$, which is represented by the solid line in Fig. 5b.

Let us look at the behavior of the signal from the Rogowski coil as a function of the microwave pulse power (Fig. 6a). The Rogowski coil was placed between the charged particle analyzer and the focus (Fig. 1a) at a distance of ~5 cm from the analyzer. The solid line in Fig. 6a parallel to the horizontal axis indicates a zero signal level. Two peaks above this zero level can be seen in the presented oscillograms. The first peak has a half-width of 0.15 μs at $P \sim 50$ W and narrows to approximately 0.8 μs at 5 kW. As can be seen, the front of the pulse from the Rogowski coil at a power of several kilowatt is much sharper than the fronts of the incident microwave pulse (40 ns) and the current pulse from the charged particle analyzer (100 ns). The second peak is smaller, but its position depends on the applied power: it virtually merges with the first peak at $P = 5$ kW and is shifted to the end of the pulse at $P < 160$ W. The first peak of the signal from the Rogowski coil is formed by the electrons accelerated during the wave breaking in the initial plasma density profile, while the second peak takes place after the significant ionization deformation of the profile, which results in the formation of a plasma waveguide channel. The ionization plasma dynamics under the action of a microwave pulse requires a separate analysis.

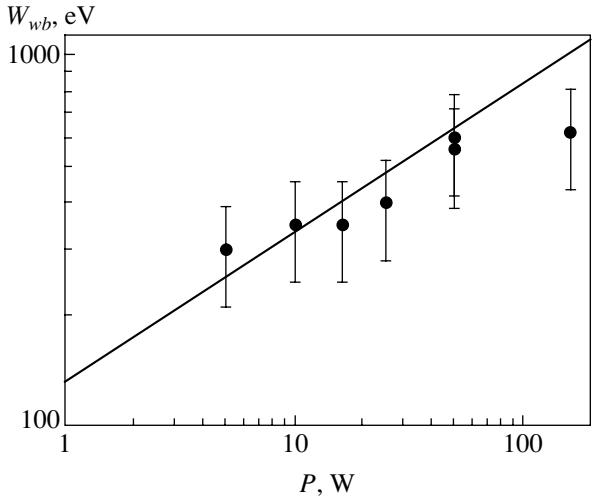

Fig. 7. Limiting energy of the accelerated electrons during wave breaking versus pumping power: the dots and the solid line represent the experimental and calculated data, respectively.

Let us consider how the signal from the Rogowski coil is formed in more detail. Figure 6b shows oscillograms for the current pulses from the charged particle analyzer (2) and the signal from the Rogowski coil (1). These pulses differ significantly in shape. On time scales shorter than 100 ns, the signal from the Rogowski coil reproduces the shape of the leading edge of the current pulse from the fast particle analyzer. In contrast, at longer times, it is most likely proportional to the derivative of the current pulse from the analyzer. This behavior is attributable to the insufficiently large time constant of the loop, $\tau = 40$ ns, as confirmed by the recalculation of the signal from the Rogowski coil using the electrical formula

$$F_{\text{out}}(t) = \int_0^t f'_{\text{in}}(t') \exp\frac{t'-t}{\tau} dt',$$

where $f'_{\text{in}}(t)$ is the derivative of the function of the input current that is encircled by the Rogowski coil, F_{out} is the output signal from the loop, and τ is the time constant of the loop. The results of our calculations with this formula using the current pulse from the analyzer as $f_{\text{in}}(t)$ are presented in Fig. 6b (solid curve). As our calculations indicate, for the Rogowski coil to satisfactorily reproduce the entire current pulse, its time constant must be an order of magnitude larger. Nevertheless, the Rogowski coil used appears to correctly reflect the behavior of the current on time scales shorter than 100 ns. It is particularly useful at high powers (several kilowatts), when the current rise times are so short that the charged particle analyzer does not resolve them because of its slow response.

Figure 5a (asterisks) shows a plot of the maximum amplitude of the first current peak from the loop against microwave pulse power. In the segment where the power changes from 50 W to 5 kW, this dependence closely corresponds to the power dependence of the current from the charged particle analyzer and to the dependence $P^{0.6}$ that follows from the above estimates based on the breaking model.

4. DISCUSSION

As follows from the presented experimental data, accelerated electrons whose effective temperature T_h depends on pulse power are formed near the focus under the action of a microwave pulse. Electrons are accelerated via wave damping, which, as was noted above, increases with the plasma wave field and becomes nonlinear. The damping reaches the largest values during the wave breaking, when the wave captures electrons that move with velocities lower than the thermal velocity. Our estimates using formula (5) show that the wave breaking must occur at a power of ~1 W. If we look at the experimental dependence $T_h(P)$ in Fig. 5b, we will then see that its slope changes sharply precisely at a power of a few watts. This may suggest that the pattern of electron acceleration changes or, more specifically, the wave breaking begins. It follows from our theoretical analysis (formula (8)) that the current of the accelerated electrons as they are captured depends on pulse power as $P^{0.6}$. This dependence is confirmed by the experimentally measured dependence of the current of accelerated electrons on microwave pulse power shown in Fig. 5a.

The maximum energy of the electrons accelerated via wave breaking can be estimated by using formula (6). Thus, for $P = 50$ W, $W_{wb} \approx 630$ eV. This value corresponds to a retarding potential of ~600 V, at which the current–voltage characteristic of the analyzer begins to decrease more sharply with increasing retarding potential (Fig. 4a). This suggests a deficit of electrons with energies above 600 eV. The maximum energy W_{wb} of the accelerated electrons determined from the current–voltage characteristics and calculated using formula (6) at various powers is shown in Fig. 7. There is satisfactory agreement between the theoretical and experimental data not only in the power dependence but also in absolute value.

The number of accelerated electrons escaping from the region of the focus can account for an appreciable fraction of the electron density at the focus ($n_c \sim 10^{11}$ cm^{-3} for $f_0 = 2840$ MHz). It can be estimated by using the current of the accelerated electrons recorded by the charged particle analyzer (Fig. 4b). Thus, at a pulse power $P = 50$ W, the current of the accelerated electrons at a retarding potential of –50 eV is about 20 mA, while their effective temperature is $T_h \sim 385$ eV. The electron current may be defined as $I = \bar{n}_h Se\langle v\rangle$,

where $\bar{n}_h$ is the mean electron density over an area $S \approx 0.07$ cm^2 of the analyzer inlet, and $\langle v \rangle = \sqrt{2T_h/m_e}$ is the mean velocity of the electrons corresponding to their effective temperature T_h. The mean density of the accelerated electrons is then $\bar{n}_h \approx 1.5 \times 10^9$ cm^{-3}.

Actually, the diameter of the accelerated electron beam is smaller than the diameter of the inlet to the analyzer. Accelerated electrons cause a significant increase in the intensity of optical plasma radiation; the diameter of its distribution probably corresponds to the diameter of the electron beam. The cross-sectional area of the accelerated electron beam estimated in this way is $S' \approx 0.02$ cm^2. The maximum density of the accelerated electrons that reached the analyzer can then be defined as $n_h = \bar{n}_h S/S' \approx 5 \times 10^9$ cm^{-3}. Thus, we have $n_h/n_c \sim 0.05$. Our calculation using formula (7) for $P_0 = 50$ W and $b = 0.4$ cm yields $n_h/n_c \approx 0.1$. Given the losses of electrons as they propagate from the focus region to the analyzer (~25 cm), there is satisfactory agreement between the estimates obtained. At high powers, $P = 5000$ W, as estimates similar to those for $P = 50$ W show, the fraction of the accelerated electrons increases by an order of magnitude and their density is comparable to the electron density near resonance, $n_h/n_c \sim 0.5$.

The energy of the accelerated electrons is much higher than the ionization energy of argon atoms, $I_i = 15.76$ eV. The accelerated electrons lose their energy during their collisions with argon atoms by ionizing and exciting them, which causes the electron density to increase. The newly formed electrons probably have low energies, as suggested by the presence of segments where the current decreases sharply near a zero retarding potential in the current–voltage characteristics (Fig. 4b). In turn, the increase in density at the focus causes a change in the propagation and absorption conditions for an oblique Langmuir wave near the focus and a displacement of the focus to lower densities. This is probably the reason why the generation of accelerated electrons subsequently decreases.

5. CONCLUSIONS

In conclusion, note that in the model experiments described in this paper, we investigated the electron acceleration dynamics during the breaking of an intense plasma wave near resonance at the plasma frequency (focus) in an inhomogeneous magnetized plasma. We experimentally determined the breaking threshold and showed its correspondence to the theoretically expected value. We showed that the derived experimental dependences of the current and energy of fast electrons on the intensity of the incident wave at various times are in good agreement with theoretical estimates. We showed that when the breaking threshold is significantly exceeded, up to 50% of the electrons at plasma resonance are captured and accelerated by the wave. The inferred good agreement between the experimental data and the simple theoretical model of one-dimensional breaking of a potential wave strongly suggests that we created a potential of several thousand volts in a plasma via the charge separation by ponderomotive forces and makes it possible to plan model experiments aimed at observing accelerated ions of the corresponding energies.

ACKNOWLEDGMEMTS

This work was supported in part by the Belorussian and Russian Foundations for Basic Research (project nos. F02R-092 and 02-02-81033 Bel 2002_a) and INTAS (grant no. AS-01-0233).

REFERENCES

1. S. V. Bulanov and L. M. Kovrizhnykh, Fiz. Plazmy **1**, 1016 (1975) [Sov. J. Plasma Phys. **1**, 555 (1975)].
2. L. M. Kovrizhnykh and A. S. Sakharov, Tr. Inst. Obshch. Fiz., Akad. Nauk SSSR **16**, 80 (1988).
3. C. G. Durfee III and H. M. Milchberg, Phys. Rev. Lett. **71**, 2409 (1993).
4. H. Ito, Y. Nishida, and N. Yugami, Phys. Rev. Lett. **76**, 4540 (1996).
5. M. P. Brizhinev, A. L. Vikharev, G. Yu. Golubyatnikov, *et al.*, Zh. Éksp. Teor. Fiz. **98**, 434 (1990) [Sov. Phys. JETP **71**, 242 (1990)].
6. A. L. Vikharev, J. Tech. Phys. (Warsaw) **41**, 485 (2000).
7. H. Ito, T. Fuji, N. Handa, *et al.*, in *Proceedings of IV International Workshop on Strong Microwaves in Plasmas* (Nizhni Novgorod, 1999), p. 550.
8. P. Mora, Plasma Phys. Controlled Fusion **43**, A31 (2001).
9. V. I. Arkhipenko, V. N. Budnikov, I. A. Romanchuk, and L. V. Simonchik, Fiz. Plazmy **7**, 396 (1981) [Sov. J. Plasma Phys. **7**, 216 (1981)].
10. V. I. Arkhipenko, V. N. Budnikov, E. Z. Gusakov, *et al.*, Fiz. Plazmy **13**, 693 (1987) [Sov. J. Plasma Phys. **13**, 398 (1987)].

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1154–1158.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1291–1297.

PLASMA, GASES

Formation of Nanostructured Carbon Films in Gas-Discharge Plasmas

A. A. Zolotukhin, A. N. Obraztsov, A. O. Ustinov, and A. P. Volkov

Moscow State University, Moscow, 119992 Russia
e-mail: obraz@acryst.phys.msu.ru

Received April 22, 2003

Abstract—Deposition of carbon materials from methane–hydrogen gas mixtures in a DC gas discharge is investigated. Parameters ensuring stable discharge conditions and synthesis of diamond and graphite-like films are determined. Optical emission spectroscopy is used to analyze the composition of the activated gas phase in the course of carbon film deposition. Synthesis of graphite-like carbon nanotubes and nanocrystallites is shown to correlate with the presence of C_2 dimers in the plasma. A noncatalytic mechanism of synthesis of nanostructured graphite in a carbon-containing gas phase is proposed. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Various nanostructured carbon materials are the subject of current interest because of their unique physical and chemical properties. In particular, carbon nanotubes and other graphite-like nanostructured materials can exhibit field emission at anomalously low electric field strengths [1–4]. In addition to graphite-like structure, the properties that determine the field emission efficiency include the morphology of nanostructured carbon cathodes and certain characteristics of their surfaces [3, 4]. These characteristics of nanostructured carbon materials depend on the parameters of their synthesis and processing. Almost all nanostructured carbon materials are deposited from carbon-containing gases activated by various methods (e.g., see [2]). For example, the nanostructured-carbon film cathodes examined in our previous studies [3, 4] were obtained by carbon deposition from methane–hydrogen mixtures activated by a DC gas discharge. This method can also be used to produce polycrystalline diamond and other carbon thin films [5, 6]. Deposition of carbon film from a gas depends on a variety of parameters. Relationship between these parameters and characteristics of the produced materials is of great importance both for practical applications and for understanding the fundamental mechanisms of synthesis of various carbon materials.

In this paper, we present the results of an experimental study of gas-phase deposition of carbon films from methane–hydrogen mixtures activated by a DC discharge. The study was conducted to determine the gas-discharge parameters ensuring stable growth of films with required properties and the deposition parameters of key importance for synthesis of nanostructured carbon materials characterized by high field-emission efficiencies.

2. EXPERIMENTAL

A schematic description of the experimental setup used in our studies of carbon film deposition can be found in our previous papers (e.g., see [4–6]). Film deposition was conducted in a reactor evacuated to a pressure of 10^{-3} Torr and filled with methane and hydrogen mixed in various proportions. The reactor (made from stainless steel) had water-cooled walls separated from the deposition zone by a thermal shield to ensure local thermodynamic equilibrium in the deposition zone. Both the internal diameter and height of the reactor were 400 mm. The current source used to initiate and sustain the DC gas discharge provided sufficient power to deposit films on substrates of diameters up to 50 mm. The gas inlet system ensured a prescribed continuous-flow rate of the mixture at a constant pressure inside the reactor. The gas discharge was initiated by applying voltage to a 50-mm gap between two electrodes. The substrate used in carbon film deposition was mounted on the anode. Diamond films are generally deposited on silicon substrates. Films consisting of nanotubes and other nanostructured carbon materials can also be

Parameters of gas-phase deposition of films of various composition and structure

Type of carbon film	Substrate temperature during growth, °C	Methane concentration, %	Gas pressure, Torr
Diamond	850–900	0.5–2	60–90
Nanocrystalline diamond	900–1000	2–5	60–100
Graphite-like	1000–1100	5–10	60–100
Soot	1100–1250	above 15	50–100

1063-7761/03/9706-1154$24.00

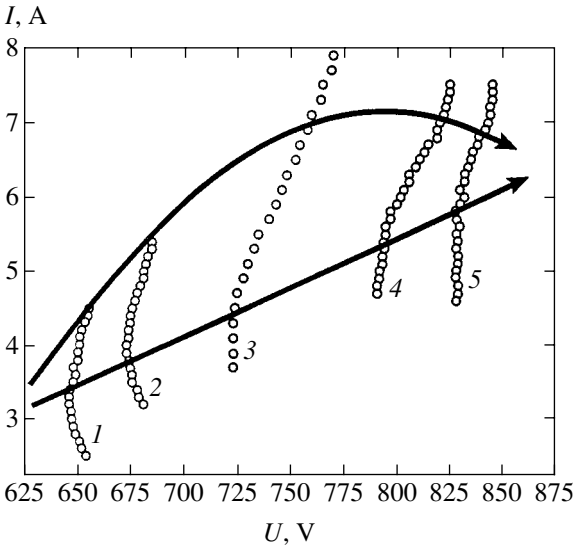

Fig. 1. Current–voltage characteristics of methane–hydrogen gas-discharge plasmas for a methane concentration of 8% at gas-mixture pressure 50 (*1*), 60 (*2*), 80 (*3*), 100 (*4*), and 110 Torr (*5*).

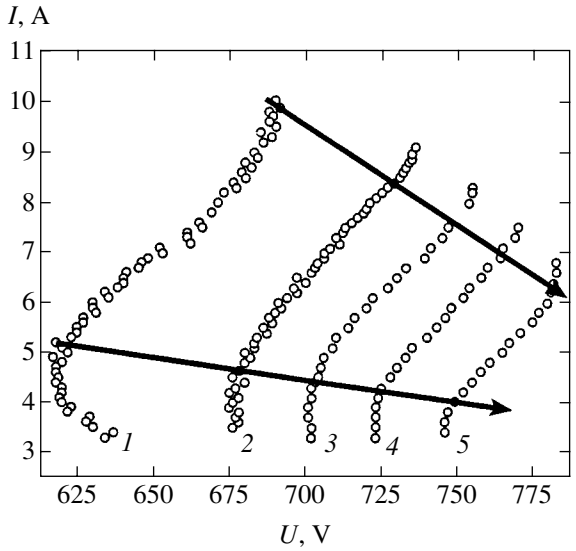

Fig. 2. Current–voltage characteristics of methane–hydrogen gas-discharge plasmas at a gas-mixture pressure of 80 Torr for methane concentrations 0 (*1*), 2% (*2*), 4% (*3*), 8% (*4*), and 15% (*5*).

deposited on nickel, tungsten, molybdenum, steel, and other substrates [4]. In the present study, we used silicon substrates 50 mm in diameter. We examined the dependence of deposition characteristics on electrical parameters of the discharge and on the gas composition and pressure. The experiments were conducted at a constant substrate temperature of 950°C maintained by simultaneously heating and water-cooling the substrate holder.

The electrical discharge parameters (voltage U and total current I) were set by using an adjustable current source. The gas-discharge plasma was monitored visually through quartz windows in the reactor walls and by recording its optical emission spectra (OES). To obtain OES, we focused the radiation emitted by the plasma through the quartz windows by a system of lenses onto the entrance slit of a monochromator in such a manner that different regions of the plasma column could be analyzed by using a high-sensitivity silicon-photodiode–based light detector. The signal acquisition system included a logarithmic amplifier for measuring spectral line intensities in a wide dynamic range. Current–voltage characteristics of the discharge and plasma OES were measured at gas-mixture pressures varying from 10 to 150 Torr for methane concentrations between 0 and 25%.

3. RESULTS AND DISCUSSION

Figure 1 shows the discharge current–voltage characteristics obtained for a methane concentration of 8% at pressures varying from 50 to 110 Torr. One common feature of all characteristics is their negative slope at relatively low voltages. Visual observation of the gas-discharge plasma revealed that the substrate (anode) surface was not completely spanned by the luminous region of glow discharge. With increasing voltage (and current), the luminous region at the substrate grew larger and the slope of the discharge current–voltage characteristic became positive after the substrate was completely spanned by the luminous region. As a certain characteristic voltage depending on the gas pressure was reached, the glow discharge spontaneously changed into an arc (as reflected by the increasing slopes of the discharge current–voltage characteristics shown in Fig. 1). The range of discharge parameters corresponding to normal (positive) slopes of the current–voltage characteristics (indicated by arrows in Fig. 1) can be interpreted as the domain of stable discharge conditions. Carbon film deposition could be controlled when conducted under these conditions. Outside this domain, we observed either nonuniform deposition of carbon on the substrate, due to the discharge inhomogeneity at its surface, or uncontrollable discharge behavior leading to overheating and even substrate failure in the arc mode.

Figure 1 also demonstrates that the domain of stable gas-discharge plasma corresponds to the widest range of current at a pressure of about 80–100 Torr when the methane concentration in the gas mixture is 8%. This particular combination of discharge parameters characterizes a previously found optimal regime of deposition of nanostructured carbon materials [3–6]. Similar trends were observed at different methane concentrations. Figure 2 shows the discharge current–voltage characteristics obtained at a gas pressure of 80 Torr inside the reactor for methane concentrations between 0 and 15%. As in Fig. 1, arrows indicate the domain of stable discharge where controlled deposition of carbon

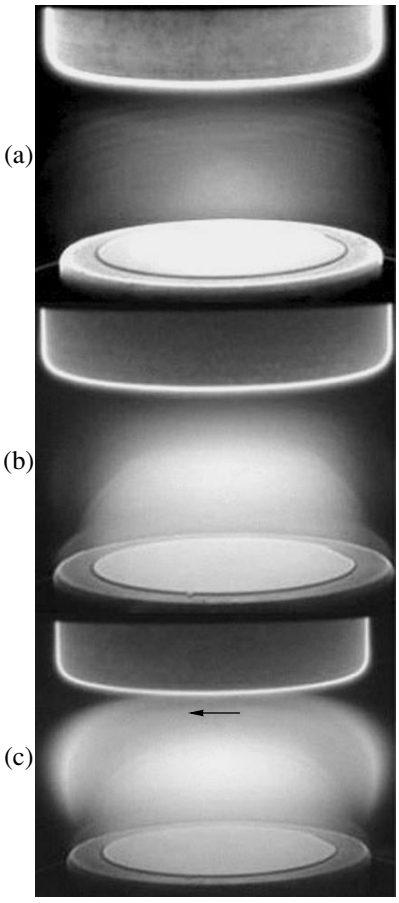

Fig. 3. Typical gas discharges in pure hydrogen (a) and hydrogen–methane mixtures with methane concentrations 10% (b) and 25% (c) at a pressure of 60 Torr. The substrate is a silicon plate of diameter 50 mm. The photographs were taken at voltages 650 V (a), 750 V (b), and 850 V (c) and currents of 7 A (a), 6 A (b), and 5 A (c).

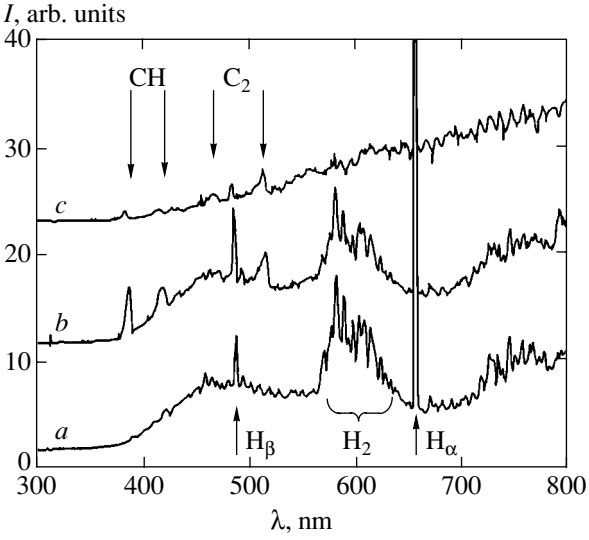

Fig. 4. Typical optical emission spectra of gas-discharge plasmas for pure hydrogen (*a*) and methane concentrations of 10% (*b*) and 25% (*c*). Gas pressure is 80 Torr. Discharge voltages are 650 V (*a*), 750 V (*b*), and 850 V (*c*) and currents are 7 A (*a*), 6 A (*b*), and 5 A (*c*).

films can be conducted. It should be noted that a stable glow discharge could be sustained at methane concentrations of up to 25%. However, the substrate quickly overheated under such conditions, and its temperature could not be reduced by water-cooling the substrate holder.

The composition of the gas phase from which carbon was deposited on the substrate was determined by analyzing the OES of the gas-discharge plasma. The geometry and color of the luminous region of the discharge varied substantially with the electrical parameters of the discharge and the composition and pressure of the mixture. As an example, Figs. 3a–3c show photographic images of the discharge taken through a quartz window in the reactor wall at a gas pressure of 60 Torr and methane concentration of 0, 10, and 25%, respectively. Figure 4 shows the OES of the gas-discharge plasma recorded for these methane concentrations near the substrate surface (curves *a* and *b* corresponding to 0 and 10%, respectively) and at the periphery of the luminous region (curve *c* corresponding to 25%).

According to the discharge spectra measured for pure hydrogen at a discharge voltage of 650 V and a current of 7 A (curve *a* in Fig. 4), plasma emission was dominated by the recombination transition lines of hydrogen atoms (486 nm for H_β and 656 nm for H_α) and molecules (550 to 650 nm for H_2). With the addition of methane, the color of the discharge region changed to yellowish green, and the typical spectrum (*b*) contained the recombination lines of CH radicals (386 and 422 nm) and C_2 dimers (515 and 560 nm). The corresponding discharge region in Fig. 3b looks much brighter as compared to the discharge in pure hydrogen, whose emission lies in the blue spectral range (Fig. 3a). The spectrum represented by curve *b* in Fig. 4 was obtained at a voltage of 750 V and a current of 6 A. The shapes and spectral positions of the lines are consistent with previously reported observations of methane–hydrogen plasmas (e.g., see [7, 8]).

The line intensities of hydrogen atoms and molecules were almost constant in all plasma-column regions examined in this study, whereas those of carbonaceous compounds (CH and C_2) increased substantially from periphery toward the substrate surface. These carbonaceous compounds were found in the plasmas with methane concentrations varying from 0.5 to 25%. At methane concentrations above 15%, we observed a region of intense yellowish orange light emission at the periphery of the plasma column. Figure 3c shows a photographic image of the discharge taken at a methane concentration of 25%, a voltage of 850 V, and a current of 5 A. The arrow indicates the region where the plasma emission spectrum represented by curve *c* in Fig. 4 was observed. The intensities of individual spectral lines are relatively low, and the intensity of the structureless background spectrum gradually increases with wavelength. These features suggest that the yel-

lowish orange light was emitted by a high-temperature condensed phase—most likely, the soot that formed by direct condensation of carbon in the gas phase. Condensation of this kind can occur at high methane concentrations in the presence of excessive carbon in regions of relatively cool plasma [9].

As an illustration of the difference between the carbon materials deposited under different conditions, Fig. 5 shows typical Raman scattering spectra (RSS) obtained for a polycrystalline diamond film (curve *1*), a nanocrystalline diamond (curve *2*), a graphite-like nanostructured material (curve *3*), and a sooty material (curve *4*). These spectra include lines characteristic of diamond-like nanocrystallites smaller than 2 nm at 1140 and 1470 cm^{-1} and the line at 1330 cm^{-1} corresponding to the "common" diamond with a substantially larger crystallite size [5]. The RSS lines at 1350 cm^{-1} and around 1580 cm^{-1} (from 1550 to 1620 cm^{-1}) correspond to various disordered graphites. Note that, since the line at 1580 cm^{-1} is also characteristic of multilayered carbon nanotubes [10], the Raman scattering technique cannot be used to identify nanostructured carbon components of such films.

An analysis of the experimental results obtained by using the Raman scattering technique, scanning and electron tunnel spectroscopy, atomic force spectroscopy, cathode luminescence microscopy, and other methods showed that polycrystalline diamond films were deposited at a gas-mixture pressure of about 80 Torr and methane concentrations between 0.1% and 2.0% (depending on the substrate temperature). When the methane concentration was between 2% and 5%, nanocrystalline diamond was produced. Formation of graphite-like carbon nanotube material and nanocrystallites was observed at methane concentrations between 5% and 10%. When the methane concentration exceeded 15%, soot-like disordered carbon was produced. Moreover, as pointed out above, the use of such excessive concentrations reduced the stability of film deposition because of substrate overheating.

A comparative analysis of the films obtained at various methane concentrations and the optical emission spectra of the plasmas unambiguously points to the existence of a correlation between the presence of C_2 dimers in the activated gas phase and synthesis of nanostructured carbon materials, such as nanocrystalline diamond or graphite and carbon nanotubes. The key role played by C_2 dimers in the synthesis of nanocrystalline diamonds was noted previously (e.g., see [11]). It was shown that the most efficient process in terms of energy is the clustering of C_2 dimers into linear chains of atoms with acetylene bonding (carbene structures). After a certain critical cluster size is reached, such a cluster can transform into a planar graphite-like layer of carbon atoms oriented perpendicular to the substrate. These layers can make up plate-like graphite crystallites several atoms thick. Alternatively, they can roll up spontaneously (or under the influence of some factors depending on the deposition process) into structures analogous to carbon nanotubes or into nuclei that subsequently grow into nanotubes. This process can be conducted without a catalyst, unlike other techniques used to produce carbon nanotubes [2, 10].

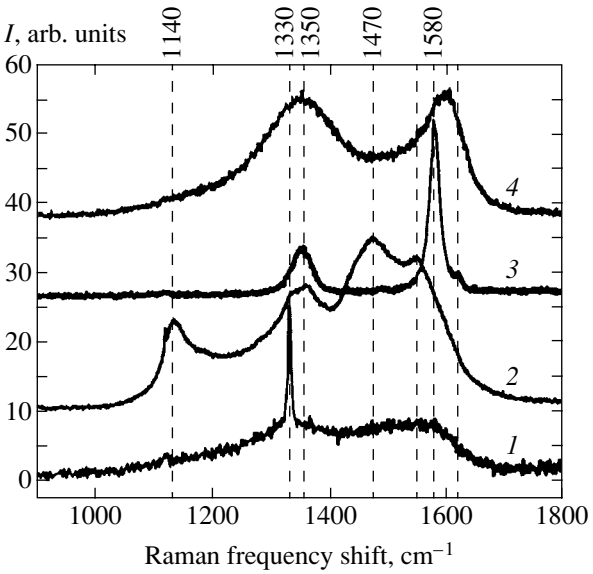

Fig. 5. Raman scattering spectra (RSS) of carbon materials produced by gas-phase deposition: (*1*) polycrystalline diamond film, (*2*) nanocrystalline diamond film, (*3*) nanographite film, (*4*) sooty material. The weak signal at about 1120 cm^{-1} is due to the spurious effect of luminescent light sources.

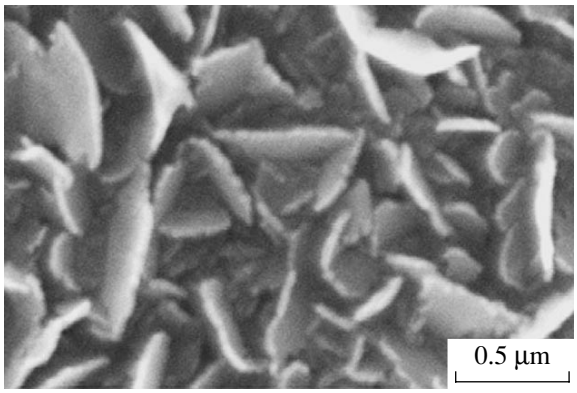

Fig. 6. Fragments of SEM images of nanographite films obtained after deposition during 15 min.

This model of noncatalytic synthesis of nanostructured carbon materials is consistent with the results obtained by means of scanning electron microscopy (SEM). Figures 6 and 7 show, respectively, the images of carbon structures taken after deposition for 15 and 60 min. One can clearly see that the characteristic size of carbon structures increases with deposition time, while their geometry changes substantially. The initial structures are planar plate-like graphite nanocrystallites

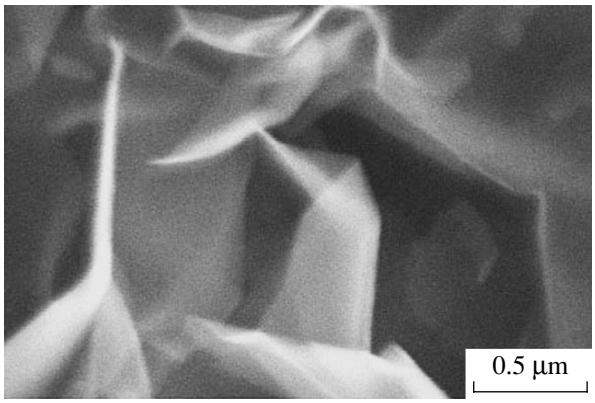

Fig. 7. Fragments of SEM images of nanographite films obtained after deposition during 60 min. The underlying parts of conically rolled-up fragments are seen through the above-lying ones, which implies a very small thickness of nanographite sheets.

oriented mainly perpendicular to the substrate surface (Fig. 6). They can transform into conically rolled-up structures (Fig. 7). According to the results obtained by using an RSS technique, electron microscopy, a diffraction technique, and Auger electron spectroscopy [5], both the films and the microscopic structures are well-ordered graphites. The fact that the inner part of a rolled-up sheet in Fig. 7 is clearly seen through its outer layer suggests that its thickness is small, since the secondary electrons forming an SEM image can penetrate a graphite layer only a few atoms thick.

4. CONCLUSIONS

In the present study, a correlation is established between the parameters of gas discharge in methane–hydrogen mixtures and the phase composition and structural properties of the carbon films obtained by deposition. The range of glow-discharge parameters ensuring stable deposition is determined. Carbon thin films characterized by various properties and fractions of diamond and graphite-like components have been produced by deposition at discharge current densities between 0.2 and 0.5 A/cm^2. By analyzing the OES of the discharge plasma recorded during deposition, both CH and C_2 were detected in the gas phase near the substrate surface. The presence of C_2 dimers in the gas phase is found to correlate with synthesis of nanostructured carbon materials (nanocrystalline diamond, carbon nanotubes, graphite nanocrystallites). A mechanism of synthesis of films consisting of graphite-like forms of nanostructured carbon is suggested. An instance of direct carbon condensation has been observed in the gas phase at the periphery of the plasma column at methane concentrations above 15%.

ACKNOWLEDGMENTS

This work was supported by the International Association for Cooperation with Scientists from the former Soviet Union (INTAS), grant no. 01-254.

REFERENCES

1. J.-M. Bonard, H. Kind, Th. Stöckli, and L.-O. Nilsson, Solid-State Electron. **45**, 893 (2001).
2. A. V. Eletskiĭ, Usp. Fiz. Nauk **172**, 401 (2002) [Phys. Usp. **45**, 369 (2002)].
3. A. N. Obraztsov, A. P. Volkov, A. I. Boronin, and S. V. Koshcheev, Zh. Éksp. Teor. Fiz. **120**, 970 (2001) [JETP **93**, 846 (2001)].
4. A. N. Obraztsov, A. P. Volkov, K. S. Nagovitsyn, *et al.*, J. Phys. D: Appl. Phys. **35**, 357 (2002).
5. A. N. Obraztsov, A. P. Volkov, and I. Yu. Pavlovskiĭ, Pis'ma Zh. Éksp. Teor. Fiz. **68**, 56 (1998) [JETP Lett. **68**, 59 (1998)].
6. A. N. Obraztsov, I. Yu. Pavlovsky, A. P. Volkov, *et al.*, Diamond Relat. Mater. **8**, 814 (1999).
7. F. Zhang, Y. Zhang, Y. Yang, *et al.*, Appl. Phys. Lett. **57**, 1467 (1990).
8. T. Vandevelde, M. Nesladek, C. Quaeyhaegens, and L. Stals, Thin Solid Films **290**–**291**, 143 (1996).
9. B. V. Spitsyn, in *Handbook of Crystal Growth*, Ed. by D. T. J. Hurle (Elsevier, Amsterdam, 1994), Vol. 3, p. 403.
10. *Carbon Nanotubes*, Ed. by M. S. Dresselhaus, G. Dresselhaus, and P. Avouris (Springer, Berlin, 2000).
11. D. M. Gruen, Annu. Rev. Mater. Sci. **29**, 211 (1999).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1159–1167.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1298–1307.

FLUIDS

The Influence of Anchoring Energy on the Prolate Shape of Tactoids in Lyotropic Inorganic Liquid Crystals

A. V. Kaznacheev, M. M. Bogdanov, and A. S. Sonin

Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences, ul. Vavilova 28, Moscow, GSP-1, 117813 Russia

e-mail: son@pmc.ineos.ac.ru

Received April 17, 2003

Abstract—A model was constructed to describe the prolate shape of anisotropic regions, tactoids, coexisting with the isotropic phase in lyotropic inorganic liquid crystals. The elastic energy of the tactoid, the surface energy, and the interaction energy between the director field and the boundary of the tactoid were taken into account. Large-sized tactoids were shown to be prolate because of the competition between the elastic energy of the nematic phase of the tactoid and the surface energy. Small-sized tactoids were prolate because of the competition of the surface energy with the anchoring energy between the director and the boundary of the tactoid. The suggested model was applied to experimental data to determine the ratio of the elastic constants K_3/K_1 and the ratio between the anchoring energy W and the surface tension σ depending on the "time of aging" of vanadium pentoxide sols in water. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

At present, the macroscopic physics of thermotropic liquid crystals has been constructed in outline [1–3]. The focus of studies is currently shifting to lyotropic liquid crystals that form mesophases when certain substances are dissolved in certain solvents [4–7]. According to the chemical classification, lyotropic liquid crystals include the class of inorganic liquid crystals. They are formed in dispersions of inorganic substances [7–10]. These systems have been known since the 1920s, but their physics has not been studied at all.

In recent years, we have initiated systematic studies of the elastic properties of inorganic mesophases for the example of a typical liquid-crystalline aqueous sol of vanadium pentoxide [11–13]. Of special interest is the so-called tactoid phase formed at low (~1 wt %) concentrations of vanadium pentoxide. This is a two-phase system in which one of the phases is isotropic and the other is an anisotropic mesophase. The mesophase is interspersed in the isotropic phase in the form of prolate spindle-shaped droplets; precisely these droplets are called tactoids (Fig. 1). Until recently, studies of the tactoid phase have largely been descriptive in character [14–17]. One of the most important questions concerning tactoid sols is that of their molecular structure and the reasons for the formation of a two-phase system in the form of a dispersion. This question still remains open. There is, however, one more no less important question: why do tactoids have prolate rather than spherical shapes, as is usually observed for droplets of thermotropic liquid crystals [18].

In our preceding work [13], we suggested a theoretical description of the prolate shape of tactoids and obtained an equation for their free energy as the sum of the elastic energy of the nematic phase and surface energy. A study of the free energy at its extremum allowed us to obtain a dependence of the geometric size of tactoids on the material constants of the mesophase. The experimental data on the size of tactoids were compared with the theoretical dependences to evaluate the ratio between the elastic constants and between the elastic constants and the surface tension. The ratio between the elastic constants K_3/K_1 in the system under consideration was found to reach 100 in certain instances. This is much larger than the ratio characteristic of thermotropic liquid crystals, which does not exceed 3 [1, 19].

In this work, we continue studies of this unusual mesophase. In some instances, a comparison of tactoid sizes with theoretical dependences was found to lead to certain contradictions. To remove them, we had to take into account the anchoring energy between the director field and the tactoid–isotropic phase boundary in the equation for the free energy of tactoids obtained earlier. This approach allowed us to elucidate the reasons for the prolate shapes of both large-sized tactoids, for which the condition of strong director–boundary binding is met, and small-sized tactoids, for which this condition does not hold. A comparison of the theoretical results with the sizes measured experimentally for tactoids of different volumes allowed us to determine the ratio between the elastic constants K_3/K_1 and between the anchoring energy W and surface tension σ depending on the "time of aging" of vanadium pentoxide sols in water.

1063-7761/03/9706-1159$24.00

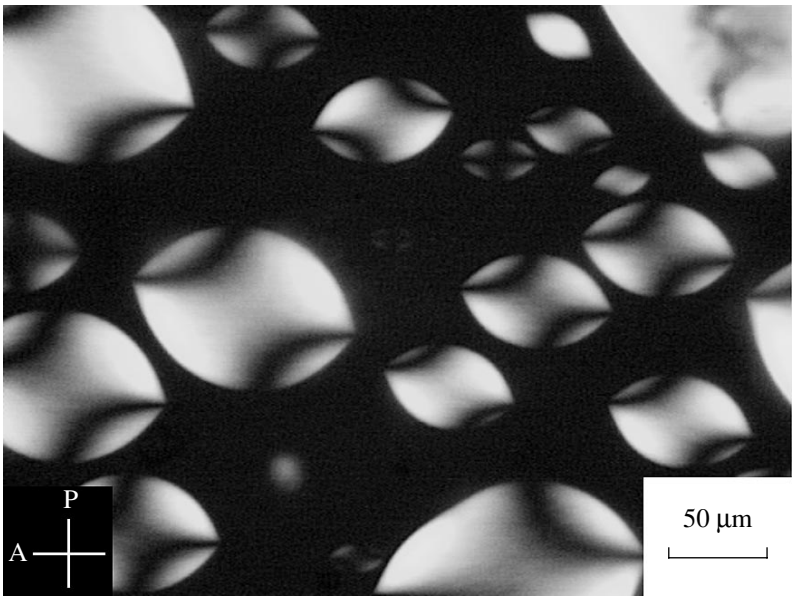

Fig. 1. Texture of the tactoid phase in a vanadium pentoxide (V_2O_5)–water sol at a V_2O_5 concentration of 1.1 wt %. Cell thickness is 200 μm.

In Section 2, we describe a model of the director shape and field for the nematic phase of a tactoid based on experimental data. The problem is formulated taking into account the elastic energy of the tactoid, the surface energy, and the interaction energy between the director field and the boundary of the tactoid. The problem is analyzed to consider situations that admit analytic solutions and allow the physical reasons for the prolate shape of tactoids to be elucidated. In conclusion, the problem is solved numerically. In Section 3, we describe a procedure for measuring the size of tactoids of different volumes in the vanadium–water lyotropic inorganic liquid crystal and the results of such measurements. The experimental results are compared with the suggested model of the shape of tactoids. As a result, dependences of the physical properties of the tactoid phase on the time of its aging were obtained. The principal results are summarized in Section 4.

2. THE INFLUENCE OF ANCHORING ENERGY AND ELASTICITY OF THE NEMATIC PHASE OF TACTOIDS ON THEIR PROLATE SHAPE

2.1. Problem Statement

In the preceding paper [13], we used experimental data to suggest a model of the director shape and field for the nematic phase of a tactoid. It was shown that the boundary of the tactoid was the surface of revolution of an arc of angle 2α of a circle of radius R about its span. Because of strong binding, the director field at the tactoid boundary has a tangential orientation. Provided the anchoring energy is finite, the director at the boundary can deviate from the tangential orientation. To take this possibility into account, we here use the director field given by the basis vector $\mathbf{e}_{\eta 1}$ of the bispherical system of coordinates.

A model of the director shape and field for the tactoid nematic phase is shown in Fig. 2. The variables of the problem are R and α, which describe the shape of the tactoid, and R_1 and α_1, which describe the director **n** field in it. The unit vector **n** coincides with the basis vector $\mathbf{e}_{\eta 1}$ of the bispherical coordinate system [20]. The angle α can take on values from zero (for needle-shaped tactoids) to $\pi/2$ (for spherical tactoids). The angle α_1 changes from zero (for a uniform director

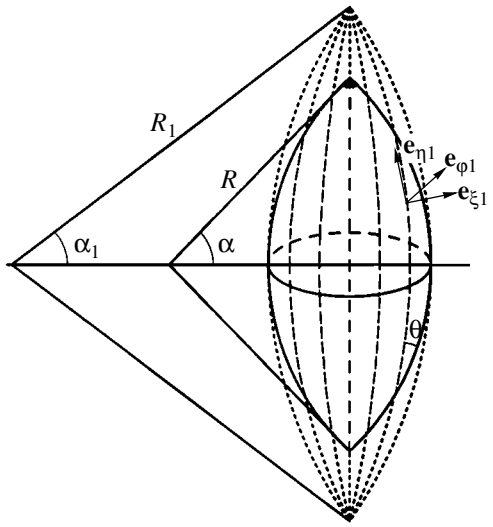

Fig. 2. Model of the director shape and field for the nematic phase of a tactoid.

field) to α (for a director with a tangential orientation at the boundary). These variables are related by two constraint equations,

$$V = R^3\Psi(\alpha) = \text{const}, \quad (1)$$

$$R(1-\cos\alpha) = R_1(1-\cos\alpha_1), \quad (2)$$

where $\Psi(\alpha) = 2\pi(\sin\alpha - \alpha\cos\alpha - (\sin^3\alpha)/3)$. Equation (1) is the condition of a constant volume V of the tactoid, and (2) follows from the geometry of the problem. Because of the presence of the constraint equations, there remain two independent variables. We can select α and α_1 as such variables. However, further calculations are more convenient to perform using the independent variables α and $\gamma = (\tan(\alpha_1/2)/\tan(\alpha/2))^2$, where $0 \le \gamma \le 1$.

The equilibrium shape of a tactoid with a constant volume is determined by the condition of minimum total energy. The total energy Φ is the sum of the elastic energy Φ_{el} of the nematic phase, the surface energy Φ_S, and the interaction energy Φ_W between the director field and the surface. The elastic energy is related to the distortion of the director field [1] and can be written directly from dimension considerations in the form

$$\Phi_{el} = K_1R\Phi_{el}^{(1)}(\alpha,\gamma) + K_3R\Phi_{el}^{(3)}(\alpha,\gamma), \quad (3)$$

where K_1 and K_3 are the elastic constants of splay and bend deformations, respectively [1], and $\Phi_{el}^{(1)}(\alpha,\gamma)$ and $\Phi_{el}^{(3)}(\alpha,\gamma)$ are the dimensionless functions of the variables α and γ related to these deformations. These functions are calculated in Appendix I. The surface energy is $\Phi_S = \sigma S$, where σ is the surface tension and S is the area of the tactoid surface. The equation for Φ_S is obtained by calculating S, which gives

$$\Phi_S = \sigma R^2\Phi_S(\alpha), \quad (4)$$

where $\Phi_S(\alpha) = 4\pi(\sin\alpha - \alpha\cos\alpha)$. The interaction energy between the director field and the surface Φ_W is calculated as [21]

$$\Phi_W = \int_S w(\theta)\,dS, \quad (5)$$

where $w(\theta) = (W/2)\sin^2\theta$ is the Rapini potential [21], W is the anchoring energy, and θ is the angle between the director and the surface of the tactoid (see Fig. 2). An expression for Φ_W follows immediately from dimensional analysis:

$$\Phi_W = WR^2\Phi_W(\alpha,\gamma), \quad (6)$$

where $\Phi_W(\alpha,\gamma)$ is the dimensionless function of the variables α and γ. This function is calculated in Appendix II. Substituting R given by constraint equation (1) into (3), (4), and (6) yields the equation for the total energy of a tactoid of a constant volume

$$\tilde\Phi = \frac{K_i}{\sigma V^{1/3}}\Psi_i(\alpha,\gamma) + \Psi_S(\alpha) + \frac{W}{\sigma}\Psi_W(\alpha,\gamma), \quad (7)$$

where $\tilde\Phi = \Phi/\sigma V^{2/3}$ is the dimensionless energy; $\Psi_i(\alpha,\gamma) = \Phi_{el}^{(i)}(\alpha,\gamma)/\Psi^{1/3}$ ($i = 1, 3$) (here and throughout, the summation over the repeating index i is implied); $\Psi_S(\alpha) = \Phi_S(\alpha)/\Psi^{2/3}$; and $\Psi_W(\alpha,\gamma) = \Psi_W(\alpha,\gamma)/\Psi^{2/3}$.

Equation (7) for the total energy contains the dimensionless parameter W/σ, characteristic problem sizes $C_i = K_i/\sigma$ ($i = 1, 3$), and tactoid volume V. Given these values, we can find the equilibrium α and γ values that correspond to minimum energy (7). Further, (1) and (2) can be used to calculate the equilibrium R and R_1 values. That is, we obtain complete information about the equilibrium director shape and field for a tactoid of a constant volume. A change in V changes the equilibrium values of the problem variables. Among various dependences that can be obtained, $R(\alpha)$ is the most important one, for it can be compared with experimentally measured sizes of tactoids of various volumes. Because of the complexity of the $\Psi_i(\alpha,\gamma)$ ($i = 1, 3$) and $\Psi_W(\alpha,\gamma)$ functions, the extremum of energy (7) and the $R(\alpha)$ dependences are found numerically. The corresponding results are given in subsection 2.3.

To understand the physical reasons for the prolate shape of tactoids and guess in advance the results of the numerical solution, the situations that admit analytic solutions to the problem under consideration are considered in the next subsection.

2.2. Analysis of the Problem

The first situation corresponds to the strong binding condition. If $W/\sigma \gg 1$ and $W/\sigma \gg K_i/\sigma V^{1/3}$, then $\Psi_W(\alpha,\gamma) = 0$ and $\gamma = 1$, because the largest director deviations from the tangential orientation cause a sharp increase in the last term in equation (7) for the energy, which then takes the simpler form

$$\tilde\Phi = \frac{K_i}{\sigma V^{1/3}}\Psi_i(\alpha,1) + \Psi_S(\alpha), \quad (8)$$

where $\Psi_i(\alpha, 1)$ are increasing functions of angle α, which can be written analytically (see Appendix I). As $\Psi_S(\alpha)$ is a decreasing function, the competition between the elastic and surface terms in (8) results in the appearance of a minimum of $\tilde\Phi$. Examination of

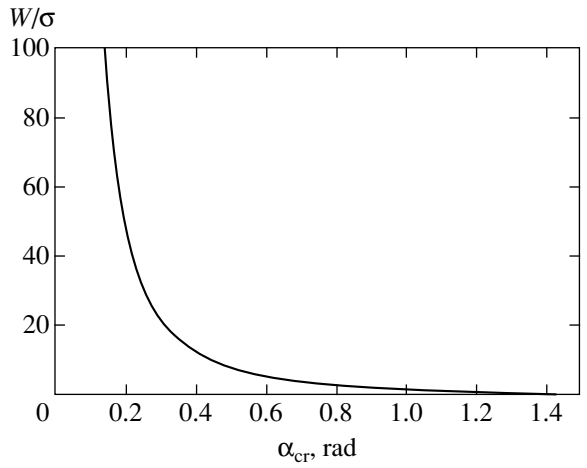

Fig. 3. Dependence of α_{cr} on W/σ.

energy (8) for an extremum yields the $R(\alpha)$ dependence, which, as in [13], has the form

$$R = C_1 f_1(\alpha) + C_3 f_3(\alpha), \tag{9}$$

where $C_i = K_i/\sigma$ ($i = 1, 3$) are the characteristic sizes of the problem and

$$f_1(\alpha) = \frac{\alpha\sin\alpha - 2\alpha^2\cos\alpha + \sin^2\alpha\cos\alpha}{\cos\alpha[\alpha(\alpha + \sin\alpha\cos\alpha) - 2\sin^2\alpha]},$$

$$f_3(\alpha) = \frac{(\alpha^2 - \sin^2\alpha)[\alpha(1 + 2\cos^2\alpha) - 3\sin\alpha\cos\alpha]}{4\sin\alpha\cos\alpha[\alpha(\alpha + \sin\alpha\cos\alpha) - 2\sin^2\alpha]}$$

are the increasing functions of angle α.

The conclusion can be drawn from the above that, under strong binding conditions, tactoids are prolate because of the competition between the elastic and surface energies.

The strong binding condition becomes invalid as $V \longrightarrow 0$; that is, we then have $W/\sigma \ll K_i/\sigma V^{1/3}$. Arbitrary small perturbations of the uniform director field then sharply increase the elastic terms in (7). Therefore, $\Psi_i(\alpha, \gamma) = 0$ and $\gamma = 0$. The equation for the energy then takes the simpler form

$$\tilde{\Phi} = \Psi_S(\alpha) + \frac{W}{\sigma}\Psi_W(\alpha, 0), \tag{10}$$

where $\Psi_W(\alpha, 0))$ is an increasing function of the angle α, which can be written analytically (see Appendix II). The competition between the surface and anchoring energies in (10) results in the appearance of a minimum of $\tilde{\Phi}$. Examination of energy (10) for an extremum yields the dependence of the angle α_{cr}, which characterizes small-volume ($V \longrightarrow 0$) tactoids, on the dimensionless parameter W/σ,

$$\frac{W}{\sigma} = -\frac{\Psi_S'(\alpha)}{\Psi_W'(\alpha, 0)} = \frac{24\alpha_{cr}}{\sin 2\alpha_{cr} - 2\alpha_{cr}} + \frac{8(\sin\alpha_{cr} - \alpha_{cr}\cos\alpha_{cr})}{\sin\alpha_{cr} - \alpha_{cr}\cos\alpha_{cr} - \frac{1}{3}\sin^3\alpha_{cr}}. \tag{11}$$

Function (11) is plotted in Fig. 3. If $W/\sigma = 0$, then $\alpha_{cr} = \pi/2$ and the tactoids have the shape of a sphere. If $W/\sigma > 0$, then $\alpha_{cr} < \pi/2$ and the tactoids are prolate because of the competition between the surface and anchoring energies. As $W/\sigma \longrightarrow \infty$, α_{cr} tends to zero.

The conclusion can be drawn that the $R(\alpha)$ dependence should tend to $(\alpha_{cr}, 0)$ as $V \longrightarrow 0$. When V increases, this dependence becomes similar to function (9), which corresponds to infinitely strong binding.

2.3. Numerical Solution

Examination of (7) for an extremum requires knowledge of the universal dimensionless functions $\Psi_i(\alpha, \gamma)$ ($i = 1, 3$) and $\Psi_W(\alpha, \gamma)$. They were obtained by numerical integrations (I.4), (I.5), and (II.6). The level lines and the position of the minimum of energy (7) at $K_1/\sigma = 10$ μm, $K_3/\sigma = 100$ μm, $W/\sigma = 10$, and $V = 10^3$ μm^3 are shown in Fig. 4. Changes in the volume of a tactoid cause changes in the position of the minimum along the dashed line. As $V \longrightarrow 0$, we obtain $\alpha \longrightarrow \alpha_{cr}$ and $\gamma \longrightarrow 0$. When $V \longrightarrow \infty$, we have $\alpha \longrightarrow \pi/2$ and $\gamma \longrightarrow 1$.

To summarize, setting the K_1/σ, K_3/σ, and W/σ parameters and varying volume V, we calculated the equilibrium values of the variables α and γ corresponding to minimum energy $\tilde{\Phi}$. Further, constraint equations (1) were used to obtain the $R(\alpha)$ dependences shown in Fig. 5. In this figure, curve *1* corresponds to the infinitely strong binding condition, which leads to (9). At finite W/σ values in the region of large volumes, all curves approach curve *1*, because the strong binding condition is then satisfied. In the region of small volumes, deviations from curve *1* are observed. Stretching of tactoids is limited by the angle α_{cr}, which decreases according to (11) as the W/σ ratio increases.

3. EXPERIMENTAL RESULTS AND DISCUSSION

A comparison of the theoretical $R(\alpha)$ dependences with experimentally measured sizes of tactoids of various volumes allows us to obtain the K_1/σ, K_3/σ, W/σ, and K_3/K_1 values. For this purpose, we prepared vanadium pentoxide sols in water by the Biltz method [22]. The concentrations C of the sols were determined by evaporation immediately after their preparation. Sols

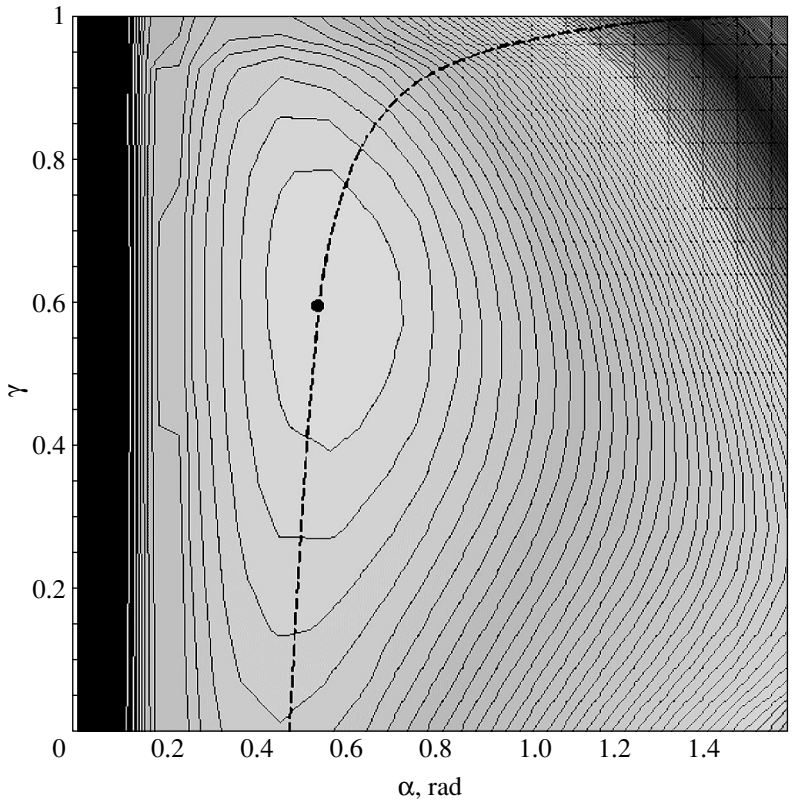

Fig. 4. Level lines and the energy (7) minimum position at $C_1 = 10$ μm, $C_3 = 100$ μm, $W/\sigma = 10$, and $V = 10^3$ μm^3. The dashed line shows how the minimum position changes with changes in the volume of the tactoid.

nos. 1 and 2 had $C = 0.5$ and 0.6 wt %, respectively, and both had pH ≈ 3. One day after the preparation, the sols were optically isotropic. Sols obtained by the Biltz method are nonequilibrium systems, and the tactoid phase appears in them as time passes. Our observations showed that the time required for the tactoid phase to appear depended on the concentration and pH of the medium. Increasing C and pH decreased this time. The concentrations and pH used in this work allowed us to study the appearance and development of the tactoid phases for half a year. Approximately two months after the preparation of the sols, they began to stratify. A turbid phase began to form in the lower parts of the vessels with the sols. A distinct interface between the upper and lower phases was observed. The amount of the lower phase grew as time passed. Polarization-optical observations showed that the lower phase was optically anisotropic, its texture corresponded to a nematic phase, and magnetohydrodynamic domains formed in it under the action of a magnetic field. The upper transparent phase remained optically isotropic. The tactoid phase was prepared by mixing the upper and lower phases. The mixing ratio influenced the number of tactoids formed.

The $R(\alpha)$ dependences were measured for samples loaded into plane-parallel capillaries about 200 μm thick, which were sealed by picein. The thickness of the capillaries was determined by Teflon lining and measured interferometrically prior to loading. Immediately after loading, the texture of the substance in the capillaries was anisotropic. We observed substance stratification into the isotropic and anisotropic regions in the time (from an hour to a day) that depended on the time of aging. Further, the anisotropic regions acquired the shape of tactoids. This was accompanied by a particle-size distribution.

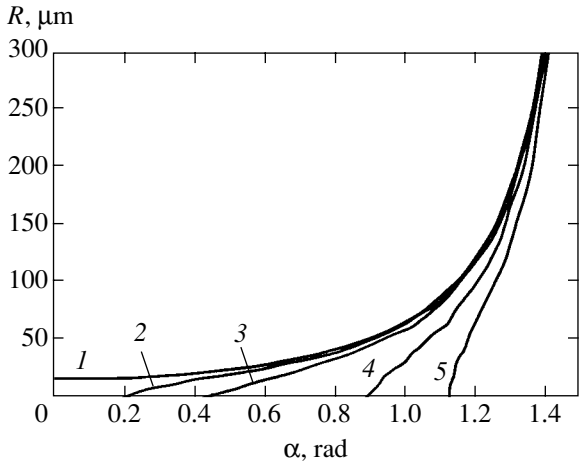

Fig. 5. Functions $R(\alpha)$ at $C_1 = 4$ μm, $C_3 = 40$ μm, and W/σ of (*1*) ∞, (*2*) 50, (*3*) 10, (*4*) 2, and (*5*) 1.

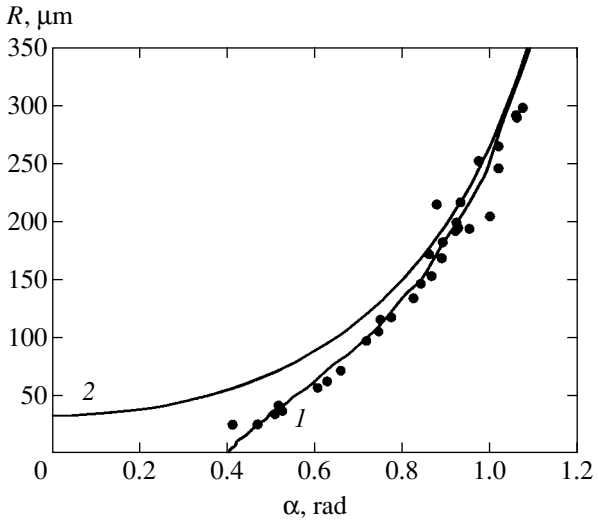

Fig. 6. Experimental $R(\alpha)$ dependence obtained for sol no. 2 (C = 0.6 wt %) 83 days after its preparation. Line *1* is the theoretical dependence calculated for $C_1 = 8 \pm 2$ μm, $C_3 = 230 \pm 5$ μm, and $W/\sigma = 12 \pm 2$, and line *2* is the plot of function (9) constructed for the same C_1 and C_3 values.

To see if the prolate shape of tactoids was equilibrium, flows were mechanically excited in capillaries with the substance. This caused smearing of tactoid boundaries and induced birefringence in the isotropic phase. The texture became anisotropic as a whole. As time passed, the same process as that directly after substance loading occurred. The anisotropic regions again acquired the shape of tactoids.

For the system to attain equilibrium, all measurements were taken in a day or later (depending on the aging time) after loading of the substance into capillaries. Measurements were taken with an Axiolab Pol (Zeiss) polarization microscope. The images were transmitted to a monitor with the use of a video camera. Tactoid sizes were determined with a Linkam VTO 232 attachment. Only tactoids whose size was smaller than the thickness of the cell were measured.

The measurement results were the R and α parameters for tactoids of various volumes. A typical experimental $R(\alpha)$ dependence obtained for sol no. 2 is shown in Fig. 6. The dependence was measured 83 days after preparation. A comparison of the theoretical and experimental $R(\alpha)$ dependences allowed us to calculate the C_1, C_3, and W/σ parameters by the method of least squares. The approximating curve is shown in Fig. 6 by solid line *1*. Line *2* is function (9) plotted for the same C_1 and C_3 values as line *1*.

Measurements of $R(\alpha)$ after various times of aging gave the dependences of C_1, C_3, W/σ, and $K_3/K_1 = C_3/C_1$ on this time shown in Figs. 7–10. It follows from Fig. 7 that $C_1 \sim 8$ μm was independent of the time and the sol number. Using the $K_1 = 4 \times 10^{-7}$ dyn value obtained from the data on the Freedericksz transition in the nematic phase of the system under consideration [12] and $C_1 = 8$ μm, we found that surface tension is $\sigma = 5 \times 10^{-4}$ erg/cm^2 at the tactoid boundary. Such low surface tension values are responsible for the prolate shape of macroscopic tactoids. It follows from Fig. 8 that C_3 decreases for both sols as time passes.

The dependence of the ratio between elastic constants K_3/K_1 on the time of aging is shown in Fig. 9, according to which this ratio has a tendency to decrease in both sols and changes in the range 30–10. The $K_3/K_1 \sim 10$ ratios obtained in this work are in agreement with similar data on another lyotropic nematic phase, the tobacco mosaic virus (TMV)–water system. For this phase, the $K_3/K_1 = 8.8$ value was obtained by studying magnetohydrodynamic domains [23]. Note that, although the TMV–water system does not belong to

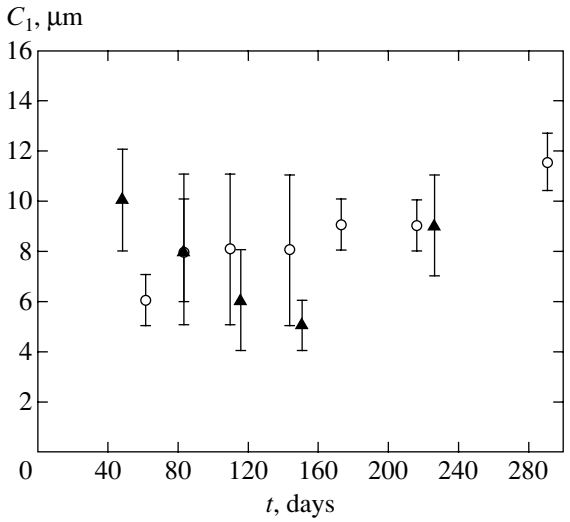

Fig. 7. Experimental dependence of $C_1 = K_1/\sigma$ on the time of aging t of sols nos. 1 (open circles) and 2 (triangles).

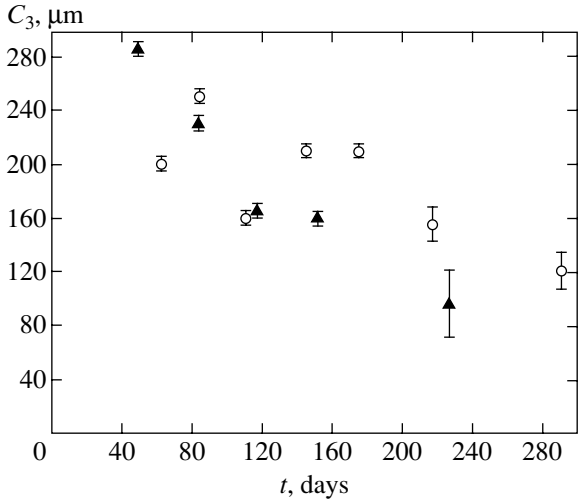

Fig. 8. Experimental dependence of $C_3 = K_3/\sigma$ on the time of aging t of sols nos. 1 (open circles) and 2 (triangles).

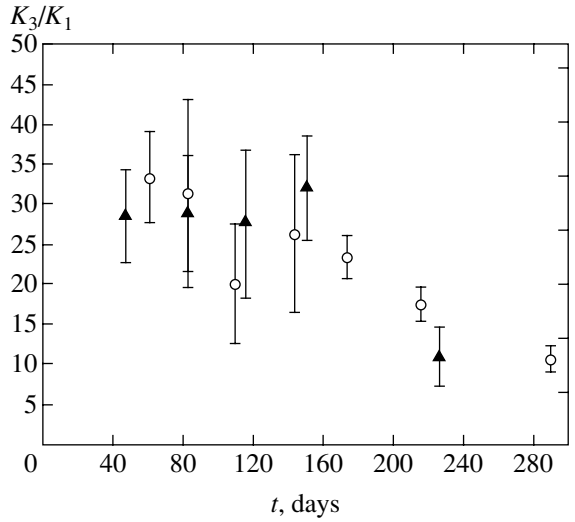

Fig. 9. Experimental dependence of the ratio between elastic constants $K_3/K_1 = C_3/C_1$ on the time of aging t of sols no. 1 (open circles) and 2 (triangles).

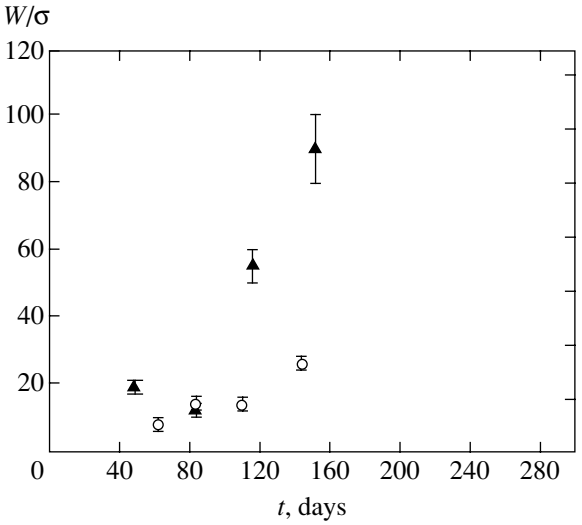

Fig. 10. Experimental dependence of the W/σ ratio on the time of aging t of sols no. 1 (open circles) and 2 (triangles).

inorganic lyotropic liquid crystals, it also contains a tactoid phase [24]. In addition, the $K_3/K_1 \sim 10$ ratio was also measured in the nematic phase N_1 of a lyotropic liquid crystal in the tetrapalladium organyl–pentadecane system [25]. These measurements were based on studying the Freedericksz transition. It appears that the large K_3/K_1 values are related to the large ratios between length L and diameter D of structural elements. For instance, for TMV, we have $L/D \approx 17$ [23], whereas for paraazoxyanisole, which is a typical representative of thermotropic liquid crystals, $L/D \approx 4$ [1].

The dependence of the ratio W/σ on the time of aging is shown in Fig. 10, according to which this ratio tends to increase with the time for both sols and changes in the range 10–100. At times of aging longer than 150 days, all experimental $R(\alpha)$ dependences are well described by function (9), which corresponds to infinitely strong binding. In order to determine W/σ, we must then measure the size of tactoids of a very small volume. Such measurements involve difficulties because of instrumental limitations. The σ estimate obtained earlier and the range of W/σ ratio variations allow W to be estimated at $W \sim 5 \times 10^{-3}$–5×10^{-2} erg/cm^2. These anchoring energy W values closely agree with similar data on thermotropic liquid crystals [21].

4. CONCLUSIONS

In this work, we suggested an explanation for the prolate shape of tactoids observed in lyotropic inorganic liquid crystals. Our approach is based on the equation for the free energy of a tactoid that includes the elastic energy of the nematic phase of the tactoid, the surface energy related to surface tension, and the interaction energy between the director field and the tactoid boundary related to the anchoring energy. The examination of the energy of the tactoid for an extremum allows its equilibrium shape to be determined. This shape depends on the characteristic problem dimensions K_1/σ and K_3/σ, the dimensionless parameter W/σ, and the volume V of the tactoid. It has been shown that large-sized tactoids, for which the strong binding condition is satisfied, are prolate because of the competition between the elastic and surface energies. The stretching of small-sized tactoids as $V \longrightarrow 0$ is limited by the competition between the surface tension and anchoring energy. The conclusion can be drawn that small surface tension and large (compared with σ) W values are required to observe the tactoid phase experimentally.

A comparison of the theoretical and experimental $R(\alpha)$ dependences obtained for a typical lyotropic inorganic liquid crystal in the vanadium pentoxide–water system allowed us to determine the K_1/σ, K_3/σ, K_3/K_1, and W/σ values as functions of the time of system aging. It was found that the K_3/K_1 ratio varied from 30 to 10, which was in agreement with similar data obtained for some lyotropic liquid crystals. The W/σ ratio varied from 10 to 100. The estimates $W \sim 5 \times 10^{-3}$–5×10^{-2} erg/cm^2 and $\sigma \sim 5 \times 10^{-4}$ erg/cm^2 were obtained. The W values closely agree with similar data obtained for thermotropic liquid crystals. The σ value is, however, exceedingly small. For this reason, nematic droplets, or tactoids, become prolate. Usually, the situation is reverse with thermotropic liquid crystals, for which $\sigma \sim 10$ erg/cm^2 and $W \sim 10^{-2}$ erg/cm^2 [21], and we do not observe prolate droplets in such systems. It appears that the inequality $W > \sigma$ obtained in this work is a manifestation of one of the special features of inorganic liquid crystals.

APPENDIX I

The equation for the nematic phase elastic energy density F_{el} has the form [1]

$$F_{el} = \frac{K_1}{2}(\mathrm{div}\mathbf{n})^2 + \frac{K_2}{2}(\mathbf{n}\cdot\mathrm{rot}\mathbf{n})^2 + \frac{K_3}{2}(\mathbf{n}\times\mathrm{rot}\mathbf{n})^2, \quad \text{(I.1)}$$

where K_1, K_2, and K_3 are the elastic constants of splay, twist, and bend deformation, respectively, and $\mathbf{n}$ is the director. The unit vector $\mathbf{n}$ coincides with the basis vector $\mathbf{e}_{\eta_1}$ of the bispherical coordinate system ξ_1, η_1, φ_1. In this system of coordinates, $\mathbf{n} = (0, 1, 0)$. The bispherical coordinates are related to the Cartesian coordinates x, y, z by the equations [20]

$$x = \frac{a_1\sin\xi_1\cos\varphi_1}{\cosh\eta_1 - \cos\xi_1},$$

$$y = \frac{a_1\sin\xi_1\sin\varphi_1}{\cosh\eta_1 - \cos\xi_1},$$

$$z = \frac{a_1\sinh\eta_1}{\cosh\eta_1 - \cos\xi_1},$$

where $a_1 = R_1\sin\alpha_1$ is the transformation parameter (Fig. 2). In bispherical coordinates, the elastic energy density (I.1) takes the form

$$F_{el} = \frac{2K_1\sinh^2\eta_1}{a_1^2} + \frac{K_3\sin^2\xi_1}{2a_1^2}. \quad \text{(I.2)}$$

The elastic energy Φ_{el} of the nematic phase is obtained by integrating (I.2) over the tactoid volume V. The integration can conveniently be performed in the ξ, η, φ coordinates with the parameter $a = R\sin\alpha$, which describes the shape of the tactoid. The limits of integration then have the simple form $\pi - \alpha \le \xi \le \pi$, $-\infty < \eta < \infty$, $0 \le \varphi \le \pi$, and the Jacobian of the transformation $D(\xi_1, \eta_1, \varphi_1/\xi, \eta, \varphi)$ is written as

$$D = \frac{4\gamma}{(\cosh\eta - \cos\xi)^2 - 2\gamma(\cosh^2\eta + \cos^2\xi - 2) + \gamma^2(\cosh\eta + \cos\xi)^2},$$

where $\gamma = (\tan(\alpha_1/2)/\tan(\alpha/2))^2$. The elastic energy Φ_{el} eventually takes the form

$$\Phi_{el} = \frac{a\gamma}{2}\int_0^{2\pi} d\varphi \int_{-\infty}^{\infty} d\eta \times \int_{\pi-\alpha}^{\pi} \frac{4K_1\sinh^2\eta\sin\xi + K_3\sin^3\xi}{(\cosh\eta - \cos\xi)^3} D\, d\xi. \quad \text{(I.3)}$$

A comparison of (I.3) with (3) yields

$$\Phi_{el}^{(1)}(\alpha, \gamma) = 4\pi\gamma\sin\alpha \times \int_{-\infty}^{\infty} d\eta \int_{\pi-\alpha}^{\pi} \frac{\sinh^2\eta\sin\xi}{(\cosh\eta - \cos\xi)^3} D\, d\xi, \quad \text{(I.4)}$$

$$\Phi_{el}^{(3)}(\alpha, \gamma) = \pi\gamma\sin\alpha \times \int_{-\infty}^{\infty} d\eta \int_{\pi-\alpha}^{\pi} \frac{\sin^3\xi}{(\cosh\eta - \cos\xi)^3} D\, d\xi. \quad \text{(I.5)}$$

If $\gamma = 1$, integrals (I.4) and (I.5) are calculated analytically to obtain

$$\Phi_{el}^{(1)}(\alpha, 1) = 4\pi(\sin\alpha - \alpha\cos\alpha),$$

$$\Phi_{el}^{(3)}(\alpha, 1) = \pi(3\sin\alpha - 3\alpha\cos\alpha - \alpha^2\sin\alpha).$$

APPENDIX II

The interaction energy Φ_W between the director field and the surface is calculated by (5). The integration in (5) is performed over the surface of the tactoid. The Φ_W energy can conveniently be calculated in bispherical coordinates. The dS surface element of the surface of the tactoid can then be written in the form

$$dS = \frac{a^2\sin\alpha}{(\cosh\eta + \cos\alpha)^2} d\eta\, d\varphi. \quad \text{(II.1)}$$

Here, it is taken into account that $\xi = \pi - \alpha$ at the boundary of the tactoid. The integration limits are $-\infty < \eta < \infty$, $0 \le \varphi \le 2\pi$. The $\sin^2\theta$ value is calculated by the formula

$$\sin^2\theta = 1 - (\mathbf{e}_{\eta_1}\cdot\mathbf{e}_\eta)^2, \quad \text{(II.2)}$$

where $\mathbf{e}_{\eta_1}$ and $\mathbf{e}_\eta$ are the normalized basis vectors of the bispherical coordinates ξ_1, η_1, φ_1 and ξ, η, φ, respectively. The Cartesian components of $\mathbf{e}_\eta$ are

$$e_{\eta x} = -\frac{\sin\xi\sinh\eta\cos\varphi}{\cosh\eta - \cos\xi},$$

$$e_{\eta y} = -\frac{\sin\xi\sinh\eta\sin\varphi}{\cosh\eta - \cos\xi}, \quad \text{(II.3)}$$

$$e_{\eta z} = \frac{1 - \cosh\eta\cos\xi}{\cosh\eta - \cos\xi}.$$

The Cartesian components of $\mathbf{e}_{\eta_1}$ are written similarly with the replacements $\xi \longrightarrow \xi_1$, $\eta \longrightarrow \eta_1$, and $\varphi \longrightarrow \varphi_1$. The use of the Cartesian components of $\mathbf{e}_{\eta_1}$ and $\mathbf{e}_\eta$ allows the scalar product $(\mathbf{e}_{\eta_1} \cdot \mathbf{e}_\eta)$ to be calculated. The result depends on ξ_1, η_1, φ_1 and ξ, η, φ. Writing ξ_1, η_1, φ_1 through ξ, η, φ and taking into account that $\xi = \pi - \alpha$ at the tactoid boundaries, we obtain

$$\sin^2\theta = \frac{2(\gamma-1)^2\sin^2\alpha\sinh^2\eta}{2(\gamma+1)^2 - 4(\gamma^2-1)\cos\alpha\cosh\eta + (\gamma-1)^2(\cos 2\alpha + \cosh 2\eta)}. \quad \text{(II.4)}$$

Taking these results into account, we can write Φ_W in the form

$$\Phi_W = \frac{WR^2\sin^3\alpha}{2}\int_0^{2\pi} d\varphi \int_{-\infty}^{\infty} \frac{\sin^2\theta}{(\cosh\eta + \cos\alpha)^2} d\eta, \quad \text{(II.5)}$$

where $\sin^2\theta$ is given by (II.4). A comparison of (II.5) with (6) yields

$$\Phi_W(\alpha, \gamma) = \pi\sin^3\alpha \int_{-\infty}^{\infty} \frac{\sin^2\theta}{(\cosh\eta + \cos\alpha)^2} d\eta. \quad \text{(II.6)}$$

If $\gamma = 0$, integral (II.6) is calculated analytically to obtain

$$\Phi_W(\alpha, 0) = \pi\left(\sin\alpha - \alpha\cos\alpha - \frac{1}{3}\sin^3\alpha\right).$$

REFERENCES

1. P. de Gennes, *The Physics of Liquid Crystals* (Clarendon Press, Oxford, 1974; Mir, Moscow, 1977).
2. L. M. Blinov, *Electro-Optical and Magneto-Optical Properties of Liquid Crystals* (Nauka, Moscow, 1978; Wiley, New York, 1983).
3. S. A. Pikin, *Structural Transformation in Liquid Crystals* (Nauka, Moscow, 1981).
4. A. A. Vedenov and E. B. Levchenko, Usp. Fiz. Nauk **141**, 3 (1983) [Sov. Phys. Usp. **26**, 747 (1983)].
5. A. S. Sonin, Usp. Fiz. Nauk **153**, 273 (1987) [Sov. Phys. Usp. **30**, 875 (1987)].
6. A. S. Vasilevskaya, É. V. Generalova, and A. S. Sonin, Usp. Khim. **58**, 1575 (1989).
7. A. S. Sonin, Kolloidn. Zh. **60**, 149 (1998).
8. A. S. Sonin, J. Mater. Chem. **8**, 2557 (1998).
9. P. Davidson, C. Bourgaux, L. Schoutteten, *et al.*, J. Phys. II **5**, 1577 (1995).
10. J.-C. P. Gabriel and P. Davidson, Adv. Mater. **12**, 9 (2000).
11. A. V. Kaznacheev, A. Yu. Kovalevskiĭ, I. A. Ronova, *et al.*, Kolloidn. Zh. **62**, 606 (2000).
12. É. V. Generalova, A. V. Kaznacheev, and A. S. Sonin, Kristallografiya **46**, 127 (2001) [Crystallogr. Rep. **46**, 111 (2001)].
13. A. V. Kaznacheev, M. M. Bogdanov, and S. A. Taraskin, Zh. Éksp. Teor. Fiz. **122**, 68 (2002) [JETP **95**, 57 (2002)].
14. H. Zocher and K. Jacobsohn, Kolloid Beih. **28**, 167 (1929).
15. A. Szegvari, Z. Phys. Chem. (Leipzig) **112**, 295 (1924).
16. H. Zocher, Kolloid Z. **139**, 81 (1954).
17. W. Heller, in *Polymer Colloids II*, Ed. by R. M. Fitch (Plenum, New York, 1980), p. 153.
18. M. V. Kurik and O. D. Lavrentovich, Usp. Fiz. Nauk **154**, 381 (1988) [Sov. Phys. Usp. **31**, 196 (1988)].
19. W. H. de Jen, *Physical Properties of Liquid Crystalline Materials* (Gordon and Breach, New York, 1980; Mir, Moscow, 1982).
20. G. Arfken, *Mathematical Methods for Physicists*, 2nd ed. (Academic, New York, 1970; Atomizdat, Moscow, 1970).
21. L. M. Blinov, E. I. Kats, and A. A. Sonin, Usp. Fiz. Nauk **152**, 449 (1987) [Sov. Phys. Usp. **30**, 604 (1987)].
22. W. Biltz, Ber. Dtsch. Chem. Ges. **37**, 109 (1904).
23. A. J. Hurd, S. Fraden, F. Lonberg, *et al.*, J. Phys. (Paris) **46**, 905 (1985).
24. J. Bernal and J. Fankuchen, Nature **139**, 923 (1937).
25. A. V. Kaznacheev, K. Praefcke, A. S. Sonin, *et al.*, Kolloidn. Zh. **64**, 468 (2002).

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1168–1185.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1308–1328.

FLUIDS

Bubble Motion in Inclined Pipes

N. A. Inogamov[a] and A. M. Oparin[b]

[a]*Landau Institute for Theoretical Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia*
[b]*Institute for Computer-Aided Design, Russian Academy of Sciences, Vtoraya Brestskaya ul. 19/18, Moscow, 123056 Russia*
e-mail: a.oparin@icad.org.ru

Received June 11, 2002

Abstract—Highly nonlinear free-surface flows in vertical, inclined, and horizontal pipes are analyzed. The problem of bubble motion in a vertical pipe is closely related to the Rayleigh–Taylor instability problem. Inclined pipe flows are intensively studied as related to gas and oil transportation. A new theory of motion of large bubbles in pipes is developed. As distinct from previous approaches, which relied on semiempirical methods or numerical fitting, analytical methods of potential theory and complex analysis are used. A careful comparison of 2D and 3D solutions is presented. It is shown that a higher dimensionality may not correspond to a higher bubble velocity. For the first time, free-surface flows in inclined pipes are analyzed by means of direct numerical simulation, which makes it possible to develop a new approach to the Rayleigh–Taylor instability problem (bubbles with wedge- and cone-shaped noses). © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The problem of two-phase pipe flow frequently arises in applications of physical fluid dynamics. One common type of flow regimes is slug flow, in which large bubbles moving through the pipe are separated by "slugs" with low gas content [1]. Flows of this type are very complicated. No satisfactory mechanistic model of slug flow has been proposed to this day. The drift velocity of the lighter phase relative to the mean flow is evaluated in terms of the velocity obtained by solving the problem of the rise of a solitary elongated bubble driven by buoyancy forces in a stagnant fluid. This first step toward understanding the physics of slug flow regimes and related problems were discussed extensively in the literature [1–15].

In this paper, we present both theoretical and numerical analyses of the rise of elongated bubbles in a stagnant liquid. Our studies are restricted to high Reynolds and Weber numbers (low-viscosity flow and pipe diameter much greater than the capillary length scale). Bubble rise has been the subject of intensive studies in the theory of Rayleigh–Taylor instability (RTI) as applied in astrophysics and high energy-density physics (see reviews in [16–18]). Bubbles of this kind develop in the course of nonlinear RTI development.

Comparing studies of two-phase pipe flows and RTI, one can notice a regrettable lack of communication between specialists in hydraulics and RTI, whereas the phenomena in question are analogous (shear turbulence, gravity waves and Richardson number, baroclinic vorticity generation, etc.).

In this study, an approximate analytical solution to the problem of bubble motion in a 2D inclined pipe is found by invoking functions of a complex variable and a hodograph method. We obtain a simple analytical dependence of the velocity U of bubble motion in a pipe of diameter D on the inclination angle α between the pipe and the horizontal axis. Previous analyses relied on empirical correlations derived from experiment [3, 6–9], or the model of elliptical bubble [2], or the empirical formula [3]

$$U(\alpha) = U_h \cos\alpha + U_v \sin\alpha,$$

where U_h and U_v are the velocities of bubble motion in horizontal and vertical pipes respectively. In the case of a circular cylindrical pipe, $U_h = 0.54\sqrt{gD}$ [19] and $U_v = 0.35\sqrt{gD}$ [16, 20–22].

In the case of an elliptical bubble, the free boundary makes a right angle with the wall at the apex of the bubble. However, our results show that the boundary and the wall make a wedge with a different angle in a small neighborhood of the apex. In the 2D case, the angle θ_c between the wall and the boundary of the liquid is 120°. In a 3D inclined cylindrical pipe, the tangent plane to the wall makes a right angle with the vertical centerplane. Thus, the three-dimensional flow pattern is locally similar to the two-dimensional one in a small neighborhood of the apex. However, the value of θ_c is slightly greater than 120° in the 3D case. The vertical centerplane is defined as the plane spanned by the pipe axis and the gravity vector **g**. The projection of the bubble surface onto the vertical centerplane is an ellipse in [2]. Accordingly, the generators of the bubble surface are perpendicular to the plane.

1063-7761/03/9706-1168$24.00

We develop a model in which the actual geometry of the boundary (wedge, not ellipse) is taken into account. The boundary geometry calculated by using this model is compared with that obtained by direct simulation of the flow.

One interesting trend in the behavior of $U(\alpha)$ is its increase with the deviation of the pipe from the vertical, which disagrees with an expected decrease. Since the liquid bounded by the free surface is "falling" inside the pipe, the "speed" of its "fall" would seem to decrease with the component of **g** along the pipe. This strange behavior of the velocity as a function of the angle was noted by all specialists who studied the problem [2, 3, 6–9].

The model developed here provides a simple quantitative explanation of the increase in bubble velocity with deviation of the pipe from the vertical. It is shown that the velocity of the bubble rise along the wall depends on the projection of **g** onto the tangent plane to the boundary at the apex, while the boundary makes an angle $\theta_c > 90°$ with the wall (as noted above).

First of all, this implies that the function $\theta_c > 90°$ has a maximum. The existence of a maximum was noted in many experimental studies [2, 3, 6–9]. Second, the angle α_{max} corresponding to the velocity maximum can be evaluated. Note also that qualitative predictions of the influence of the decrease in gravitational potential along a curved boundary on the bubble velocity can be found in an earlier study [7].

We also present some new results that are of interest for the theory of RTI. A wedge-shaped bubble with $\theta_c = 120°$ exists in the entire range of α. Consequently, the corresponding solution remains valid as $\alpha \longrightarrow 90°$ (vertical pipe). Thus, the 2D RTI problem has two steady solutions corresponding to blunted and wedge-shaped bubbles.

The velocity and shape of the blunted bubble ($\theta_c = 90°$) are well known [16–18]:

$$U_{90} = (0.33–0.34)\sqrt{gD}.$$

The radius of curvature at its apex is

$$R = (0.80 \pm 0.06)D,$$

where D is the width of the strip in which a symmetrical half-bubble is considered. A good estimate is

$$\bar{U}_{90} = U_{90}/\sqrt{gD} = (3\pi)^{-1/2}$$

[22] (see also [16–18, 23–25]). In this study, theoretical estimates for the velocity and shape of a wedge-shaped bubble consistent with experiment are obtained for the first time. The hodograph method predicts

$$\bar{U}_{120} = (2\pi)^{-1/2} \approx 0.40;$$

the force balance method,

$$\bar{U}_{120} = 0.42.$$

A direct numerical simulation (DNS) yields

$$\bar{U}_{120} = 0.42 \pm 0.02.$$

The estimate $\bar{U}_{120} = 0.51$ obtained previously [26] exceeds the value obtained in our numerical experiment by 23%. We estimate the difference between the point solutions, U_{120}/U_{90}, as 22%. The corresponding bubble shapes are compared below.

One key problem in the theory of RTI is which of the two steady flow regimes develops under smooth initial conditions (which steady state is the attractor of trajectories). In [17, 18, 25], it was hypothesized that the trajectories of the dynamical system modeled by a Cauchy problem approach the steady 120° regime. This conjecture was based on results obtained both analytically [25] and numerically by analyzing the behavior of the dynamical system that approximates instability development from a slightly perturbed hydrostatic equilibrium [17, 18, 25, 27] by the method of asymptotic collocations (MAC). Analysis of the system makes it possible to simulate the instability development to degrees of nonlinearity much higher than those attained by integrating time-dependent integral equations by means of a Fourier transform technique [28, 29] or by using weakly nonlinear series expansions [30–32].

It is also important that the same system can be used to describe flow evolution near the steady 90° regime [23] by taking an initial point lying sufficiently close to this regime. The steady 120° regime cannot be reached by starting from near-hydrostatic conditions by means of MAC, because the corresponding trajectory is "blocked" by a singularity. The MAC approximation "works" near the steady 90° regime. The explanation may not lie in the fact that the method fails for large amplitudes (as do weakly nonlinear expansions or integral equations). This may imply that the approximated system approaches the steady 120° regime (as suggested by the aforementioned analytical solution). However, the steady 120° regime cannot be reached by using MAC, because approximation by means of MAC "incorporates" the structure of the stagnation point at the bubble apex (the first term in the expansion of the potential is quadratic).

The steady 90° regime is a very interesting one. There exists a one-parameter ($1d$) family of fictitious or formal steady-state solutions containing a point ($0d$) solution [17, 18, 23]. Generally, the steady 90° regime is interpreted as the $0d$ point. The $1d$ steady-state solutions (except for the $0d$ point) are fictitious in that they cannot be approached in the course of time [17, 18, 23] (even starting from an arbitrarily small neighborhood of any $1d$ steady-state solution).

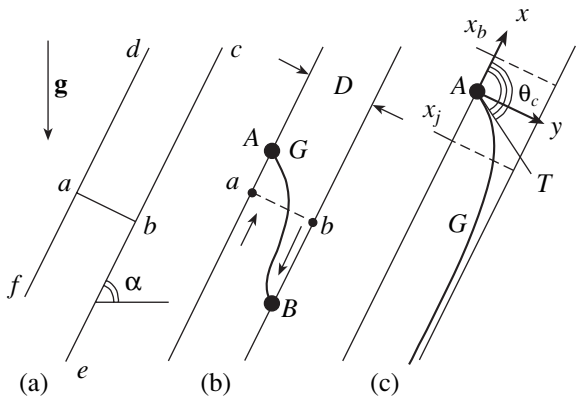

Fig. 1. Emptying of a closed pipe. (a) The initial configuration ($t = 0$): heavy liquid is held in a gravity field (with acceleration **g**) by pipe walls *bc* and *ad* and diaphragm *ab*. (b) The intermediate stage: $t \sim \sqrt{D/g}$ (*D* is the pipe width. (c) Steady state: $t \gg \sqrt{D/g}$. In the plane flow, tangent *AT* to boundary *G* at the apex *A* makes the angle $\theta_c = 120°$ with the wall.

The DNS results presented in this paper show that the RTI development initiated by an infinitely differentiable near-hydrostatic disturbance can evolve into a steady 120° regime.

Consider the Rayleigh–Taylor instability in a 3D geometry. As in the 2D case, we consider the rise of Rayleigh–Taylor bubbles, i.e., $\alpha = 90°$ (we tentatively set aside analysis of the 3D flow in an inclined pipe). There exists a 1*d* family of steady bubbles with rounded noses ($\theta_c = 90°$) containing an exceptional 0*d* point.

In this study, a 3D analog of the wedge-shaped bubble is found for the first time and shown to have a cone-shaped nose. We calculate the cone angle $\pi - \theta_c$ ($\theta_c = 114.799° \approx 115°$) and the bubble-rise velocity. The difference between the axially symmetric steady states, U_{115}/U_{90}, is 10% (less than in the 2D geometry): $U_{115} \approx 0.54\sqrt{gR}$ (by the force balance method) and $U_{90} \approx (0.48–0.50)\sqrt{gR}$ [1–8, 16, 20, 21], where *R* is the radius of a circular cylindrical pipe.

However, one may consider the flow not only in a circular pipe, but also in pipes with hexagonal, square, or triangular cross sections. Each of these cross sections can be combined into a tessellated plane, i.e., a plane lattice. The velocities and shapes of the blunted bubbles (with $\theta_c = 90°$) that make up such lattices were calculated in [23]. Steady bubbles with cone-shaped noses can exist not only in circular pipes, but also in pipes of arbitrary cross section. To evaluate the rise velocity for a lattice of cone-shaped bubbles, one may neglect the dependence of the gap width U_{115}/U_{90} on the lattice type in the first approximation, using the values of U_{90} obtained in [23].

The paper is organized as follows. In Section 2, the flow geometry is described. In Sections 3 and 4, we present the hodograph method used in the cases of blunted and wedge-shaped bubbles, respectively. Note that a simple analytical expression for $U(\alpha)$ is obtained in Section 4 for the velocity of a wedge-shaped bubble in an inclined pipe flow. The momentum of a wedge-shaped bubble is calculated in Sections 5 and 6 for 2D and 3D geometries, respectively. In Sections 7 and 8, we analyze a 3D flow in the gravity field with a free surface having a conical singularity. In Sections 9–11, we compare the theory with numerical and physical experiments.

2. WEDGE GEOMETRY

The geometry of the problem is depicted in Fig. 1. The 2D pipe *fecd* contains a diaphragm *ab* (see Fig. 1a). The region *abcd* is occupied by an incompressible inviscid liquid. The vertical direction defined by the gravity force is indicated by the arrow **g**. At $t = 0$, the diaphragm is removed. The interface *G* separates the liquid from a gas of negligible density (see Figs. 1b and 1c). The liquid moves downwards, and a "tongue" or a jet develops. The flow must satisfy certain conditions at infinity (as $x \longrightarrow +\infty$). The far end of the pipe *abcd* is plugged by a wall *cd*, at which the liquid is at rest. The liquid cannot slip freely along the walls *ec* and *fd*, because the gas pressure pushes it toward the plug. This situation is referred to as the emptying of a closed pipe. Consider the evolution of the interface *G*. The liquid penetrates the gas in the form of a jet *B*, while the gas penetrates the liquid in the form of a bubble with its apex at point *A* (see Figs. 1b and 1c). A gas–liquid two-phase flow develops as a result. The velocities of the phases in the laboratory frame (where the walls are at rest) are shown by arrows (see Fig. 1b). Vorticity (a jump in the velocity component tangential to *G*) is concentrated on the interface *G*. Outside *G*, both gas and liquid flows are potential.

Asymptotically (as $t \longrightarrow \infty$), the flow near the apex *A* tends to a steady state in which the shape of the interface *G* remains invariant in a frame tied to point *A*. We restrict our analysis to the important case of "ideal" walls, low-viscosity flow (high Reynolds number), and a wide pipe (high Weber number, i.e., low surface tension). In the 2D geometry, the steady free surface *G* makes an obtuse angle $\theta_c = 120°$ with the *x* axis (see Fig. 1c). The intersection angle θ_c is independent of the pipe inclination angle α (see Fig. 1a). The angle between the downward vertical direction and the tangent line *AT* is $\alpha - 30°$ (see Fig. 1c).

The rise of bubbles with rounded noses and $\theta_c = 90°$ (see Figs. 2a and 2b) in vertical pipes ($\alpha = 90°$) was analyzed in numerous studies (see reviews in [16–18]). The development of RTI under monochromatic initial conditions in an unbounded liquid results in the formation of a periodic array of bubbles. The symmetric flow

pattern spanning a period λ of the array is illustrated by Fig. 2a; one-half of it, by Fig. 2b. Wedge-shaped bubbles with $\theta_c = 120°$ (Fig. 1c) exist at any inclination angle α. By continuity, this entails the existence of such a solution for the vertical pipe flow (Fig. 2c).

Thus, when $\alpha = 90°$, there exist two solutions to the time-independent boundary value problem: with $\theta_c = 90°$ (Fig. 2b) and $\theta_c = 120°$ (Fig. 2c). It is obvious that the symmetrical half of the solution illustrated by Fig. 2c generates a periodic array of wedge-shaped bubbles in an unbounded liquid (solution to the RTI problem with $\theta_c = 120°$). The interface G can make a right angle with a straight pipe wall only in the vertical case. This explains why only wedge-shaped bubbles with $\theta_c = 120°$ exist when $0 \le \alpha < 90°$.

3. HODOGRAPH METHOD FOR $\theta_c = 90°$

The ideas underlying our quantitative analysis of wedge-shaped bubbles are easily explained by considering the well-studied case of round-nosed bubbles. Generally, this example is analyzed in physical coordinates (see [16–18, 20–25]). In these variables, the potential of the steady flow illustrated by Fig. 2b is

$$f = \varphi + i\psi = -U\left(\sum_{n=1}^{N}\frac{a_n e^{-nz}}{n} + z\right), \qquad \sum_{n=1}^{N} a_n = 1, \quad z = x + iy. \tag{3.1}$$

The flow velocity is determined by the real part of the potential,

$$\mathbf{v} = (u, v) = \nabla\varphi,$$

and the boundary condition

$$\mathbf{v}|_{x\to+\infty} = (-U, 0), \quad U > 0.$$

The potential is written in (3.1) in a frame tied to the apex at $z = 0$ for $D = \pi$. The interface G is aligned with the streamline $\psi = 0$, since the amplitudes a_n are real numbers.

Figure 3 shows the physical plane and the hodograph plane

$$\zeta = \frac{df}{dz} = u - iv,$$

where u and v are the x and y velocity components, respectively (see Fig. 3a). It should be noted that the apex at $z = 0$ (see Fig. 2b) is the stagnation point, where $\mathbf{v} = 0$ and $\zeta = 0$. The region $U0G\infty U$ occupied by the liquid in Fig. 3a is conformally mapped by $\zeta(z)$ into the region $0G\infty(-U)0$ in the ζ plane (see Fig. 3b). These

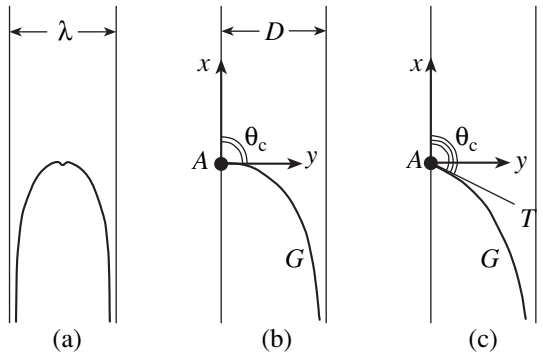

Fig. 2. Special case of the vertical position ($\alpha = 90°$). In contrast to the case of $\alpha \neq 0$, there exist two steady-state solutions: (a, b) round-nosed bubble and (c) wedge-shaped bubble. When $\alpha \neq 0$, only the wedge-shaped bubble exists (Fig. 1c).

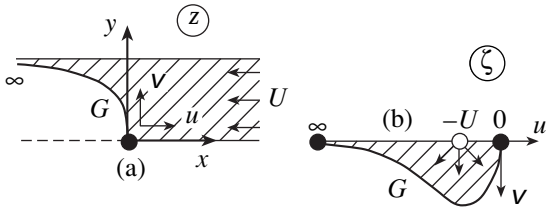

Fig. 3. Hodograph ζ for round-nosed bubbles.

regions are bounded by line segments and the gas–liquid interface G. The points U, 0, and ∞ in the z plane correspond to the points $-U$, 0, and ∞ in the ζ plane. The tangent lines to curve G at the points 0 make right angles with the x and u axes in the z and ζ planes, respectively.

It is well known that even the first term in expansion (3.1) provides a good approximation of the required steady-state solution (with $N = 1$) [17, 18]. The velocity U obtained with $N = 1$ agrees with experimental results within an experimental error of a few percent. Originally, the approximation with $N = 1$ was used in [21] for a cylindrical pipe flow (see also [16–18, 20, 22–25]). The Fourier expansion of expression (3.1) satisfies both the lateral (periodic) boundary conditions and the conditions at infinity (as $x \longrightarrow +\infty$). The velocity U is determined by the conditions on G. These include a kinematic condition dictating that G be a streamline $\psi|_G = \text{const}$ (with const = 0) and a dynamical (isobaric) condition of independence of pressure $p|_G$ on the interface of coordinates (with $p|_G = 0$). Expanding (3.1) in powers of z about $z = 0$, one finds (see [16–18, 20–25])

$$U = \frac{\sqrt{gD}}{\sqrt{3\pi}}. \tag{3.2}$$

Here, we derive (3.2) by the hodograph method with a view to proceed from the case of $\theta_c = 90°$ to the case of $\theta_c = 120°$ (see next section). We write an approximation of the complex potential f in the hodograph plane as

$$f = U\ln(\zeta + U) - \zeta \approx \frac{\zeta^2}{2U} + \frac{\zeta^3}{3U^2}. \quad (3.3)$$

The first term in (3.3) is a source of intensity $2\pi U$ located at $(-U, 0)$ (see Fig. 3b). It corresponds to the flow denoted by U in Fig. 3a and the term $-Uz$ in (3.1). The term $-Uz$ plays the predominant role at infinity in the z plane, while the source in (3.3) dominates near the point $(-U, 0)$ in the ζ plane. Other terms become essential in the neighborhood of the stagnation point: the term containing an exponential in (3.1) and the term $-\zeta$ in (3.3). The corresponding coefficients are such that the stagnation point is located at the origin. It is for this reason that the expansion in powers of ζ in (3.3) begins with a quadratic term.

To determine U in approximation (3.3) up to the first two terms of an expansion in ζ, we differentiate (3.3) and substitute the result into the equation $\zeta = df/dz$ to obtain

$$\frac{dz}{d\zeta} = -\frac{1}{U} + \frac{\zeta}{U^2}.$$

Integrating this expression, we find

$$z(\zeta) = -\frac{\zeta}{U} + \frac{\zeta^2}{2U^2}.$$

Determining $\zeta(z)$ as the inverse of $z(\zeta)$ under the condition $\zeta(0) = 0$, we have

$$\zeta = \frac{df}{dz} = -Uz + \frac{Uz^2}{2}.$$

Hence, we derive expressions for the streamfunction $\psi(x, y)$ and the squared velocity $v^2(x, y) = \zeta\zeta^*$ and use the boundary conditions on G to obtain (3.2).

4. HODOGRAPH METHOD FOR $\theta_c = 120°$

The hodograph method can be used to obtain an approximate analytical expression for the velocity of wedge-shaped bubbles (Fig. 2c). The only difference between the z and ζ planes corresponding to the case illustrated by Fig. 2c and the cases illustrated by Figs. 2b and 3 lies in the angle θ_c between the tangent line to the curve G at the origin and the x or u axis (90° and 120°, respectively). Accordingly, we rely on Fig. 3 in our explanations, assuming that G makes the appropriate angle at the bubble apex. We write

$$f = U\ln\left(1 + \frac{\zeta}{U}\right) - \zeta + \frac{\zeta^2}{2U} \approx \frac{\zeta^3}{3U^2}. \quad (4.1)$$

In (4.1), the term containing a logarithm is a source located at $(-U, 0)$. The remaining terms are adjusted to meet the following requirements. First of all, the stagnation point must lie at $\zeta = 0$. Furthermore, the first term in the expansion must be cubic (i.e., the quadratic term must be compensated). This is dictated by the requirement that the zero streamline $\psi(\mathrm{Re}\,\zeta, \mathrm{Im}\,\zeta) = 0$ must emanate from the origin ($\zeta = 0$) at an angle of $-120°$ to the $\mathrm{Re}\,\zeta$ axis.

To calculate the velocity U in (4.1), we take the first nonvanishing term in the expansion of the logarithm. Using the equation $\zeta = df/dz$, we obtain

$$dz = \frac{\zeta}{U^2}d\zeta.$$

Consequently,

$$z = \frac{\zeta^2}{2U^2}$$

or

$$\zeta = \frac{df}{dz} = -\sqrt{2}U\sqrt{z},$$

where the use of the minus sign is dictated by an analysis of the mapping $\zeta \longrightarrow z$ (see Fig. 3). Integrating the last equation, we find the complex potential in terms of physical variables near the apex:

$$f = -\frac{2\sqrt{2}U}{3}z^{3/2}.$$

Setting $\psi(x, y) = \mathrm{Im}f = 0$, we write an equation for the zero streamline

$$y|_{\psi = 0} = -\sqrt{3}x$$

in the neighborhood of $z = 0$ (kinematic condition). The velocity magnitude squared is

$$\zeta\zeta^* = 2U^2\sqrt{x^2 + y^2}.$$

On the zero streamline, it is

$$(\zeta\zeta^*)|_{\psi = 0} = 2U^2\sqrt{x^2 + [y(x)|_{\psi = 0}]^2} = 4U^2(-x). \quad (4.2)$$

Expressions (4.1) and (4.2) are valid in the general case, remaining invariant for any inclination angle α (Fig. 1). Let us now write out the dynamical condition. The angle α is contained in the expression for U by virtue of this condition. By Bernoulli's theorem, we have

$$\frac{v^2}{2} = \frac{\zeta\zeta^*}{2} = g(y\cos\alpha - x\sin\alpha). \quad (4.3)$$

In Eq. (4.3), the point (x, y) lies on the interface G. Substituting the function

$$y(x)|_{\psi = 0} \approx -\sqrt{3}x$$

for y in (4.3), combining (4.2) with (4.3), and performing some simple algebra, we obtain

$$U(\alpha) = \sqrt{\frac{\cos(\pi/6 - \alpha)}{\pi}}\sqrt{gD}. \qquad (4.4)$$

Expression (4.4) is the desired approximate analytical one characterizing wedge-shaped rising bubbles in inclined pipes at arbitrary α. When $\alpha = 90°$, we have

$$U/\sqrt{gD} = 1/\sqrt{2\pi}$$

(cf. (3.2)). It is clear that the velocity of a wedge-shaped bubble is higher than that of a round-nosed one by a factor of $\sqrt{3/2} \approx 1.225$ (by 23%).

The function $U(\alpha)$ in (4.4) has a shallow maximum at $\alpha = 30°$ (the pipe makes 60° with the vertical):

$$\frac{U(30°)}{U(90°)} = \sqrt{2}\text{TS} \quad \frac{U(30°)}{U(0°)} = \frac{\sqrt{2}}{\sqrt{\sqrt{3}}} \approx 1.075$$

The projection of the velocity of bubble motion along the pipe onto the vertical line is

$$U_{\text{vert}}(\alpha) = U(\alpha)\sin\alpha \sim \sin\alpha\sqrt{\cos(\pi\text{oxy}6 - \alpha)}.$$

The highest velocity $U_{\text{vert}}(\alpha)$ is reached when the deviation from the vertical position is 21.6°:

$$\frac{(U_{\text{vert}})_{\max}}{U_{\text{vert}}(90°)} \approx 1.16.$$

Thus, our analytical study shows that the highest $U(\alpha)$ and $U_{\text{vert}}(\alpha)$ are reached at nonvertical positions. Note that the maximum of the function $U(\alpha)$ lies far from the vertical position. The fact that the maximum of $U(\alpha)$ is reached far from the vertical is well known to experimentalists (see [2, 3, 6–9]). In [7], a qualitative explanation of the nature of this maximum was suggested. The analysis presented above exposes the underlying mechanism. Consider the displacement of a liquid particle to a certain distance along the streamline ψ_0 emanating from the stagnation point in the neighborhood of the stagnation point. The gravitational energy

$$\mathbf{g} \cdot \mathbf{r} = -g\sin\alpha x + g\cos\alpha y$$

decreases the most rapidly when the streamline ψ_0 (not the pipe) is aligned with the vertical. The orientation of the coordinate system (x, y) (along and across the pipe) relative to the horizontal and vertical directions is shown in Fig. 1. In a small neighborhood of the apex, $y = -\sqrt{3}x$ on the curve of ψ_0. Hence,

$$\mathbf{g} \cdot \mathbf{r} \propto \cos(\pi/6 - \alpha).$$

This explains why the maximum velocity is attained when the pipe is not in the vertical position.

The value $\alpha_{\max} = 30°$ is obtained in the first approximation, when only the first term in retained in expansion (4.1) of the potential f. The corresponding maximum of the bubble velocity U is reached when the tangent line AT to the free boundary at the apex is parallel to the vector $\mathbf{g}$. One may expect that the maximum of U corresponding to higher order expansions is reached when a certain secant of the free surface making a small angle with AT is parallel to $\mathbf{g}$. Accordingly, the angle $\alpha_{\max}$ must be slightly greater than 30°.

5. BUBBLE MOMENTUM: 2D GEOMETRY

An alternative to the hodograph method described above relies on conservation laws. The mass, momentum, and energy conservation laws can be used to determine the function $U(\alpha)$ approximately. An important advantage over the hodograph method is the possibility of extending the analysis to 3D flows. It can also be shown that exact values of $U(0)$ and volume fraction of the liquid can be obtained in the horizontal case ($\alpha = 0$) in both 2D and 3D geometries. The corresponding calculations were presented in [19].

First, consider the 2D flow between the lines $x = x_j$ and $x = x_b$ (see Fig. 1c). The coordinate system (x, y) is tied to the bubble apex (Figs. 1–3). The free surface makes an angle of 120° with the wall. It can be shown that energy conservation is equivalent to Bernoulli's theorem. The key role is played by the invariance of the steady-flow momentum, which entails the equation

$$-\bar{U}^2\left(1 - \frac{1}{1 - N_j}\right) + \frac{C_\alpha}{2}(1 - N_j)^2 - \Pi_b - \frac{C_\alpha}{2} - S_\alpha X_b - S_\alpha \int_{-L_j}^{0} [1 - N(X)]dX = 0. \qquad (5.1)$$

Hereinafter, we use the dimensionless notation

$$\bar{U} = \frac{U}{\sqrt{gD}}, \quad N(X) = \frac{\eta(x)}{D}, \quad N_j = \frac{\eta_j}{D},$$
$$\eta_j = \eta(x_j), \quad x_j = -l_y, \quad \Pi_b = \frac{p_b}{\rho g D}, \qquad (5.2)$$
$$C_\alpha = \cos\alpha, \quad S_\alpha = \sin\alpha, \quad X = \frac{x}{D},$$
$$X_j = \frac{x_j}{D} = -L_j, \quad X_b = \frac{x_b}{D}.$$

In (5.2), p_b denotes the pressure at the uppermost point of the cross section x_b (where $x = x_b$ and $y = 0$), the curve $y = \eta(x)$ is the free boundary, and D is the pipe width.

The first term in Eq. (5.1) represents convective momentum transfer across the cross sections at x_b and x_j. The point x_b is located far from the bubble, and the flow across the corresponding cross section is parallel. The flow across the cross section at x_j can approximately be treated as parallel as well. The second term in (5.1) is the positive momentum due to pressure forces at the cross section x_j. The next two terms express the momentum due to pressure forces at the cross section x_b. The last two terms represent the momentum due to the gravity force (bulk acceleration). The former of these corresponds to the momentum of the rectangle bounded by the cross sections at $x = 0$ and $x = x_b$. The term containing the integral corresponds to the momentum due to the weight component parallel to the pipe in the domain bounded by the cross sections at $x = x_j$ and $x = 0$ and the interface G (see Fig. 1c). The interface G does not contribute to the pressure force, since $p|_G = 0$. The left-hand side of Eq. (5.1) is the change in the total momentum per unit time. Since the flow is steady, the change is zero.

We write Bernoulli's integral over the intervals between the points ($x = x_j$, $y = \eta_j$), ($x = 0$, $y = 0$), and ($x = x_b$, $y = 0$) lying on the same streamline:

$$\frac{\bar{U}^2}{2} + \Pi_b + S_\alpha X_b = 0,$$
$$\frac{\bar{U}^2}{2(1-N_j)^2} = S_\alpha L_j + C_\alpha N_j. \tag{5.3}$$

Equation (5.1) involves the unknown function $\eta(x)$ describing the geometry of G (in N_j and the integral). We approximate it as follows:

$$N(X) = \tan\theta_c X - X^2/2r, \quad \theta_c = 120°. \tag{5.4}$$

The approximation reflects the interface geometry at the apex of the bubble. It contains an additional (quadratic) term of an expansion in powers of X. This approximation makes the present model different from the model with an ellipse developed in an enlightening study [2]. In the case of an ellipse, $\theta_c = 90°$. When surface tension is low, the angle θ_c is close to 120° in the 2D geometry. Wedge-shaped bubbles have been observed experimentally in a 3D inclined pipe flow [6]. The large gas bubbles that occur in slug flows in wide pipes also have wedge-shaped noses of this kind.

Substituting (5.4) into Eqs. (5.1)–(5.3), we obtain a set of three equations for the unknown r, $\bar{U}$, and Π_b. Before we briefly describe its solution, let us make a remark about the possibility of refining the proposed approximation (i.e., the use of higher order expansions). The method of asymptotic collocations [17, 18] is advantageous in that it provides a tool for analyzing terms of order higher than one in the expansion. The method can be applied to blunted 2D and 3D bubbles in both steady and unsteady flows. The solution for the pipe flow is expanded in a Fourier series to take into account the decay of velocity disturbances away from the bubble (as $x \longrightarrow +\infty$, see Fig. 2). However, an analogous expansion cannot be obtained in the case of a wedge-shaped bubble because of a singularity at its apex. The force balance method may help to circumvent this difficulty by using momentum conservation instead of the decay condition at $x \longrightarrow +\infty$. In this method, both $\eta(x)$ and $\varphi(x, y)$ are represented by the Taylor series expansions about the origin (the bubble's apex). Equations for the Taylor coefficients are derived from the kinematic and dynamical boundary conditions (in the first approximation, we have only the coefficient containing r in (5.4)).

We restrict the present analysis to first-order terms. We do not need the power series expansion of the velocity potential about the origin. Consider system (5.1), (5.3). We use (5.3) to eliminate the unknown Π_b and $\bar{U}$. As a result, we have the equation

$$(1-N_j^2)(S_\alpha L_j + C_\alpha N_j) - \frac{C_\alpha}{2} N_j(2-N_j) - S_\alpha \int = 0, \tag{5.5}$$

where

$$N_j = -\tan\theta_c L_j - \frac{L_j^2}{2r}, \quad \int = L_j + \tan\theta_c \frac{L_j^2}{2} + \frac{L_j^3}{6r}.$$

We see that (5.5) is a cubic equation in r with an essential parameter α and an auxiliary parameter L_j. The cross section at $x = -l_j = -L_j D$ (see (5.2)) should be sufficiently far from the apex (otherwise, the outgoing flow would be poorly approximated by a parallel flow). However, this cross section should be relatively close to the apex (otherwise, higher order terms of the expansion in powers of X should be retained in (5.4)). Our computations show that the result is weakly affected when L_j varies from 0.5 to 0.8 (greater than half the pipe diameter, but less than the diameter). The required root r is readily found numerically and is used to determine the desired velocity $\bar{U}$ by solving the second equation in (5.3). The results thus obtained (graphs of $U(\alpha)$) are presented below, because they should be described together with numerical results.

In the case of a horizontal pipe ($\alpha = 0$), the vector **g** is parallel to the y axis (perpendicular to the pipe walls). The momentum due to the gravity force vanishes in the bulk of the flow, and the integral in Eq. (5.5) drops out. Equation (5.5) becomes independent of $\eta(x)$ (containing only the final level N_j) and exact. In contrast to cases

with $\alpha > 0$ (see Fig. 1), the jet approaches a constant-velocity flow regime far from the apex when $\alpha = 0$. Its width tends to a constant L_jD. When $\alpha > 0$, the vector **g** has a nonzero component parallel to the pipe. Therefore, the jet flow accelerates, while the jet width decreases. In the ζ plane (Fig. 3b), the outgoing jet is represented by a sink located at the point $(-U/(1 - N_j), 0)$ when $\alpha = 0$. When $\alpha > 0$, the sink lies at infinity. Solving Eq. (5.5) in the case of $\alpha = 0$, we find the exact values ([19], see also [33])

$$N_j = \frac{1}{2}, \quad U(0) = \frac{\sqrt{gD}}{2}. \tag{5.6}$$

The approximate value

$$\bar{U}(0) = 3^{1/4}(2\pi)^{-1/2}$$

predicted by (4.4) is 5% higher than the exact one given by (5.6), whereas the approximate value obtained in [34] ($\bar{U} = 0.43$) is less by 16%. The numerical result obtained by direct simulation, $\bar{U} = 0.49 \pm 0.01$, is in very good agreement with (5.6). This provides solid evidence of the reliability of numerical simulation.

6. BUBBLE MOMENTUM: 3D GEOMETRY

Now, we present the solution of the 3D problem. We consider a circular cylindrical pipe as a case of primary importance for applications. We use the Cartesian coordinate system (x, y, z) in which the x and y axes are set as shown in Fig. 1c, i.e., lying in the vertical midplane spanned by the pipe axis and the vector **g**. The corresponding z axis is horizontal. Let us write out Bernoulli's integrals for the same points as in the 2D case. Now, these points lie on the streamline extending along the crest of the pipe and through the uppermost point on the free surface (the bubble apex). We set the cross sections $x = -l_j$ and x_b perpendicular to the pipe axis as in Fig. 1c. Making use of the mass conservation law, we write Bernoulli's integrals as

$$\bar{U}^2/2 = -\Pi_b - S_\alpha X_b, \tag{6.1}$$

$$\bar{U}^2 = 2\Gamma_j^2(S_\alpha L_j + C_\alpha N_j). \tag{6.2}$$

We use the notation defined in (5.2) (D is the pipe diameter). Note that Eq. (6.1) is identical to the first equation in (5.3).

The 3D equations are qualitatively similar to those corresponding to the 2D case, but are more cumbersome. This is explained by the stereometry of the problem. In (6.2), Γ_j is the volume fraction of the liquid in the cross section $-l_j$. We assume that the free-surface generators are parallel to the (horizontal) z axis. Then, the free surface is defined by a function η depending only on x: $\eta = \eta(x)$. This is a fairly good approximation as long as α is not too close to a right angle. In this approximation, the liquid occupies the segment of the circle with $y = \eta(x_{cut})$ in the cross section $x = x_{cut}$ $(x < 0)$. The segment is defined by the chord

$$y = \eta(x_{cut}) = N_{cut}D.$$

In the cross section $x = -l_j$, the relative area occupied by the liquid is

$$\Gamma_j = 1 - \frac{\gamma_j}{\pi} + \frac{1 - 2N_j}{\pi}\sin\gamma_j, \quad \cos\gamma_j = 1 - 2N_j. \tag{6.3}$$

Expression (6.3) is obviously valid for any cross section x (without the subscript "j"). The chord of G subtends an angle of 2γ with vertex at the pipe axis, where $\gamma = 0$ at cross sections $x \geq 0$ (the liquid occupies the entire pipe cross section), $\gamma = \pi$ as $x \longrightarrow -\infty$ if $\alpha > 0$ (the entire cross section is occupied by the gas), and $\gamma \longrightarrow \gamma_j$ as $x \longrightarrow -\infty$ if $\alpha = 0$. In any particular cross section, the configuration is determined by one of three related geometric parameters: N, γ, or Γ. The function $Y = N(X)$ defines the free surface in dimensionless notation (5.2):

$$N_j = N(X_j) = N(-L_j), \quad L_j = l_j/D.$$

In (6.3), N is either less or greater than 1/2. Accordingly, either $0 < \gamma < \pi/2$ or $\pi/2 < \gamma < \pi$.

Let us balance the forces. Equating the sum of the pressure force and weight to the momentum increment, we obtain

$$-\bar{U}^2 + \frac{\bar{U}^2}{\Gamma_j} + \frac{C_\alpha}{\pi}\phi_j - \Pi_b - \frac{C_\alpha}{2} - S_\alpha X_b - S_\alpha \int_{-L_j}^{0} \Gamma(X)dX = 0, \tag{6.4}$$

$$\phi_j = \frac{\pi}{4}(1 - 2N_j) + \sqrt{N_j - N_j^2}\,\frac{3 - 4N_j + 4N_j^2}{3} + \frac{1 - 2N_j}{2}\arcsin(1 - 2N_j).$$

The terms contained in Eq. (6.4) are analogous to those in Eq. (5.1): convective momentum transfer (under mass conservation), pressure impulse on the cross sections x_j and x_b, and the weight pressure exerted by the two volumes bounded by the cross sections $x = 0$, $x = x_b$ and $x = x_j$ and $x = 0$. The pressure on the free surface is zero: $p|_G = 0$.

The pressure force in the cross section x_b is

$$f_b = 2R\int_0^{2R}(p_b + \rho g G_\alpha)\sqrt{1 - (1 - y/R)^2}\,dy = \pi R^2(p_b + \rho g C_\alpha D/2),$$

where $R = D/2$ is the pipe radius. An analogous force in the cross section x_j is

$$f_i = 2\rho g C_\alpha R \int_{\eta_j}^{2R} (y - \eta_j)\sin\gamma_j dy = 2\rho g C_\alpha R^3 \phi_j,$$

where $D\sin\gamma_j$ is the length of the chord (belonging to the boundary η) that separates the liquid and gas in the cross section x_j and the expression for ϕ_j is given above.

Eliminating the unknown Π_b and $\bar{U}$ from balance equation (6.4) by using (6.1) and (6.2), we obtain

$$\Gamma_j(2-\Gamma_j)(S_\alpha L_j + C_\alpha N_j) + \frac{C_\alpha}{\pi}\phi_j - \frac{C_\alpha}{2} - S_\alpha \int_{-L_j}^{0} \Gamma(X)dX = 0. \quad (6.5)$$

Equation (6.5) is analogous to Eq. (5.5).

To solve Eq. (6.5), one must define the boundary $y = \eta(x)$. We define it by Eq. (5.4), where the angle θ_c is a parameter. In the 2D geometry, this angle is 120°. In [2], a planar ellipse (lying the xy plane) was used as a boundary of the 3D flow (in which case $\theta_c = 90°$). Equation (6.5) contains the following quantities:

$$\Gamma_j = \Gamma(N_j), \quad N_j = N(-L_j), \quad N(X), \quad \Gamma(X) = \Gamma[N(X)].$$

Here, $N(X)$ is defined by (5.4). Note that Eq. (6.5) is to be solved for the unknown r with parameters α, θ_c, and L_j. The integral in (6.5) has to be calculated numerically. This is one difference between Eqs. (6.5) and (5.5). The function $U(\alpha)$ obtained as a result of the solution is discussed below.

When $\alpha = 0$, Eq. (6.5) (derived from conservation laws) is exact, the integral containing the function $N(X)$ vanishes and the result is independent of the boundary geometry. Equation (6.5) for $\alpha = 0$ was derived and solved in [19]. The corresponding values are

$$N_j = 0.43719, \quad \bar{U} = 0.54213, \quad \Gamma_j = 0.57977. \quad (6.6)$$

The liquid reaches well above the pipe midplane (cf. the 2D values in (5.6)). The velocity $\bar{U}$ is higher by 8%. The experimental value $\bar{U} = 0.54$ [6] obtained in cylindrical pipes of diameter greater than the capillary scale is very close to that in (6.6). This means that surface tension was negligible in that experiment.

7. CONE-SHAPED BUBBLES

The formation of a wedge-shaped bubble with an angle of 120° is a common phenomenon, attributed to two factors. First, the two-phase flow involves an interface G separating the phases. As the phases move toward one another, a stagnation point appears on the boundary, where the surface vorticity vanishes. Second, since the gravity field (i.e., the acceleration $\mathbf{g}$) is uniform, the gravitational potential (per unit mass) $\mathbf{g}\cdot\mathbf{r}$ is a linear function of the coordinates (recall Eq. (4.3)). Accordingly, the velocity relative to the stagnation point squared is also a linear function of the coordinates.

In terms of a function of a complex variable, we have

$$f \propto \zeta^n$$

near the stagnation point. Hence,

$$\zeta \propto \zeta^{n-1} d\zeta/dz, \quad dz \propto d\zeta^{n-1},$$
$$\zeta \propto z^{1/(n-1)}, \quad f \propto z^{n/(n-1)}.$$

Since the gravity force is uniform, it holds that

$$\zeta\zeta^* \propto |z|.$$

Therefore, $n = 3$. The imaginary part of the complex potential is

$$\psi \propto \sin\left(\frac{n\theta}{n-1}\right),$$

where θ is the polar angle in a coordinate system tied to the stagnation point. This leads to a three-ray streamline pattern and an angle of 120°.

Note that there exist other three-ray patterns:

I. Trihedral junctions of the "walls" (liquid sheets separating bubbles) arise in random lattices of densely packed bubbles [23]. In [23], a simple explanation was proposed for the nature of this typical singularity (collision of the "heads" generated by two sources to which a third source is added).

II. The junction of three "cracks" is a basic feature of a flame front [35].

III. An important example can be found in cosmology (the large-scale structure of the Universe): the formation of caustics, "pancakes," and surfaces of concentrated matter when the initial velocity field is nonuniform [36, 37] (adhesion dynamics or gravitational clustering). Collision of surfaces leads to trihedral structures with filaments of higher density extending along the lines of their intersection (cf. [35, 23]).

To facilitate an analysis of a 3D geometry, let us write out in terms of a real variable the complex expressions given in Section 4 for a wedge-shaped bubble. Instead of the complex potential f, we use the velocity potential φ. It is governed by the Laplace equation

$$\Delta\varphi = \frac{(r\varphi_r)_r}{r} + \frac{\varphi_{\theta\theta}}{r^2} = 0,$$

where r and θ are the polar coordinates of the stagnation point, and the angle θ is measured from the x axis (see Fig. 3a). Near the origin, we have

$$\varphi = r^{3/2}\Phi(\theta)$$

(in a uniform gravity field). The angular function satisfies the equation

$$\Phi_{\theta\theta} + (3/2)^2\Phi = 0.$$

Hence,

$$\Phi = a\cos(3\theta/2) + b\sin(3\theta/2).$$

The angular velocity is

$$v^{(\theta)} = \sqrt{r}\Phi_\theta.$$

By virtue of the boundary condition on the x axis, it holds that

$$v^\theta(0) = 0,$$

which entails $b = 0$. Therefore,

$$\varphi \propto \cos(3\theta/2). \tag{7.1}$$

This implies that $v^{(\theta)} = 0$ on the ray $\theta_c = 120°$.

Consider a 3D flow in a vertical pipe. Let us show that there exists a 3D analog of the wedge-shaped bubble. In spherical coordinates, the harmonic equation is

$$\Delta\varphi = \frac{(r^2\varphi_r)_r}{r^2} + \frac{(S\varphi_\theta)_\theta}{r^2 S} = 0,$$

where $S \equiv \sin\theta$ (no azimuthal dependence). Consider the neighborhood of the stagnation point. Since the gravity field is uniform, it holds that

$$v \sim \sqrt{r}, \quad \varphi = r^\nu\Phi(\theta), \quad \nu = 3/2.$$

The equation for Φ is

$$\frac{(S\Phi_\theta)_\theta}{S} + \nu(\nu+1)\Phi = 0.$$

This is the Legendre equation of degree ν. Its solution regular on the axis $\theta = 0$ is the Legendre polynomial $P_{3/2}(\cos\theta)$ of degree 3/2 (cf. (7.1)). The function $P_{3/2}$ is regular at $0 \le \theta < \pi$ [38]. The polar axis $\theta = 0$ points toward the heavier fluid along the x axis in Fig. 3a. Let us consider the axially symmetric solution near the bubble apex.

We represent $P_{3/2}(C)$ with $C \equiv \cos\theta$ as a Taylor series expansion on the polar axis $\theta = 0$, where $C = 1$:

$$P_{3/2}(C) = F\left(-\frac{3}{2}, \frac{5}{2}; 1; \frac{1-C}{2}\right),$$

$$F(a, b; 1; \xi) \tag{7.2}$$

$$= 1 + \frac{ab}{(1!)^2}\xi + \frac{a(a+1)b(b+1)}{(2!)^2}\xi^2 + \ldots,$$

where F is the hypergeometric function [38]. Expansion (7.2) is convergent on the circle $|1 - C| < 2$. On the ray $\theta = \pi$ inside the bubble ($C = -1$), expansion (7.2) has a logarithmic singularity:

$$P_{3/2}(C) \longrightarrow -(\pi)^{-1}\ln(1 + C), \tag{7.3}$$

which is due to a line of sources (see below).

The polar velocity component is

$$v^{(\theta)} = \sqrt{r}\Phi_\theta = -\sqrt{r}\frac{dP_{3/2}(C)}{dC}\sin\theta. \tag{7.4}$$

By virtue of the symmetry, it vanishes on the polar axis $\theta = 0$. Moreover, the component $v^{(\theta)}$ given by (7.4) vanishes on the θ_c cone (whose generators make an angle of θ_c with the polar axis). The root $C_c = \cos\theta_c$ of the equation

$$\frac{dP_{3/2}(C)}{dC} = 0$$

is determined numerically:

$$\theta_c \approx 114.799° \approx 114.8°. \tag{7.5}$$

To find (7.5) up to the third or fourth decimal place, we used about 20 terms of the power series in (7.2).

8. VELOCITY OF A BUBBLE WITH A CONE-SHAPED NOSE

Figure 4 shows the flow patterns with wedge-shaped and conical singularities. They include the limit case of a Stokes (gravity) wave (Fig. 4a) and a wedge- or cone-shaped bubble (Fig. 4b). The period of the gravity wave is λ, and its crests make up a chain of crests. There exists a "soliton" solution (a single 120° crest on the entire horizontal axis in the case of a finite depth). The wedge- or cone-shaped bubble is confined in a pipe of diameter λ (see Fig. 4b), with the apex on the pipe axis.

In the wave solution, there is a cut in the complex potential above the crest (see Fig. 4a). It is a vorticity line (jump in the tangential velocity component). There is also a cut inside the wedge-shaped bubble (a line of sources 0∞, see Fig. 4b). The liquid mass flux generated by this line expands the bubble and "neutralizes" the downward mass flux from infinity. The upper endpoint

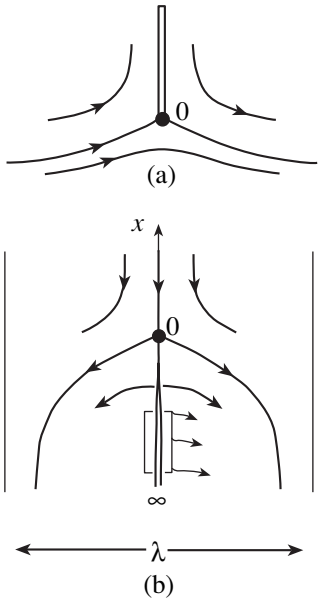

Fig. 4. Wedge and conical singularities on the free surface of the heavier fluid: (a) 2D water wave and (b) 2D or 3D rising bubble.

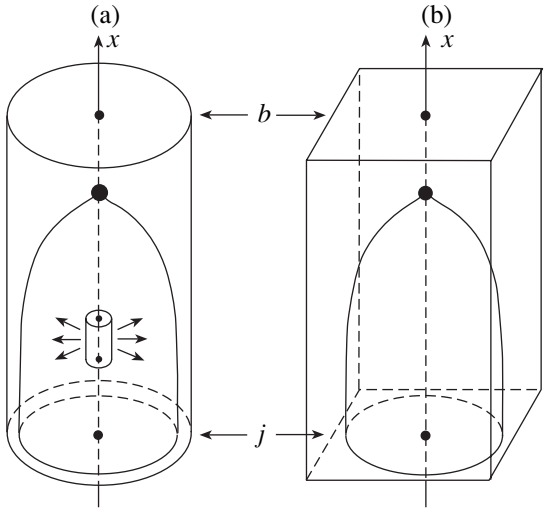

Fig. 5. Cone-shaped bubbles in vertical pipes having different cross sections.

of the cut lies on the free surface. A similar line is confined within the cone-shaped bubble (see Fig. 5).

The analytic continuation of the velocity potential into the bubble has logarithmic singularity (7.3) on the line of sources, which is due to the decrease in the lateral area $2\pi(\delta r)\delta x$ of a cylinder surrounding an infinitesimal segment δx on the x axis (Fig. 5a). The cylinder's radius is δr, and its height is δx. In Figs. 4 and 5, we denote the vertical axis by x by analogy with the inclined pipe flow (Figs. 1–3). Since the lateral area decreases (see Figs. 4a and 4b), the velocity diverges, $v^{(\theta)} \propto 1/\delta r$ (the mass flux generated by the segment δx of the line of sources is invariant) or $v^{(\theta)} \propto 1/\delta\theta$, where $\delta\theta = \theta - \pi$ (cf. (7.3)). In the 2D geometry, the lateral portion of the perimeter of an infinitesimal rectangle $\delta y \times \delta x$ is constant (see Fig. 4b), there is no singularity, and potential (7.1) is not singular on $\theta = \pi$.

Now, let us calculate the rise velocity for cone-shaped bubbles (Fig. 5) by applying the force balance method (see Sections 5 and 6). Restricting our analysis to the case of a circular pipe (Fig. 5a), we consider a flow characterized by global azimuthal symmetry. Denoting the cylindrical coordinates z and r (cylindrical radius) by x and y, respectively, we write the momentum equation

$$W^2 - \frac{W^2}{1 - N_j^2} + \Pi_b + X_b + \int_{-L_j}^{0} [1 - [N(X)]^2]dX = 0. \tag{8.1}$$

It is analogous to Eqs. (5.1) and (6.4). The liquid in the jet j flows through an annular region (Fig. 5a). Equation (8.1) involves the cross-sectional area of the annulus. As in Sections 5 and 6, we denote by b and j cross sections located above and below the bubble apex (see Fig. 5). Here, we use the dimensionless quantities obtained by using the pipe radius $R = D/2$ as a reference length (as $W = U/\sqrt{gR}$ and $L_j - l_j/R$) instead of the quantities defined in (5.2).

Following Sections 5 and 6 (eliminating the unknown W and Π_b by invoking Bernoulli's theorem), we obtain

$$L_j(1 - N_j^4) - \int_{-L_j}^{0} (1 - N^2)dX = 0. \tag{8.2}$$

We use (5.4) to approximate the bubble's boundary $N(X)$ and set θ_c equal to the value given by (7.5). This approximation makes the integral in (8.2) easy to calculate. We obtain a fourth-order equation for the unknown r defined by (5.4). We solve it numerically to find both r and the bubble-rise velocity $U = W/\sqrt{gR}$.

Performing calculations for a cone-shaped bubble in a circular pipe of radius R (Fig. 5a), we obtain $r \approx 0.32R$ and

$$U \approx 0.54\sqrt{gR} = 0.38\sqrt{gD}. \tag{8.3}$$

For a planar wedge-shaped bubble in a strip of width $\lambda = 2R$ (Fig. 4b), $r \approx 0.43R$ and

$$U \approx 0.42\sqrt{gR} = 0.297\sqrt{g\lambda}. \tag{8.4}$$

The results given by (8.3) and (8.4) can be compared with the available values of the velocity of round-nosed bubbles[1] [16–18]. For an axially symmetric bubble in a circular pipe of radius R,

$$U \approx 0.35\sqrt{gD} = 0.495\sqrt{gR}$$

(see [20–22]). In a strip of width $\lambda = 2R$, a round-nosed bubble that is symmetric about its centerline rises with the velocity

$$U \approx 0.24\sqrt{g\lambda} = 0.34\sqrt{gR}.$$

Thus, a singular 3D bubble (with a cone-shaped nose) moves faster than a singular 2D one (with a wedge-shaped nose) by 29% (0.54/0.42 = 1.286), whereas a regular 3D bubble (with a spherical nose) moves faster than a round-nosed 2D bubble (with a circular nose) by 46% (0.35/0.24 = 1.46). Note that the velocity of the singular 3D bubble is higher than that of the round-nosed 3D bubble only by 9% (0.54/0.495 = 1.09). For the 2D bubbles, the corresponding ratio is much greater: 0.42/0.34 = 1.235.

Three-dimensional pipes may have various cross sections (see Fig. 5). For example, even multiply connected (annular) cases were considered in [6]. Pipes that can be packed into space-filling arrays are of interest for analysis of RTI. Such arrays are associated with spatially periodic solutions [23]. Figure 5b shows a pipe of rectangular cross section corresponding to a rectangular bubble lattice.

Of special interest for analysis of RTI are square and hexagonal honeycomb structures (see [23] for explanation). The corresponding elementary cells (pipes) have square and hexagonal cross sections, respectively. In the rectangular case, the cone at the nose is not circular (the flow near the apex is asymmetric). In the case of a regular lattice, the flow is azimuthally symmetric in a small neighborhood of the apex and a circular cone with the angle given by (7.5) can be inscribed in the nose. We choose sides of the square (k_4) and hexagon (k_6) such that the growth exponents characterizing linear RTI development, $\sqrt{gk_{4,6}}$, are equal. The appropriate sides and areas were compared in [23]. Under this choice, the rise velocity for steady round-nosed bubbles is

$$U^{90}_{4,6} = (1.00 \pm 0.02)\sqrt{g/k_{4,6}}$$

for both square and hexagonal solutions [23]. The velocity of a blunted 3D bubble moving along the centerline of a $\lambda \times \lambda$ square pipe is substantially higher (by 68%, 1/0.595 = 1.68) than that of a blunted 2D bubble moving along the centerline of a strip of width λ (Fig. 2a).

[1] An inscribed (2D) circle or (3D) sphere tangent to the boundary at the apex.

The force balance method developed in Sections 5, 6, and 8 can be extended to the case of a pipe with square or hexagonal cross section. The corresponding cumbersome calculations are omitted here. The rise velocity $U^{114.8}_{4,6}$ for cone-shaped bubbles in such pipes can be estimated assuming that the ratio $U^{114.8}_{4,6}/U^{90}_{4,6}$ is approximately equal to that for a circular pipe.

9. COMPARISON WITH NUMERICAL SIMULATION

Now, we compare our theoretical results with those obtained by direct numerical simulation. We are unaware of any previous numerical analysis of slug flow and motion of elongated bubbles in inclined pipes. On the one hand, numerical experiments have revealed many useful facts concerning RTI. On the other hand, the RTI problem (viewed as bubble motion in a vertical pipe) and the problem of bubble motion in an inclined pipe have much in common. Therefore, application of numerical methods is a promising approach.

The simulation was conducted by the method described in [39–41]. We ran a gas-dynamics code at a very low Mach number Ma ~ 10^{-2}. The computations were performed on an $N_y \times N_x$ square grid in a $\Delta y \times \Delta x$ slender rectangle, where N_y is the number of cells spanning the pipe width D ($N_y \sim 10^2$). To eliminate the effects due to the boundary, we considered the case of a large aspect ratio, $\Delta x/\Delta y \sim 10$. At the initial moment, the "gas" (fluid with density ρ_l) occupied 70% of the pipe length.

Front-capturing computations were performed [39–41]. The density ratio was $\mu = \rho_l/\rho_h = 1/20$. At $t = 0$, the boundary ab of the liquid was straight: $x = \eta(y, t = 0) \equiv 0$ (Fig. 1a). The density difference was created by a temperature difference. The initial density was a piecewise constant function with a jump across the boundary ab (Fig. 1a). Since the pressure was high while the Mach number was low, the temperature gradient caused by density stratification was low ($T/|\nabla T| \gg D$). The resulting flow is almost identical to the incompressible, piecewise continuous flow (with densities ρ_l and ρ_h) in a gravity field.

The initial velocity field was defined by the potential

$$\varphi(x, y, t = 0) = -u_0 \cos y e^{-x},$$
$$u = \varphi_x, \quad v = \varphi_y, \quad u_0 = 0.2\text{–}0.5.$$

We set

$$D = \pi, \quad g = 1.$$

The flow evolution was computed from the initial state shown in Fig. 1a to the steady state illustrated by Fig. 1c. The intermediate stage of development from the initial to the final state is particularly important. We performed numerous computations for various values

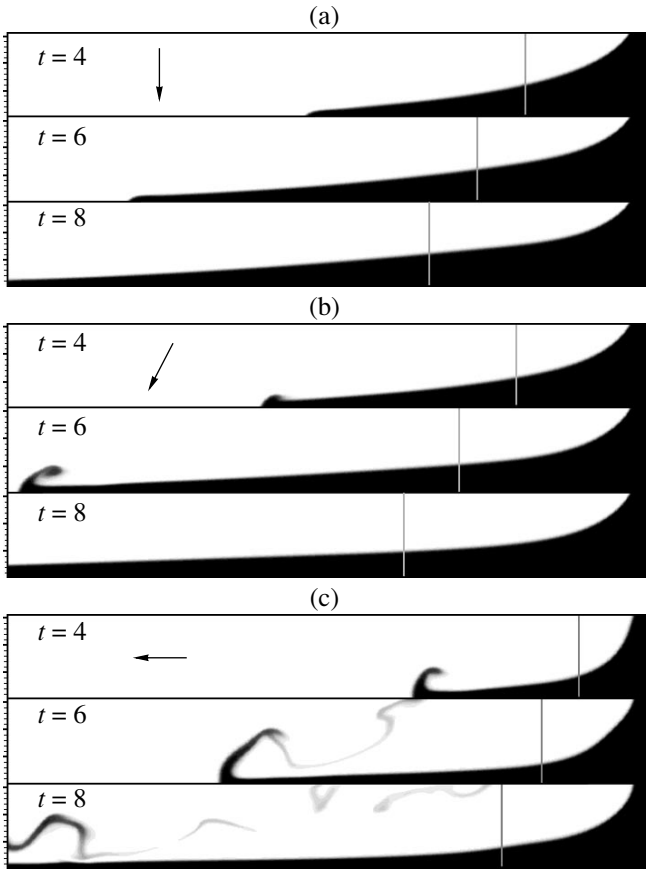

Fig. 6. Transition to a steady bubble in pipes with various inclination angles: (a) $\alpha = 0$; (b) 30°; (c) 90°. The arrow points in the direction of the vector **g**. Vertical line indicates the initial location of the interface. Computations were performed for $D = \pi$ and $g = 1$.

of N_y (grid density), $\Delta x/\Delta y$ (pipe aspect ratio), *Ma* (compressibility), and α (inclination). Figure 6 shows several examples illustrating the evolution toward asymptotic regimes. The numerical experiments confirm the existence of a wedge with an angle of 120° on the free surface.

In the horizontal case, the steady jet has the width

$$h_j = D - \eta_j = D/2$$

(see (5.6)). The steady jet flow is approached slowly, whereas the bubble quickly reaches a steady state. The steady-state jet velocity is $U_j = \sqrt{gD}$ (see (5.6)) in a reference frame tied to the apex. The leading edge j_{end} of the jet remains unsteady. Driven by an additional release of gravitational energy, it moves with a velocity U_{end} higher than U_j. On the interval between the bubble apex and the leading edge j_{end}, the jet velocity increases almost linearly[2] from U_j to U_{end} along the x axis, while the jet width decreases almost linearly (as in a uniformly strained flow) from 1/2 to zero (see Fig. 6). The unsteady behavior extends over the entire jet rather than concentrates near the leading edge j_{end}. This explains the slow relaxation to a steady jet flow. The leading edge must extend very far for the free-surface slope $|\eta_x|$ to become small.

On the contrary, the bubble reaches a steady state relatively quickly. The slow jet relaxation to a steady state does not affect the relaxation of a bubble, because the jet propagates at a "supersonic" velocity (see also [19]): $U_j > c_s$, where $c_s = \sqrt{gh_j} = \sqrt{gD/2}$ is the speed of sound in the shallow-water approximation and $h_j = D - \eta_j$ is the jet width. Outside the intermediate flow region, a solution for the jet flow can be obtained by invoking the shallow-water approximation. The intermediate flow region adjoining the apex is a few diameters D long. Since the free-surface slope $|\eta_x|$ is small outside this region, we can apply the long-wavelength (shallow-water) approximation. A shallow-water flow is equivalent to a one-dimensional (x, t) gas flow. In this sense, the jet j is analogous to the jet exhausting from a jet engine. It can be represented as a centered rarefaction wave with $\gamma = 2$. The shallow-water equations can be extended to describe the case of $\alpha \neq 0$.

Figure 7a compares the computed free-surface shape η with that predicted by an analytical steady-state solution [19] for $\alpha = 0$ obtained by using an approximate conformal mapping. The figure shows that the analytical and numerical results are in very good agreement in the bubble domain, demonstrating the high accuracy of the latter.

Figures 7b and 7c compare the numerical and analytical free surfaces η obtained near the bubble apex for arbitrary inclination angles α. The analytical results were obtained by the method described in Section 5. The figures demonstrate fair agreement between the analytical and numerical results.

10. VELOCITY BEHAVIOR

The analysis presented above concerned the relaxation of a bubble to a steady state in an inclined pipe (Figs. 1 and 6) and compared theoretical results with simulations in terms of the free-surface shape η (Fig. 7). Now, we consider the behavior of velocity in space and time. Of primary interest are the asymptotic values approached as $t \longrightarrow \infty$. Figure 8 shows a map of longitudinal velocity $u(x, y, t)$ at $t = 6$ (with a step of 0.5) in the laboratory frame tied to the walls. The zero-velocity isopleth is labeled 0. The lightest and darkest areas correspond to $u = +7$ and $u = -7$, respectively (recall that $D = \pi$ and $g = 1$).

Figures 6 and 8 demonstrate that the intensity of vortex motion at the leading edge of the jet increases with α. This may be attributed to a decrease in the acceleration component $g_\perp = g\cos\alpha$ pressing the liquid

[2] If $\alpha \neq 0$, then the jet velocity increases as $\sqrt{-x}$.

against the wall. The vortices resemble halves of mushroomlike structures. The gas density ρ_l is 20 times lower than the liquid density ρ_h. Nevertheless, the gas inertia (aerodynamic drag [17, 18]) leads to the formation of mushrooms. This is explained by acceleration of jets to a high velocity (Fig. 8). No mushroom is observed when the pipe is set horizontally (Fig. 6a).

In a steeply inclined pipe (when $\alpha \sim 90°$), the pressing acceleration $g_\perp$ is small. The leading-edge vortex separates from the lower wall ($y = D$) underlying the jet and adheres to the upper wall ($y = 0$) (cf. Figs. 6b and 6c with Fig. 8 for $\alpha = 30°$ and 90°). The adhering vortex makes up a vortex dipole with its mirror image across the upper wall. In accordance with the sense of circulation (clockwise in Figs. 6 and 8), the dipole moves counter to the direction of the jet that gives rise to the vortex (cf. the collimation effect in Rayleigh–Taylor turbulence [41]). The velocity of this motion is so high that the dipole tends to approach the bubble apex (see the selection of images in Fig. 6c). Note that when the angle lies between $\alpha = 0$ and $\alpha = 90°$, both the bubble velocity $U(\alpha)$ and the jet length exceed those corresponding to the vertical pipe position (see Fig. 8).

Figure 9a shows the velocity profile $u(x, y = 0, t = 6)$ in the direction of bubble rise. When $\alpha = 90°$, the velocity near the vortex is very high. As noted above, this is explained by the formation of fast-moving vortex dipoles at the wall.

Consider the bubble dynamics. While the vortex remains far from the bubble apex, we can apply the analysis developed in Sections 3–6. Figure 9b shows portions of the velocity profiles at $t = 6$ (see Fig. 9a) near the bubble apex. The interface locations corresponding to the average density $\rho_{av} = (1/2)(\rho_l + \rho_h)$ and the asymptotic velocity $U(\alpha)$ of wedge-shaped bubbles are shown by vertical and horizontal bars, respectively. These locations were determined as follows. A profile $\rho(x)$ (corresponding to specific y and t) was used to find x_ρ such that $\rho(x_\rho) = \rho_{av}$, and a vertical line segment (bar) crossing the curve $u(x)$ at $x = x_\rho$ was drawn. Similarly, a horizontal line segment (bar) $u = U(\alpha)$ crossing the curve $u(x)$ at $x = x_u$ was drawn for a specific α.

The values of x_ρ and x_u are very close. This means that the bubble boundary at x_ρ (determined by the jump in ρ) has the required velocity. Thus, first, the theory is sufficiently accurate ($U(\alpha)$ is the theoretical value here), and second, the bubble is almost steady at $t = 6$.

Figure 9b shows the theoretical curves $u(x)$ illustrating the longitudinal-velocity distribution near the wedge-shaped nose for several α. The distribution,

$$u(x, y = 0, t = \infty) = -\sqrt{2\pi}\bar{U}(\alpha)\sqrt{gx}, \qquad (10.1)$$

corresponds to the asymptotic steady state (at $t = \infty$) in the reference frame tied to the bubble apex. Expression (10.1) can be obtained by the hodograph method described in Section 4. The value of $\bar{U}(\alpha)$ in (10.1) is

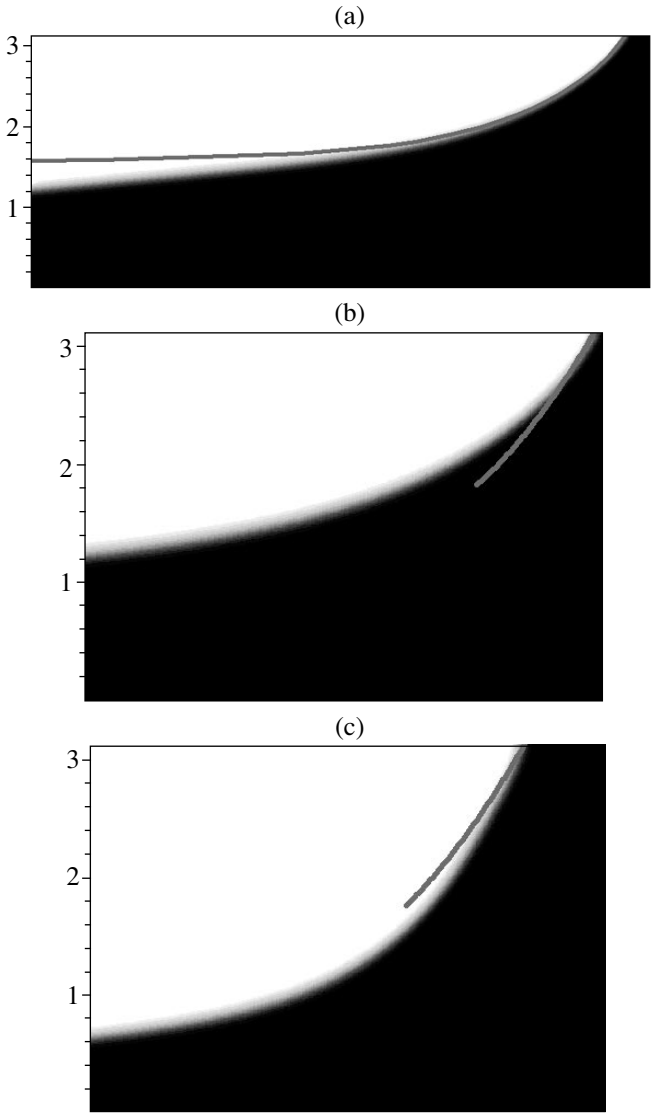

Fig. 7. Comparison of steady bubble geometries calculated numerically at $t = 8$ (Figs. 7a and 7b) and $t = 5$ (Fig. 7c) and analytically for various α: (a) 0; (b) 30°; (c) 90°.

given by (4.4). In Fig. 9b, asymptotic behavior is represented by thin solid curves extending from points x_ρ rightwards. Expression (10.1) can be used to estimate the velocity gradient near the bubble apex.

For comparison, Fig. 9b also shows the dashed curve

$$u(x) = -\sqrt{\frac{\pi}{3}}\sqrt{\frac{g}{D}}x \qquad (10.2)$$

corresponding to $\alpha = 90°$ (vertical motion) in the case of a blunted 2D half-bubble ($\theta_c = 90°$) in a strip of width D (see Fig. 2a), whose velocity is

$$U_{90} = (0.33\text{–}0.34)\sqrt{gD}.$$

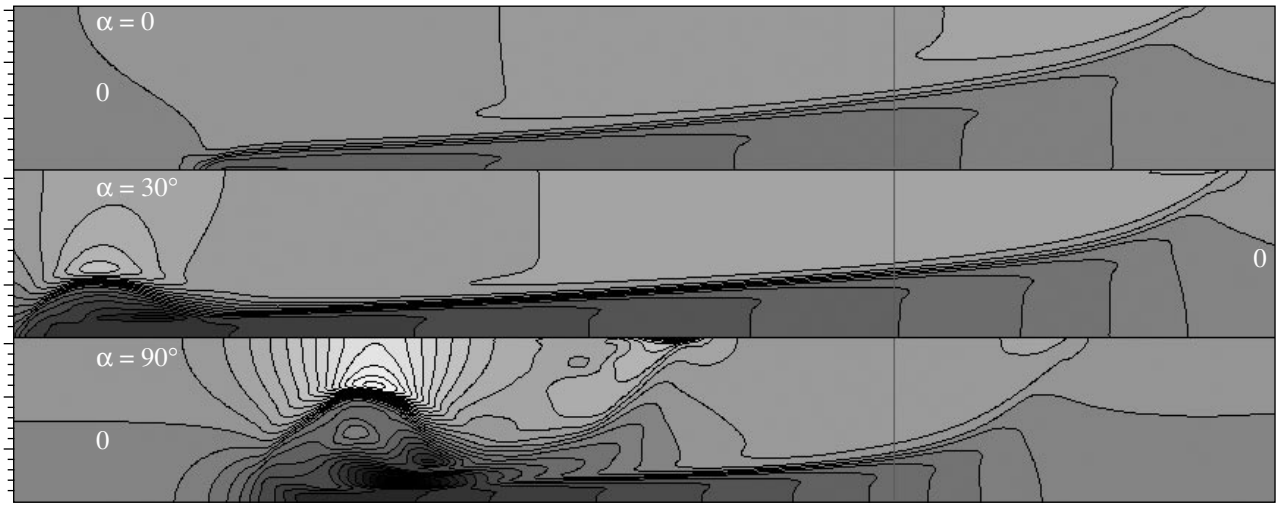

Fig. 8. Map of the longitudinal velocity $u = 0$ at $t = 6$. The curve of $u = 0$ is labeled by 0. Black areas correspond to leftward motion (with the jet), $u < 0$; white areas, to rightward motion (with the bubble), $u > 0$.

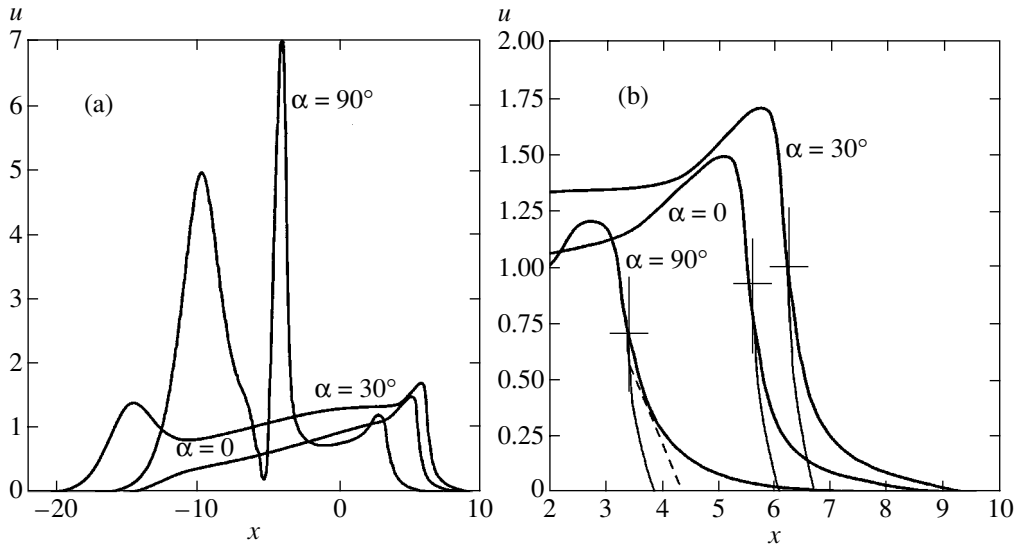

Fig. 9. Velocity distribution in the laboratory frame (a) over the upper wall inside the bubble and beyond its apex and (b) near its apex at $t = 6$. The velocity peaks for $\alpha = 90°$ are associated with vortex dipoles.

The dashed curve extends rightwards from the point (x_ρ, U_{90}) lying on a vertical bar. As the wedge angle increases (the bubble's nose widens), the rate of velocity decay away from the nose decreases (cf. (10.1) and (10.2)).

Since the gas inside the bubble is in motion, the gas velocity inside the bubble is not equal to the velocity of the bubble as a whole. Accordingly, the velocity continues to increase to the left of the horizontal bars in Fig. 9b (the gas moves toward the apex). If the gas were at rest relative to the bubble, then $u(x)$ on the left of x_u would tend to a constant $U(\alpha)$ and the graph of $u(x)$ would contain a plateau level with the horizontal bar.

When $\alpha = 30°$, the bubble penetrates the most deeply into the liquid (see Fig. 9): $U(30°) > U(0) > U(90°)$. This bubble is characterized by the highest velocity $u(x)$ in the neighborhood of the apex.

At considerable distances from the apex, the velocity decreases exponentially, since the incompressible flow is governed by elliptic equations (expressions (10.1) and (10.2) are valid near the apex):

$$u \longrightarrow a_1 \exp(-\pi x/D).$$

In either case (a wedge-shaped or blunted bubble), the behavior of the velocity is dominated by the Fourier harmonic having the longest wavelength (with ampli-

tude a_1), which decays slower than other harmonics. If $\theta_c = 90°$, then $a_1 \approx U_{90}$ [17, 18].

Now, let us consider the time-dependent behavior of the velocity. Figure 10 shows how the asymptotic bubble velocity is approached in the course of transition from the initial state to the asymptotic regime. The curves were obtained by numerical simulation for $g = 1$ and $D = \pi$. At long times, the velocity tends to the limit values $U(\alpha; t \longrightarrow \infty)$ of interest in this study. The time required to approach the asymptotic regime depends on the initial velocity u_0 and the inclination angle α. The curves shown in Fig. 10 can be used to find the asymptotic velocities $U(\alpha)$ for particular α and the corresponding rms deviations $\delta U(\alpha)$ due to the errors of numerical velocity calculation.

Figure 11 shows the longitudinal bubble velocity $U(\alpha)$ as a function of inclination. The solid curve is function (4.4) calculated by the hodograph method in Section 4. The dashed curve represents the results obtained by using the invariance condition for momentum (Section 5). The closed circles with error bars are numerical results. The largest error corresponds to $\alpha = 90°$ (vertical pipe). Open circles I and II on lines $\alpha = 0$ and $\alpha = 90°$ were obtained by applying the models developed in [34] and [26], respectively. Circle III represents the velocity of the blunted half-bubble in a strip of width D (Fig. 2b). Wedge-shaped bubbles adhere to one side of the strip and therefore resemble the half-bubble (compare Figs. 2b, 2c, and 1c with Fig. 2a).

The dependences $U(\alpha)$ obtained by the two theoretical methods and a numerical method are close to each other. The discrepancy between them, including the simulation error, does not exceed 4%. For the wedge-shaped bubble in the case of $\alpha = 90°$, theory and simulation predict

$$U(90) = (0.41\text{–}0.42)\sqrt{gD}.$$

The value obtained in [26] (symbol II) is higher than this value by 23%. The value obtained in [34] (symbol I) is lower than the exact value by 16% (see Section 5).

The functions $U(\alpha)$ calculated by the hodograph and force balance methods or by simulation have maxima at $\alpha_{max} = 30°$ and $\alpha_{max} \approx 40°$, respectively.

11. COMPARISON WITH EXPERIMENT

Let us compare the theoretical results obtained for a 3D flow (Section 6) with experimental data [6]. Figure 12 shows the free surface η. Curve *1* was obtained in experiment with a circular pipe [6], whereas curve *2* was calculated for $\theta_c = 125°$ by the method described in Section 6. Near the apex, the functions η are in good agreement. This means that the corresponding bubble-rise velocities must also agree. The experimental results obtained for low surface tension (Fig. 12) provide strong evidence against the model with η having the form of an ellipse (in which case $\theta_c = 90°$).

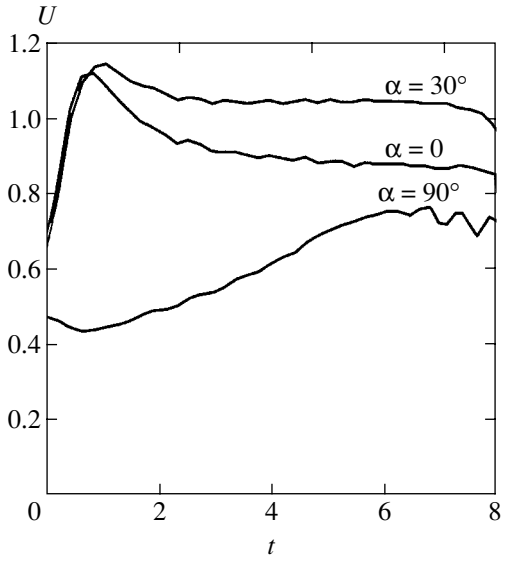

Fig. 10. Velocity of the bubble apex versus time in the laboratory frame.

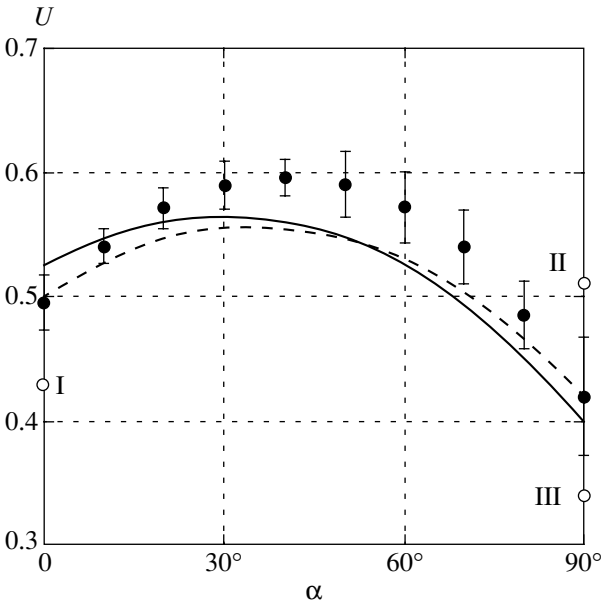

Fig. 11. Effect of inclination angle α on the limit (or asymptotic) velocity of a 120° bubble: numerical simulation (closed circles with error bars), Eq. (4.4) (solid curve), calculation by the force balance method of Section 5 (dashed curve), and various theoretical predictions (open circles).

Figure 13 shows the final 3D results as compared to some 2D results. Closed circles with error bars represent the experimental results obtained in [6]. The curve labeled *2D DNS* is taken from Fig. 11 to illustrate the influence of flow geometry (2D as compared to 3D) on velocity. The results calculated by the force balance method (Section 6) for the wedge angles $\theta_c = 125°$, 115°, and 110° are shown by dashed curves starting from the exact value given by (6.6) for $\alpha = 0$. Open cir-

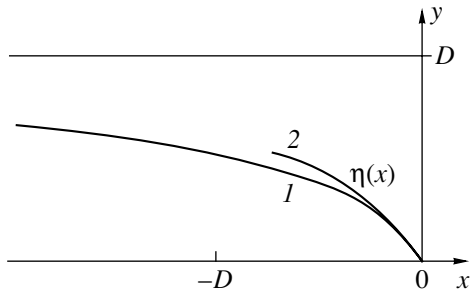

Fig. 12. Wedge-shaped bubble in a circular pipe with $\alpha = 45°$: experiment [6] (curve *1*) and calculation in Section 6 (curve *2*).

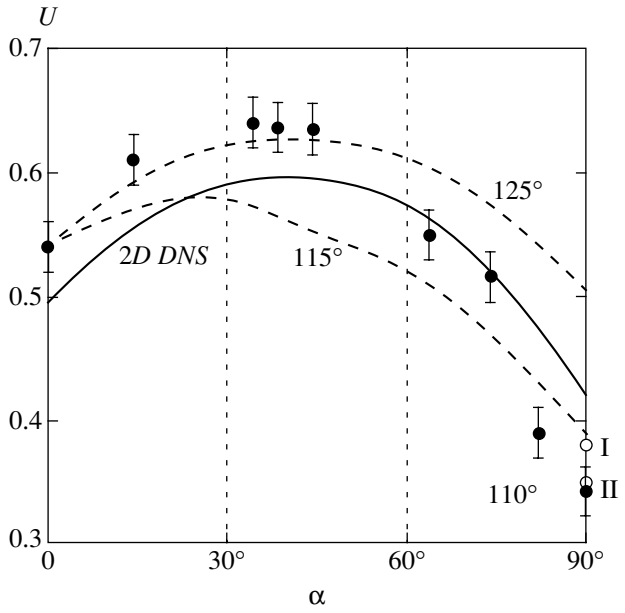

Fig. 13. 3D flow in an inclined circular pipe (except for curve *2D DNS*): experiment [6] (closed circles with error bars), calculation by the force balance method of Section 6 (dashed curve), and various theoretical predictions for $\alpha = 90°$ (open circles).

cles I and II on the line $\alpha = 90°$ correspond to the velocities of the axially symmetric cone-shaped and blunted bubbles. The experimental value obtained in [6] for $\alpha = 90°$ (closed circle) is approximately equal to the velocity of the blunted bubble (open circle II).

The calculations based on conservation laws (Section 6) are sensitive to the nose geometry (cf. dashed curves in Fig. 13). Slender bubbles are characterized by higher velocities, and even slight variations of the angle θ_c are important.

Now, we can gave an overview of the results obtained. Two features are exhibited by the experimental dependence (closed circles in Fig. 13). First, there exists a maximum when the pipe is not in the vertical position. The theoretical model captures the maximum. Second, a substantial decrease in velocity $U(\alpha)$ is observed as the vertical position is approached ($\alpha \longrightarrow 90°$). The velocity becomes comparable to the velocity of a symmetric bubble moving along the axis. This is impossible in the 2D geometry, because the velocity of a bubble moving along the wall is always higher (by a factor of $\sqrt{2}$) than that of a symmetric bubble moving along the centerline (by virtue of the obvious difference in transverse size). It should be noted that the velocity begins to decrease when the position is not vertical. Thus, we should find the limit approached by $U(\alpha)$ as $\alpha \longrightarrow 90°$ and compare it with the value of U for symmetric bubbles. When the pipe is in vertical position, the nose of a bubble may separate from the wall and lie on the pipe axis.

The velocity of 3D bubbles in a circular pipe of diameter D is even lower than that of 2D bubbles in a strip of width D! This can be interpreted as a manifestation of a decrease in the angle θ_c. The bubble's nose widens as the vertical pipe position is approached.

Thus, the velocity of an asymmetric 3D bubble rising along the upper wall (when $\alpha < 90°$) is almost equal to that of an axially symmetric bubble moving along the axis. In computations based on the model of elliptical bubble [2], approximately equal velocities were obtained for asymmetric and symmetric 3D bubbles.

Let us summarize the comparative analysis of bubble dynamics in 2D and 3D geometries. The naive notion is as follows. Since gravitational energy seems to be released at a higher rate in the 3D geometry (additional "paths" are available for motion), a 3D bubble moves substantially faster than a 2D bubble of equal size. This conjecture is based on a comparison of bubbles moving in the $\lambda \times \lambda$ square pipe and a 2D pipe (the ratio of respective velocities is 1/0.6 = 1.7). The ratio of velocities in circular and 2D pipes of equal diameter is 0.35/0.24 = 1.5.

However, the analysis presented here shows that the behavior of bubbles in an inclined pipe is much more complicated. The ratio of near-maximum velocities in 3D and 2D pipes at intermediate inclination angles is slightly greater than unity (≈1.1). For horizontal pipes, the exact ratio is 0.54/0.5 = 1.08. Moreover, it becomes even markedly lower than unity when the pipe is close to the vertical position.

12. CONCLUSIONS

The classical problem of free-surface flow in a gravity field is analyzed. The analysis is essential both for the theory of Rayleigh–Taylor instability and for technological applications of two-phase flows. An attempt is made to bridge a certain gap between the RTI theory and hydraulics. This is a necessary step toward combining the extensive databases amassed in these areas of knowledge.

The following new results have been obtained in this study.

1. Both the cone angle and velocity are found for cone-shaped bubbles.

2. The velocity of rising wedge-shaped bubbles is calculated by two theoretical methods and numerical simulation. The use of momentum invariance makes it possible to overcome the lack of simple Fourier expansions analogous to (3.1) for periodic flows with singularities on their boundary of the type analyzed here. This method is used to find the velocity of cone-shaped bubbles.

3. The angle dependence $U(\alpha)$ is calculated for the entire range of inclinations.

4. A numerical experiment on two-phase pipe flow offers a promising tool for analyzing technological problems.

ACKNOWLEDGMENTS

We thank S.I. Anisimov and O.M. Belotserkovskii for their support of this study. This work was supported by the Russian Foundation for Basic Research, project nos. 02-02-17499 and 03-01-00700. The numerical simulations were supported by Presidium of the Russian Academy of Sciences under the program "Mathematical Modeling, Intelligent Systems, and Control of Nonlinear Mechanical Systems." One of us (A.M.O.) is also grateful to the Russian Foundation for Support of Domestic Science.

REFERENCES

1. G. B. Wallis, *One-Dimensional Two-Phase Flow* (McGraw-Hill, New York, 1969).
2. I. N. Alves, O. Shoham, and Y. Taitel, Chem. Eng. Sci. **48**, 3063 (1993).
3. K. H. Bendiksen, Int. J. Multiphase Flow **10**, 467 (1984).
4. K. H. Bendiksen, Int. J. Multiphase Flow **11**, 797 (1985).
5. R. I. Nigmatulin, in *Proceedings of International Conference on Dynamics of Multiphase Systems, ICMS 2000*, Ed. by M. Ilgamov, I. Akhatov, and S. Urmancheev (Ufa, 2000), p. 1.
6. E. E. Zukoski, J. Fluid Mech. **25**, 821 (1966).
7. R. H. Bonnecaze, W. Eriskine, Jr., and E. J. Greskovich, Am. Inst. Chem. Eng. J. **17**, 1109 (1971).
8. C. C. Maneri and N. Zuber, Int. J. Multiphase Flow **1**, 623 (1974).
9. A. R. Hasan and C. S. Kabir, SPE Prod. Facil. **14**, 56 (1999).
10. G. É. Odishariya and A. A. Tochigin, *Applied Hydrodynamics of Gas–Liquid Mixtures* (Vseross. Nauchno-Issled. Inst. Prirod. Gazov Gazovykh Tekhnol., Ivanov. Gos. Énerg. Univ., 1998).
11. P. Vigneaux, P. Chenais, and J. P. Hulin, Am. Inst. Chem. Eng. J. **34**, 781 (1988).
12. R. I. Nigmatulin, *Dynamics of Multiphase Media* (Nauka, Moscow, 1987).
13. L. Mattar and G. A. Gregory, J. Can. Petr. Technol. **13**, 69 (1974).
14. M. K. Nicholson, K. Aziz, and G. A. Gregory, Can. J. Chem. Eng. **56**, 653 (1978).
15. D. M. Maron, N. Yacoub, and N. Brauner, Lett. Heat Mass Transfer **9**, 333 (1982).
16. H.-J. Kull, Phys. Rep. **206**, 197 (1991).
17. N. A. Inogamov, A. Yu. Dem'yanov, and É. E. Son, *Hydrodynamics of Mixing* (Mosk. Fiz.–Tekh. Inst., Moscow, 1999).
18. N. A. Inogamov, Astrophys. Space Phys. Rev. **10**, 1 (1999).
19. T. B. Benjamin, J. Fluid Mech. **31**, 209 (1968).
20. D. T. Dumitrescu, Z. Angew. Math. Mech. **23**, 139 (1943).
21. R. M. Davies and G. I. Taylor, Proc. R. Soc. London, Ser. A **200**, 375 (1950).
22. D. Layzer, Astrophys. J. **122**, 1 (1955).
23. N. A. Inogamov and A. M. Oparin, Zh. Éksp. Teor. Fiz. **116**, 908 (1999) [JETP **89**, 481 (1999)].
24. J. Hecht, U. Alon, and D. Shvarts, Phys. Fluids **6**, 4019 (1994).
25. N. A. Inogamov and A. Yu. Dem'yanov, Prikl. Mekh. Tekh. Fiz. **37**, 93 (1996).
26. J.-M. Vanden-Broeck, Phys. Fluids A **5**, 2454 (1993).
27. N. A. Inogamov, M. Tricottet, A. M. Oparin, and S. Bouquet, Phys. Lett. A (2003) (in press); physics/0104084.
28. A. I. D'yachenko, Dokl. Akad. Nauk **376**, 27 (2001).
29. V. E. Zakharov and A. I. Dyachenko, Physica D (Amsterdam) **98**, 652 (1996).
30. A. L. Velikovich and G. Dimonte, Phys. Rev. Lett. **76**, 3112 (1996).
31. Q. Zhang and S.-I. Sohn, Z. Angew. Math. Phys. **50**, 1 (1999).
32. M. Berning and A. M. Rubenchik, Phys. Fluids **10**, 1564 (1998).
33. C.-S. Yih, *Stratified Flows* (Academic, New York, 1980).
34. V. V. Bychkov, Phys. Rev. E **55**, 6898 (1997).
35. E. A. Kuznetsov and S. S. Minaev, Phys. Lett. A **221**, 187 (1996).
36. S. F. Shandarin and Ya. B. Zeldovich, Rev. Mod. Phys. **61**, 185 (1989).
37. L. Kofman, D. Pogosyan, and S. Shandarin, Mon. Not. R. Astron. Soc. **242**, 200 (1990).
38. H. Bateman and A. Erdelyi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953, Nauka, Moscow, 1973).
39. O. M. Belotserkovskii, *Numerical Simulation in Mechanics of Continuous Media* (Fizmatlit, Moscow, 1994).
40. O. M. Belotserkovskii and A. M. Oparin, *Numerical Experiment in Turbulence* (Nauka, Moscow, 2000).
41. N. A. Inogamov, A. M. Oparin, A. Yu. Dem'yanov, *et al.*, Zh. Éksp. Teor. Fiz. **119**, 822 (2001) [JETP **92**, 715 (2001)].

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1186–1200.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1329–1344.

FLUIDS

Generation of Large-Scale Coherent Structures in the KPZ Equation and Multidimensional Burgers Turbulence

S. N. Gurbatov and A. Yu. Moshkov

Nizhni Novgorod State University, Nizhni Novgorod, 603950 Russia
e-mail: gurb@rf.unn.ru

Received June 23, 2003

Abstract—The development of multiple-scale random fluctuations in the problem of interface growth is analyzed. The evolution of the interface is described by the Kardar–Parisi–Zhang (KPZ) equation, and the gradient vector field satisfies the multidimensional Burgers equation. It is shown that the nonlinear effects in the evolution of statistically inhomogeneous multiple-scale fluctuations lead to the generation of large-scale coherent structures. Due to combined effects of nonlinearity and dissipation, localized disturbances tend to exhibit universal long-time behavior. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The nonlinear diffusion equation

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \nu\frac{\partial^2 v}{\partial x^2} \qquad (1)$$

was originally proposed by Burgers as a model of hydrodynamic turbulence [1, 2]. Indeed, It has much in common with the classical Navier–Stokes equation, including the type of nonlinearity, invariants, and frequency dependence of energy losses [3]. The distinctions between the Burgers and Navier–Stokes equations are as interesting as their similarity [4], even more so with regard to the multidimensional Burgers equation. It was shown that Eq. (1) describes a variety of wave phenomena in acoustics, plasma physics, dynamics of flame front propagation, etc. [5–8]. In particular, Eq. (1) supplemented with random initial conditions describes the evolution of intense acoustic noise. Accordingly, such solutions to the Burgers equations are referred to as acoustic turbulence.

The multidimensional Burgers equation with random forcing is widely used as a model of Navier–Stokes hydrodynamic turbulence without pressure [9–13]. The three-dimensional Eq. (1) combined with the continuity equation is used to analyze the development of large-scale structures in the Universe via nonlinear gravitational instability at the stage when pressure forces are negligible. This approach is known in astrophysics as the adhesion model. The model describes the formation of highly inhomogeneous structures in the distribution of matter initiated by a random disturbance [14–17]. Other phenomena described by the multidimensional Burgers equations or its modifications include the interface growth due to random deposition of substance on a surface and flame front propagation [18]. In these problems, the potential ψ ($\mathbf{v} = -\nabla\psi$) represents the surface profile, and its evolution is governed by an equation equivalent to the Kardar–Parisi–Zhang (KPZ) equation [8, 18–20]. The mean square gradient

$$E(t) = \langle(\nabla\psi(\mathbf{x}, t))^2\rangle = \langle\mathbf{v}^2(\mathbf{x}, t)\rangle,$$

which characterizes surface roughness, may either decrease or increase with time.

The dynamical and statistical characteristics of both one-dimensional and (more recently) multidimensional Burgers equations have been analyzed in numerous studies (e.g., see the bibliography in [7, 8]). Even though the Burgers equation has an exact (Cole–Hopf) solution [21, 22], investigation of statistical properties of this equation is a formidable mathematical problem. In particular, the first significant results for a Brownian initial potential were published in [2] after over thirty years had passed since the equation was introduced in [1], and an exact statistical treatment of this special case was presented only recently [23]. The power spectral density of this signal at zero frequency is preserved, whereas its energy decreases as $t^{-2/3}$. If the initial power spectrum does not contain large-scale components, then turbulence decay follows a different scenario. In the case of a Gaussian initial distribution, energy decreases as t^{-1} up to a logarithmic correction factor [24–27]. A comprehensive statistical description of decaying turbulence can also be obtained in this case. In particular, it can be shown that its statistical characteristics (spectra, correlation functions, probability distributions) become asymptotically self-similar [25, 27].

In this paper, we analyze the evolution of modulated waves, such as multiple-scale fluctuations. Distur-

bances of this kind arise, for example, in problems of flame front propagation [28], when the domain of initial disturbance is localized in space and the length scale characterizing its intrinsic structure is much less than the domain size. By introducing certain approximations, the evolution of the flame front can be reduced to the two-dimensional Burgers equation. The present study shows that the nonlinear effects in the evolution of statistically inhomogeneous fluctuations lead to the generation of large-scale coherent structures. Due to combined effects of nonlinearity and dissipation, localized disturbances tend to exhibit universal long-time behavior.

The paper is organized as follows. In Section 2, the multidimensional Burgers equation is solved and the asymptotic behavior of its solution is considered in the long-time and low-viscosity limit, when nonlinear effects are negligible. It is shown that nonlinear effects lead to local self-similarity of the velocity and potential fields in the limit of vanishing viscosity. We also discuss the evolution of the basic types of disturbances described by the one-dimensional Burgers equation. In Section 3, we analyze the evolution of multiple-scale localized disturbances governed by the multidimensional Burgers equation in the limit of vanishing viscosity. Section 4 focuses on the long-time behavior of localized disturbances in the case of a finite viscosity.

2. BASIC EQUATIONS AND APPROACHES

2.1. Local Self-Similarity and Long-Time Asymptotics of the Multidimensional Burgers Equation

We consider the vector Burgers equation without external forcing,

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = \nu\Delta\mathbf{v}, \tag{2}$$

and seek potential solutions of the form

$$\mathbf{v}(\mathbf{x}, t) = -\nabla\psi(\mathbf{x}, t). \tag{3}$$

The velocity potential $\psi(\mathbf{x}, t)$ satisfies the nonlinear equation

$$\frac{\partial \psi}{\partial t} = \frac{1}{2}(\nabla\psi)^2 + \nu\Delta\psi. \tag{4}$$

It is identical to the KPZ equation [8, 18, 19], which is commonly written in terms of $h = \lambda^{-1}\psi$, where the local interface growth velocity λ has the dimension of velocity and the surface height $h(\mathbf{x}, t)$ has the dimension of length. In problems of interface growth, the coefficient ν represents surface tension and the corresponding term on the right-hand side represents linear effects responsible for smoothing of the interface. The measure of surface roughness is the mean-square gradient

$$E(t) = \langle(\nabla\psi(\mathbf{x}, t))^2\rangle = \langle\mathbf{v}^2(\mathbf{x}, t)\rangle = \sum_i E_i(t), \tag{5}$$

$$E_i(t) = \left\langle\left(\frac{\partial\psi}{\partial x_i}\right)^2\right\rangle = \langle v_i^2\rangle. \tag{6}$$

The angle brackets denote ensemble averages or integrals over the spatial coordinates (for localized disturbances). In the one-dimensional case, the energy $E(t)$ of turbulence in a dissipative medium is a monotone decreasing function of time. In the limit of vanishing viscosity ($\nu \longrightarrow 0$), the energy $E(t)$ is conserved until the wave profile becomes discontinuous and then decreases through dissipation in infinitely thin shocks.

Before the shocks begin to develop, the multidimensional Burgers equation with vanishing viscosity corresponds to free motion of particles. In the Lagrangian representation, the velocity $\mathbf{V}(t; \mathbf{y})$ of a particle is constant, depending only on its initial (Lagrangian) coordinate $\mathbf{y}$. In the one-dimensional case, an increase in segment length, $\Delta x = \Delta y + t\Delta V$, is balanced in the Eulerian representation by a decrease in the length of an adjoining interval, $\Delta x = \Delta y - t\Delta V$, so that the energy of disturbances is conserved. When shocks appear, the energy begins to decrease with time. In the multidimensional case, a change in volume element depends on the initial curvature of a surface disturbance in the Eulerian representation, and there is no balance between the contracting and expanding volumes observed. Accordingly, the surface roughness measured by $E(t)$ can either increase or decrease with time when $d > 1$ [29, 30], since no conservation law applies in this case. Nevertheless, we call $E(t)$ the turbulence energy and $E_i(t)$ the energy of the ith velocity component here.

Changing to the Cole–Hopf variables [21, 22],

$$\begin{aligned}\psi(\mathbf{x}, t) &= 2\nu\ln U(\mathbf{x}, t),\\ \mathbf{v}(\mathbf{x}, t) &= -2\nu\nabla\ln U(\mathbf{x}, t),\end{aligned} \tag{7}$$

we reduce (4) to the linear diffusion equation

$$\frac{\partial U}{\partial t} = \nu\Delta U, \tag{8}$$

$$U(\mathbf{x}, 0) = U_0(\mathbf{x}) = \exp\frac{\psi_0(\mathbf{x})}{2\nu}. \tag{9}$$

In this paper, we analyze the evolution of coherent and stochastic signals at high Reynolds numbers, when nonlinear interaction between spatial harmonics plays an important role, while dissipation is essential only for high-wavenumber components. Energy is dissipated in space only in small neighborhoods of the shocks. Thus, the solution to problem (2) obtained in the limit of van-

ishing viscosity can be used at intermediate times. At a later stage, when nonlinear interaction becomes unimportant, the evolution of the field is controlled by linear dissipation only and we have a linearized Cole–Hopf solution.

As $\nu \longrightarrow 0$, the use of a saddle point method in the Cole–Hopf solution leads to the so-called "maximum method" for the potential field [7, 17, 21]:

$$\psi(\mathbf{x}, t) = \max_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y}, t), \tag{10}$$

$$\Phi(\mathbf{x}, \mathbf{y}, t) = \psi_0(\mathbf{y}) - \frac{(\mathbf{x}-\mathbf{y})^2}{2t}, \tag{11}$$

$$\mathbf{v}(\mathbf{x}, t) = \frac{\mathbf{x}-\mathbf{y}(\mathbf{x}, t)}{t} = \mathbf{v}_0(\mathbf{y}(\mathbf{x}, t)). \tag{12}$$

Here, $\psi_0(\mathbf{y})$ is the initial potential: $\mathbf{v}_0(\mathbf{x}) = -\nabla\psi_0(\mathbf{x})$. The function $\mathbf{y}(\mathbf{x}, t)$ in (12) is the Lagrangian coordinate at which $\Phi(\mathbf{x}, \mathbf{y}, t)$ attains its absolute maximum for some particular values of Eulerian coordinate $\mathbf{x}$ and time t. It can readily be shown that $\mathbf{y}$ is the initial coordinate of the particle that reaches $\mathbf{x}$ at the instant t [7].

After a sufficiently long time has elapsed, the second term on the right-hand side of (11) varies much more slowly than the initial potential $\psi_0(\mathbf{y})$. Therefore, the absolute maximum of $\Phi(\mathbf{x}, \mathbf{y}, t)$ is one of the maximums of $\psi_0(\mathbf{y})$. In the neighborhood of its local maximum point $\mathbf{y}_k$, the initial potential can be represented as

$$\psi_0(\mathbf{x}) = \psi_{0,k}\left[1 - \sum_i \frac{(x_i - y_{i,k})^2}{2L_i^2}\right], \tag{13}$$

where the set of x_i defines the basis of principal axes for expanding a local quadratic form about the point $\mathbf{y}_k$. Using solutions (11) and (12), we obtain

$$\psi(\mathbf{x}, t) = \psi_{0,k}\left[1 - \sum_i \frac{(x_i - y_{i,k})^2}{2L_i^2(1 + \psi_{0,k}t/L_i^2)}\right], \tag{14}$$

$$v_i(\mathbf{x}, t) = \frac{\psi_{0,k}(x_i - y_{i,k})}{L_i^2(1 + \psi_{0,k}t/L_i^2)}. \tag{15}$$

It follows from (15) that nonlinear effects cause the velocity field to become locally isotropic and locally self-similar in the neighborhood of a maximum of $\psi_0(\mathbf{x})$. Its behavior is determined by the particles initially located at points $y_i(\mathbf{x}, t)$ lying in a small neighborhood of the maximum,

$$y_i(\mathbf{x}, t) = \frac{x_i - y_{i,k}}{1 + \psi_{0,k}t/L_i^2}. \tag{16}$$

Thus, the Lagrangian coordinate $\mathbf{y}(\mathbf{x}, t)$ becomes a discontinuous function of $\mathbf{x}$ at long times that is constant within the domain (cell) that corresponds to a maximum and has jumps on its boundaries [7, 17]. The velocity field $\mathbf{v}(\mathbf{x}, t)$ is discontinuous, and the derivatives of the potential $\psi(\mathbf{x}, t)$ are discontinuous, across the cell boundaries. It is clear from (14) and (15) that both the potential and the velocity fields have a universal self-similar structure within the cells:

$$\psi(\mathbf{x}, t) = \psi_0(\mathbf{y}_k) - \frac{(\mathbf{x}-\mathbf{y}_k)^2}{2t}, \tag{17}$$

$$\mathbf{v}(\mathbf{x}, t) = \frac{\mathbf{x}-\mathbf{y}_k}{t}. \tag{18}$$

The longitudinal component of the velocity vector $\mathbf{v}(\mathbf{x}, t)$ is a sawtooth pulse train, as in the one-dimensional case. The transverse velocity component is constant within a cell. According to (14) and (15), this universal behavior manifests itself at earlier times in the directions of steeper initial gradients (smaller L_i).

At later stages, the evolution of the velocity and potential fields is determined by the characteristics of local peaks of $\psi_0(\mathbf{y}_k)$. When the initial profile is periodic, the developing universal structure has a periodic invariant form and the velocity amplitude decreases as t^{-1}. In the case of a random initial profile, the surface continues to change, because cells (or one-dimensional shocks) merge as the structure develops, and the integral length scale $L(t)$ of Burgers turbulence tends to increase accordingly.

Now, let us discuss the long-time limit solution to the Burgers equation in the case when time increases indefinitely while $\nu \neq 0$ remains constant. Consider the class of initial disturbances with a bounded potential, $\langle\psi_0(\mathbf{x})^2\rangle < \infty$, and assume that $\psi_0(\mathbf{x})$ is a localized disturbance or a statistically homogeneous random field. Any scalar profile $U(\mathbf{x}, t)$ of this kind contains a constant component $\bar{U}$,

$$U(\mathbf{x}, t) = \bar{U} + \tilde{U}(\mathbf{x}, t) = \bar{U}(1 + u(\mathbf{x}, t)), \tag{19}$$

where $u(\mathbf{x}, t) = \tilde{U}(\mathbf{x}, t)/\bar{U}$ is s the relative fluctuation of the field $U(\mathbf{x}, t)$. The fields $\tilde{U}_0(\mathbf{x})$ and $u_0(\mathbf{x})$ are assumed to have zero-mean distributions. In the course of time, the field $U(\mathbf{x}, t)$ described by Eq. (8) is smoothed by diffusion and the amplitude of $\tilde{U}(\mathbf{x}, t)$ decreases. When $|\tilde{U}| \ll \bar{U}$ ($|u| \ll 1$), the Cole–Hopf solution (7) can be linearized:

$$\mathbf{v}(\mathbf{x}, t) = -2\nu\nabla u(\mathbf{x}, t). \tag{20}$$

Since both $\tilde{U}(\mathbf{x}, t)$ and $u(\mathbf{x}, t)$ satisfy linear diffusion equations, the field $\mathbf{v}(\mathbf{x}, t)$ is also governed by a linear equation. This means that the linear stage of evolution is reached. Nonlinear effects contribute to this solution

only through the nonlinear integral relation between the initial velocity field $\mathbf{v}_0(\mathbf{x})$ and the fields $\tilde{U}(\mathbf{x}, 0)$ and $\bar{U}$ (see (3) and (9)). They are characterized by the initial Reynolds number $\mathrm{Re}_0 \sim |\Delta\psi_0|/\nu$, where $\Delta\psi_0$ is the characteristic amplitude of ψ_0.

Relation (20) leads to a well-known result concerning the asymptotic behavior of a harmonic disturbance governed by the one-dimensional Burgers equation [5, 6]. When $\mathrm{Re}_0 \gg 1$, a harmonic wave transforms into a sawtooth wave. However, its harmonic form is restored in the long-time limit via dissipation, with an amplitude that is independent of the initial one. When the initial Reynolds number is high, a statistically homogeneous Gaussian field $v_0(x)$ also transforms into a sawtooth pulse train that has essentially non-Gaussian properties at the stage of nonlinear development [7, 27]. Nevertheless, any random field $\mathbf{v}(\mathbf{x}, t)$ with statistically homogeneous initial potential $\psi_0(\mathbf{x})$ weakly converges to a homogeneous zero-mean Gaussian field [29]. This stage is known as the Gaussian scenario in the theory of Burgers turbulence. This scenario applies to the multidimensional Burgers equation. In the absence of long-range correlations, the distributions of initial velocity potential and potential field are characterized by a universal covariance function whose amplitude depends nonlinearly on its initial value and is proportional to $\exp(\mathrm{Re}_0^2)$ [7, 29]. When the initial potential has long-range correlations,

$$\langle \psi_0(\mathbf{x})\psi_0(0)\rangle = |\mathbf{x}|^{(-\alpha)}F(\mathbf{x}/x), \quad 0 < \alpha < 3,$$

both the anisotropy of $F(\mathbf{x}/x)$ and the long-range correlations persist at the linear stage [29].

2.2. Evolution of the Basic Types of One-Dimensional Disturbances

Let the initial potential $\psi_0(\mathbf{x})$ be the sum of one-dimensional functions $\psi_{0,i}(x_i)$:

$$\psi_0(\mathbf{x}) = \sum_i \psi_{0,i}(x_i), \quad v_{0,i}(\mathbf{x}) = v_{0,i}(x_i). \qquad (21)$$

In this case, it follows directly from (2) that the field components $v_{0,i}$ do not interact and the evolution of each $v_{0,i}(\mathbf{x}, t) = v_{0,i}(x_i, t)$ is governed by the one-dimensional Burgers equation (1). Here, we briefly consider the evolution of the basic types of disturbances described by the one-dimensional Burgers equation [5–8].

First, consider the evolution of a localized disturbance. Suppose that the initial potential

$$\psi_0(x) = m\left(1 - \frac{x^2}{2L_0^2} + \ldots\right)$$

has a single maximum $m = \psi_0(y_k)$ localized within an interval of length on the order of L_* ($L_0 \approx L_*$) around the point $x = y_k$ and assume that $\psi_0(x) = 0$ if $|x - y_k| > L_*$. As $\nu \longrightarrow 0$, this disturbance eventually transforms into the so-called N wave [5], with the gradient $\partial v/\partial x = 1/t$ and the shock-front location determined by the equation $|x_s - y_k| = (2mt)^{1/2}$. As the interval occupied by the disturbance increases, its amplitude decreases according to the equation

$$\frac{x_s(t)}{t} \sim m^{1/2}t^{-1/2}$$

and its energy decreases as

$$\frac{x_s^3}{t^2} \sim m^{3/2}t^{-1/2}.$$

Therefore, the asymptotic nonlinear behavior of a localized disturbance is determined only by the maximum $m = \psi_0(y_k)$ of the initial potential and is independent of its profile. At a finite Reynolds number $\mathrm{Re}_0 = m/2\nu \gg 1$, the shocks have finite widths and the location of the discontinuity is

$$x_s(t) = \left[2t\left(m - \nu\ln\frac{4\pi\nu t}{L_{\mathrm{eff}}^2}\right)\right]^{1/2},$$

where $L_{\mathrm{eff}}^2 \sim L_0^2/\mathrm{Re}_0$. Using (20), one readily finds that the field $v(x, t)$ has universal structure at the stage of its linear evolution (at $t \gg L_0^2\exp(2\mathrm{Re}_0)\mathrm{Re}_0\nu$):

$$v(x, t) = B\frac{x}{\sqrt{4\pi\nu t^3}}\exp\left(-\frac{x^2}{4\nu t}\right) \qquad (22)$$

with the constant

$$B \approx L_0\left(\frac{2\pi}{\mathrm{Re}_0}\right)^{1/2}\exp(\mathrm{Re}_0).$$

Localized multiple-scale random fluctuations also exhibit universal asymptotic behavior, with a large-scale mean component evolving from a zero-mean initial through nonlinear effects [31].

Consider the evolution of the harmonic disturbance $v_0(x) = k_0\psi_0\sin(k_0x)$ associated with the potential $\psi_0(x) = \psi_0\cos(k_0x)$. Its initial energy is

$$\langle v^2\rangle = \frac{\psi_0^2k_0^2}{2} = \frac{2\pi^2\psi_0^2}{L_0^2}, \quad k_0 = \frac{2\pi}{L_0}.$$

When the Reynolds number is infinite ($\nu \longrightarrow 0$), the velocity field transforms into a periodic sawtooth pulse train with the gradient $\partial v/\partial x = 1/t$ at $t \gg t_{\mathrm{nl}}$, where $t_{\mathrm{nl}} =$

$1/k_0^2\psi_0$ is the characteristic time of nonlinear development. It should be noted that both amplitude $a = L_0/t$ and energy $E(t) = L_0^2/12t^2$ are independent of the initial amplitude at this stage. Therefore, if the components of a two-dimensional initial disturbance (21) have equal amplitudes $\psi_0(x)$, but widely different periods L_i ($L_1 \ll L_2$), then the initial energy of the component with the larger length scale is much greater: $E_1(0)/E_2(0) = L_2^2/L_1^2$. However, a reverse relation is true at long times when $t \gg t_{nl,2} = 1/k_2^2\psi_0$: $E_1(t)/E_2(t) \longrightarrow L_1^2/L_2^2$. When the Reynolds number $\mathrm{Re}_0 = \psi_0/2\nu$ is moderately high, a linear stage of the field evolution is observed at $t \gg t_{nl}\mathrm{Re}_0$, when

$$v(x, t) = 4\nu k_0 \sin(k_0 x)\exp(-\nu k_0^2 t).$$

The evolution of a two-dimensional disturbance (21) at this stage is also dominated by the component with the larger period L_0.

A continuous random field also transforms into an array of cells with equal gradients $\partial v/\partial x = 1/t$, which are separated by randomly distributed shocks. As the cells merge, the integral turbulence length scale $L(t)$ increases. Therefore, the energy of a random field, $E(t) \propto L^2(t)/t^2$, decreases at a slower rate as compared to the energy of a periodic wave. The scenario of turbulence is determined by the behavior of the large-scale component of the initial power spectrum

$$E_0(k) = \frac{1}{2\pi}\int \langle v_0(x)\, v_0(x+z)\rangle \exp(ikz)\, dz, \qquad (23)$$

$$E_0(k) = \alpha^2 k^n b_0(k). \qquad (24)$$

Here, the function $b_0(k)$ rapidly decreases at $k > k_0 \sim l_0^{-1}$. When $n < 1$, the initial potential is a Brownian or fractal distribution, and a scaling method can be applied [2, 7, 17, 32]. In this case, we deal with self-similar turbulence characterized by the length scale

$$L(t) = (\alpha t)^{2/(3+n)},$$

which is independent of the length scale l_0 of the initial power spectrum [7]. Moreover, the behavior of individual realizations of the random field is determined by the large-scale components of the initial disturbance, weakly depending on small-scale fluctuations [33].

When $n > 1$, the law of energy decay strongly depends on the statistical characteristics of the initial field [8, 30]. For a Gaussian initial profile, the mean potential $\langle\psi\rangle$, length scale $L(t)$, and turbulence energy $E(t)$,

$$L(t) = (t\sigma_\psi)^{1/2}\ln^{-1/4}\left(\frac{t\sigma_\psi}{2\pi l_0^2}\right),$$
$$E(t) = t^{-1}\sigma_\psi \ln^{-1/2}\left(\frac{t\sigma_\psi}{2\pi l_0^2}\right), \qquad (25)$$

are completely determined by the variances of the initial potential and velocity, $\sigma_\psi^2 = \langle\psi_0^2\rangle$ and $\sigma_v^2 = \langle v_0^2\rangle$ [24–27, 34]. Here, $l_0 = \sigma_\psi/\sigma_v$ is the characteristic length scale of the initial fluctuation. Thus, the energies of the two components of a two-dimensional initial disturbance (21) characterized by equal variances σ_ψ and different length scales $l_{0,i}$ ($l_{0,1} \ll l_{0,2}$) at $t = 0$ differ substantially:

$$\frac{E_1(0)}{E_2(0)} = \frac{l_2^2}{l_1^2} \gg 1.$$

However, they are almost equal at long times: $E_1(t)/E_2(t) \approx 1$ (up to a small logarithmic correction). In the case of a finite Reynolds number $\mathrm{Re}_0 = \sigma_\psi/2\nu$, the linear regime is reached at $t \gg t_{nl}\exp(\mathrm{Re}_0^2)/\mathrm{Re}_0$ as a result of multiple shock collisions ($t_{nl} = \sigma_\psi/l_0^2$ is the characteristic time of nonlinear development). At the linear stage, energy decreases as $Ct^{-3/2}$, where $C \sim l_0\exp(\mathrm{Re}_0^2)/\mathrm{Re}_0$.

3. EVOLUTION OF LOCALIZED DISTURBANCES IN THE LIMIT OF VANISHING VISCOSITY

3.1. Evolution of Simple Disturbances

Consider the evolution of a localized anisotropic disturbance governed by the multidimensional Burgers equation. To simplify analysis, we begin with the two-dimensional case ($d = 2$) (extension to $d > 2$ is mostly straightforward). First, we analyze the special case when the initial potential $\psi_0(\mathbf{x})$ is a quadratic function of coordinates (as in (13) with $y_{i,k} = 0$) within the domain S_0 defined by the relation

$$\frac{x_1^2}{2L_1^2} + \frac{x_2^2}{2L_2^2} \le 1,$$

and $\psi_0(\mathbf{x}) = 0$ outside S_0. The evolution of the potential $\psi(\mathbf{x}, t)$ and velocity $\mathbf{v}(\mathbf{x}, t)$ within the expanding ellipse

$S(t)$ defined as

$$\frac{x_1^2}{2L_1^2(t)} + \frac{x_2^2}{2L_2^2(t)} \le 1, \tag{26}$$

$$L_i(t) = L_i\left(1 + \frac{\psi_0 t}{L_i^2}\right)^{1/2} = L_i\left(1 + \frac{t}{t_{\mathrm{nl},\,i}}\right)^{1/2} \tag{27}$$

is described by (14) and (15), whereas $\mathbf{v} = 0$ and $\psi = 0$ outside the ellipse. Here, $t_{\mathrm{nl},\,i} = L_i^2/\psi_0$ is the characteristic time of nonlinear development for the ith velocity component. Each velocity component $v_i(x_i, t)$ varies independently; in particular,

$$v_2(x_2, t) = \frac{\psi_0 x_2}{L_2^2(1 + \psi_{0,\,k} t/L_2^2)}$$

is independent of the coordinate x_1. However, the length of the expanding interval on the x_1 axis,

$$\Delta x_1(x_2, t) = 2L_1(t)\left[2\left(1 - \frac{x_2^2}{2L_2^2(t)}\right)\right]^{1/2},$$

where $v_2(x_2, t)$ is independent of x_1, is determined by both $L_2(t)$ and $L_1(t)$. In other words, the velocity components are coupled by strong interaction. The energies of the velocity components are

$$E_1(t) = \frac{\pi\psi_0^2 L_2(t)}{L_1(t)}, \quad E_2(t) = \frac{\pi\psi_0^2 L_1(t)}{L_2(t)}. \tag{28}$$

Consider the evolution of a highly anisotropic fluctuation ($L_1 \ll L_2$), using the energy ratio

$$\kappa(t) = \frac{E_1(t)}{E_2(t)} = \frac{L_2^2(t)}{L_1^2(t)}$$

as a measure of anisotropy. It follows from (28) that the energy $E_1(t)$ of the small-scale component is a monotonically decreasing function, whereas the energy $E_2(t)$ of the large-scale one is a monotonically increasing one. Accordingly, the anisotropy parameter $\kappa(t)$ monotonically decreases from $\kappa(0) = L_2^2/L_1^2 \gg 1$, approaching 1 at $t \gg t_{\mathrm{nl},\,2}$. When $t_{\mathrm{nl},\,1} \ll t \ll t_{\mathrm{nl},\,2}$, the nonlinear self-interaction of the large-scale component is negligible, and the component energies are

$$E_1(t) \approx \frac{\pi\psi_0^2 L_2}{(\psi_0 t)^{1/2}}, \quad E_2(t) \approx \frac{\pi\psi_0^2 (\psi_0 t)^{1/2}}{L_2}. \tag{29}$$

Thus, the decrease in the energy of the small-scale component follows the same law as in the one-dimensional case. The energy of the large-scale component increases with time, because the energy of the preserved velocity component $v_2(x_2, t) \approx v_2(x_2, 0)$ is transferred in space by the component v_1. Both the total energy $E(t)$ and the component energies remain invariant over a long time interval. The invariance of energy (mean surface roughness) over long times is explained by the effect of increase in volume, $V(t) \propto t^{1/2}$, which compensates for the decrease in the steepest gradients, $v_i(t) \sim t^{-1/2}$. Figure 1 shows the dimensionless energies $e_i(t) = E_i(t)\psi_0^2/\pi$ of the components of an anisotropic fluctuation ($L_2/L_1 = \sqrt{10}$) as functions of the dimensionless time $\tau = t/t_{\mathrm{nl},\,1}$.

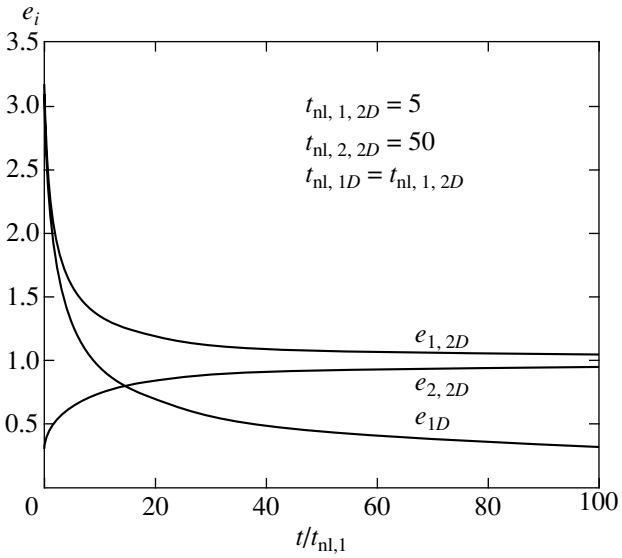

Fig. 1. Evolution of the dimensionless energy $e_{i,\,2D} = E_i\psi_0^2/\pi$ of the components of a two-dimensional anisotropic disturbance ($L_2/L_1 = \sqrt{10}$) as a function of dimensionless time $\tau = t/t_{\mathrm{nl},\,1}$. Evolution of the energy of a one-dimensional localized disturbance, e_{1D}.

In the three-dimensional case, it can readily be shown that the energy of the velocity component $v_1(t)$ of an initial potential $\psi_0(\mathbf{x})$ described by expression (13) within the corresponding ellipsoid varies as follows:

$$E_1(t) \propto \frac{\psi_0^2 L_2(t) L_3(t)}{L_1(t)}, \tag{30}$$

where $L_i(t)$ is given by (27). When the fluctuation is highly anisotropic, the component energies may either increase or decrease at the initial stage ($\min t_{\mathrm{nl},\,i} \ll t \ll \max t_{\mathrm{nl},\,i}$), but the field becomes isotropic and its energy increases as $t^{1/2}$ at $t \gg \max t_{\mathrm{nl},\,i}$.

Suppose that the initial anisotropic localized disturbance can be represented as

$$\psi_0(\mathbf{x}) = \psi_0 f_1(x_1) f_2(x_2), \tag{31}$$

where each f_i reaches a maximum at $\mathbf{x} = 0$ ($f_1(0) = f_2(0) = 1$) and is characterized by a respective length scale L_i ($L_1 \ll L_2$). For such a disturbance, $E_1(0) \gg E_2(0)$.

At the stage when

$$\frac{t_{\mathrm{nl},1}}{f_2(x_2)} \ll t \ll t_{\mathrm{nl},2}, \quad t_{\mathrm{nl},i} = \frac{L_i^2}{\psi_0},$$

the small-scale component $v_1(\mathbf{x}, t)$ transforms into an *N*-shaped pulse while the self-interaction of the large-scale component $v_2(\mathbf{x}, t)$ is still negligible. Accordingly, at any point in the spatial interval

$$|x_1| \le L_s(t) = (2\psi_0 f_2(x_2) t)^{1/2}, \tag{32}$$

both field components exhibit universal behavior:

$$v_1(\mathbf{x}, t) \approx \frac{x_1}{t}, \quad v_2(\mathbf{x}, t) \approx \psi_0 f_1(0) \frac{\partial f_2(x_2)}{\partial x_2}. \tag{33}$$

In other words, $v_1(\mathbf{x}, t)$ is independent of x_2 and the initial amplitude, and $v_2(\mathbf{x}, t)$ is the initial field on the axis $x_1 = 0$ (independent of x_1): $v_2(\mathbf{x}, t) = v_2(0, x_2, 0)$. Combining (32) with (33), we obtain the component energies

$$E_1(t) \approx \frac{2^{5/2}\psi_0^{3/2}}{3t^{1/2}} \int f_2^{3/2}(x_2) dx \propto E_1(0) \frac{L_1}{(\psi_0 t)^{1/2}}, \tag{34}$$

$$\begin{aligned} E_2(t) &\approx 2^{3/2}\psi_0^{3/2} t^{1/2} \int f_2^{1/2}(x_2) \left(\frac{\partial f_2(x_2)}{\partial x_2}\right)^2 dx \\ &\propto E_2(0) \frac{(\psi_0 t)^{1/2}}{L_1}. \end{aligned} \tag{35}$$

Again, we see that the small-scale component v_1 decays, whereas the large-scale one v_2 grows. When $t \gg t_{\mathrm{nl},2}$, the disturbance becomes isotropic and the velocity profile has universal structure (18) within the domain of $|\mathbf{x}| < (2\psi_0 t)^{1/2}$.

3.2. Evolution of Multiple-Scale Disturbances

Representation (12) of the solution to the Burgers equation implies that asymptotic solution (17), (18) of the form

$$\psi(\mathbf{x}, t) = \psi_0(\mathbf{y}_*) - \frac{(\mathbf{x} - \mathbf{y}_*)^2}{2t}, \quad \mathbf{v}(\mathbf{x}, t) = \frac{\mathbf{x} - \mathbf{y}_*}{t}, \tag{36}$$
$$|\mathbf{x} - \mathbf{y}_*| < L_s(t) = (2Ht)^{1/2}$$

is valid for any localized initial profile with a single maximum $H = \psi_0(\mathbf{y}_*)$. At this stage, the energy of a *d*-dimensional field is expressed as

$$E(t) = \frac{2^{(d+4)/2} H^{(d+2)/2} t^{(d-2)/2}}{\Gamma(d/2)(d+2)}, \tag{37}$$

where $\Gamma(z)$ is the gamma function. According to (37), the energy $E(t)$ decreases with time when $d = 1$, remains constant in the two-dimensional case, and increases with time when $d \ge 3$. Recall that

$$E(t) = \langle (\nabla\psi(\mathbf{x}, t))^2 \rangle = \langle \mathbf{v}^2(\mathbf{x}, t) \rangle$$

is the mean square gradient characterizing the surface roughness in the multidimensional case.

In the case of a multiple-scale localized initial disturbance, the evolution toward an isotropic state may involve several stages. Hereinafter, the evolution of a multiple-scale localized disturbance is analyzed under the assumption that the initial potential can be represented as

$$\psi_0^M(\mathbf{x}) = M(\mathbf{x})\psi_0(\mathbf{x}), \quad M(\mathbf{x}) = 1 - \sum \frac{x_i^2}{2L_{M,i}^2} + \ldots. \tag{38}$$

Here, $\psi_0(\mathbf{x})$ is a statistically homogeneous Gaussian random field characterized by the correlation function

$$\langle \psi_0(\mathbf{x})\psi_0(\mathbf{x} + \rho) \rangle = B_\psi(\rho) = \sigma_\psi^2 \prod_{i=1}^{d} R_i(\rho_i), \tag{39}$$

$$R_i(\rho_i) = 1 - \frac{\rho_i^2}{2! l_{0,i}^2} + \frac{\rho_i^4}{4! l_{1,i}^4} + \ldots. \tag{40}$$

Furthermore, we assume that if $B_\psi(|\rho| > l_{\mathrm{st}}) \approx 0$, where $l_{\mathrm{st}} \sim l_0, l_1$, then the values of the Gaussian field $\psi_0(\mathbf{x})$ at points separated by a distance $|\mathbf{x}_1 - \mathbf{x}_2| > l_{\mathrm{st}}$ are statistically independent. The envelope $M(\mathbf{x})$ attains its maximum at $\mathbf{x} = 0$, and the characteristic lengths $L_{M,i}$ are much greater than the intrinsic length scales $l_{0,i}$.

It can be shown (see [7]) that, when the initial field $\psi_0(\mathbf{x})$ is statistically homogeneous, both velocity field $v(\mathbf{x}, t)$ and potential $\psi(\mathbf{x}, t)$ are isotropic at $t \gg \max(l_{0,i})^2/\sigma_\psi$. The statistical characteristics of the developing turbulence become self-similar, with an integral length $L(t)$ expressed as

$$L(t) = (\sigma_\psi t)^{1/2} d^{-1/4} \left[\ln \frac{\sigma_\psi t}{l_{\mathrm{eff}}^2 (2\pi)^{1/d}}\right]^{-1/4}, \tag{41}$$

where

$$l_{\mathrm{eff}}^d = \prod_{i=1}^{d} l_{0,i}$$

is the effective length scale of the initial field. The energy of each component and the turbulence energy are

$$E_i(t) = \frac{L^2(t)}{t^2} = \frac{\sigma_\psi}{t} d^{-1/2}\left[\ln\frac{\sigma_\psi t}{l_{\text{eff}}^2(2\pi)^{1/d}}\right]^{-1/2}, \qquad E(t) = \sum_i E_i(t). \tag{42}$$

The corresponding mean potential (mean surface height) increases according to the logarithmic law

$$\langle \psi(\mathbf{x}, t)\rangle = d^{1/2}\left[\ln\frac{\sigma_\psi t}{l_{\text{eff}}^2(2\pi)^{1/d}}\right]^{1/2}. \tag{43}$$

At the stage when the integral turbulence scale $L(t)$ is much less than the modulation scale $L_{M,i}$, statistically inhomogeneous field (38) can be analyzed in a quasi-static approximation. In this approximation, when

$$\frac{\min(L_{0,i})^2}{\sigma_\psi} \gg t \gg \frac{\max(l_{0,i})^2}{\sigma_\psi},$$

the integral scale $L(\mathbf{x}, t)$ and energy $E(\mathbf{x}, t)$ of turbulence are expressed by (41) and (42), respectively, whereas the variance of the initial potential contained in these expressions is a slowly varying function of the coordinates: $\sigma_\psi^2 = \sigma_\psi^2 M^2(\mathbf{x})$. This means that the intrinsic structure becomes locally isotropic. The rate of increase in the integral scale is higher in regions of higher field amplitude, $L(\mathbf{x}, t) \approx (\sigma_\psi M(\mathbf{x})t)^{1/2}$. The modulation is partly eliminated by nonlinearity, $E(\mathbf{x}, t) \approx \sigma_\psi M(\mathbf{x})t^{-1}$, whereas $E(\mathbf{x}, 0) = \sigma_\psi^2 M^2(\mathbf{x})$.

At this stage, the field has cellular structure and exhibits universal behavior described by (18) inside each cell. The boundaries of a cell are determined as the intersections of the surface profiles dominated by adjacent local maximums of the function $\psi_0(\mathbf{x})$, $\mathbf{y}_k$ and $\mathbf{y}_m$. These boundaries are planes orthogonal to the vector $\Delta\mathbf{y}_{k,m} = \mathbf{y}_k - \mathbf{y}_m$. Each boundary moves with a constant velocity parallel to the corresponding vector $\Delta\mathbf{y}_{k,m}$, whose magnitude is proportional to the potential difference $\psi(\mathbf{y}_k) - \psi(\mathbf{y}_m)$ between adjacent maximums, toward the cell associated with the lower maximum. When initial field (38) is statistically inhomogeneous, the mean value of a local maximum slowly decreases away from the center of the disturbance. Accordingly, both the cell-boundary velocity and the velocity field $\mathbf{v}(\mathbf{x})$ involve mean components directed away from the point $\mathbf{x} = 0$, and the outer boundary of a localized disturbance has a bubblelike structure. The boundaries between outer and inner cells are planes, whereas the outer boundary consists of spherical segments, $|\mathbf{x} - \mathbf{y}_k| < (2\psi_0(\mathbf{y}_k)t)^{1/2}$. Ultimately, the field consists of a surviving single cell, which is associated with the absolute maximum of the potential, and both potential and velocity have universal structure described by (36).

Consider a velocity field consisting of a large-scale component $\mathbf{v}_l$ and a small-scale one $\mathbf{v}_s$:

$$\mathbf{v}(\mathbf{x}, t) = \mathbf{v}_l(\mathbf{x}, t) + \mathbf{v}_s(\mathbf{x}, t), \quad \mathbf{v}_l(\mathbf{x}, t) = \langle \mathbf{v}(\mathbf{x}, t)\rangle. \tag{44}$$

The angle brackets here denote statistical averages. The Burgers equation for a potential field can be written as

$$\frac{\partial \mathbf{v}}{\partial t} = -\frac{1}{2}(\nabla\mathbf{v})^2 + \nu\Delta\mathbf{v}. \tag{45}$$

Suppose that the evolution of the small-scale component $\mathbf{v}_s$ can be described in the quasi-static approximation. Assuming that both nonlinear distortion and dissipation of the large-scale component are negligible, we average (45) to obtain

$$\frac{\partial \mathbf{v}_l(\mathbf{x}, t)}{\partial t} = -\frac{1}{2}\nabla\langle \mathbf{v}_s^2(\mathbf{x}, t)\rangle = -\frac{1}{2}\nabla E_s(\mathbf{x}, t). \tag{46}$$

Thus, we find that a large-scale coherent component with a nonzero mean value develops in an inhomogeneous random field. Before the small-scale component is distorted by nonlinear effects (at $t \ll t_{\text{nl},s} = \min(l_{1,i}^2/\sigma_\psi)$), the coherent component is determined by the initial energy of the velocity field:

$$\mathbf{v}_l(\mathbf{x}, t) = \frac{t\sigma_v^2}{2}\nabla M^2(\mathbf{x}).$$

At the stage of strongly nonlinear development, when the intrinsic structure becomes locally isotropic, Eqs. (44) and (43) predict logarithmic growth of the large-scale component:

$$\mathbf{v}_l(\mathbf{x}, t) = -\nabla\langle \psi(\mathbf{x}, t)\rangle \approx -\nabla|M(\mathbf{x})|\sigma_\psi d^{1/2}\left(\ln\frac{\sigma_\psi t}{l_{\text{eff}}^2}\right)^{1/2}. \tag{47}$$

When the function $M(\mathbf{x})$ is anisotropic, the mean velocity is also anisotropic at this stage. When the intrinsic length scale $L(\mathbf{x}, t)$ becomes comparable to that of the modulating function $M(\mathbf{x})$, nonlinear distortion of the coherent component $\mathbf{v}_l(\mathbf{x}, t)$ must also be taken into account. Ultimately, isotropic velocity and potential fields develop as a result of collisions (merger) of cells. In the long-time limit, the cell associated with the absolute maximum of the potential absorbs other cells, and the asymptotic solution is described by Eq. (36). For the resulting structure, the value of $L_s(t)$ in (36) and the energy given by (37) are determined only by the absolute maximum H of the initial potential $\psi_0^M(\mathbf{x})$ defined by (38). Figures 2a–2c show the contour maps of

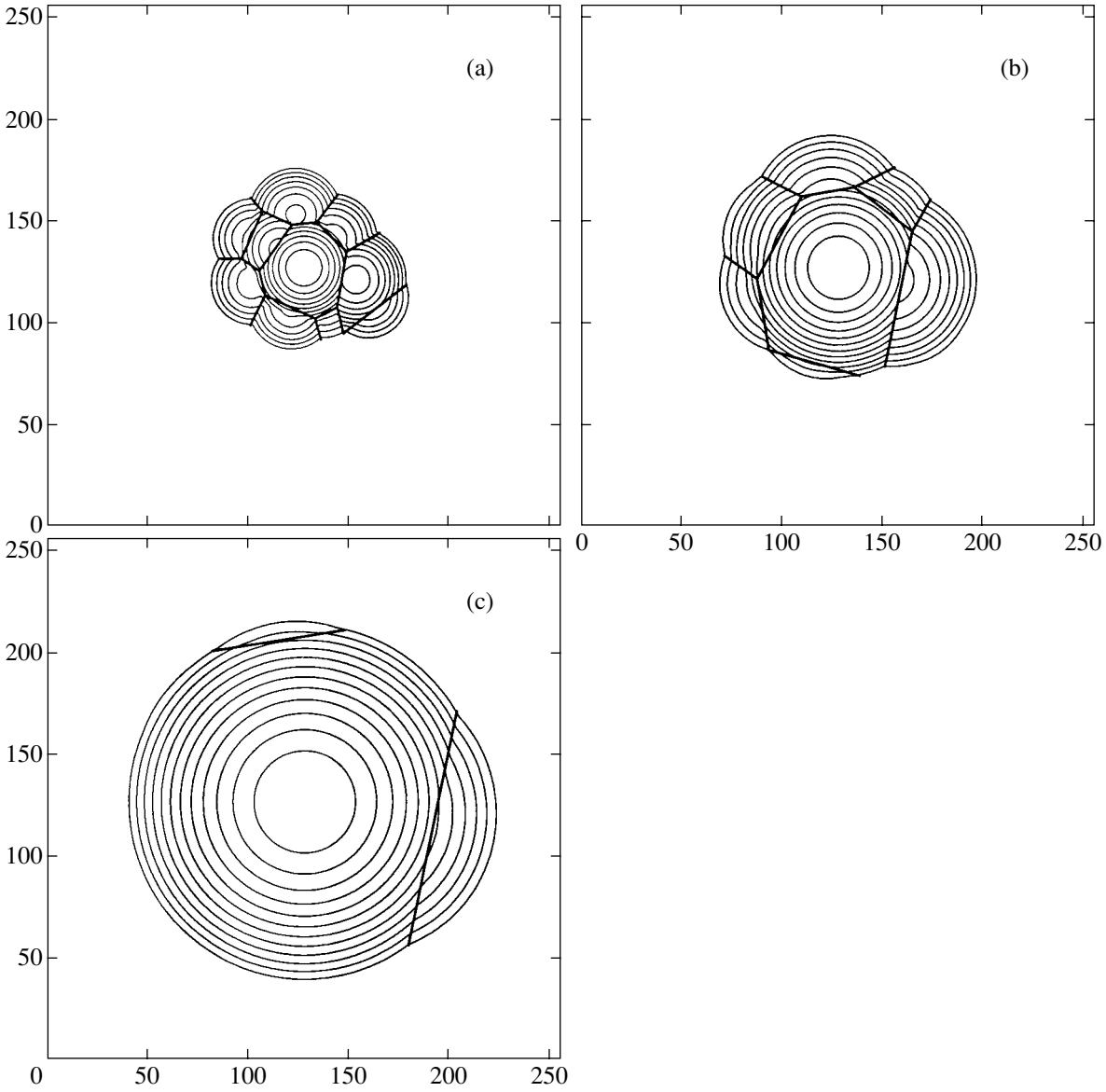

Fig. 2. Contour maps illustrating the nonlinear evolution of a localized disturbance: $t/t_{\rm nl}$ = 10 (a), 30 (b), and 80 (c).

$\psi(\mathbf{x}, t)$ corresponding to $t/t_{\rm nl}$ = 10, 30, and 80, respectively, to illustrate the nonlinear evolution of a two-dimensional localized disturbance. The maps demonstrate the development of isotropic field structure and the geometry of cell boundaries.

Let us now analyze the statistical characteristics of the absolute maximum H of a multiple-scale disturbance, assuming that the length scale characterizing its intrinsic structure is much less than the length scale of the envelope. In the Appendix, it is shown that integral probability distribution function $Q(H, V)$ for the absolute maximum in a volume V can be expressed in terms of the mean number $N(H; V)$ of peaks of the potential ψ_0 that exceed the level H (see (81)):

$$Q(H; V) = \exp(-N(H; V)). \tag{48}$$

The density n of peaks of a statistically homogeneous field is given by (78) (see Appendix). When $L_{M,i} \gg l_i$, we can write a similar expression for the local density $n_{\rm loc}(\mathbf{x})$ in a statistically inhomogeneous field $\psi_0^M(\mathbf{x}) = M(\mathbf{x})\psi_0(\mathbf{x})$ described by (38), with σ_ψ replaced by $M(\mathbf{x})\sigma_\psi$. Then, the average total number $N_\infty(H)$ of peaks of the potential ψ_0^M is

$$N_\infty(H) = \frac{1}{(2\pi)^{(d+1)/2} l_{\rm eff}^d} \times \int \left(\frac{H}{M(\mathbf{x})\sigma_\psi}\right)^{d-1} \exp\left(-\frac{H^2}{2M^2(\mathbf{x})\sigma_\psi^2}\right) d^d\mathbf{x}. \tag{49}$$

If H is sufficiently large, we can apply Laplace's method to Eqs. (38) and (49) to obtain

$$N_\infty(H) = \frac{1}{(2\pi)^{1/2}}\left(\frac{H}{\sigma_\psi}\right)^{d-1}\left(\frac{L_{\text{eff}}^M}{l_{\text{eff}}}\right)^d \exp\left(-\frac{H^2}{2\sigma_\psi^2}\right), \quad (50)$$

where

$$l_{0,\text{eff}}^d = \prod_{i=1}^d l_{0,i}, \quad (L_{\text{eff}}^M)^d = \prod_{i=1}^d L_{M,i}.$$

The ratio $(L_{\text{eff}}^M/l_{\text{eff}})^d = N_{\max}$ in (50) is the number of independent local maximums of an initial disturbance described by (38). In the case considered here ($N_{\max} \gg 1$), we can define the dimensionless potential

$$h = \frac{H}{\sigma_\psi} = h_0\left(1 + \frac{z}{h_0^2}\right), \quad (51)$$

with $h_0 = H_0/\sigma_\psi$, where H_0 solves the equation $N_\infty(H_0) = 1$:

$$H_0 \approx \sigma_\psi\left[2d\ln\left(\frac{L_{\text{eff}}^M}{l_{\text{eff}}}\right)\right]^{1/2}. \quad (52)$$

The dimensionless potential h has a double exponential distribution:

$$Q(z) = \exp[-\exp(-z)],$$
$$Q_h(h) = \exp\{-\exp[-(h-h_0)h_0]\}. \quad (53)$$

When $N_{\max} \gg 1$, integral distribution function (48) for the absolute maximum is localized in a small neighborhood of the point $H_0 = h_0\sigma_\psi$, $\Delta H/H \approx 1/h_0^2 \ll 1$. Thus, expressions (36) and (37) imply that the relative fluctuations of the length scale,

$$\frac{\Delta L_s(t)}{L_s(t)} \approx \frac{1}{2h_0^2}$$

and energy,

$$\frac{\Delta E(t)}{E(t)} \approx \frac{d+2}{2h_0^2}$$

of an isotropic velocity field are weak. By following the calculations presented above in this section, it can be shown that the probability distribution of $\mathbf{y}_k$ for an isotropic field described by (36) is a Gaussian one with the mean square value $\langle y_{k,i}^2\rangle = L_{M,i}^2/h_0^2$. At the stage when $L_s(t) \gg L_{M,i}$, field variations are determined only by relatively small changes in the shock locations $L_s(t)$. Combining Eqs. (36), (51), and (53), we obtain the mean velocity $\mathbf{v}_l(\mathbf{x}, t)$ (see (44)) and variance $\sigma_\mathbf{v}^2(\mathbf{x}, t) = \langle \mathbf{v}_s^2(\mathbf{x}, t)\rangle$:

$$\mathbf{v}_l(\mathbf{x}, t) = \langle \mathbf{v}(\mathbf{x}, t)\rangle = \frac{\mathbf{x}}{t}\left[1 - Q_h\left(\frac{\mathbf{x}^2}{2\sigma_\psi t}\right)\right], \quad (54)$$

$$\sigma_\mathbf{v}^2(\mathbf{x}, t) = \frac{\mathbf{x}^2}{t^2}Q_h\left(\frac{\mathbf{x}^2}{2\sigma_\psi t}\right)\left[1 - Q_h\left(\frac{\mathbf{x}^2}{2\sigma_\psi t}\right)\right]. \quad (55)$$

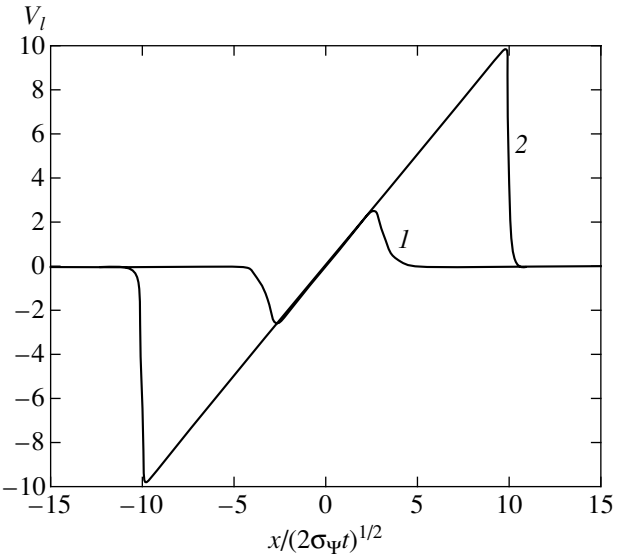

Fig. 3. Dimensionless self-similar mean field $\mathbf{V}_l(x)$ for h_0 = 3 (*1*) and 10 (*2*).

These expressions show that both the mean field and the variance are self-similar functions,

$$\mathbf{v}_l(\mathbf{x}, t) = (2\sigma_\psi t)^{1/2}\mathbf{V}_l\left(\frac{\mathbf{x}}{(2\sigma_\psi t)^{1/2}}\right),$$

$$\sigma_\mathbf{v}^2(\mathbf{x}, t) = 2\sigma_\psi t D_\mathbf{v}\left(\frac{\mathbf{x}}{(2\sigma_\psi t)^{1/2}}\right),$$

completely defined by the parameter $h_0 = H_0/\sigma_\psi$ determined by (52). As the ratio of the characteristic length of the disturbance to the length scale of its intrinsic structure increases, the mean field transforms into an N wave and its dispersion concentrates near the shocks, as illustrated by Figs. 3 and 4. Expression (54) shows that the energy of the coherent component $\mathbf{v}_l(\mathbf{x}, t)$ is given by (37) with $H = h_0\sigma_\psi$ when $N_{\max} \gg 1$. The variance $\sigma_\mathbf{v}^2(\mathbf{x}, t)$ does not vanish only in a small neighborhood ($\Delta L_s(t)/L_s(t) \approx 1/2h_0^2$) of the shock located at $L_s(t) = (2\sigma_\psi h_0 t)^{1/2}$. The energy $E_s(t)$ of the stochastic compo-

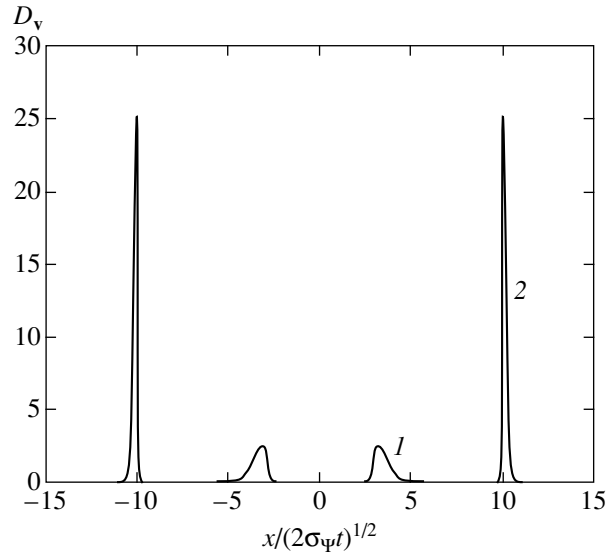

Fig. 4. Dimensionless self-similar variance $D_\mathbf{v}(x)$ for h_0 = 3 (*1*) and 10 (*2*).

nent is much lower than the energy $E_l(t)$ of the mean component $\mathbf{v}_l(\mathbf{x}, t)$ of the field:

$$\frac{E_s(t)}{E_l(t)} \approx \frac{d+2}{h_0^2} \approx \frac{d+2}{d}\left[2\ln\left(\frac{L_{\text{eff}}^M}{l_{\text{eff}}}\right)\right]^{-1} \ll 1. \qquad (56)$$

Thus, nonlinear effects lead to generation of large-scale isotropic coherent structure in the limit of vanishing viscosity.

4. LONG-TIME ASYMPTOTIC BEHAVIOR OF THE FIELD

When the Reynolds number is high, but finite, we use the Cole–Hopf solution to analyze the behavior of the field. Assuming that the correlation length of the Green function of linear diffusion equation (8) is much greater than the length scale of the initial localized disturbance, we solve problem (8), (9) to obtain

$$U(\mathbf{x}, t) = 1 + \frac{B}{(4\pi\nu t)^{d/2}}\exp\left(-\frac{\mathbf{x}^2}{4\nu t}\right), \qquad (57)$$

where

$$B = \int\left(\exp\frac{\psi_0(\mathbf{y})}{2\nu} - 1\right)d^d\mathbf{y}. \qquad (58)$$

When the initial Reynolds number is high ($\text{Re}_0 \sim H/2\nu$, where H is the global maximum of $\psi_0(y)$), the constant B can be expressed as

$$B = L_{\text{eff}}^d \exp\frac{H}{2\nu}. \qquad (59)$$

where L_{eff} is a length scale associated with the integrand in (58). Then, (57) becomes

$$U(\mathbf{x}, t) = 1 + \exp\left[-\frac{1}{2\nu}\left(\frac{\mathbf{x}^2}{2t} - H + \nu d\ln\frac{4\pi\nu t}{L_{\text{eff}}^2}\right)\right]. \qquad (60)$$

In the low-viscosity limit ($\nu \longrightarrow 0$), Eqs. (7) and (60) can be combined to obtain solution (36). However, when the Reynolds number is finite, expression (36) is valid only within a limited time interval. When t is limited and x is sufficiently small, solution (36) holds within the d-sphere

$$|\mathbf{x}| \le L_s(t) = \left[2t\left(H - d\nu\ln\frac{4\pi\nu t}{L_{\text{eff}}^2}\right)\right]^{1/2}. \qquad (61)$$

At finite Reynolds numbers, shock fronts have finite widths: $\delta \sim t^{1/2}$. However, it should be noted that the ratio of the shock width to the shock coordinate, $\delta(t)/L_s(t)$, increases with time because of a logarithmic correction to the shock coordinate in (36). This leads to a decrease in the Reynolds number and dissipation of the nonlinear structure. In the long-time limit, when $B/(4\pi\nu t)^{d/2} \ll 1$, the solution exhibits linear behavior, and the velocity field is found by using expressions (20) and (57):

$$\mathbf{v}(\mathbf{x}, t) = -\frac{B}{(4\pi\nu t)^{d/2}}\nabla\exp\left(-\frac{\mathbf{x}^2}{4\nu t}\right). \qquad (62)$$

In the long-time limit, the evolution of the surface described by $\psi(\mathbf{x}, t)$ is governed by a linear diffusion equation, and the surface has a Gaussian profile with the height $B(4\pi\nu t)^{-d/2}$ and the length scale $(2\nu t)^{1/2}$.

If the initial potential has a single maximum and representation (13) holds, then Eq. (59) becomes

$$B = \left(\frac{4\pi\nu}{\psi_0}\right)^{(d+1)/2}\prod_{i=1}^{d} L_i \exp\frac{\psi_0}{2\nu}.$$

If the initial random fluctuation is Gaussian and representation (38) holds, then the mean value of B defined by (58) is

$$\langle B\rangle = \int\left(\exp\frac{\text{Re}_0^2 M^2(\mathbf{y})}{2} - 1\right)d^d\mathbf{y}, \quad \text{Re}_0 = \frac{\sigma_\psi}{2\nu}. \qquad (63)$$

This result implies that $\langle B\rangle$ is independent of the intrinsic structure of $\psi_0(\mathbf{x})$ and is positive definite, because a

mean field is generated at the stage of nonlinear evolution. The variance $\sigma_B^2 = \langle (B - \langle B \rangle)^2 \rangle$ is expressed as

$$\sigma_B^2 = \iint \exp\left(\frac{\sigma_\psi^2 (M^2(\mathbf{y}) + M^2(\mathbf{y}'))}{8\nu^2}\right) \times \left[\exp\frac{B_\psi(\mathbf{y}-\mathbf{y}')M(\mathbf{y})M(\mathbf{y}')}{4\nu^2} - 1\right] d^d\mathbf{y}\, d^d\mathbf{y}', \quad (64)$$

where $B_\psi(\mathbf{z})$ is the correlation function of a homogeneous initial potential $\psi_0(\mathbf{x})$. In the case of a high initial Reynolds number,

$$\langle B \rangle = \left(\frac{2\pi}{\mathrm{Re}_0}\right)^{(d+1)/2} \prod_{i=1}^{d} L_{M,i} \exp\frac{\mathrm{Re}_0^2}{2} \propto \left(\frac{L_{\mathrm{eff}}^M}{\mathrm{Re}_0^{1/2}}\right)^d \exp\frac{\mathrm{Re}_0^2}{2}, \quad (65)$$

$$\sigma_B^2 \propto \left(\frac{l_{\mathrm{eff}} L_{\mathrm{eff}}^M}{\mathrm{Re}_0}\right)^d \exp(2\mathrm{Re}_0^2). \quad (66)$$

Thus, the field has universal structure (62) with a random amplitude B at the stage of linear evolution. The corresponding ratio of the energies of the random and coherent field components is determined by the relative variance of B:

$$\frac{E_s(t)}{E_l(t)} = \frac{\sigma_B^2}{(\langle B \rangle)^2} \propto \left(\frac{l_{\mathrm{eff}}}{L_{\mathrm{eff}}^M}\right)^d \exp(\mathrm{Re}_0^2). \quad (67)$$

At the stage of linear evolution, this variance is greater by a factor of $\exp(\mathrm{Re}_0^2)$ as compared to that characteristic of the nonlinear stage (see (56)). Nevertheless, when $(l_{\mathrm{eff}}/L_{\mathrm{eff}}^M)^d \ll \exp(-\mathrm{Re}_0^2)$, the scatter in B is relatively narrow and the distribution of B gradually approaches a Gaussian one as the ratio $(l_{\mathrm{eff}}/L_{\mathrm{eff}}^M)^d$ increases. This behavior is analogous to the long-time behavior of a homogeneous field [29].

5. CONCLUSIONS

An analysis of the evolution of multiple-scale random disturbances in interface growth is presented. The interface dynamics is described by the KPZ equation for a potential ψ, and the gradient vector field ($\mathbf{v} = -\nabla\psi$) satisfies the multidimensional Burgers equation. The mean square gradient

$$E(t) = \langle (\nabla\psi(\mathbf{x}, t))^2 \rangle = \langle \mathbf{v}^2(\mathbf{x}, t) \rangle$$

characterizing the surface roughness can decrease or increase with the time elapsed. The relative importance of nonlinearity and diffusion (dissipation) is characterized by the Reynolds number Re_0.

In the limit of vanishing viscosity, nonlinear effects lead to local self-similarity of the velocity and potential fields characterized by partition of a random disturbance into cells in which the fields exhibit universal behavior. As a result of absorption of cells by other cells, a single cell having the highest initial potential survives in the long-time limit. In this limit, the surface developing in a d-dimensional sphere is a paraboloid. Its height H and radius L_s are determined by the maximum H of the initial potential of a localized disturbance. The velocity field is discontinuous across the domain boundary, and the shock amplitude decreases with time. In the case when the length scale characterizing the intrinsic structure is substantially smaller than the disturbance size, the parameters H and L_s of the asymptotic structure are shown to vary slowly between random realizations. This means that a multiple-scale zero-mean random localized disturbance evolves into a virtually deterministic coherent structure. It is shown that both mean field and variance are characterized by self-similar dynamics in the long-time limit.

When the Reynolds number is finite, nonlinear development is eventually followed by linear dynamics. At intermediate times, the shock width is $\delta \sim t^{1/2}$. However, it should be noted that the ratio $\delta(t)/L_s(t)$ increases with time because of a logarithmic correction to the shock coordinate. This leads to decrease in the effective Reynolds number and ensuing dissipation of the nonlinear structure. In the long-time limit, the evolution of the surface is governed by a linear diffusion equation, and the surface has a Gaussian profile with a height $B(4\pi\nu t)^{-d/2}$ such that the mean $\langle B \rangle$ is independent of the intrinsic structure of $\psi_0(\mathbf{x})$ and is positive definite. This is explained by mean-field generation at the stage of nonlinear development. The height fluctuation at the linear stage is much greater as compared to the nonlinear stage, because the time of transition from nonlinear to linear development depends exponentially on the maximum height of the initial disturbance.

ACKNOWLEDGMENTS

We thank U. Frisch, A. Noullez, and A.I. Saichev for helpful discussions. This work was supported by grant Nsh-838.2003.2 under the Program of State Support of Leading Scientific Schools; by the Russian Foundation for Basic Research, project no. 02-02-17374; and under the program "Universities of Russia."

APPENDIX

Statistical Characteristics of Maximums of Inhomogeneous Gaussian Fields

Since the long-time asymptotic behavior of a field is determined by the maximum of the random field ψ_0 whose amplitude exceeds the variance σ_ψ of the initial potential, one can make use of some results borrowed from the theory of extremal processes [7, 27, 35].

Statistics of Gaussian peaks have been thoroughly analyzed in the isotropic, statistically homogeneous case (see [36]). However, analysis of Burgers turbulence requires knowledge of the statistical behavior of the absolute maximums of a statistically inhomogeneous field $\Phi(\mathbf{x}, \mathbf{y}, t)$. First, let us consider the statistics of peaks of a statistically homogeneous field $S(\mathbf{x})$. In the case of a relatively smooth field, it is obvious that the number of peaks exceeding a certain level is asymptotically equal to the number of maximums and extremums higher than this level. Therefore, one should consider the properties of extremums of the field $S(\mathbf{x})$. Assuming that the equation $\nabla S(\mathbf{x}) = 0$ has a unique root $\mathbf{x}_\mathbf{r}$, one can write the following expression for the integral of a multidimensional delta function:

$$\int \delta(\nabla S(\mathbf{x}))d\mathbf{x} = \frac{1}{|J(\mathbf{x}_\mathbf{r})|}, \tag{68}$$

where J is the Jacobian

$$J = J(a_{ij}) = \det(a_{ij}), \quad a_{ij} = \frac{\partial S(\mathbf{x})}{\partial x_i \partial x_j}. \tag{69}$$

By using the properties of the delta function, the following expression can be obtained for the mean number $N(H) = \langle N_{\text{ext}}(H)\rangle$ of extremums that lie in a domain V and exceed H:

$$N(H) = \left\langle \int_V \delta(\nabla S)|J(a_{ij})|E(S-H)d\mathbf{x}\right\rangle, \tag{70}$$

where $E(S)$ is the unit function.

For a statistically homogeneous field, the density of extremums, $n(H) = N(H)/V = \langle N_{\text{ext}}\rangle/V$, is determined by an integral probability distribution function depending on S, its gradient $v_i = \partial S/\partial x_i$, and the tensor $a_{ij} = \partial S/\partial x_i \partial x_j$. For a statistically homogeneous Gaussian field, it holds that

$$W_{S, v_i, a_{ij}} = W_v(v_i)W_{S, a_{ij}}(S, a_{ij}). \tag{71}$$

It follows from (70) and (71) that

$$n(H) = W_v(0)\int_H^\infty \int dS J(a_{ij})W_S(S, a_{ij})da_{ij}. \tag{72}$$

Using the conditional probability density function

$$W_{\text{con}}(a_{ij}/S) = W_S(S, a_{ij})/W_S(S),$$

one obtains

$$n = W_v(0)\int_H^\infty dS W_S(S) \times \int J(a_{ij})W_{\text{con}}(a_{ij}/S)da_{ij}. \tag{73}$$

Suppose that the correlation function of the field $S(\mathbf{x})$ can be represented as

$$B_S(\rho) = \langle S(\mathbf{x})S(\mathbf{x}+\rho)\rangle = \sigma_S^2 \prod_{i=1}^d R_i(\rho_i), \tag{74}$$

$$R_i(\rho_i) = 1 - \frac{\rho_i^2}{2! l_{0,i}^2} + \frac{\rho_i^4}{4! l_{1,i}^4} + \ldots. \tag{75}$$

Then, the Gaussian distribution in (72) is determined by the following constant parameters:

$$\langle a_{ij}\rangle_S = \frac{S\langle a_{ij}S\rangle}{\sigma_S^2}, \quad \langle a_{ij}S\rangle = -\frac{\delta_{ij}\sigma_S^2}{l_{0,i}},$$

$$\langle a_{ij}^2\rangle = \frac{\sigma_S^2}{l_{0,i}^2 l_{0,j}^2}, \quad i \neq j, \quad \langle a_{ii}^2\rangle = \frac{\sigma_S^2}{l_{1,i}^4}, \tag{76}$$

$$W_v(0) = \left(\frac{l_{0,\text{eff}}^2}{2\pi\sigma_\psi^2}\right)^{d/2}, \quad l_{0,\text{eff}}^d = \prod_{i=1}^d l_{0,i},$$

where $l_{0,\text{eff}}$ is an effective length scale.

The asymptotic behavior of $n(H)$ corresponding to high H is approximately characterized by

$$\langle J(a_{ij})\rangle_S \approx J(\langle a_{ij}\rangle_S) \approx \prod_{i=1}^d \langle a_{ii}\rangle_S = \frac{S^d}{l_{0,\text{eff}}^{2d}}. \tag{77}$$

Combining Eqs. (73)–(77), one obtains a final expression for the density of maximums:

$$n_{\text{ext}}(H) = W_v(0)\int_H^\infty dS W_S(S)\frac{S^d}{l_{\text{eff}}^{2d}} = \frac{1}{(2\pi)^{(d+1)/2} l_{\text{eff}}^d}\int_{H/\sigma_S}^\infty S^d e^{-S^2/2}dS \approx \left(\frac{H}{\sigma_S}\right)^{d-1}\frac{1}{(2\pi)^{(d+1)/2} l_{\text{eff}}^d}\exp\left(-\frac{H^2}{2\sigma_S^2}\right). \tag{78}$$

This expression implies that the mean number of maximums of an anisotropic field $S(x)$ depends on the effective length scale l_{eff} defined as the geometric mean of $l_{0,i}$ (see (76)). When H is relatively high, the density of extrema given by (78) is equivalent to the density of upward crossings of H by the random field $S(\mathbf{x})$.

The probability density function for the absolute maximum of S in a domain V can be found by a method analogous to that used in the one-dimensional case [7]. For a constant volume V and high values of H, one can assume that the possibility of repeated upward crossings of H by $S(\mathbf{x})$ is negligible and there is a single extremum of this kind in the domain. Then, the following expression can be obtained for the mean number $\langle N(H, V)\rangle$ of upward crossings of H by the field $S(\mathbf{x})$ or the mean number $\langle N_{\text{ext}}(H, V)\rangle$ of extrema higher than H:

$$\langle N(H, V)\rangle \approx \langle N_{\text{ext}}(H, V)\rangle \approx P(1; H, V), \qquad (79)$$

where $P(M; H, V)$ is the probability that the number of upward crossings of H by S in V is M.

Assuming that $V \gg l_{\text{st}}^d$, where l_{st}^d is the correlation distance for $S(\mathbf{x})$, one can partition the domain V into physically small subdomains dV_k ($l_{\text{st}}^d \approx dV_k$). The probability that a function $S(\mathbf{x})$ does not exceed H in dV_k can be expressed as

$$dP_k = 1 - P(1; H, dV_k) = 1 - \langle N(H, dV_k)\rangle, \quad (80)$$

where $\langle N(H, dV_k)\rangle \ll 1$. Since $l_{\text{st}}^d \approx dV_k$, the events taking place in different subdomains dV_k are statistically independent. Therefore, the integral probability distribution for the absolute maximum of $S(\mathbf{x})$ in V is

$$\begin{aligned} Q(H, V) &= P(S(\mathbf{x}) < H, \mathbf{x} \in V) \\ &= \prod_k dP_k = \prod_k (1 - \langle N(H, dV_k)\rangle) \\ &= \exp(-N(H, V)). \end{aligned} \qquad (81)$$

The function $Q(H, V)$ is equivalent to the joint probability of the absence of upward crossings of H by $S(x)$ in V and the absence of extremums exceeding H for high H. Expression (81) reflects a well-known fact in the theory of Poisson distribution of peaks [35]. When H is high, $N(H, V) = Vn(H)$ for a statistically homogeneous field, where n is given by (78). When the field is inhomogeneous, the expression for N is much more complicated even in the one-dimensional case. However, one may assume that the inhomogeneity is due to a bias $\Phi_\alpha(\mathbf{x}) = S(\mathbf{x}) - \alpha(\mathbf{x})$ or a modified variance $\Phi_M(\mathbf{x}) = S(\mathbf{x})M(\mathbf{x})$, where both $\alpha(\mathbf{x})$ and $M(\mathbf{x})$ are sufficiently smooth functions (over the scales characterizing $S(\mathbf{x})$). Then, a quasi-static approximation can be invoked. In particular, the mean number of upward crossings of H by $\Phi_M(\mathbf{x})$ in a domain V is expressed as

$$N(H, V) = \int_V n_{\text{ext}}\left(\frac{H}{M(\mathbf{x})}\right)dV, \qquad (82)$$

where n is the density of extremums of a uniform function $S(\mathbf{x})$ given by (78).

REFERENCES

1. J. M. Burgers, K. Ned. Akad. Wet. Verh. **17**, 1 (1939).
2. J. M. Burgers, *The Nonlinear Diffusion Equation* (Reidel, Dordrecht, 1974).
3. U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ. Press, Cambridge, 1995).
4. R. Kraichnan, Phys. Fluids Mech. **11**, 265 (1968).
5. G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974; Mir, Moscow, 1977).
6. O. V. Rudenko and S. I. Soluyan, *Theoretical Foundations of Nonlinear Acoustics* (Nauka, Moscow, 1975; Consultants Bureau, New York, 1977).
7. S. N. Gurbatov, A. N. Malakhov, and A. I. Saichev, *Nonlinear Random Waves in Nondispersion Media* (Nauka, Moscow, 1990).
8. W. A. Woyczynski, *Burgers–KPZ Turbulence. Gottingen Lectures* (Springer, Berlin, 1998).
9. A. Cheklov and V. Yakhot, Phys. Rev. E **52**, 5681 (1995).
10. A. M. Polyakov, Phys. Rev. E **52**, 6183 (1995).
11. W. E. Khanin, A. Mazel, and Ya. G. Sinai, Phys. Rev. Lett. **78**, 1904 (1997).
12. S. A. Boldyrev, Phys. Rev. E **59**, 2971 (1999).
13. J. Davoudi, A. A. Masoudi, M. R. R. Tabar, *et al.*, Phys. Rev. E **63**, 056308 (2001).
14. S. N. Gurbatov and A. I. Saichev, Izv. Vyssh. Uchebn. Zaved., Radiofiz. **27**, 4 (1984).
15. S. N. Gurbatov, A. I. Saichev, and S. F. Shandarin, Mon. Not. R. Astron. Soc. **236**, 385 (1989).
16. S. F. Shandarin and Ya. B. Zeldovich, Rev. Mod. Phys. **61**, 185 (1989).
17. M. Vergassola, B. Dubrulle, U. Frisch, and A. Noullez, Astron. Astrophys. **289**, 325 (1994).
18. A.-L. Barabási and H. E. Stanley, *Fractal Concepts in Surface Growth* (Cambridge Univ. Press, Cambridge, 1995).
19. M. Kardar, G. Parisi, and Y. C. Zhang, Phys. Rev. Lett. **56**, 889 (1986).
20. J.-P. Bouchaud, M. Mézard, and G. Parisi, Phys. Rev. E **52**, 3656 (1995).
21. E. Hopf, Commun. Pure Appl. Mech. **3**, 201 (1950).
22. J. D. Cole, Q. Appl. Math. **9**, 225 (1951).
23. L. Frachebourg and Ph. A. Martin, J. Fluid Mech. **417**, 323 (2000).
24. S. Kida, J. Fluid Mech. **93**, 337 (1979).

25. S. N. Gurbatov and A. I. Saichev, Zh. Éksp. Teor. Fiz. **80**, 689 (1981) [Sov. Phys. JETP **53**, 347 (1981)].
26. J. D. Fournier and U. Frisch, J. Méc. Théor. Appl. **2**, 699 (1983).
27. S. A. Molchanov, D. Surgailis, and W. A. Woyczynski, Commun. Math. Phys. **168**, 209 (1995).
28. E. A. Kuznetzov and S. S. Minaev, Phys. Lett. A **221**, 187 (1996).
29. S. Albeverio, A. A. Molchanov, and D. Surgailis, Probab. Theory Relat. Fields **100**, 457 (1994).
30. S. N. Gurbatov, Phys. Rev. E **61**, 2595 (2000).
31. S. N. Gurbatov, B. O. Enflo, and G. V. Pasmanik, Acta Acust. (China) **85**, 181 (1999); Acta Acust. (China) **87**, 16 (2001).
32. Z. S. She, E. Aurell, and U. Frisch, Commun. Math. Phys. **148**, 623 (1992).
33. S. N. Gurbatov and G. V. Pasmanik, Zh. Éksp. Teor. Fiz. **115**, 564 (1999) [JETP **88**, 309 (1999)].
34. S. N. Gurbatov, S. I. Simdyankin, E. Aurel, *et al.*, J. Fluid Mech. **344**, 339 (1997).
35. M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes* (Springer, Berlin, 1983).
36. J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay, Astrophys. J. **304**, 15 (1986).

Translated by A. Betev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1201–1211.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1345–1357.

SOLIDS
Structure

Investigation of Roughness Cross Correlation in a Ni/C Multilayer Mirror by X-ray Diffuse Scattering Method

N. V. Kovalenko[a], S. V. Mytnichenko[b], and V. A. Chernov[c]

[a]*Budker Institute of Nuclear Physics, Siberian Division, Russian Academy of Sciences, pr. Akademika Lavrent'eva 11, Novosibirsk, 630090 Russia*

[b]*Institute of Solid-State Chemistry and Mechanochemistry, Siberian Division, Russian Academy of Sciences, Novosibirsk, 630128 Russia*

[c]*Boreskov Institute of Catalysis, Siberian Division, Russian Academy of Sciences, pr. Akademika Lavrent'eva 5, Novosibirsk, 630090 Russia*

e-mail: s.v.mytnichenko@inp.nsk.su

Received April 24, 2003

Abstract—The possibility of applying X-ray diffuse scattering for studying roughness in multilayer X-ray mirrors, including the correlation of roughnesses of neighboring interfaces (roughness cross-correlation) is considered. It is shown that the reliability and informativeness of this method can be improved by rejecting the classical experimental schemes and using alternative schemes in which not only the intensity of diffuse scattering itself, but also its dependence on certain experimental parameters (conditions), vary. Such parameters can be the spatial coherence of incident radiation, the direction of the momentum transfer relative to the specular diffraction plane, or the X-ray wavelength. In the framework of this approach, the results of comparative measurements of diffuse scattering from a Ni/C multilayer X-ray mirror prepared by laser ablation are considered for two close values of photon energy: below (8.325 keV) and above (8.350 keV) the *K* absorption edge for nickel. It is shown that, in view of effective screening of deep layers in the hard photoabsorption mode, this method provides more reliable (as compared to the standard diffuse scattering method) information on the evolution of interfaces between the layers. It is found that the smoothing of roughness in the experimental sample occurs over large spatial scales such as the micrometer scale. Only large-scale defects with a size exceeding 10 μm are replicated well from layer to layer. Possible physical reasons for the observed effect are considered. It is shown that effective smoothing on the micrometer and submicrometer spatial scales is of fundamental importance for preparing multilayer X-ray mirrors with high reflectances. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

X-ray diffuse scattering (XRDS) is widely used at present as a method for studying the structure of the surface and internal interfaces in multilayer X-ray mirrors. This nondestructive experimental technique, which makes it possible, in particular, to carry out in situ investigations, indeed provides structural information averaged over the sample surface in wide (from atomic to macroscopic) limits of lateral spatial dimensions of roughnesses. In addition, deep penetration of X-ray photons to the bulk of a multilayer X-ray mirror makes it possible to determine the degree of correlation between the roughnesses of neighboring interfaces from the angular distributions of the diffuse scattering intensity.

Unfortunately, the advantages listed above have a reverse side. As a rule, the transverse coherence length of an incident X-ray beam is much smaller than its physical size, and the XRDS data turn out to be averaged over an infinitely large ensemble of spatially coherent beams. In the process of such a total averaging, the information on individual parameters of the structure of interfaces between the layers in the sample is completely lost. As a result, the Gaussian functions are found to be quite suitable for describing roughness correlation in lateral directions and are widely used for calculating the angular distribution of the XRDS intensity [1–3],

$$C(r) = \sigma^2 \exp[-(r/\xi)^{2h}], \qquad (1.1)$$

where σ is the roughness dispersion, ξ is the characteristic correlation length, and h ($0 \le h \le 1$) is a parameter characterizing the fractal properties of interfaces between the layers [3]. It is difficult to enumerate all assumptions (which are often unrealistic) used in substantiating the possibility of application of expression (1.1). First, roughness properties are assumed to be isotropic in the lateral directions. Although this assumption can be justified for averaging over the entire surface area of the sample, it may be invalid on the spatial scale of coherently illuminated areas. Second, roughness defects are assumed to be pointlike; i.e., $C(r) \longrightarrow 0$ as

$r \longrightarrow \infty$. This may also be incorrect in the presence of extended scratches, roughness waves, terraced steps, or mosaic blocks [4]. Third, it is assumed that the spatial frequency spectrum of roughness has a uniform distribution. In fact, this spectrum may have singularities associated with concrete physicochemical or technological factors [5]. All these singularities are disregarded when expression (1.1) is used for describing XRDS.

Parameter h from expression (1.1) weakly affects the angular distribution of the XRDS intensity unless it assumes its limiting values. If quartz substrates are used, the value of this parameter, $h \approx 1/2$, describes the experimental data quite satisfactorily [6]. The only integrated parameter ξ that can be obtained only from the XRDS data[1] characterizes the rough structure of interfaces between the layers in lateral directions. However, this parameter is also difficult for interpretation. As a matter of fact, statistical quantities σ and ξ are determined not only by the actual properties of the interfaces, but also by the sizes of the area elements over which averaging is carried out in determining the values of these quantities. These sizes are in turn determined by spatial coherence of incident X-rays. As spatial coherence of the incident wave increases (when synchrotron sources are used), a situation may occur when parameter ξ reflects not the real properties of roughness, but rather the experimental conditions of measurements [7]. Thus, expression (1.1) describes the method of averaging over the ensemble of coherent beams rather than the actual properties of roughness of interlayer interfaces. In this case, only relative changes in parameter ξ, as compared to a certain standard sample under the same experimental conditions, provide more reliable information. Thus, the interest in XRDS is due not to the high informativeness of this method, but rather due to the lack of alternative methods of investigation. Indeed, microscopy of the cross cut cannot provide reliable data on the roughness of a multilayer X-ray mirror with periods smaller than 5 nm. Atomic force microscopy may provide extensive and important structural information only on the surface, but not on internal interfaces.

Nevertheless, the difficulties mentioned above are not fundamental. The XRDS method can be made more informative by using more sophisticated setups for measurements, in which not only the XRDS intensity itself, but also its dependence on certain experimental conditions, is measured. Such conditions include the spatial coherence of incident radiation [7]; the direction of the momentum transfer relative to the specular diffraction plane, which is determined by the wave vectors of incident and specularly reflected radiation [4]; and the wavelength of X-rays. Another example is a combination of the XRDS method and the extended X-ray absorption fine structure (EXAFS) method, which involves the measurement of the extended fine structure of the XRDS intensity as a function of the photon energy beyond the absorption edge of the atoms constituting a multilayer X-ray mirror and provides information on the atomic structure (the atomic radial distribution function) at the interfaces [8].

The evolution of the profiles of interlayer interfaces during the growth of a multilayer X-ray mirror is a structural problem of considerable interest. This is due to the fact that structural information of this kind gives an idea of the mechanism of physicochemical phenomena occurring during the growth of the mirror and, hence, makes it possible to optimize the growth technology to improve the optical parameters of such a mirror. In addition, the possibility of smoothing the interlayer roughness during the growth of a multilayer X-ray mirror is of considerable interest also. It is generally accepted that such smoothing occurs in the case of magnetron sputtering [9] and laser ablation [10] as well as during thermal sputtering with ionic polishing [11–14] since it is these methods that make it possible to prepare multilayer X-ray mirrors with acceptable optical parameters. It should be noted that processes of smoothing or, conversely, increasing roughness were observed experimentally with the help of electron microscopy of the transverse cut only for multilayer thin films with a period larger than 10 nm. However, this method is useless as a rule for mirrors with a period smaller than 5 nm.

In spite of the fact that potentialities of XRDS as a method for studying the cross-correlation of roughnesses in a multilayer X-ray mirror through simulation of angular intensity distributions are widely discussed in the literature, real potentialities of this method are quite limited (at least, when conventional experimental setups are used). The difficulties arising in this method are due to the fact that a coherent replication of rough interfaces from layer to layer leads to resonant enhancement of XRDS, generating the so-called quasi-Bragg band [15–18] under the modified Wulf–Bragg condition,

$$\lambda = \Lambda(\sin\theta_0 + \sin\theta_1) = 2\Lambda\sin\theta_B, \qquad (1.2)$$

where λ is the X-ray photon wavelength, Λ is the period of a multilayer X-ray mirror, θ_0 and θ_1 are the angles of incidence and scattering relative to lateral planes, respectively (Fig. 1), and θ_B is the Bragg angle. Condition (1.2) is precisely the condition for the emergence of a diffraction maximum for scattering from a grating whose reciprocal vector coincides with the reciprocal vector of the multilayer X-ray mirror [19]. Indeed, the emergence of quasi-Bragg XRDS is associated with the fact that a roughness defect, being replicated from layer to layer, forms a "grating" (Fig. 2). The intensity of XRDS from roughnesses correlated from layer to layer is proportional to N^2, where N is the number of bilayers in the X-ray mirror; the intensity of XRDS from non-

[1] The roughness dispersion can be estimated more easily from the data on specular reflection.

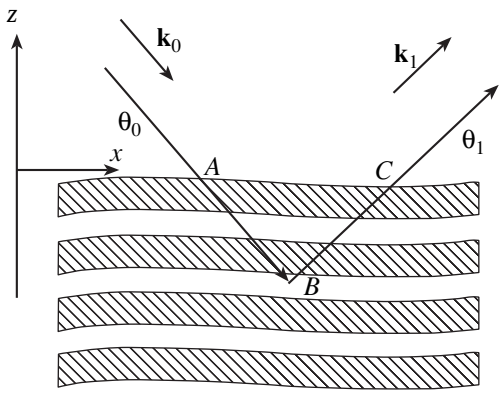

Fig. 1. XRDS experimental geometry. The *z* axis is directed along the normal to the lateral planes, the *x* axis is parallel to the lateral planes and the plane of specular diffraction determined by the wave vectors of the incident and specularly reflected waves, and the *y* axis is perpendicular to the specular diffraction plane; $\mathbf{k}_0$ and $\mathbf{k}_1$ are the wave vectors of the incident wave and the diffusely scattered wave, respectively; θ_0 and θ_1 are the angles between the wave vectors and the lateral planes. The azimuth angle, i.e., the angle between the projection of wave vector $\mathbf{k}_1$ onto the lateral plane and the *x* axis, will be denoted by φ.

correlated roughnesses making an equivalent contribution to roughness dispersion σ is proportional to N.

At first glance, such an enhancement of the intensity due to the coherent replication of roughnesses makes it possible to study directly the extent of their cross-correlation. Indeed, if a roughness defect is replicated not throughout the entire stack, but only through several layers, the quasi-Bragg scattering width Δq_z must increase as compared to the width of specular Bragg reflection (Fig. 3). By measuring the quasi-Bragg scattering width Δq_z as a function of lateral momentum transfer q_x, one can obtain information on the extent of cross-correlation depending on the spatial size of a roughness defect [18]. However, this approach is inapplicable, as a rule, in the case of a multilayer X-ray mirror with a small total thickness ($T \le 1$ μm). Indeed, the characteristic cross-correlation length ξ_z can be determined from the width of the quasi-Bragg band only if $\xi_z \ll T$. However, in the case of a multilayer X-ray mirror, the opposite situation is observed as a rule; i.e., $\xi_z \ge T$, and the quasi-Bragg bandwidth is approximately equal to the width of the Bragg peak irrespective of the momentum transfer q_x (at least, for moderate values of this quantity) [20]. Consequently, violations of complete cross-correlation of roughness affect the XRDS intensity relatively weakly not only in the case of a multilayer X-ray mirror, but also for other multilayer thin films with a small total thickness (of fractions of a micrometer).

Attempts to overcome these difficulties were made, for example, in [21, 22], where XRDS was studied in the vicinity of Kiessig modulations. Since the maximal difference in the behavior of interfaces is observed between the surface of an X-ray mirror and the interface between the mirror and the substrate, such an approach improves the sensitivity of the XRDS method to violations of complete roughness cross-correlation. Rendering this approach its due, we must note that both the surface of a multilayer X-ray mirror and the interface between this mirror and the substrate are unique by nature and their behavior may differ considerably from the behavior of internal interfaces. More reliable data on the nature of cross-correlation of roughness can also be obtained by studying an X-ray mirror with a small number of layers, in which the enhancement of XRDS due to coherent replication of the interfaces is insignificant [23]. Naturally, the class of objects accessible for investigation becomes much smaller in this case.

In our previous study [24], we used another modification of the XRDS method, which makes it possible to considerably increase the potential for studying the cross-correlation of roughness. The proposed method is based on comparative measurement of the XRDS intensity at two photon energies: slightly lower and slightly higher than the photoabsorption edge for atoms constituting a multilayer X-ray mirror. In the former case, the amplitudes of diffuse scattering from rough interfaces are approximately identical over the mirror volume, while in the latter case the lower interfaces are effec-

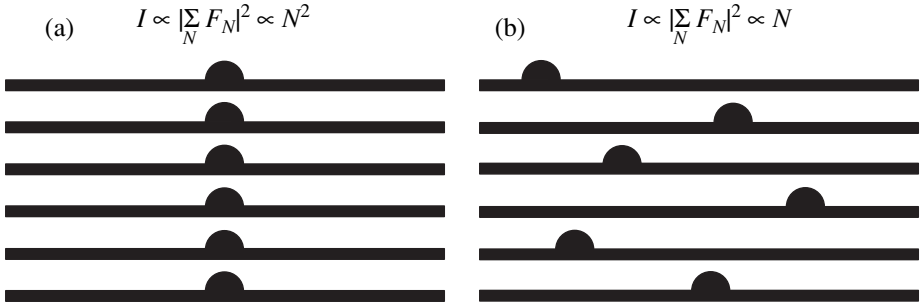

Fig. 2. Coherent replication of roughnesses from layer to layer (a) increases the XRDS intensity in proportion to N^2 in accordance with condition (1.2), while the XRDS intensity from the same number of noncorrelated roughnesses (b) is proportional to N. Both cases are characterized by identical contributions to roughness dispersion σ.

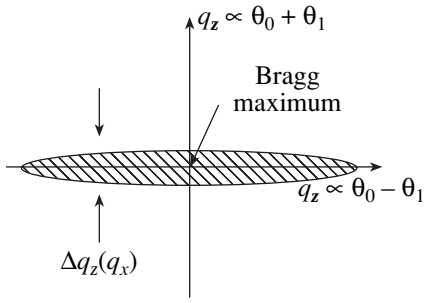

Fig. 3. In the diffraction space, the quasi-Bragg band is arranged along the q_x axis and passes through the Bragg maximum. The q_z axis in the figure corresponds to specular θ–2θ measurements. The quasi-Bragg bandwidth along this axis must increase with the modulus of q_x since the smaller the lateral size of the roughness, the worse its replication from layer to layer [18].

tively screened due to strong photoabsorption. The measurements of the relative difference in the XRDS intensities in the former and latter cases make it possible to considerably improve the sensitivity of experiment to violations of complete cross-correlation of roughnesses. Moreover, by carrying out such measurements for various values of the lateral component of the momentum transfer, it is possible to study the behavior of cross-correlation depending on the spatial sizes of roughnesses. We used this method to observe experimentally the smoothing of roughnesses on the micrometer spatial scale in a Ni/C multilayer X-ray mirror prepared with the help of laser ablation. In this article, the theory and experimental results obtained by using this method are represented in greater detail.

2. THEORY

Under the conditions of hard photoabsorption, it is necessary to take into account the attenuation of an X-ray wave in the sample. In addition, small values of incidence angles necessitate the inclusion of the refraction effect. All this can be taken into account by using the distorted wave Bohr approximation (DWBA) [3, 25] instead of the conventional Born approximation. However, we will disregard specular reflection since the angles of incidence and scattering will be assumed to differ from Bragg's angles. In this approximation, the expression for the XRDS amplitude has the form

$$f(\mathbf{Q}) = r_0 \int \Delta\rho(\mathbf{r}) T_0(z) T_1(z) \exp(-i\mathbf{Q}\cdot\mathbf{r}) d\mathbf{r}, \quad (2.1)$$

where integration is carried out over the volume of the X-ray mirror; r_0 is the classical electron radius; $\Delta\rho(\mathbf{r})$ is the correction to the electron density of the X-ray mirror at point $\mathbf{r}$ due to the presence of roughnesses; $T_0(z)$ and $T_1(z)$ are the amplitudes of transmitted waves for angles of incidence θ_0 and θ_1, respectively; and $\mathbf{Q} = \mathrm{Re}(\mathbf{K}_0 - \mathbf{K}_1)$, $\mathbf{K}_0$ and $\mathbf{K}_1$ being the wave vectors of the incident and scattered waves, respectively. Strictly speaking, wave vectors $\mathbf{Q}$, $\mathbf{K}_0$, and $\mathbf{K}_1$ are functions of coordinate z; however, in the approximation of slowly varying amplitudes [26], we assume that these quantities are constant and can be calculated proceeding from the average refractive index $n = 1 - \delta$ of the X-ray mirror,

$$K_{x0,1} = k_{x0,1}, \quad K_{y0,1} = k_{y0,1},$$

$$K_{z0,1} = k\sqrt{\sin^2\theta_{0,1} - 2\delta} \approx k_{z0,1}\left(1 - \frac{\delta}{\theta_{0,1}}\right),$$

where $\mathbf{k}_0$ and $\mathbf{k}_1$ are the corresponding wave vectors in vacuum. Taking into account the corrections for the refraction effect, we find that condition (1.2) is transformed into the condition

$$\lambda = \Lambda\left(\sqrt{\sin^2\theta_0 - 2\delta} + \sqrt{\sin^2\theta_1 - 2\delta}\right) \approx \Lambda(\theta_0 + \theta_1)\left(1 - \frac{\delta}{\theta_0\theta_1}\right). \quad (2.2)$$

It can easily be seen that the effect of screening of interlayer interfaces in the vicinity of the substrate due to photoabsorption is described in expression (2.1) by the factor

$$T_0(z)T_1(z) \approx \exp\left[\frac{\mu}{2}\left(\frac{1}{\Theta_0} + \frac{1}{\Theta_1}\right)z\right],$$

where $\mu = 2\mathrm{Im}K$ is the linear photoabsorption coefficient and angles Θ_0 and Θ_1 correspond to angles θ_0 and θ_1 on account of the refraction effect. Indeed, the amplitude of scattering from the roughness located at point B in Fig. 1 decreases due to the factor $\exp[-\mu(r_{AB} + r_{BC})/2]$, where r_{AB} and r_{BC} are the distances from point A to point B and from point B to point C, respectively.

We assume that the electron density at the interfaces varies smoothly rather than at a jump. Such an assumption is justified at least in the case of Ni/C multilayer X-ray mirrors prepared using laser ablation [20, 27–29]. It should be noted that, in the approximation of slowly varying amplitudes, the computational method used here is valid irrespective of the form of variation of the electron density at the interfaces, including sharp interfaces. In the latter case, however, the amplitudes can be calculated more accurately by using the Fresnel coefficients.

Suppose that the interlayer roughness is completely conformal. Then the behavior of all interlayer interfaces can be described by the same function $\Delta z(x, y)$, and the electron density $\Delta\rho(\mathbf{r})$ can be represented in the form

$$\Delta\rho(x, y, z) = \rho(z + \Delta z(x, y)) - \rho(z),$$

where $\rho(z)$ is the average electron density along the z axis. It is natural to expand this expression in the small parameter $\Delta z(x, y)$ using the condition of smallness of the roughness dispersion as compared to the period of the mirror:

$$\Delta\rho(x, y, z) \approx \Delta z(x, y)\frac{d\rho(z)}{dz}. \quad (2.3)$$

Proceeding from the general physical considerations, we can describe the evolution of the profiles of the interfaces by using the replication factor $a(x, y)$, i.e., a function whose convolution makes it possible to determine the profiles of the interfaces of the next bilayer from those of the previous one [18, 30]. In the most general case, this can be written in the form

$$\Delta\rho_n(\mathbf{r}) = h_n(\mathbf{r}) + a_n(x, y) * \Delta\rho_{n-1}(\mathbf{r}),$$

where $\Delta\rho_n(\mathbf{r})$ is the roughness electron density in the nth bilayer, $h_n(\mathbf{r})$ is its own roughness, and $a_n(x, y)$ is the replication factor corresponding to this bilayer. However, we will use several simplifying assumptions in the subsequent analysis.

First, we will neglect the quantity $h_n(\mathbf{r})$ leading to an increase in the roughness during the growth of the multilayer X-ray mirror. This simplification may obviously become unjustified for large values of transferred momenta q_x corresponding to small lateral dimensions of the roughness. The criterion for inapplicability of this assumption is the increase in the quasi-Bragg bandwidth in the direction of q_z as compared to the width of specular reflection. It was mentioned above that this effect is rather weak in the case of a multilayer X-ray mirror.

Second, we will assume that the behavior of function $\Delta\rho_n(\mathbf{r})$ within a bilayer is completely conformal, which allows us to use expression (2.3) in the integration. At first glance, this assumption appears unjustified. As a matter of fact, continuous smoothing is replaced by stepwise smoothing, which obviously cannot lead to a large error. The application of formula (2.3) automatically suggests that the "structural factor" for roughnesses in a bilayer coincides with the structural factor of the bilayer to within a phase factor. On the other hand, there are indications that roughnesses on the alternative interfaces A/B and B/A may differ [29, 31, 32]; i.e., replication factors $a_{A/B}(x, y)$ and $a_{B/A}(x, y)$ may be different. In fact, these differences can be significant only for microscopic roughnesses, which lead to the formation of a mixed layer. It should be emphasized, however, that quantity $\Delta\rho(\mathbf{r})$ is a correction to the electron density due to the violation of translational symmetry in lateral directions and not due to the difference from the electron density of a certain ideal multilayer X-ray mirror with sharp interfaces. Thus, the fact that the alternative interfaces may have mixed layers of different thickness does not rule out the application of formula (2.3).[2]

The third assumption is that roughnesses are smoothed uniformly over the entire thickness of the mirror, i.e.,

$$a_n(x, y) \equiv a(x, y). \quad (2.4)$$

This assumption may be erroneous. A situation is possible where smoothing occurs just in a few layers near the substrate. Nevertheless, the application of formula (2.4) is justified in the sense that this effect changes the XRDS intensity insignificantly (see above).

The above assumptions make it possible to represent the additional electron density within a bilayer with number n in the following form:

$$\Delta\rho_n(x, y, z) = [\underbrace{a(x, y) * \ldots * a(x, y)}_{n} * \Delta z_0(x, y)]\frac{d\rho(z)}{dz}. \quad (2.5)$$

In our subsequent computations, we represent the integral in formula (2.1) as the sum of two integrals over bilayers, removing slowly varying factors $T_0(z)$ and $T_1(z)$ from the integrand,

$$f(\mathbf{Q}) = r_0\sum_n T_0(z_n)T_1(z_n)\int_{V_n}\Delta\rho_n(\mathbf{r}) \times \exp(-i\mathbf{Q}\cdot\mathbf{r})d\mathbf{r}, \quad (2.6)$$

where integration is carried out over the corresponding volumes V_n of bilayers. Using the properties of the Fourier transform (namely, the fact that the Fourier transform of the convolution of functions is equal to the product of their Fourier transforms and that the Fourier transform of the derivative of a function is equal to its own Fourier transform to within a constant factor), we can simplify expression (2.6) by substituting expression (2.5) into it:

$$f(\mathbf{Q}) = -ir_0Q_zF(Q_z)F_{xy}(\mathbf{s}) \times \sum_n T_0(z_n)T_1(z_n)a^n(\mathbf{s}), \quad (2.7)$$

where

$$F(Q_z) = \int_{-\Lambda/2}^{\Lambda/2}\rho(z)\exp(-iQ_zz)dz,$$

$$F_{xy}(\mathbf{s}) = \iint\Delta z_0(x, y)\exp(-iQ_xx - iQ_yy)dxdy,$$

[2] It can be seen from the above discussion that the only possible way of studying alternative layers by the XRDS method is to use the effect of standing waves.

$\mathbf{s} = (Q_x, Q_y)$ is the projection of the momentum transfer onto lateral planes and $F(Q_z)$ is the structural factor of a bilayer of an X-ray mirror; we also assume that resonance condition (2.2) is satisfied and the corresponding phase factors in the sum are omitted. If we set $T_0(z_n) = T_1(z_n) = a(\mathbf{s}) = 1$, it can easily be seen that scattering amplitude f becomes proportional to N; i.e., the roughnesses repeated coherently from layer to layer are scattered in phase.

Using expressions (2.7) and the identity

$$|F_{xy}(\mathbf{s})|^2 = L_x L_y C_0(\mathbf{s}) = \frac{S}{\sin\theta_0} C_0(\mathbf{s}),$$

where L_x and L_y are the dimensions of the coherently illuminated area element, S is the cross section of the incident coherent beam, and $C_0(\mathbf{s})$ is the correlation function of the "substrate" in the reciprocal space, we can easily obtain the differential cross section of diffuse scattering:

$$\frac{d\sigma}{d\Omega} = |f(\mathbf{Q})|^2 = S\frac{r_0^2 Q_z^2}{\sin\theta_0}|F(Q_z)|^2 C_0(\mathbf{s}) \times \left|\sum_n T_0(z_n) T_1(z_n) a^n(\mathbf{s})\right|^2. \quad (2.8)$$

In the standard experimental geometry, cross section (2.8) is measured to a certain degree of integration with respect to the azimuth scattering angle φ or, which is the same, with respect to momentum transfer q_y in the direction perpendicular to the plane of specular diffraction. Integration of cross section (2.8) with respect to momentum q_y substantially complicates the computational problem when exact calculations are required. Integration can be carried out analytically only for special forms of functions $C_0(\mathbf{s})$ and $a(\mathbf{s})$. Quantitative computations require knowledge of function $C_0(\mathbf{s})$, which is obviously difficult for experimental determination. Indeed, function $C_0(\mathbf{s})$ is not a correlation function of the substrate roughness in the proper sense. It was introduced as a limit of the correlation function of roughness of a multilayer X-ray mirror for $n \longrightarrow 0$. Thus, direct measurements of the correlation function of the substrate (e.g., with the help of atomic force microscopy) do not solve the problem.

Integration with respect to the azimuth angle requires knowledge of the angular resolution in the azimuthal direction. Although momenta q_x and q_y appear symmetrically in expression (2.8) for the cross section, their positions are not quite equivalent from the geometrical point of view (with respect to scattering angles):

$$q_x = k\cos\theta_1\cos\varphi - k\cos\theta_0 \approx k(\theta_0^2 - \theta_1^2 - \varphi^2),$$

$$q_y = k\cos\theta_1\sin\varphi \approx k\varphi.$$

The XRDS intensity is mainly concentrated in a narrow region of small angles φ. Thus, if the angular resolution of an experimental setup is not high (fractions of a degree or even worse), the limits of integration with respect to q_y can be taken as infinitely large:

$$I_{\exp}(q_x) \propto \int_{-\infty}^{\infty} \frac{d\sigma}{d\Omega} dq_y.$$

The actual properties of cross section (2.8) and replication factor $a(\mathbf{s})$ make it possible to radically simplify the expression for the cross section, avoiding integration with respect to the azimuth angle in the case when only an approximate solution is required. Indeed, in accordance with formula (2.8), the amplitude of scattering from a roughness defect is determined by the size of the defect not only along the x axis, but also by its size in the perpendicular direction along the y axis. Accordingly, smoothing of roughnesses occurs also in two directions. The integration in question can be reduced to evaluation of the average scattering amplitude taking into account the effect of smoothing along the y axis. It should be noted above all that, obviously, the larger the spatial size of roughness defects, the more precisely these defects must be replicated from layer to layer [18]; i.e., $a(\mathbf{s}) \longrightarrow 1$ for $\mathbf{s} \longrightarrow 0$. On the other hand, the contribution of large roughness defects dominates in the XRDS cross section [33].

In deriving expression (2.8), we presumed that the size of coherently illuminated area elements is much larger than the characteristic size of roughness defects in a multilayer X-ray mirror. However, when synchrotron sources with a high degree of spatial coherence of incident X-rays are used, roughnesses of a large spatial scale start participating in diffraction, and the above assumption becomes invalid [7]. On account of spatial coherence of incident radiation, expression (2.8) for the cross section has a noticeably more complicated form [34, 35]. In addition, a situation can be realized when the far-field approximation (Fraunhoffer diffraction) becomes inapplicable. In this case, the corresponding corrections should also be introduced into the expression for the XRDS cross section [34, 35]. However, that such complications do not play any significant role in the sense that the parameters of a coherent wave packet vary insignificantly upon a transition from one value of photon energy to the other if these energies are close. Thus, the corrections to cross section (2.8) due to the inclusion of coherent properties of X-rays do

not play a decisive role if we measure not the absolute XRDS intensity, but the ratio of the intensities for two close photon energies.

The inclusion of spatial coherence is important since the form of diffraction is basically different for different directions of the momentum transfer relative to the specular diffraction plane, i.e., for q_x and q_y. Indeed, even if the transverse dimensions of coherence in the specular diffraction plane and in the perpendicular direction are approximately equal,[3] the value of L_x increases strongly (in proportion to θ_0^{-1}) as compared to L_y in view of the smallness of angle of incidence θ_0. It follows hence that roughness defects with a size larger than L_y but smaller than L_x cause nonspecular diffuse scattering in the specular diffraction plane, while their presence in the transverse direction is imperceptible. We repeatedly observed such a concentration of XRDS in the specular diffraction plane in experiments [4, 7], including those with a Ni/C multilayer X-ray mirror. For roughnesses of the size considered here, the effect of smoothing along the y axis obviously does not reduce the scattering amplitude; on the other hand, it is these roughnesses that ensure the main contribution to the XRDS cross section.

Thus, under certain conditions, the effect of smoothing along the y axis can be disregarded, and the problem changes from two-dimensional to one-dimensional:

$$I_{\exp}(q_x) \sim S\frac{r_0^2 Q_z^2}{\sin\theta_0}|F(Q_z)|^2 C_0(Q_x) \times \left|\sum_n T_0(z_n)T_1(z_n)a^n(Q_x)\right|^2.$$

If we compare the XRDS intensities for two close photon energies, E_0 and E_1, the quantity $\eta(q_x)$ being measured here and defined as

$$\eta(q_x) = \sqrt{\frac{I(E_1, q_x)}{I(E_0, q_x)}},$$

where $I(E_0, q_x)$ and $I(E_1, q_x)$ are the intensities of quasi-Bragg scattering for the corresponding photon energies, can be calculated using the expression

$$\eta(q_x) = \left|\frac{\delta\rho(E_1)}{\delta\rho(E_0)}\right| \times \left|\frac{\sum_n T_0(E_1, z_n)T_1(E_1, z_n)a^n(Q_x)}{\sum_n T_0(E_0, z_n)T_1(E_0, z_n)a^n(Q_x)}\right|, \quad (2.9)$$

where $\delta\rho(E)$ is the electron density contrast at interfaces for photon energy E. In the expression derived above, we disregard the change in the angles due to the difference in refractive indices $n(E_0)$ and $n(E_1)$.

In order to express the replication factor $a(Q_x)$ analytically in terms of $\eta(q_x)$, additional simplification of expression (2.9) is required; in the general case, such simplification may not be valid. Nevertheless, numerical calculations of replication factor $a(Q_x)$ with the help of expression (2.9) do not require any analytic form of its dependence on q_x.

3. EXPERIMENTAL CONDITIONS

The Ni/C multilayer X-ray mirror with 30 bilayers ($N = 30$) studied here was prepared with the help of laser ablation [27] on a quartz substrate polished thoroughly with nanodiamonds [36]. The roughness dispersion obtained from preliminary X-ray reflectometry ($\lambda = 0.154$ nm) of the substrate was 0.5–0.6 nm. Optical parameters of the mirror, which were obtained by simulating X-ray reflectometric data in the dynamic approximation [37], were as follows: the period was $\Lambda \approx 5.2$ nm; the ratio of the thickness of Ni layers to the period was $\beta \approx 0.4$; the densities of nickel and carbon layers were $\rho_{Ni} \approx 8.2$ g/cm^3 and $\rho_C \approx 2.3$ g/cm^3; and the roughness dispersion was $\sigma \approx 0.4$–0.5 nm. It should be noted that the roughness dispersion of the mirror turned out to be smaller than for the initial substrate, which indirectly indicates the existence of smoothing processes.[4] Moreover, our previous studies [27–29] showed that the above value of roughness dispersion in a multilayer mirror reflects both the true roughness and the presence of mixed layers. The true roughness was estimated as 0.1–0.2 nm [20].

Diffraction experiments were made on a three-crystal diffractometer using synchrotron radiation of a VEPP-3 storage ring [38]. A single channel-cut Si(111)

[3] In actual practice, vertical geometry is used, as a rule, in diffraction setups with synchronous sources (the specular diffraction plane is arranged vertically). This is for two reasons. First, synchronous radiation is usually polarized in the horizontal direction. Second, and more important, the size of a synchronous source in the vertical direction is always much smaller than in the horizontal direction. For this reason, for equivalent angular resolutions, the optical efficiency in the vertical geometry is higher. Accordingly, the vertical component of spatial coherence turns out to be much larger than the horizontal component.

[4] Our experience in preparing Ni/C multilayer mirrors with the help of laser ablation revealed the following interesting fact. Although the roughness dispersion of the initial substrates can vary in relatively wide limits in accordance with X-ray reflectometric data, the optical quality of the prepared mirror turns out to be approximately the same. This visually indicates a strong smoothing of roughnesses during the multilayer growth.

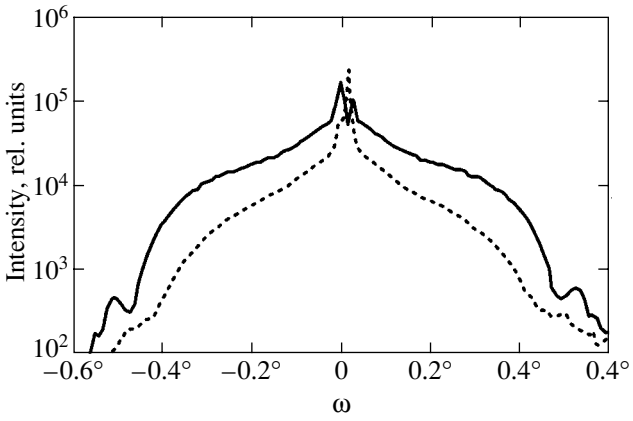

Fig. 4. Experimental ω profiles for photon energies E_0 = 8.325 keV (solid curve) and E_1 = 8.350 keV (dashed curve). The vertical axis corresponds to the XRDS intensity normalized to the intensity of the incident beam.

crystal was used as a monochromator, while a Ge(111) single crystal was used as a secondary crystal-collimator. The monochromator, the multilayer X-ray mirror, and the crystal-collimator were arranged in the (+, +, +) geometry. The experimentally measured angular resolution of the diffractometer was 15–18 angular seconds for an X-ray photon energy of 8 keV. Measurements were made at two photon energies (E_0 = 8.325 keV and E_1 = 8.350 keV) using transverse scanning through the first Bragg reflection; i.e., the XRDS intensity was measured as a function of angle $\omega = (\theta_0 - \theta_1)/2 \approx q_x/2k\theta_B$ under condition (2.2).

It should be noted that the use of the secondary crystal-collimator in the measurements above the *K* absorption edge of nickel atoms enabled us to avoid distortions of experimental data due to excitation of fluorescence. The experimentally measured value of the fluorescence background was found to be at a level of 10–20 Hz, while the desired signal was not weaker than 1 kHz.

4. DISCUSSION OF EXPERIMENTAL RESULTS

Figure 4 shows the experimentally obtained ω profiles. It was found that these profiles have different angular widths, which corresponds to different "characteristic lateral correlation lengths": approximately 0.35 μm for a photon energy of E_0 = 8.325 keV and about 0.40 μm for E_1 = 8.350 keV. It should be emphasized that the absolute values of the above quantities reflect the statistical properties of an ensemble of spatially coherent wave packets rather than the actual properties of roughness. Nevertheless, considering that the contribution to the XRDS comes from the roughness of the interfaces in the entire volume of the multilayer mirror in the former case and predominantly from the interfaces of the upper layers in the latter (Fig. 5), the difference in the obtained results unambiguously indicates the smoothing of roughnesses in the sample. Indeed, a decrease in the effective number of reflecting layers due to screening may increase the width of quasi-Bragg scattering along q_z (see Fig. 3), which is a weak but experimentally observable effect. However, this does not lead to any changes in the XRDS cross section in the q_x direction.

Figure 6 shows the same results (circles) in the form of the dependence of quantity η from expression (2.9) on ω. It can be clearly seen that the experimental points noticeably deviate from the theoretical curve (dashed line) calculated under the assumption of complete cross-correlation. The deviation sign corresponds to a decrease in the roughness amplitudes during the growth of the multilayer mirror, while the magnitude of deviation is larger, the higher the value of the momentum transfer $q_x \approx 2k\theta_B\omega$.

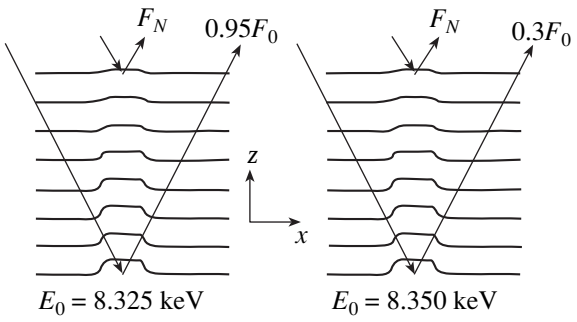

Fig. 5. Comparison of the amplitudes of diffuse scattering from a Ni/C multilayer X-ray mirror for photon energies lower and higher than the *K* absorption edge for nickel atoms; in the second case, the contribution to the amplitude of diffuse scattering from roughness defects of lower layers is noticeably smaller due to the effective screening under conditions of hard photoabsorption.

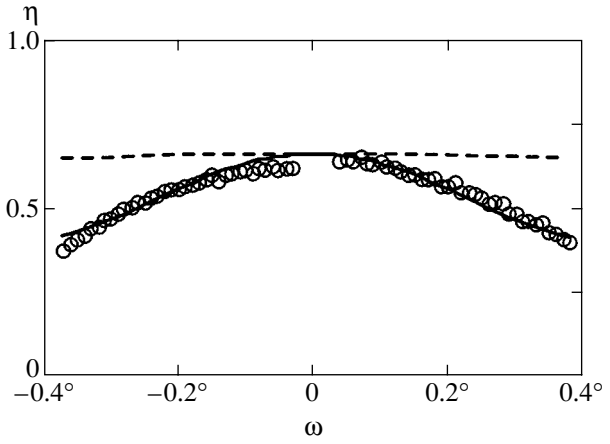

Fig. 6. Experimental data (circles) and the results of theoretical calculations in the DWBA for function η from expression (2.9) for complete cross-correlation (dashed line) and for smoothing of roughnesses (solid curve).

The solid curve in Fig. 6 shows the results of theoretical calculations made by the least-squares technique with the replication factor in the form

$$a(s) = \exp(-\nu t s^2), \qquad (4.1)$$

as in [18, 39, 40]. Here, t is the film (bilayer) thickness and ν is a coefficient of the dimension of length, which characterizes the rate of smoothing. The value of this coefficient was found to be equal approximately to 0.5 μm.[5]

It should be emphasized that the analytic form of the replication factor is immaterial for determining its value when approximate expression (2.9) is used. On the other hand, knowledge of the analytic form is required if calculations are based on exact expression (2.8) for the cross section. We calculated the replication factor by using both methods. In the calculation based on exact expression (2.8), we used the roughness correlation function in form (1.1), which was calculated on the basis of experimental data, but we took into account the fact that the size of the coherently illuminated area element along the y axis under our experimental conditions is equal approximately to 0.5 μm, i.e., much smaller than its size along the x axis (about 300 μm).[6] As a result of such a difference in the sizes, XRDS is predominantly concentrated in the specular diffraction plane [4, 7] and, hence, the error due to the transition to approximate expression (2.9) is additionally reduced. A comparison of the results of calculations revealed that the rejection of averaging of the smoothing effect along the y axis leads to an insignificant error (smaller than 1%). However, this small value leads to an appreciable error in determining the absolute value of the replication factor. At the same time, the error in question weakly affects the dependence of the replication factor on the momentum transfer q_x. Moreover, an experimental error of 1–2% in the determination of the XRDS intensity ratio is inevitable, which is equivalent to double the error in determining the absolute value of the replication factor. Thus, the error associated with the application of approximate expression (2.9) in our case is smaller than the experimental error and this approach cannot lead to a considerable distortion of the results of model calculations.

Figure 7 shows the dependence of the total decrease in the amplitudes of roughnesses during the evolution of interfaces (from the substrate to the surface) on their lateral size. It can be clearly seen that micrometer-scale roughnesses are effectively smoothed during the multilayer growth. At the same time, roughnesses of about 10 μm in size are replicated from layer to layer with an appreciable accuracy. It should be noted that the method used here makes it possible to determine the replication factor, but does not provide absolute values of roughness amplitudes in the bulk of a multilayer X-ray mirror. This fact should not be astonishing since we model not the absolute values of XRDS intensities, but their ratio.

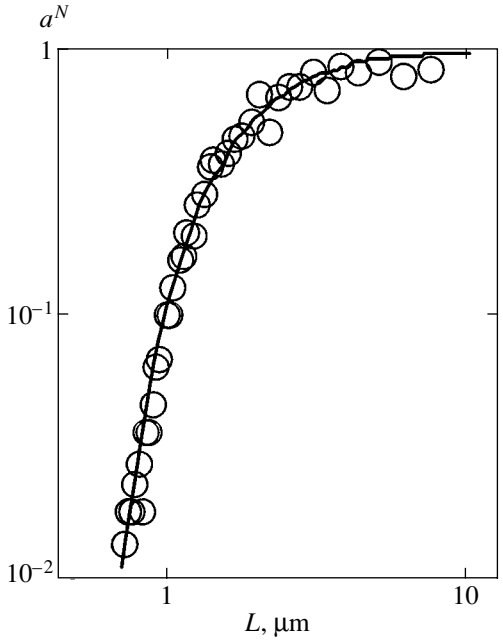

Fig. 7. Total decrease in the amplitudes a^N of roughnesses during the evolution of interfaces from the substrate to the surface as a function of their lateral size $L \sim 2\pi/q_x$; circles are the calculated values of a^N corresponding to experimental points; the solid curve is the best approximation in the framework of model (4.1).

In this study, we assumed that smoothing occurs evenly through the stack of layers of a multilayer X-ray mirror. At the same time, it is reasonable to suppose that effective smoothing may take place even in a few first layers of the mirror near the substrate. Obviously, this effect will be relatively weakly pronounced in the framework of the method used here. It should be noted in this connection that simultaneous application of our method and the method used in [21, 22], where XRDS was studied in the vicinity of Kiessig beats, may provide a complete pattern of the behavior of roughnesses during the growth of a multilayer X-ray mirror.

It should be emphasized that effective smoothing in the micrometer range of roughness defect sizes is of fundamental importance for preparing high-quality multilayer mirrors [7]. In the absence of such smoothing, mirrors with a high reflection coefficient could hardly be manufactured. Indeed, we can easily estimate that the path length of a photon experiencing Bragg

[5] It should be noted that, in spite of the obvious good agreement between the experimental data and such a form of the replication factor, the value of parameter ν obtained by us is at least three orders of magnitude higher than the value predicted in [18, 39, 40].

[6] The size along the y axis is equal to the corresponding component (in the plane of the electron beam orbit) of the transverse coherence of incident X-rays. The size along the x axis can be estimated by using the formula L_z/θ_0 where L_z is the spatial coherence component perpendicular to the orbit. In our case, $L_z \approx 5$ μm.

reflection from a multilayer mirror in the lateral direction is on the order of 1 μm. Thus, roughnesses with a size exceeding 1 μm do not affect Bragg diffraction,[7] while roughnesses with a size smaller than 1 μm effectively suppress the reflection of an X-ray wave.

In conclusion, we must consider possible physical reasons for smoothing of roughnesses on a large spatial scale such as the micrometer scale. It was taken into account in [18, 30, 39, 40] that an atom can be displaced during deposition by a distance coinciding in order of magnitude with the atomic size. In the framework of this model, the replication factor was proposed in a form similar to expression (4.1). It is obvious, however, that such a mechanism cannot be responsible for smoothing over the micrometer scale. Schlatmann *et al.* [41] analyzed the effect of viscous flow during polishing by ions with a high kinetic energy (200–1300 eV), as a result of which smoothing can occur on a large spatial scale. Although the kinetic energy of atoms deposited in the course of laser ablation is considerably lower than the energy of ions during polishing, an analogous process of viscous flow is still possible. Bushuev and Kozak [31, 32] took into account possible diffusion processes, i.e., the fact that adsorbed atoms can move over very large distances on the surface, which leads to uniform smoothing in all regions of the spatial frequency spectrum. According to these authors, the replication factor does not tend to unity upon an increase in the momentum transfer. Another explanation of smoothing of roughnesses on the microscopic spatial scale can be given on the basis of possible processes of reevaporation during the deposition of atoms. "Splashing" of atoms over the surface of a multilayer X-ray mirror may lead, on the one hand, to healing of "valleys" and, on the other hand, to effective leveling out of "hills" on the surface.

Summarizing the result obtained, we can state that the method used here for studying the cross-correlation of roughnesses in a multilayer X-ray mirror allowed us to observe the smoothing of roughnesses on the micrometer spatial scale. It is shown that smoothing is of fundamental importance since the presence of roughnesses of such a size inevitably reduces the reflectance in the case of Bragg diffraction.

ACKNOWLEDGMENTS

The authors are grateful to V.A. Bushuev for useful discussions, to the staff of the Siberian Research Center headed by G.N. Kulipanov, and to experimenters from the VEPP-3 storage ring for their attention and support.

This study was supported financially by the Russian Foundation for Basic Research (project no. 03-02-16259).

[7] The presence of such roughnesses causes phase shifts at the wave packet front, leading to additional XRDS; however, the total coefficient of reflection including the specular (coherent) reflection as well as diffusely scattered radiation does not decrease in this case. At a high spatial coherence of the incident beam of photons, the entire reflection may become diffuse, although it appears as specular reflection [7].

REFERENCES

1. A. V. Vinogradov, N. N. Zorev, I. V. Kozhevnikov, *et al.*, Zh. Éksp. Teor. Fiz. **89**, 2124 (1985) [Sov. Phys. JETP **62**, 1225 (1985)].
2. A. V. Andreev, Usp. Fiz. Nauk **145**, 113 (1985) [Sov. Phys. Usp. **28**, 70 (1985)].
3. S. K. Sinha, E. B. Sirota, S. Garoff, *et al.*, Phys. Rev. B **38**, 2297 (1988).
4. V. A. Chernov, V. I. Kondratiev, N. V. Kovalenko, *et al.*, Nucl. Instrum. Methods Phys. Res. A **470**, 145 (2001).
5. N. V. Vostokov, S. V. Gaponov, V. L. Mironov, *et al.*, Poverkhnost, No. 1, 38 (2001).
6. E. L. Church, Appl. Opt. **27**, 1518 (1988).
7. V. A. Chernov, V. I. Kondratiev, N. V. Kovalenko, *et al.*, J. Appl. Phys. **92**, 7593 (2002).
8. V. A. Chernov, N. V. Kovalenko, and S. V. Mytnichenko, Nucl. Instrum. Methods Phys. Res. A **470**, 210 (2001).
9. T. W. Barbee, Opt. Eng. **25**, 893 (1986).
10. S. V. Gaponov, F. V. Garin, S. A. Gusev, *et al.*, Nucl. Instrum. Methods Phys. Res. **208**, 227 (1983).
11. E. Spiller, A. Segmuller, J. Rife, *et al.*, Appl. Phys. Lett. **37**, 1048 (1980).
12. E. Spiller, Appl. Phys. Lett. **54**, 2293 (1989).
13. M. P. Bruijn, P. Chakraborty, H. W. van Essen, *et al.*, Proc. SPIE **563**, 36 (1985).
14. E. J. Puik, M. J. van der Wiel, H. Zeijlemarker, *et al.*, Rev. Sci. Instrum. **63**, 1415 (1992).
15. A. V. Andreev, A. G. Michette, and A. Renwick, J. Mod. Opt. **35**, 1667 (1988).
16. A. Bruson, C. Dufour, B. George, *et al.*, Solid State Commun. **71**, 1045 (1989).
17. D. E. Savage, N. Schimke, Y.-H. Phang, *et al.*, J. Appl. Phys. **71**, 3283 (1992).
18. D. G. Stearns, J. Appl. Phys. **71**, 4286 (1992).
19. A. V. Vinogradov, private communication.
20. V. A. Chernov, E. D. Chkhalo, N. V. Kovalenko, *et al.*, Nucl. Instrum. Methods Phys. Res. A **448**, 276 (2000).
21. H. Laidler, I. Pape, C. I. Gregory, *et al.*, J. Magn. Magn. Mater. **154**, 165 (1996).
22. I. Pape, T. P. A. Hase, B. K. Tanner, *et al.*, Physica B (Amsterdam) **253**, 278 (1998).
23. V. E. Asadchikov, A. Yu. Karabekov, V. V. Klechkovskaya, *et al.*, Kristallografiya **43**, 119 (1998) [Crystallogr. Rep. **43**, 110 (1998)].
24. N. V. Kovalenko, S. V. Mytnichenko, and V. A. Chernov, Pis'ma Zh. Éksp. Teor. Fiz. **77**, 85 (2003) [JETP Lett. **77**, 80 (2003)].
25. V. Holy and T. Baumbach, Phys. Rev. B **49**, 10668 (1994).
26. A. V. Vinogradov, I. A. Brytov, A. Ya. Grudskiĭ, *et al.*, in *Mirror X-ray Optics*, Ed. by A. V. Vinogradov (Mashinostroenie, Leningrad, 1989), p. 86.

27. V. A. Chernov, N. I. Chkhalo, M. V. Fedorchenko, *et al.*, J. X-Ray Sci. Technol. **5**, 65 (1995).

28. V. A. Chernov, N. I. Chkhalo, M. V. Fedorchenko, *et al.*, J. X-Ray Sci. Technol. **5**, 389 (1995).

29. V. A. Chernov, N. I. Chkhalo, and S. G. Nikitenko, J. Phys. IV **7**, C2-699 (1997).

30. S. F. Edwards and D. R. Wilkinson, Proc. R. Soc. London, Ser. A **381**, 17 (1982).

31. V. A. Bushuev and V. V. Kozak, Kristallografiya **42**, 809 (1997) [Crystallogr. Rep. **42**, 742 (1997)].

32. V. A. Bushuev and V. V. Kozak, Poverkhnost, No. 2, 96 (1999).

33. D. K. G. de Boer, Phys. Rev. B **53**, 6048 (1996).

34. S. K. Sinha, M. Tolan, and A. Gibaud, Phys. Rev. B **57**, 2740 (1998).

35. M. Tolan and S. K. Sinha, Physica B (Amsterdam) **248**, 399 (1998).

36. A. I. Volokhov, É. P. Kruglyakov, and N. I. Chkhalo, Poverkhnost, No. 1, 130 (1999).

37. L. G. Parrat, Phys. Rev. **95**, 359 (1954).

38. *Brief Description of the SR Experimental Station*, Preprint No. 90-92, INP (Inst. of Nuclear Physics, Siberian Division, USSR Academy of Sciences, Novosibirsk, 1990).

39. D. G. Stearns, Appl. Phys. Lett. **62**, 1745 (1993).

40. E. Spiller, D. Stearns, and M. Krumrey, J. Appl. Phys. **74**, 107 (1993).

41. R. Schlatmann, J. D. Shindler, and J. Verhoeven, Phys. Rev. B **54**, 10880 (1996).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1212–1218.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1358–1366.

SOLIDS
Electronic Properties

A Tunneling Spectroscopy Study of the Temperature Dependence of the Forbidden Band in Bi_2Te_3 and Sb_2Te_3

V. A. Kul'bachinskiĭ[a], H. Ozaki[b], Y. Miyahara[b], and K. Funagai[b]

[a]*Moscow State University, Vorob'evy gory, Moscow, 119992 Russia*
[b]*Department of Electrical, Electronics, and Computer Engineering, Waseda University, Tokyo, 169-8555 Japan*
e-mail: kulb@mig.phys.msu.ru

Received February 7, 2003

Abstract—Tunneling measurements of dI/dV, d^2I/dV^2, and d^3I/dV^3 were performed along the C_3 axis (normally to layers) for Bi_2Te_3 and Sb_2Te_3 layered semiconductors in the temperature range $4.2 < T < 295$ K. Temperature dependences of the forbidden band energy E_g were obtained. The forbidden band energy in Bi_2Te_3 was 0.20 eV at room temperature and increased to 0.24 eV at $T = 4.2$ K. The E_g value for Sb_2Te_3 was 0.25 eV at 295 K and 0.26 eV at 4.2 K. The distance between the top of the higher valence band of light holes and the top of the valence band of heavy holes situated lower was found to be $\Delta E_V \approx 19$ meV in Bi_2Te_3; this distance was independent of temperature. The conduction bands of Bi_2Te_3 and Sb_2Te_3 each contain two extrema with distances between them of $\Delta E_c \approx 25$ and 30 meV, respectively. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Bi_2Te_3 and Sb_2Te_3 semiconductors are layered crystals with rhombohedral structures, space group R_{3m} (D_{3d}^5). The crystal lattice is formed by periodically ordered layers lying in the planes perpendicular to the C_3 symmetry axis. Each layer comprises five atomic planes (quintets) that form the sequence Te^1–Bi–Te^2–Bi–Te^1. Here, Te^1 and Te^2 are Te atoms in different sites. The atoms in separate layers are identical and form a planar hexagonal lattice. The atoms of each subsequent layer are situated above the centers of the triangles formed by the atoms of the preceding layer (close hexagonal packing); that is, the Te^1 and Bi atoms occupy octahedral sites in the tetradymite structure. The chemical bonds within the quintets are covalent-ionic. Between the quintets, the distance is comparatively long and bonds are van der Waals in nature and weak. This determines the anisotropy of the electrophysical properties of the single crystals [1].

The following types of defects are characteristic of Bi_2Te_3 (Sb_2Te_3) single crystals: Bi (Sb) and Te vacancies, Bi (Sb) and Te atoms in interstices, antisite defects Bi_{Te} and Sb_{Te} (a Bi or Sb atom in a Te site) and Te_{Sb} and Te_{Bi} (a Te atom in a Bi or Sb site), impurity antisite defects, impurity atoms in interstices, etc. The antisite defects are negatively charged. For this reason, bismuth and antimony tellurides grown under stoichiometric conditions always have *p*-type conductivity and a substantial concentration of holes, up to 10^{19} cm^{-3} in bismuth telluride and 10^{20} cm^{-3} in antimony telluride.

The forbidden band energy E_g in Bi_2Te_3 and Sb_2Te_3 (and also in Bi_2Se_3, Bi_2S_3, Sb_2Se_3, and Sb_2S_3) was determined in several works, mainly at room temperature, by measuring the temperature dependences of resistance [2–4] and by optical methods [5–11]. The results obtained in these works are summarized in the table. We do not consider theoretical works, because the accuracy of calculations in them is insufficient for narrow-gap materials of the type of bismuth and antimony tellurides. Apart from a large spread of the reported E_g values, data on the temperature dependences of E_g and forbidden band widths at low temperatures are lacking. As follows from the table, bismuth and antimony tellurides are narrow-gap semiconductors, whereas Sb_2Se_3, Bi_2S_3, and Sb_2S_3 are high-energy-gap materials. Note that the forbidden band widths in Bi_2Te_3 and Sb_2Te_3 are difficult to determine by optical methods because of a high concentration of free carriers, which causes additional absorption and shifts the absorption edge as a result of the Burstain–Moss effect. It follows that the E_g value can only be calculated from experimental optical data using some model of the energy spectrum of current carriers, which is the reason why different E_g values were reported in different works. Similarly, because of the high concentration of current carriers, activation conductivity can only be observed in a narrow temperature range, which also contributes to errors in forbidden band values.

In this work, the forbidden band width was for the first time determined in the temperature range 4.2–295 K by directly measuring the singularities of the first tun-

Forbidden band of $A_2^V B_3^{VI}$ semiconductors

Refs.	Material	E_g, eV	Temperature interval of measurements, K	Method of measurements
[2]	Bi_2Te_3	0.21	$100 < T < 750$	$\rho(T)$
[3]		0.20	$77 < T < 380$	$\rho(T)$
[4]		0.16	$160 < T < 650$	$\rho(T)$
[5]		0.15	300	IR transmission
		0.16	77	
[6]		0.13	300	IR transmission and reflectance
		0.17	85	
[7]		0.15–0.22	10	IR reflectance
[5]	Sb_2Te_3	0.30	300	IR transmission
[6]		0.21	300	IR transmission and reflectance
[8]		0.21	300	IR absorption
[9]	Bi_2Se_3	0.21	$80 < T < 300$	IR transmission and absorption
[10]		0.115	300	IR absorption
		0.160	77	
[5]	Sb_2Se_3	1.2	300	IR transmission
		1.3	77	
[5]	Bi_2S_3	1.3	300	IR transmission
[5]	Sb_2S_3	1.7	300	IR transmission
[11]		1.74	27	Absorption edge

nel current derivative with respect to the voltage, dI/dV, which is directly related to the density of states. We also used the second, d^2I/dV^2, and third, d^3I/dV^3, derivatives to locate the singularities more reliably.

2. SAMPLES AND THE PROCEDURE FOR MEASUREMENTS

In this work, we studied bismuth telluride Bi_2Te_3 and antimony telluride Sb_2Te_3 single crystals grown by the Bridgman method from the elements of a 99.999% purity taken in stoichiometric amounts. Prior to measurements, the single crystals were cleaved normally to the C_3 axis. A tunneling contact was created on the cleavage. It is essential that measurements be taken using freshly cleaved samples, because the surface of single crystals is fairly rapidly oxidized [12], and the resulting oxide cannot be used to create tunneling contacts.

Tunnel junctions in metal–dielectric–semiconductor systems are well known and have been considered in detail by Wolf and Solimar in monographs [13, 14]. We employed two different methods with clamping contacts used to produce metal–insulator–semiconductor tunneling contacts. In the first method, a point tunneling contact was created using tantalum oxide Ta_2O_5 as an insulator. A thin polished tantalum wire was subjected to controlled oxidation in an oxygen atmosphere. The thickness of the oxide layer was adjusted experimentally to obtain good tunneling characteristics. A scheme of the measuring cell is shown in Fig. 1a. The tunnel junction was stabilized by a thin bronze spring, which pressed the tantalum wire to the base plane of the single crystalline sample. Ohmic contacts with the sample were made of gold and soldered with indium. The four-point contact method was used.

In the second method (Fig. 1b), the tunneling contact was created from aluminum preliminarily sputtered onto a quartz rod and oxidized in an oxygen atmosphere at 300°C. The rod was pressed to the surface of the sample also using a bronze spring. The force of pressing the contact at an arbitrary temperature could be controlled by a screw fastened on the cap of a Dewar flask. This screw pressed the plate by means of a rod (shown by an arrow in Fig. 1b).

Both methods were fairly laborious but gave stable reproducible results that coincided with each other. A modulation technique was used to record the first current derivative with respect to the voltage dI/dV and also the second d^2I/dV^2 and third d^3I/dV^3 derivatives. A holder with a sample (see Fig. 1) was placed into a bronze chamber filled with gaseous helium to even out temperature. The chamber on a special support was moved above the surface of liquid nitrogen or helium or immersed into the corresponding liquid to obtain temperatures of 4.2 and 77 K. The temperature was mea-

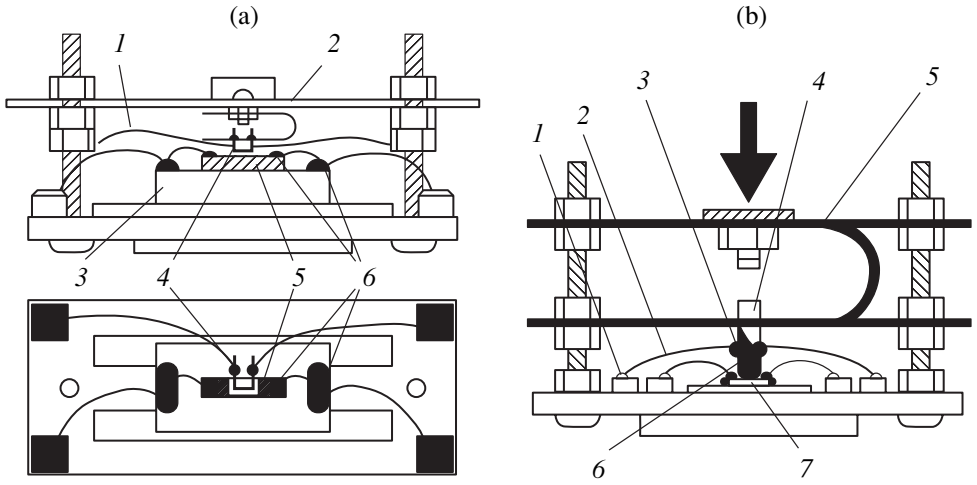

Fig. 1. Scheme of sample holders for tunnel measurements (a) with an oxidized tantalum wire: (*1*) gold wire, (*2*) thin phosphor bronze plate, (*3*) quartz plate, (*4*) tantalum wire, (*5*) sample, and (*6*) indium contact and (b) with a quartz rod coated with oxidized aluminum: (*1*) indium contact, (*2*) gold wire, (*3*) silver paste, (*4*) quartz rod, (*5*) plate and spring of phosphor bronze, (*6*) aluminum oxidized at 300°C, and (*7*) sample. The arrow shows the force with which the rod acts to press alumina to the surface of the sample.

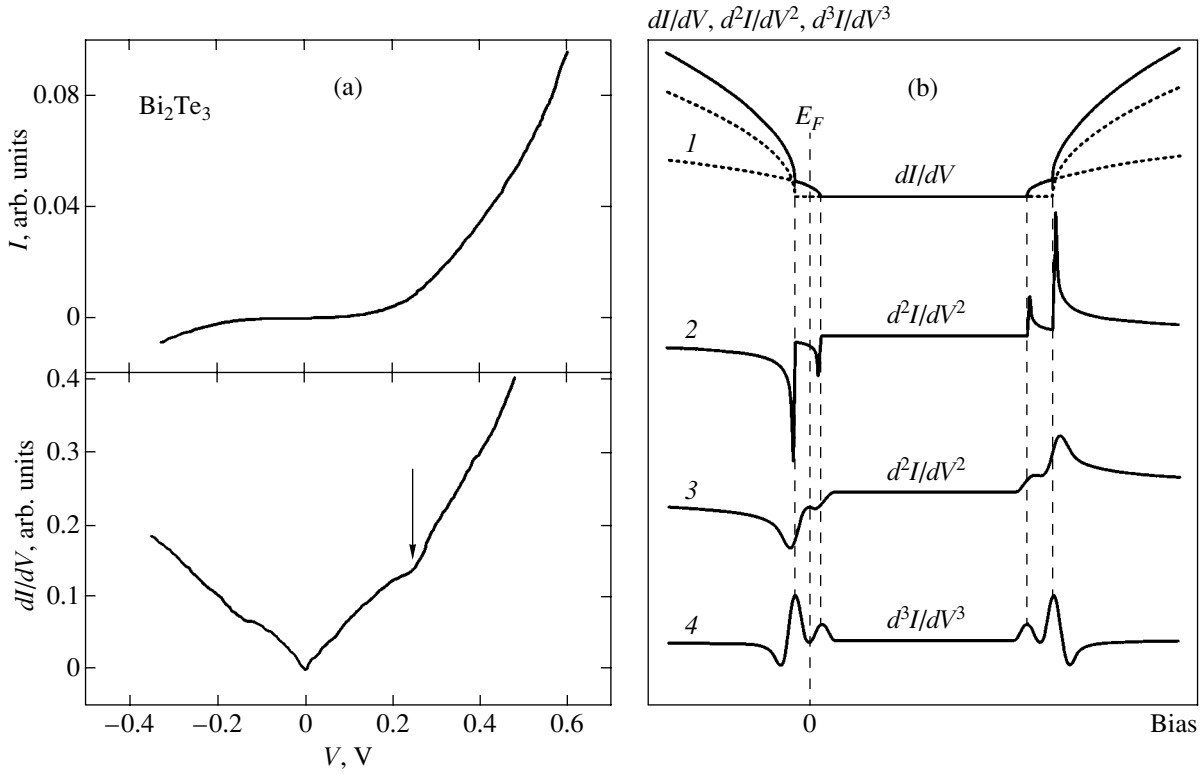

Fig. 2. (a) Voltage V dependences of current I and dI/dV for Bi_2Te_3. The arrow shows the position of the singularity at a positive bias. (b) Schematic features of the tunnel spectrum of bismuth telluride for (*1*) the first derivative dI/dV, (*2*) the second derivative d^2I/dV^2, (*3*) the second derivative d^2I/dV^2 taking into account smearing of the dependences, and (*4*) the third derivative d^3I/dV^3; E_F is the position of the Fermi level in the sample, and the dashed lines schematically show two valence band extrema (on the left) and two conduction band extrema (on the right).

sured by a calibrated germanium thermometer glued to a quartz plate close to the sample. The accuracy of controlling temperature during recording one dependence was better than 1 K.

By way of example, current–voltage characteristics of bismuth telluride obtained in the measurement cell (see Fig. 1a) at 4.2 K are shown in Fig. 2a. The special features of the tunneling spectrum of the samples with

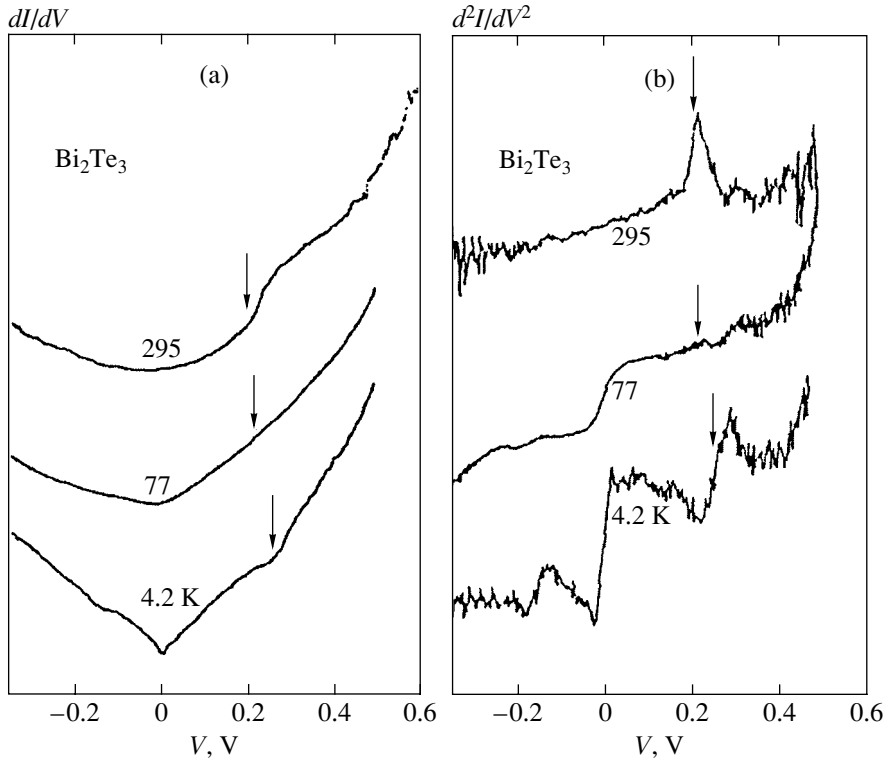

Fig. 3. (a) The first derivative of current with respect to voltage dI/dV and (b) the second derivative d^2I/dV^2 for Bi_2Te_3 at three temperatures. Zero bias corresponds to the Fermi level; arrows show the singularity corresponding to the boundary of the lower conduction band (LCB).

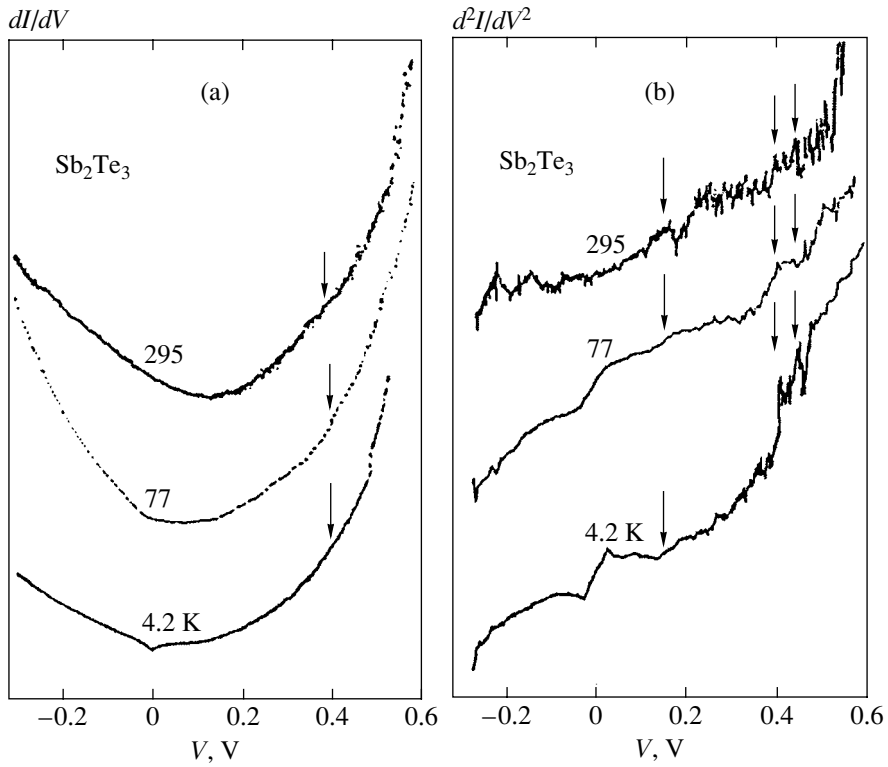

Fig. 4. (a) The first derivative of current with respect to voltage dI/dV (arrows show the singularity corresponding to the boundary of the lower conduction band LCB) and (b) the second derivative d^2I/dV^2 for Sb_2Te_3 at three temperatures. Zero bias corresponds to the Fermi level; arrows show the singularities that correspond to the top of the valence band and the boundaries of the lower (LCB) and upper (UCB) conduction bands (from left to right).

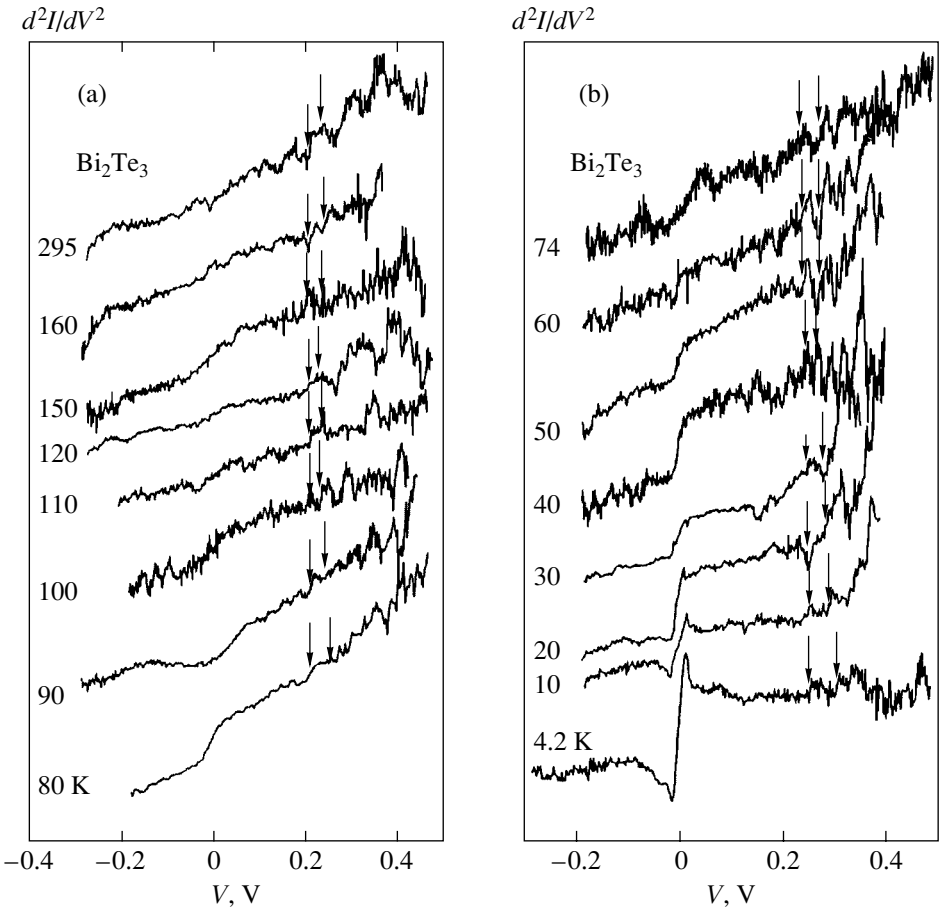

Fig. 5. Voltage V dependences of the second derivative d^2I/dV^2 for Bi_2Te_3 at various temperatures. Arrows show the singularities corresponding to the boundaries of the lower (LCB) and upper (UCB) conduction bands.

two extrema in the valence band (the upper valence band, UVB, and the lower valence band, LVB) and two extrema in the conduction band (the upper conduction band, UCB, and the lower conduction band, LCB) are illustrated by Fig. 2b. The first derivative dI/dV is determined by the density of states, which has root singularities at band boundaries. Since the grown samples were *p*-type, their Fermi levels were situated in the valence bands, as is shown in Fig. 2b, between the upper and lower valence band extrema. A scheme of the band structure of bismuth telluride (antimony telluride) is given below in Fig. 6b.

3. MEASUREMENT RESULTS

The tunnel characteristics $dI/dV(V)$ and $d^2I/dV^2(V)$ in bismuth telluride are shown in Fig. 3. A positive bias corresponds to a higher energy of electrons. As the sample has *p*-type conductivity, the Fermi level (zero bias) is in the valence band, as is shown in Fig. 2b. The singularity of the density of states, which appears with the onset of filling the lower conduction band, is shown by arrows in Fig. 3. The top of the valence band is difficult to determine from the spectra given in Fig. 3 because of interfering thermal fluctuations at 295 K and a large zero-bias anomaly at 4.2 K.

Similar dependences were observed for antimony telluride. The corresponding tunnel characteristics $dI/dV(V)$ and $d^2I/dV^2(V)$ are plotted in Fig. 4. Figure 4a demonstrates that only the singularity related to the onset of filling the lower conduction band is discernible in the $dI/dV(V)$ dependence for Sb_2Te_3. Note that this singularity is observed at much larger bias voltages, which corresponds to a much larger (in magnitude) Fermi energy E_F in this material resulting from a much higher concentration of holes. As concerns the second derivative d^2I/dV^2 (see Fig. 4b), the first singularity is observed at a bias of 0.14 eV (marked by arrows). This singularity is independent of temperature and corresponds to the top of the upper valence band. The next singularity marked by arrows, which is observed as the bias voltage increases, corresponds to the onset of filling the lower conduction band at a bias voltage of 0.39 eV (T = 295 and 77 K) and 0.4 eV (T = 4.2 K). These results imply that E_g = 0.25 eV at 295 and 77 K and E_g = 0.26 eV at 4.2 K. Further, a d^2I/dV^2 singularity marked by arrows and corresponding to the bottom of the upper conduction band is observed in the spectrum.

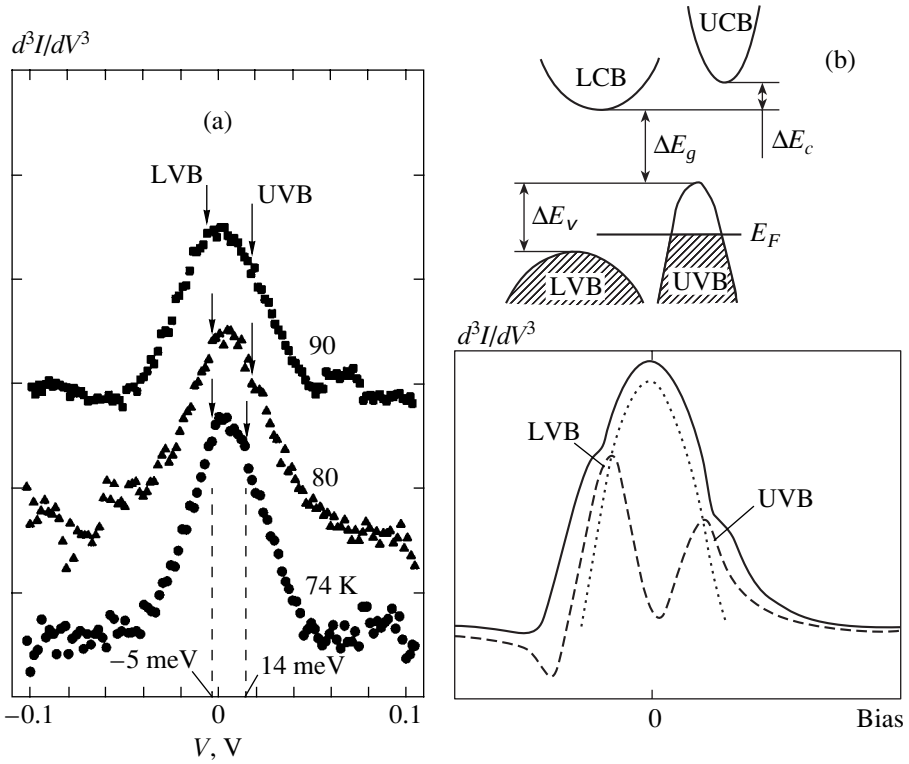

Fig. 6. (a) Voltage V dependences of the third derivative d^3I/dV^3 for Bi_2Te_3 at three temperatures. Arrows show the singularities that correspond to the edges of the lower (LVB) and upper (UVB) valence bands. (b) At the top: a scheme of the energy spectrum of Bi_2Te_3 (Sb_2Te_3); E_F is the Fermi energy, and ΔE_c and ΔE_v are the distances between two conduction and valence band extrema, respectively. At the bottom: the spectrum shown in Fig. 6a corresponds to the superposition of two peaks (the dashed line) and the anomaly at zero bias (the dotted line).

It follows that, according to the measurement results, the conduction band of Sb_2Te_3 has two extrema separated by an energy of $\Delta E_c \approx 30$ meV, which is independent of temperature. Up to now, experimental data on the conduction band of antimony telluride have been lacking, because this material always has a high initial concentration of holes, which cannot be compensated by doping.

The $d^2I/dV^2(V)$ second derivative dependences for Bi_2Te_3 obtained at different temperatures are shown in Fig. 5. Arrows show the edges of the lower and upper conduction bands. By way of example, the third derivatives $d^3I/dV^3(V)$ for Bi_2Te_3 in the region of low bias voltages are shown in Fig. 6a for several temperatures. These dependences contain singularities related to the edges of two valence bands. Schematically, the formation of these singularities is shown in Fig. 6b (bottom). Note that the Fermi level in the sample lies in the band of light holes above the top of the lower valence band, as is shown in Fig. 6b (top). It follows that one singularity is observed in the region of positive and the other in the region of negative bias voltages. The conclusion can be drawn that the distance ΔE_v between the two valence band extrema is about 19 meV and hardly changes when temperature varies. This value closely agrees with that of $\Delta E_v \approx 20$ meV measured in [15, 16] by other indirect methods. We were unable to obtain reliable data on the distance between the upper and lower valence bands from the tunneling spectra of antimony telluride.

The special features of the tunnel characteristics at various temperatures (see Fig. 5) can be used to determine the temperature-dependent positions of the edges of two conduction bands in Bi_2Te_3 with respect to the Fermi level. The corresponding temperature dependence is shown in Fig. 7a. According to this figure, the distance between the upper and lower conduction bands is around $\Delta E_c = 25$ meV and does not change as temperature varies to within the accuracy of measurements. Earlier, the ΔE_c value was estimated at 24 meV from indirect measurements [17].

The temperature dependence of the forbidden band energy E_g of bismuth telluride determined by analyzing the tunneling spectra (see Figs. 5 and 6a) is given in Fig. 7b. The figure shows that the forbidden band energy increases as temperature decreases, from 0.20 eV at $T = 295$ K to 0.24 eV at $T = 4.2$ K. It can be seen from the table that an increase in the forbidden band energy in bismuth telluride at lower temperatures was observed in several works, although the E_g values themselves were inaccurate and different in different works. The nature of the nonmonotonic $E_g(T)$ dependence is not quite clear, because the crystal lattice parameters of Bi_2Te_3 change insignificantly and mono-

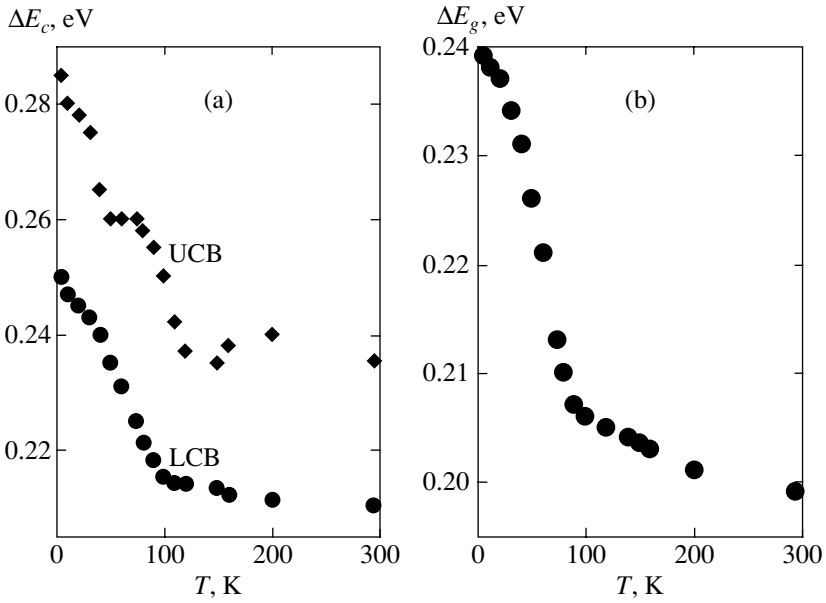

Fig. 7. (a) Temperature dependence of the positions of the edges of the lower (LCB) and upper (UCB) conduction bands counted from the Fermi level and (b) temperature dependence of the forbidden band in Bi_2Te_3.

tonically as temperature varies [18]. By way of example, consider the temperature dependence of the forbidden band energy for A_2B_6 semiconductors, in which this energy can both increase (HgTe, HgSe) and decrease (ZnTe, ZnSe) as temperature lowers [19].

To summarize, our study of the tunneling spectra allowed us to determine the complex band structure of bismuth and antimony tellurides with two closely spaced conduction band extrema. The forbidden band energy of bismuth telluride increases as temperature decreases from 0.20 eV at room temperature to 0.24 eV at $T = 4.2$ K. In antimony telluride Sb_2Te_3, the forbidden band energy is $E_g = 0.25$ eV (295 K) and 0.26 eV (4.2 K); that is, it virtually does not depend on temperature. We found that the distance between the top of the upper valence band of light holes and the top of the lower valence band of heavy holes is $\Delta E_v \approx 19$ meV in Bi_2Te_3 and does not depend on temperature. The conduction bands of Bi_2Te_3 and Sb_2Te_3 have two extrema, the upper conduction band and the lower conduction band, the distances between which are $\Delta E_c \approx 25$ and 30 meV, respectively.

REFERENCES

1. B. M. Gol'tsman, V. A. Kudinov, and I. A. Smirnov, *Semiconducting Thermoelectric Materials Based on Bi_2Te_3* (Nauka, Moscow, 1972).
2. S. Shigetomi and S. Mori, J. Phys. Soc. Jpn. **11**, 915 (1956).
3. C. B. Satterthwaite and R. W. Ure, Jr., Phys. Rev. **108**, 1164 (1957).
4. T. C. Harman, B. Paris, S. E. Miller, and H. L. Goering, J. Phys. Chem. Solids **2**, 181 (1957).
5. J. Black, E. M. Conwell, L. Seigle, and C. W. Spencer, J. Phys. Chem. Solids **2**, 240 (1957).
6. R. Serh and L. R. Testardi, J. Phys. Chem. Solids **23**, 1219 (1962).
7. G. A. Thomas, D. H. Rapkine, R. B. Van Dover, *et al.*, Phys. Rev. B **46**, 1553 (1992).
8. S. D. Shutov, V. V. Sobolev, and L. I. Smeshlivyĭ, in *Semiconductor Compounds and Their Solid Solutions*, Ed. by S. I. Radautsan (Akad. Nauk Mold. SSR, Kishinev, 1970), p. 155.
9. H. Gobreht, S. Seek, and T. Klose, Z. Phys. **190**, 427 (1966).
10. H. Kohler and J. Hartmann, Phys. Status Solidi B **63**, 171 (1974).
11. T. Fujita, K. Kurita, K. Takiyama, and T. Oda, J. Phys. Soc. Jpn. **56**, 3734 (1987).
12. H. Bando, K. Koizumi, Y. Oikawa, *et al.*, J. Phys.: Condens. Matter **12**, 5607 (2000).
13. E. L. Wolf, *Principles of Electron Tunneling Spectroscopy* (Oxford Univ. Press, New York, 1985; Naukova Dumka, Kiev, 1990).
14. L. Solymar and D. Walsh, *Lectures on the Electrical Properties of Materials*, 4th ed. (Oxford Univ. Press, Oxford, 1988; Mir, Moscow, 1991).
15. H. Kohler, Phys. Status Solidi B **74**, 591 (1976).
16. V. A. Kulbachinski, V. Inoue, M. Sasaki, *et al.*, Phys. Rev. B **50**, 16921 (1994).
17. H. Kohler, Phys. Status Solidi B **73**, 95 (1976).
18. M. H. Francombe *et al.*, Br. J. Appl. Phys. **9**, 415 (1958).
19. N. N. Berchenko, V. E. Krevs, and V. G. Sredin, *Semiconductor Solid Solutions and Their Applications: Reference Book*, Ed. by V. G. Sredin (Voenizdat, Moscow, 1982), p. 36.

Translated by V. Sipachev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1219–1230.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1367–1379.

SOLIDS
Electronic Properties

On Oscillations of Thermionic Current in Composite Systems

A. N. Starostin and M. A. Chesnokov
Troitsk Institute for Innovation and Fusion Research,
Troitsk, Moscow oblast, 142190 Russia
e-mail: a.starostin@relcom.ru
Received November 10, 2002

Abstract—Thermionic current emitted into vacuum by a composite system consisting of a metal layer of finite thickness L and a metal half-space is investigated. An explicit expression for thermionic current is obtained that takes into account the quantum phenomena of above-barrier reflection and the strong degeneracy of the electron gas in metal at sufficiently low temperatures. Actually, a generalization of the classical Richardson–Dashmen result is obtained for the case of low temperatures. A special emphasis is placed on the effect of impurities contained in the metal half-space and in the layer on the total thermionic current. It is shown that violation of the strictly periodic field of an ideal crystal due to impurities breaks the one-to-one relation between the momenta and the total energy of conductivity electrons. It is shown that the expression for the generalized distribution function depending on independent variables (energy and momentum) is naturally included in the equation for the thermionic current. Numerical analysis shows that the variation in the distribution function of electrons due to the impurity field leads to variation of the total thermionic current emitted from the system. In particular, an oscillating dependence of the thermionic current on the layer thickness and on the impurity concentration in the layer is revealed. All the calculations are performed within the formalism of nonequilibrium Green's functions.

1. INTRODUCTION

The study of the structure of normal metals (i.e., metals in the normal state) is one of the most promising directions in modern solid-state physics owing to numerous industrial applications of the results obtained. This problem arose as early as the beginning of the 20th century, when various models were proposed to describe the behavior of electrons in the periodic field of a crystal. In particular, the problem of the motion of an electron in a periodic potential was studied in detail. It was shown that, in many cases, the motion of electrons in a metal can be interpreted as a motion of free particles whose effective mass depends on the structure of the lattice and the shape of the energy spectrum of electrons in the metal. This model has been called a model of free electrons.

One of the most important results of the model of free electrons is the expression for the thermionic current emitted by a metal half-space, which is known as the Richardson–Dushman formula [1]:

$$J_{\text{therm}} = -\bar{D}\frac{M^*T^2}{2\pi^2\hbar^3}\exp\left(-\frac{W_a}{T}\right), \qquad (1)$$

where $\bar{D}$ is the averaged transmission coefficient, T is the metal temperature, W_a is the work function of the metal, and M^* is the electron effective mass. Here, one should make two important remarks. First, according to cyclotron-resonance experiments, the difference between the effective mass of an electron and its actual mass may be very small. For such metals, one can replace the effective mass M^* in (1) by the actual electron mass M to a high degree of accuracy. Second, the expression for thermionic current (1) is actually classical. It has been obtained under the assumption that the distribution function of conductivity electrons in metal is the Fermi distribution; moreover, it was assumed that all electrons with energy greater than the threshold value may leave the metal, thus making a contribution to the thermionic current. The quantum phenomena of above-barrier reflection (i.e., the reflection of the part of electrons that, according to the classical theory, should be freely emitted from the metal) are taken into consideration only in the additional normalizing factor $\bar{D}$, which is generally justified only for sufficiently high temperatures. Hence, one can expect that the accuracy of formula (1), which gives good agreement with experimental data at high temperatures, will decrease as temperature decreases.

The aim of the present study is to calculate a thermionic current with regard to various quantum phenomena such as the above-barrier reflection, possible existence of bound states, and the strong degeneracy of the electron gas in metal. Moreover, we pay special attention to the role of impurities that are inevitably contained in metals. When introducing impurities into an

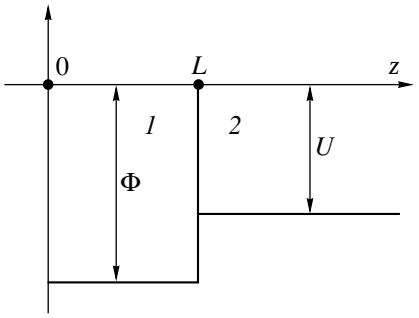

Fig. 1.

ideal lattice, we introduce an external potential in which electrons in metal move. This breaks the spatial homogeneity and, hence, leads to indeterminacy in the value of the electron momenta. In other words, the generalized energy and momentum distribution of electrons is characterized by a finite width that depends on the character of interaction between electrons and impurity ions. Further analysis is carried out within the framework of a generalized distribution function $F(E, \mathbf{p})$, where the electron momentum and the total energy are independent variables [2]. Below, we present an explicit expression for $F(E, \mathbf{p})$ and show that, in the absence of impurities (i.e., in the case of an ideal crystal lattice), the generalized distribution function F is expressed as

$$F(E, \mathbf{p}) = n(E)\delta\left(E - \frac{p^2}{2M^*}\right); \qquad (2)$$

i.e., in this case, we have a one-to-one relationship between the momentum and energy of electrons. We demonstrate that a variation in the distribution function due to the impurity field may substantially affect the thermionic current. In particular, we reveal an oscillating behavior of the current as a function of impurity concentration. Moreover, we carry out a detailed investigation of the situation when the surface of a metal half-space is coated with a layer of a different metal that also contains impurities. We show that, in this case, the thermionic current as a function of the layer thickness and the concentration of impurity ions in the layer also exhibits oscillating behavior.

The paper is organized as follows. In Section 2, we formulate the problem and obtain an expression for the thermionic current in terms of kinetic Green's function. In Section 3, we solve the Dyson equation for retarded and kinetic Green's functions and obtain explicit expressions for the contribution of each metal to the thermionic current. In Section 4, we discuss the results of the numerical analysis for the total thermionic current.

2. STATEMENT OF THE PROBLEM: THERMIONIC CURRENT IN TERMS OF KINETIC GREEN'S FUNCTIONS

We consider a composite system (see Fig. 1) consisting of a plane-parallel layer of finite thickness L ($0 < z < L$) of metal 1 and a half-space ($z > L$) of metal 2.

The metals are assumed to have different compositions, i.e., different work functions W_1 and W_2 and free-electron concentrations Ne_1 and Ne_2 (for the numerical analysis, we took silver as metal 1 and sodium as metal 2). Moreover, it is assumed that the composite system is in thermodynamic equilibrium at a finite temperature T.

As we pointed out in the Introduction, the aim of the present study is to calculate the thermionic current emitted into a vacuum (occupying the domain $z < 0$) by the system considered. By a vacuum we mean a domain with a negligible concentration of particles at a temperature close to absolute zero.

We will not restrict the analysis to an ideal metal and to the case when only quantum corrections to the classical Richardson–Dashmen result (1) are taken into account. Since the field inside a real metal may significantly differ from the field of an ideal lattice, it would be of interest to find out how seriously such a difference may affect the total thermionic current. Note that this nonideality of a real metal can be associated both with thermal vibrations of the lattice atoms about their equilibrium positions and with various defects such as vacancies, interstices, and dislocations.

Since our primary concern is the thermionic current emitted from the system at low temperatures, we will not take into account thermal vibrations of the lattice and attribute the violation of ideality solely to the defects in the metal structure. In the present paper, we will not consider the effects of dislocations and vacancies; we will restrict the analysis to the case when a metal contains impurity ions that distort the field of ideal crystal.

Thus, we assume that metal 1, just as metal 2, contains impurities. We also assume that the concentration of these impurities is known and that one can produce a metal with a prescribed concentration of impurities. In this paper, we will use a model similar to the model of free electrons; i.e., we assume that the electron gas in the metal is free and moves in a certain effective potential formed by the lattice ions and the impurity centers. The potentials U and Φ (Fig. 1) describe the potentials of the electron–electron interaction and the interactions of electrons with the metal and impurity atoms, respectively. Below, we present these potentials in the explicit form as a function of the metal work function and the concentration of electrons, lattice ions, and impurity particles.

To derive an expression for the thermionic current emitted from the system, we need a reliable mathematical formalism that would enable us to take into consid-

eration all the phenomena of interest. Below, we will show that it is convenient to take the apparatus of kinetic Green's functions [3] as such a formalism. To obtain an equation for the thermionic current, we first write out a well-known result for the current density of a system of particles, borrowed from a course in quantum mechanics:

$$\mathbf{J}(\mathbf{R}, t) = \frac{i\hbar}{2M} e\{\nabla_{\mathbf{R}}\rho(\mathbf{R}', \mathbf{R}, t) - \nabla_{\mathbf{R}'}\rho^*(\mathbf{R}', \mathbf{R}, t)\}\Big|_{\mathbf{R}' = \mathbf{R}}, \quad (3)$$

where $\hbar$ is the Planck constant (henceforth, we set it equal to unity) and M is the electron mass. The one-particle density matrix $\rho(\mathbf{R}', \mathbf{R}, t)$, expressed in terms of secondary quantized field operators, is given by

$$\rho(\mathbf{R}', \mathbf{R}, t) = \langle \Psi^+(\mathbf{R}, t)\Psi(\mathbf{R}', t)\rangle. \quad (4)$$

Here, $\hat{\Psi}$ and $\hat{\Psi}^+$ are the Heisenberg field operators acting in the space of occupation numbers. These operators satisfy the Fermi statistics and, in our case, describe the motion of electrons with regard to their interaction with the fields of the lattice and of impurity ions. It is well known that the wave functions in the Heisenberg model are independent of time. Hence, one can average (4) over the ground state of the system in which the interaction with impurity is neglected (we assume the adiabatic switching off of the interaction with impurity particles as $t \longrightarrow -\infty$).

It is important that, for simplicity, we omitted the spin indices in (4). Indeed, in the absence of external magnetic fields, the function $\rho(\mathbf{R}', \mathbf{R}, t)$ is isotropic with respect to the spin variables; i.e.,

$$\rho_{ij}(\mathbf{R}', \mathbf{R}, t) = \rho(\mathbf{R}', \mathbf{R}, t)\delta_{ij}.$$

Hence, it is very convenient to consider a model of electrons with the spin equal to zero. In the final results, we will make corrections due to the electron spin by means of appropriate numerical coefficients.

For a given density matrix $\rho(\mathbf{R}', \mathbf{R}, t)$, Eq. (3) actually answers the question posed. Hence, our aim is to obtain an explicit expression for $\rho(\mathbf{R}', \mathbf{R}, t)$. The most natural method for finding this matrix is the apparatus of nonequilibrium Green's functions. Indeed, one can easily establish a direct relation between the one-particle density matrix ρ and the kinetic Green's function G^{-+}:

$$\rho(\mathbf{R}', \mathbf{R}, t) = -iG^{-+}(\mathbf{R}'t', \mathbf{R}t)\big|_{\mathbf{R}' = \mathbf{R}, t' = t},$$

where G^{-+} is defined by

$$iG^{-+}(\mathbf{R}t, \mathbf{R}'t) = -\langle \hat{\Psi}^+(\mathbf{R}'t')\hat{\Psi}(\mathbf{R}t)\rangle. \quad (5)$$

Hence, taking into account (3) and (5), we can rewrite the expression for the z projection of the current-density vector in terms of nonequilibrium Green's functions as

$$J_z^{\text{therm}} = \frac{e}{M}\mathrm{Re}\int_0^\infty \frac{d\omega}{2\pi}\int \frac{d\mathbf{k}_\perp}{(2\pi)^2} \times\left(\frac{d}{dz} - \frac{d}{dz'}\right)G^{-+}(z, z', \mathbf{k}_\perp, \omega)\big|_{z = z'}. \quad (6)$$

Here, $G^{-+}(z, z', \mathbf{k}_\perp, \omega)$ is the Fourier image of G^{-+} with respect to the variables $\mathbf{r}_\perp - \mathbf{r}'_\perp$ (this vector lies in the plane xy) and $t - t'$. In what follows, we will omit the arguments $\mathbf{k}_\perp$ and ω to simplify the expressions. Thus, the problem of calculating the thermionic current has reduced to the determination of the kinetic Green's function G^{-+}. In the next section, we will show that such a transition from the one-particle density matrix to nonequilibrium Green's functions allows us to reduce our problem (i.e., the determination of the explicit form of G^{-+}) to a system of integrodifferential equations.

3. EQUATIONS FOR RETARDED (ADVANCED) GREEN'S FUNCTIONS

As we pointed out above, to determine an explicit expression for the thermionic current, we have to calculate the kinetic function G^{-+}. It is known from the theory of Green's functions that G^{-+} can be represented as a sum of diagrams corresponding to different orders of perturbation theory. Note that this sum can be expressed as the Dyson integrodifferential equation [4] (we omit the arguments $\mathbf{k}_\perp$ and ω),

$$G^{-+}(z, z') = -\int_{-\infty}^{\infty} dz_1 dz_2 G^{\mathrm{R}}(z, z_1)\Sigma^{-+}(z_1, z_2)G^{\mathrm{R}*}(z', z_2). \quad (7)$$

Here, Σ^{-+} is a mass operator that represents a sum of appropriate irreducible diagrams and describes the interaction between the electron gas in a metal and the field of impurity ions (as will be clear from the further analysis, it is proportional to the impurity concentration), and G^{R} is the retarded Green function defined by

$$iG^{\mathrm{R}}(\mathbf{R}'t', \mathbf{R}t) = \begin{cases} \langle \hat{\Psi}^+(\mathbf{R}'t')\hat{\Psi}(\mathbf{R}t) - \hat{\Psi}(\mathbf{R}t)\hat{\Psi}^+(\mathbf{R}'t')\rangle, & t' > t, \\ 0, & t' < t. \end{cases}$$

The retarded Green's function G^R also satisfies the Dyson equation (with appropriate mass operator Σ^R), which can be expressed as a system of partial integrodifferential equations. Solving this system with given initial and boundary conditions for G^R and for explicitly given mass operator Σ^{-+}, we can calculate the required thermionic current. Thus, our immediate task is to solve a system of equations for the retarded Green function and the operator Σ^{-+}.

Let us make a few remarks concerning the determination of Σ^{-+}. First, we stress that the metal boundary manifests itself only within a thin layer of about the thermal de Broglie wavelength of a particle. Since most collisions occur in the bulk of the metal, we can neglect the boundary effects and assume that the metal is infinite. In this case (as follows from the kinetic equation for G^{-+}), the operator Σ^{-+} determines the rate at which a particle reaches a certain state due to the interaction with an impurity. Since we consider a system in the state of thermodynamic equilibrium, the principle of detailed equilibrium yields

$$\Sigma^{-+}(\omega, \mathbf{p}) = \frac{G^{-+}(\omega, \mathbf{p})}{G^{+-}(\omega, \mathbf{p})}\Sigma^{+-}(\omega, \mathbf{p}) = i\frac{G^{-+}}{G^{+-} - G^{-+}}\Gamma. \tag{8}$$

Here, Σ^{ab} and G^{ab} are the Fourier images of the mass operator and the Green function, respectively, and $\Gamma = 2\,\mathrm{Im}\,\Sigma^R$. Below, we will show that Γ is a positive quantity that describes indeterminacy in the values of the electron momentum for a given total energy. In the first order of perturbation theory, when the functions G^{-+} and G^{+-} correspond to the free motion of conductivity electrons in metal in the absence of impurities, formula (8) can be rewritten as

$$\Sigma^{-+}(\omega, \mathbf{p}) = -in(\omega)\Gamma(\omega, \mathbf{p}) = -i\Gamma\{e^{\omega/T} + 1\}^{-1}. \tag{9}$$

Here, $n(\omega)$ are the Fermi occupation numbers of electrons in metal. One can see from this formula that all details of the interaction between an impurity and electrons are contained precisely in $\Gamma(\omega, \mathbf{p})$. We assume that this is a short-range interaction and set

$$\Sigma^{-+}(z, z') = \mathrm{const} \cdot \delta(z - z'), \tag{10}$$

where const is independent of z and z' (it is assumed that the concentration of impurity centers is independent of the z coordinate). In this case, formula (9) remains valid, but Σ^{-+} and Γ are now independent of the z component of the momentum vector $\mathbf{p}$. Comparison of (9) and (10) shows that

$$\mathrm{const} = -in(\omega)\Gamma.$$

It is important that, according to (9), the mass operator Σ^{-+} is proportional to the occupation numbers of electrons states. Since these numbers are equal to zero in vacuum, we set

$$\Sigma^{-+}_{\mathrm{vac}} = 0.$$

Thus, from perturbation theory and the principle of detailed equilibrium, we have derived an expression for the mass operator Σ^{-+}. According to Eq. (7), to obtain a final expression for the kinetic Green function G^{-+} (and, further, for the thermionic current from the composite system by (6)), we have to find an explicit expression for the retarded function G^R. First of all, we have to find out which functions G^R exactly we need. For this purpose, we note that the particle flux density j_z in (6) should be independent of the coordinate z. Therefore, without loss of generality, we can set $z = -0$ (which is the left limit from the side of vacuum). Taking into account the assumption made about the locality of interaction (10) and the fact that $\Sigma^{-+}(z, z') = 0 = 0$ in vacuum ($z < 0$), the expression for G^{-+} reduces to

$$G^{-+}(z, z') = -\int_0^L G^R(z, z_1)\Sigma^{-+}_{\mathrm{sl}}G^{R*}(z', z_1)\,dz_1 - \int_L^\infty G^R(z, z_1)\Sigma^{-+}_{\mathrm{met}}G^{R*}(z', z_1)\,dz_1, \tag{11}$$

where $\Sigma^{-+}_{\mathrm{sl}}$ and $\sigma^{-+}_{\mathrm{met}}$ are the mass operators describing the interaction between electrons and the impurity in the layer and in the metal half-space, respectively. It is clear from (11) that, for z, $z' < 0$, only the components $G^R(z < 0, 0 < z' < L)$ and $G^R(z < 0, z' > L)$ are of interest.

To obtain an expression for the thermionic current by means of Eqs. (6) and (7), we have to calculate two retarded Green functions G^R, namely, the components $G^R(z < 0, 0 < z' < L)$ and $G^R(z < 0, z' > L)$. To this end, we consider the Dyson equation for G^R. In the general case, this is an integrodifferential equation; however, taking into account the assumption about the locality of interaction (10), we can rewrite this equation as a second-order partial differential equation. Hence, taking

into account the nonhomogeneity of space, we arrive at the following system (it is assumed that $z < 0$):

$$\left[\Omega_+ - \frac{\hbar^2}{2M}\frac{d^2}{dz'^2} - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_+ + U_0\right]G^{\mathrm{R}}(z,z') = 0, \quad z' > L, \tag{12}$$

$$\left[\Omega_0 - \frac{\hbar^2}{2M}\frac{d^2}{dz'^2} - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_0 + \Phi_0\right]G^{\mathrm{R}}(z,z') = 0, \quad 0 < z' < L, \tag{13}$$

$$\left[\omega - \frac{\hbar^2}{2M}\frac{d^2}{dz'^2} - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_-\right]G^{\mathrm{R}}(z,z') = \delta(z-z'), \quad z' < 0. \tag{14}$$

Here, μ_-, μ_0, and μ_+ are the chemical potentials of the electron in vacuum, in the metal layer, and in the metal half-space, respectively. It is important that the following relation holds in the state of thermodynamic equilibrium:

$$\mu_- = \mu_0 = \mu_+,$$

which follows directly from the conservation law of the number of particles (electrons) in the system. Potentials U_0 and Φ_0 in (12)–(14) describe the fields created by the lattice in the metal and in the layer, respectively. In the present model, we do not take into account the spatial variation of U_0 and Φ_0 and consider them as certain effective, spatially averaged, potentials.

The quantities

$$\Omega_k = \omega + \Delta_k + i\Gamma_k/2, \quad k = +, 0,$$

on the left-hand sides of Eqs. (12)–(14) determine the generalized (i.e., obtained in the presence of impurities) frequencies. The subscript $k = +$ corresponds to the metal half-space, and $k = 0$ corresponds to the metal layer. It is important that, in general, the frequencies Ω_k are complex numbers. The real part of Ω_k contains the term Δ_k, which depends on the impurity concentration and describes a shift in the electron energy due to the impurity field. Note that Δ_k are related to Σ^{R} by the following formula:

$$\Delta_k = \mathrm{Re}\Sigma^{\mathrm{R}} = \frac{\Sigma^{--} + \Sigma^{++}}{2}. \tag{15}$$

The imaginary part of Ω_k, i.e., Γ_k, describes the scattering of electrons by impurities in the metal and is also related to Σ^R:

$$\Gamma_k = 2\mathrm{Im}\Sigma^{\mathrm{R}} = \Sigma^{+-} - \Sigma^{-+}. \tag{16}$$

Below, we will show that Γ_k determines the smearing of the generalized distribution function. Note that, in [5], a quantum-mechanical calculation of the shift Δ_k and the widths Γ_k was carried out for the so-called Lorentz gas, i.e., a nonideal gas in the field of external static ions. It was shown that Γ_k is directly related to the scattering amplitude on an external ion (see also [6]).

A general solution to the system of equations for the retarded Green functions G^{R} (12)–(14) can be expressed as

$$G^{\mathrm{R}}(z<0,z') = \begin{cases} -\dfrac{iM}{s}\exp(is|z-z'|) + C_1\exp(-is(z+z')), & z' < 0, \\ \{A_1\exp(i\beta z') + A_2\exp(-i\beta z')\}\exp(-isz), & 0 < z' < L, \\ C_2\exp(i\alpha z' - isz), & z' > L, \end{cases} \tag{17}$$

where A_1, A_2, C_1, and C_2 are constants that are determined from the continuity of the Green functions G^{R} and their derivatives on the boundaries between the vacuum and the metal layer and the metal layer and the metal half-space. Here, the following notations are introduced:

$$\alpha = \sqrt{\frac{2M}{\hbar^2}\left(\omega - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_+ + U + \frac{i\Gamma_+}{2}\right)},$$

$$\beta = \sqrt{\frac{2M}{\hbar^2}\left(\omega - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_0 + \Phi + \frac{i\Gamma_0}{2}\right)},$$

$$s = \sqrt{\frac{2M}{\hbar^2}\left(\omega - \frac{\hbar^2\mathbf{k}_\perp^2}{2M} + \mu_-\right)}.$$

Below, we will show that the widths Γ in the above formulas are positive. Hence, the imaginary parts of α and β are also positive and describe the attenuation of electron waves in the medium (see (17)). To obtain a final expression for G^{R}, we have to determine the constants that appear in (17). These constants can be determined from the boundary conditions for G^{R}. As the boundary conditions, we can choose the continuity of the Green functions G^{R} themselves and their first-order deriva-

tives (due to the continuity of the particle flux density) on the interfaces $z = 0$ and $z' = L$. Taking into account these conditions for the functions $G^R(z < 0, 0 < z' < L)$ and $G^R(z < 0, z' > L)$, we obtain

$$G^R(z<0, z'>L) = \frac{4iM\beta\exp(-i\alpha L)}{(\beta-\alpha)(\beta-s)\exp(i\beta L)-(\beta+\alpha)(\beta+s)\exp(-i\beta L)} \times \exp(i\alpha z' - isz), \tag{18}$$

$$G^R(z<0, 0<z'<L) = 2iM\frac{(\alpha-\beta)\exp(i\beta(L-z'))-(\alpha-\beta)\exp(-i\beta(L-z'))}{(\beta+\alpha)(\beta+s)\exp(-i\beta L)-(\alpha-\beta)(s-\beta)\exp(i\beta L)} \times \exp(-isz). \tag{19}$$

Now, we have expressions for the Green functions G^R and can obtain an explicit expression for the thermionic current emitted into vacuum from the composite system considered. Using (11), we can determine the kinetic function G^{-+} from (10), (18), and (19). Next, taking into account (6), we can directly calculate the thermionic current. For example, the current due to the right half-space (the domain $z > L$) is given by

$$j_z^+ = -\frac{8M}{\pi}\int_0^\infty d\omega n_+(\omega)\int\frac{d\mathbf{k}_\perp}{(2\pi)^2} \times \frac{|\beta|^2 \mathrm{Re}s}{|(\beta-\alpha)(\beta-s)\exp(i\beta L)-(\beta+\alpha)(\beta+s)\exp(-i\beta L)|^2}\frac{\Gamma_+}{\mathrm{Im}\alpha}. \tag{20}$$

The minus sign indicates the direction of current, namely, from metal to vacuum. Similarly, for the current due to the metal layer ($0 < z < l$), we obtain

$$j_z^0 = -\frac{4M}{\pi^2}\int_0^\infty d\omega n_0(\omega)\int\frac{d\mathbf{k}_\perp}{(2\pi)^2}\Gamma_0 \mathrm{Re}s \times \frac{|\beta+\alpha|\exp(\mathrm{Im}\beta L)+|\beta-\alpha|\exp(-\mathrm{Im}\beta L)}{|(\beta-\alpha)(\beta-s)\exp(i\beta L)-(\beta+\alpha)(\beta+s)\exp(-i\beta L)|^2} \times \frac{\sinh(\mathrm{Im}\beta L)}{\mathrm{Im}\beta L}. \tag{21}$$

The total thermionic current emitted into vacuum is given by the sum of these currents:

$$J_z^{\mathrm{therm}} = j_z^0 + j_z^+. \tag{22}$$

We should make several remarks concerning formulas (20) and (21). First, these expressions include quantum phenomena associated with the transmission of electrons through a barrier (for example, the above-barrier reflection). The quantity Res on the right-hand side of the expression for the current intensity is directly related to the transmission coefficient through the barrier. This quantity shows that not all the electrons can leave the metal. Only those electrons whose energy is greater than the work function of the metal (i.e., those with Re$s > 0$) can leave the metal. Second, formulas (20) and (21) contain the Fermi occupation numbers of electrons. This is associated with the fact that, in the temperature range of interest (up to about 1000 K), the electron gas is strongly degenerate. Third, one can see that the expressions for the thermionic current explicitly contain the widths Γ_k, which depend on the presence of impurities in the metal. For sufficiently low impurity concentrations and, consequently, for small Γ_k (recall that Γ_k are proportional to the impurity concentration), one can show that the right-hand sides of (20) and (21) contain the following function of the energy and momentum of electrons:

$$f(\omega, k_z, \mathbf{k}_\perp) = n(\omega) \times \frac{\Gamma}{\left(\omega+\Delta-\frac{\hbar^2k_z^2}{2M}-\frac{\hbar^2\mathbf{k}_\perp^2}{2M}+\mu-U\right)^2+\Gamma^2}. \tag{23}$$

This function is proportional to the occupation numbers of electron states and plays the role of the distribution of electrons over states in the presence of the impurity field. We will interpret $f(\omega, k_z, \mathbf{k}_\perp)$ as a generalized distribution function, where energy ω and momentum $(k_z, \mathbf{k}_\perp)$ are considered to be independent variables.

Note that, in general, the quantities Δ and Γ in (23) may also depend on the energy ω and momentum $(k_z, \mathbf{k}_\perp)$. It is clear from (23) that Δ represents a variation in the total energy of an electron due to the impu-

rity field and Γ describes the indeterminacy in the electron momentum for a given total energy.

This indeterminacy in the momentum for a given energy of electron is associated with violation of the homogeneity of space by impurity sites, which leads to violation of the momentum conservation law. It is important that, in the limit as $\Gamma \longrightarrow 0$ (i.e., in the case of an ideal metal, which does not contain impurities), we have

$$f(\omega, k_z, \mathbf{k}_\perp) \propto n(\omega)\delta\left(\omega + \Delta - \frac{\hbar^2 k_z^2}{2M} - \frac{\hbar^2 \mathbf{k}_\perp^2}{2M} + \mu - U\right). \quad (24)$$

In this case (as we pointed out in the Introduction), there is a one-to-one correspondence between the energy and momentum of an electron (see (2)). In the limit of high temperatures (where the classical theory is applicable), we can apply distribution function (24) to again obtain the Richardson–Dashmen result (1).

Thus, one can expect that the field of impurity centers can change the distribution function of electrons in a metal. Hence, for sufficiently high impurity concentrations, we may have a substantial deviation from the classical law (1). In the subsequent sections, we present the results of the numerical analysis of Eqs. (20)–(22).

4. NUMERICAL ANALYSIS

Consider explicit formulas (20) and (21) for the thermionic current emitted from the composite system into vacuum. We have shown that, in the presence of an impurity, the current depends on the generalized distribution function $f(E, \mathbf{p})$, where energy E and momentum $\mathbf{p}$ are independent variables, in contrast to the classical result of Fermi, where the energy is uniquely related to the momentum. It is important that, according to (23), the explicit form of the function $f(E, \mathbf{p})$ depends on the concrete form of interaction between electrons and impurity ions through $\Delta(E, \mathbf{p})$ and $\Gamma(E, \mathbf{p})$. Here, $\Delta(E, \mathbf{p})$ defines the energy shift of an electron due to the interaction with an impurity, and $\Gamma(E, \mathbf{p})$ describes the smearing of the distribution function, i.e., the indeterminacy in the momentum for a given energy.

Our primary concern is how the thermionic current depends on temperature, the thickness of the metal layer, and the impurity concentrations (in the metal half-space and in the layer). To obtain a final explicit expression for the generalized distribution function $f(E, \mathbf{p})$ (23), we have to find the dispersion laws for $\Delta(E, \mathbf{p})$ and $\Gamma(E, \mathbf{p})$; i.e., we must determine the explicit dependence of these parameters on the total energy and momentum. This problem has been studied in sufficient detail in [5, 6, 8, 9]. In particular, it was shown that, when the interaction between electrons and impurities is described by the potential $U(\mathbf{r}) \propto \delta(\mathbf{r})$, the generalized distribution function (23) behaves as $f \propto 1/p^4$ for large $\mathbf{p}$ (see [9]); in the case of the Coulomb potential $U(\mathbf{r}) \propto 1/r$, it was shown in [8] that $f \propto 1/p^8$. In [5], the case of the Debye potential was investigated in detail. In that paper, the equation for the mass operator Σ^R was solved numerically (recall that, according to Eqs. (15) and (16), Δ and Γ are expressed in terms of Σ^R) in the case of a nonideal gas in the field of external ions, and the following convenient asymptotics (as $p \longrightarrow \infty$) was proposed for the width $\Gamma(E, \mathbf{p})$:

$$\Gamma(E, \mathbf{p}) = \frac{2\sqrt{2}\pi e^4 N_p}{\sqrt{M}} \frac{\sqrt{E}}{E_p^2}; \quad (25)$$

here, M is the electron mass, N_p is the impurity concentration in the metal, and $E_p = \mathbf{p}^2/2M$ is the electron kinetic energy. In the numerical calculations, we use the following approximate expression for $\Gamma(E, \mathbf{p})$:

$$\Gamma(E, \mathbf{p}) = \frac{2\sqrt{2}\pi e^4 N_p}{\sqrt{M}} \frac{\sqrt{E}}{E_{p\perp}^2 + U^2}. \quad (26)$$

Formula (26) is analogous to (25); the only difference is that we have neglected the dependence on the momentum component p_z in (26), setting

$$E_{pz} = p_z^2/2M = U$$

(recall that, according to the classical theory, only electrons with $E_{pz} > U$) can leave the metal).

In [5], the shift $\Delta(E, \mathbf{p})$ was also calculated for a plasma with a small concentration of impurities. This analysis employs the properties of the imaginary and real parts of the mass operator Σ^R (see [6]) and is based on perturbation theory. It was found that

$$\Delta(E, \mathbf{p}) = \frac{\sqrt{\pi} e^2 n_p e_p^2}{\sqrt{4Tn_I e_I^2}}. \quad (27)$$

Here, n_I, e_I and n_p, e_p are the concentrations and charges of the ionic lattices and the impurity sites, respectively. Note that, in the limit of low impurity concentrations, $\Delta(E, \mathbf{p})$ is a linear function of n_p. Moreover, it follows from (27) that, in the first order of perturbation theory, this shift is independent of energy and momentum and depends only on the concentrations and charges of ions. We will use approximation (27) in the numerical analysis to give a qualitative illustration of the oscillating behavior of the thermionic current emitted from the composite system as a function of the impurity concentration.

Using the concentration of conductivity electrons in the metal and the work function, we can easily determine the effective potential in which electrons move in the case of an ideal metal, i.e., a metal with negligible concentration of impurity centers:

$$U_0 = E_F^0 + W. \quad (28)$$

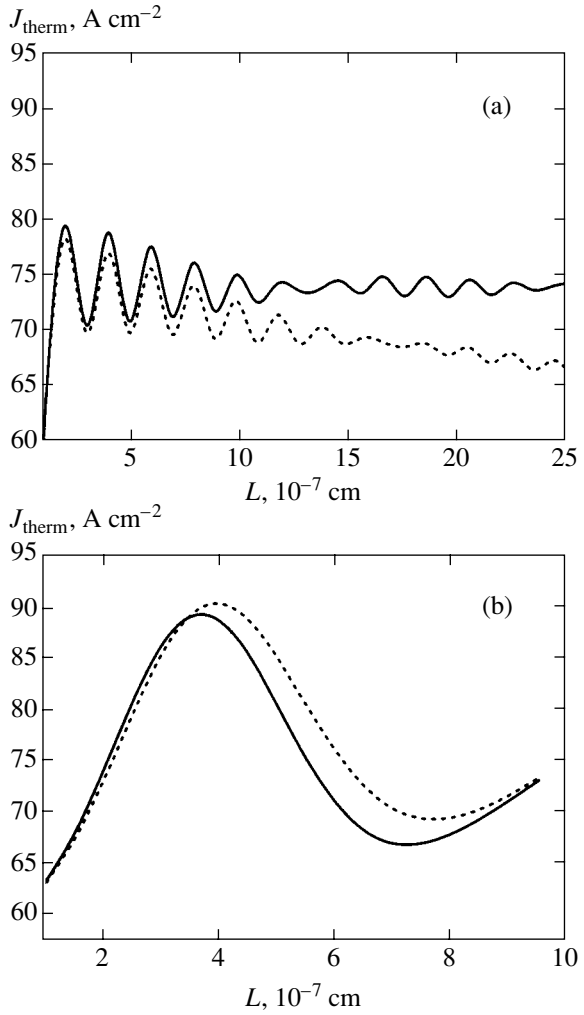

Fig. 2. Thermionic current J_{therm} as a function of the layer thickness L for various impurity concentrations $N_p^{(s)}$ in the layer; 10^{18} cm^{-3} (solid curve) and 10^{20} cm^{-3} (dashed curve); T = 1500 K, and the impurity concentration $N_p^{(M)}$ in the metal half-space is (a) 10^{19} cm^{-3} and (b) 10^{18} cm^{-3}.

Here, E_F^0 is the Fermi energy of electrons in the absence of the potential U_0 and W is the work function of the metal. Note that we borrowed the values of the work function W for our numerical calculations from experimental tables on photoemission. We can also use experimental data on the concentration of conductivity electrons in metals to determine the values of E_F^0 by the well-known formula

$$E_F^0 = \frac{\hbar^2}{2M}(3\pi^2 N)^{2/3}, \tag{29}$$

where N is the electron concentration. It is important that, strictly speaking, expression (29) is valid for an infinite metal and cannot be used for determining the chemical potential of a metal layer of finite thickness because this formula presumes that the spatial distribution of electrons is uniform. In the case of a finite thickness, this distribution is essentially nonuniform, which requires a more careful analysis. However, it can be shown that, for a metal layer of finite thickness of $L >$ 10 Å, we can neglect the effects associated with the finiteness of a sample and apply formula (29).

Using the correction $\Delta(E, \mathbf{p})$ (27) to the electron energy due to impurities, we can generalize the result (28) to the case of small impurity concentrations by adding an appropriate term to the right-hand side; then, we have

$$U = E_F^0 + W + \Delta(E, \mathbf{p}). \tag{30}$$

It should also be noted that the potential inside an isolated metal (a metal layer) may differ from the corresponding potential in the composite system. When two metals are brought into contact, part of electrons migrate from one metal to another. Then, on the one hand (in one metal), we have an excess of electrons, whereas, on the other hand (in the second metal), we have a depletion of electrons near the surface. Such a charge redistribution between two conductors gives rise to a depletion region that prevents further migration of electrons (the so-called double layer). The formula for the double-layer potential is well known (see, for example, [7]) and can be expressed in terms of the difference of the work functions of the metals:

$$\Delta\Phi = W_1 - W_2. \tag{31}$$

To take into consideration a correction due to the additional potential $\Delta\Phi$, we replace the potential Φ of an isolated metal layer in (17) by $\Phi + \Delta\Phi$ in our numerical calculations.

Thus, we numerically analyzed the following system: as metal 1 (a layer of thickness L), we took silver, and, as metal 2 (metal half-space), sodium. The work functions and the concentrations of conductivity electrons in these metals are as follows: W = 2.3 eV and $N_e = 2.5 \times 10^{22}$ cm^{-3} for silver and W = 1.8 eV and $N_e = 0.85 \times 10^{22}$ cm^{-3} for sodium (the data are borrowed from [10]). Figures 2–6 show the total thermionic current (20)–(22) emitted from the composite system versus temperature, layer thickness L, and the impurity concentrations in the metal half-space ($N_p^{(M)}$) and in the layer ($N_p^{(s)}$) for variations of other parameters.

First, we consider the dependence of the thermionic current J_{therm} on the thickness L of the metal layer. It follows from (20)–(22) that this function should exhibit oscillating behavior (with decaying amplitude, since

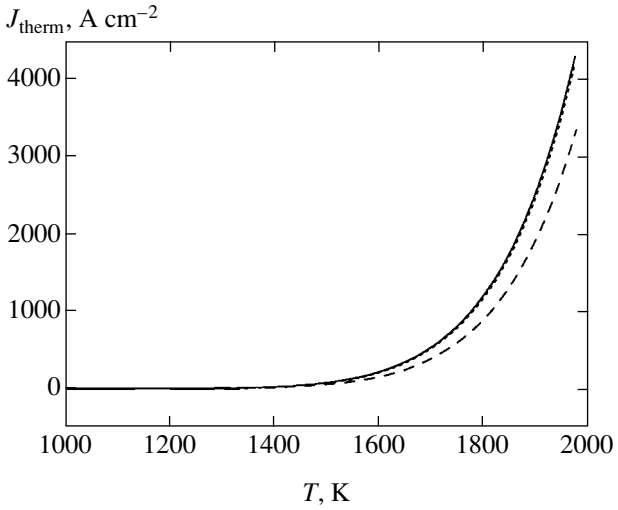

Fig. 3. Thermionic current as a function of temperature for various impurity concentrations $N_p^{(\mathrm{M})}$ in the metal half-space; 10^{18} cm^{-3} (solid curve), 10^{19} cm^{-3} (dashed curve), and 10^{20} cm^{-3} (dotted curve); the layer thickness $L = 10^{-6}$ cm, and the impurity concentration in the layer is $N_p^{(\mathrm{s})} = 10^{19}$ cm^{-3}.

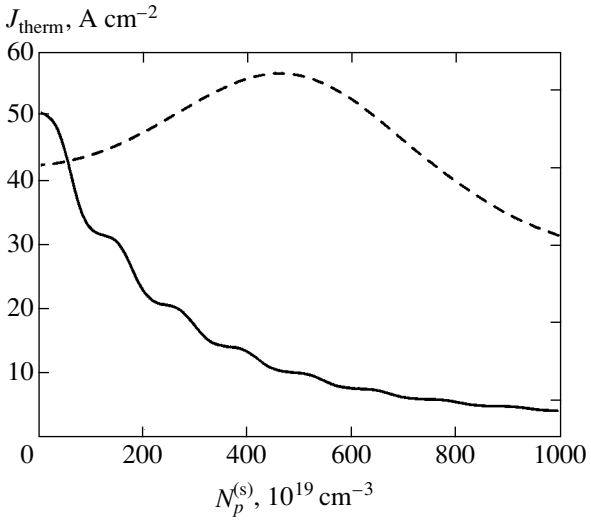

Fig. 4. Thermionic current as a function of the impurity concentrations $N_p^{(\mathrm{s})}$ in the layer for various values of the layer thickness L; 10^{-6} cm (solid curve) and 10^{-7} cm (dashed curve); the impurity concentration in the metal half-space is $N_p^{(\mathrm{M})} = 10^{20}$ cm^{-3} and $T = 1500$ K.

$\mathrm{Im}\beta > 0$) due to the presence of exponential functions $\exp(\pm i\beta L)$ in the integrand. This assumption is confirmed by the graphs shown in Figs. 2a and 2b.

Figure 2a demonstrates the behavior of J_{therm} for sufficiently large L. One can see that, as the layer thickness L increases, the thermionic current asymptotically tends to a value equal to the current emitted by a heated metal half-space; the rate of this convergence essentially depends on the impurity concentration $N_p^{(\mathrm{s})}$ in the layer. Indeed, according to (20), the probability that an electron from the domain $z < L$ passes through the layer and reaches the vacuum is proportional to $\exp(-2\mathrm{Im}\beta L)$. Thus, increasing the layer thickness L or the impurity concentration $N_p^{(\mathrm{s})}$ (recall that $\mathrm{Im}\beta \propto N_p^{(\mathrm{s})} > 0$), we thereby decrease the contribution of the half-space $z > L$ and simultaneously increase the contribution of the domain $0 < z < L$ (see (21)). Figure 2b shows that a variation in the impurity concentration in the layer leads to variations in the period and amplitude of oscillations of the thermionic current due to the presence of the exponential factors $\exp(\pm i\beta L)$ in (20) and (21).

Figure 3 represents the thermionic current as a function of temperature for various impurity concentrations $N_p^{(\mathrm{M})}$ in the metal. One can see that, for sufficiently high concentrations of $N_p^{(\mathrm{M})} \sim 10^{20}$ cm^{-3}, we can observe a deviation from the classical result, which corresponds to extremely low concentrations and reduces to the Richardson–Dashmen law (1) at high temperatures. A similar curve can be drawn in the case of variations in the impurity concentration in the layer, and a similar deviation from the classical result can be displayed for sufficiently high concentrations $N_p^{(\mathrm{s})}$.

Figure 4 represents the graph of J_{therm} as a function of $N_p^{(\mathrm{s})}$ for various values of L. One can see that this

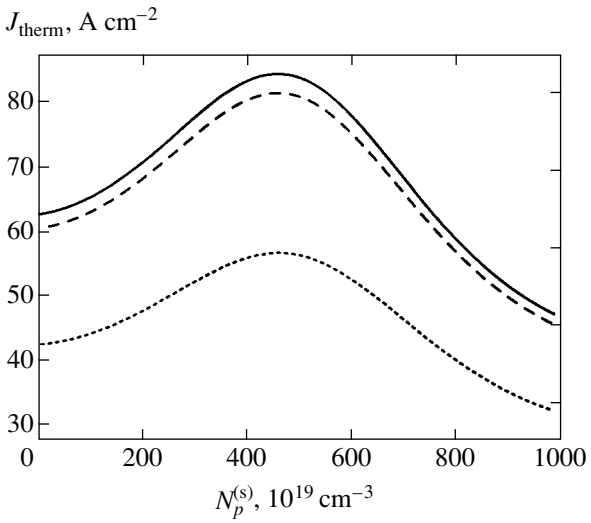

Fig. 5. Thermionic current as a function of the impurity concentrations $N_p^{(\mathrm{s})}$ in the layer for various impurity concentrations $N_p^{(\mathrm{M})}$ in the metal half-space: 10^{18} cm^{-3} (solid curve), 10^{19} cm^{-3} (dotted curve), and 10^{20} cm^{-3} (dashed curve); the layer thickness is $L = 10^{-7}$ cm, the impurity concentration in the layer is $N_p^{(\mathrm{s})} = 10^{19}$ cm^{-3}, and $T = 1500$ K.

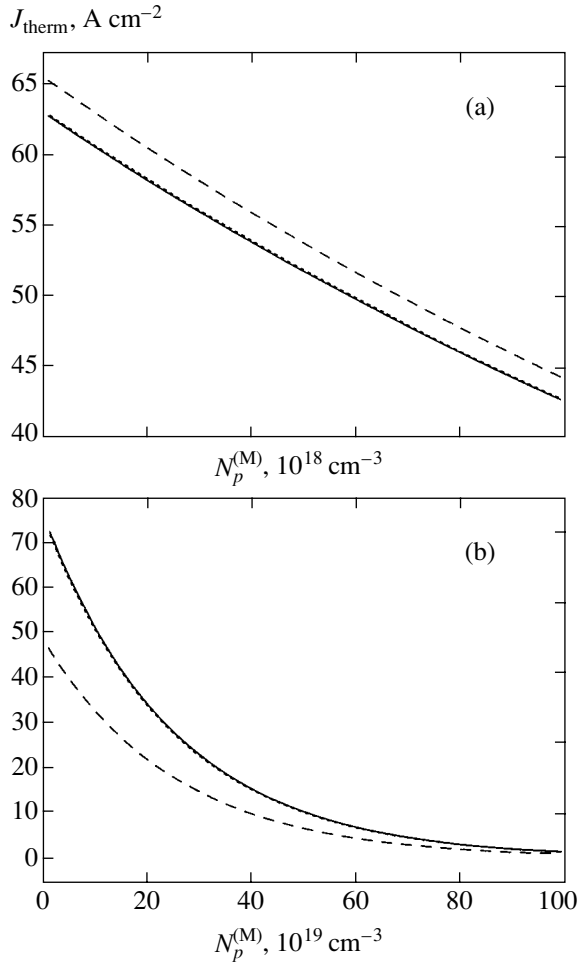

Fig. 6. Thermionic current as a function of the impurity concentrations in the metal half-space for various impurity concentrations in the layer: (a) $N_p^{(s)} = 10^{18}$ cm^{-3} (dashed curve), 10^{19} cm^{-3} (solid curve), and 10^{21} cm^{-3} (dotted curve); $L = 10^{-7}$ cm and $T = 1500$ K; (b) $N_p^{(s)} = 10^{18}$ cm^{-3} (solid curve), 10^{20} cm^{-3} (dashed curve), and 10^{21} cm^{-3} (dotted curve); $L = 10^{-6}$ cm and $T = 1500$ K;

function has an oscillating character. These oscillations (like the oscillations that occur for the variation of the layer thickness) are attributed to the exponential factors $\exp(\pm i\beta L)$ in the integrands of (20) and (21). One can see that the period and amplitude of these oscillations depend on L.

Figure 5 shows a similar graph of J_{therm} as a function of the impurity concentration $N_p^{(s)}$ in the layer, but now for various impurity concentrations in the half-space of metal. One can see that an increase in the impurity concentration in the metal half-space decreases the amplitude and the period of oscillations of J_{therm}.

Figures 6a and 6b represent the graphs of J_{therm} as functions of the impurity concentration $N_p^{(M)}$ in the metal half-space for various impurity concentrations in the layer for low (Fig. 6a) and high (Fig. 6b) concentrations $N_p^{(M)}$. For low impurity concentrations in the metal (Fig. 6a), the total thermionic current linearly decreases as the concentration $N_p^{(M)}$ increases. As $N_p^{(M)}$ increases, this curve becomes nonlinear (Fig. 6b). As we have already mentioned, an increase in the impurity concentration in the layer decreases the total thermionic current.

5. CONCLUSIONS

Before passing to discussion of the results, we will speak on the case of a single metal half-space (without the metal layer). In this case, the expression for the thermionic current can readily be derived from Eqs. (20)–(22) if we set $\beta = s$ and $n_0 = \Gamma_0 = 0$:

$$j_z^{\text{therm}} = j_z^+ = -\frac{4|e|}{\pi}\int_0^\infty dE\, n_+(E) \times \int \frac{d\mathbf{k}_\perp}{(2\pi)^2}\frac{\text{Re}\beta\,\text{Re}\alpha}{|\beta+\alpha|^2}. \tag{32}$$

This equation has a clear physical interpretation. The quantity $n_+(E)$ in the integrand represents the energy distribution of electrons. The second integral in (32) is none other than the probability that an electron with energy E leaves the metal.

Earlier, we have shown that, due to the presence of impurity ions in the metal, the distribution function of electrons may substantially differ from the classical Fermi–Dirac distribution. The generalized distribution function introduced in this paper depends on the impurity concentration in the metal via the parameters Δ and Γ (see (23)). However, one can see from (32) that, in the case of a half-space, the final expression for the current does not explicitly depend on Δ and Γ. In this case, thermionic current (32) depends on the impurity concentration $N_p^{(M)}$ only via the reflection coefficient ρ of the electron wave from the boundary of the metal (recall that $\text{Im}\alpha$ is proportional to $N_p^{(M)}$). The numerical calculations by formula (32) have shown that this dependence is weak. Thus, in the case of a single half-space, the results may be in good agreement with the Richardson–Dashmen result (1). This fact has been confirmed by numerical analysis.

Thus, we have found that the thermionic current (32) emitted by a heated metal half-space weakly depends on the impurity concentration in the metal; this agrees well with the classical result (1) despite the fact that the distribution function of electrons may differ from the

Fermi distribution. To resolve this seeming contradiction, we rewrite (32) in a somewhat different form by introducing a generalized distribution function. Then, for moderately high impurity concentrations, we obtain

$$j_z^{\text{therm}} = -\frac{8|e|}{M}\int_0^\infty dE n_+(E)\int\frac{d\mathbf{k}_\perp}{(2\pi)^2} \times\int_0^\infty\frac{dk_z}{2\pi}k_z\delta_\Gamma\left(E-\frac{\hbar^2\mathbf{k}^2}{2M}+\mu_+-U\right)(1-\rho^2), \tag{33}$$

where we introduced the auxiliary function

$$\delta_\Gamma(x) = \frac{1}{2\pi}\frac{\Gamma}{x^2+\Gamma^2/4}.$$

Recall that U in (32) is a potential generalized to the case of a doped metal (see (30)). The quantity $1-\rho^2$ defines the fraction of electrons that leave the metal, while ρ^2 defines the fraction of electrons that are reflected from the boundary of the metal. Note that the integral with respect to dk_z is

$$\int_0^\infty dk_z k_z\delta_\Gamma\left(\tilde{E}-\frac{\hbar^2k_z^2}{2M}\right)W(k_z) \propto\int_0^\infty\frac{dE_{k_z}\Gamma}{(\tilde{E}-E_{k_z})^2+\Gamma^2/4}WE_{k_z} \approx W(E)\int_0^\infty\frac{dE_{k_z}\Gamma}{(\tilde{E}-E_{k_z})^2+\Gamma^2/4}\approx 2\pi W(E),$$

where we introduced the notation

$$\tilde{E} = E-\frac{\hbar^2\mathbf{k}_\perp^2}{2M}+\mu_+-U,\quad W = 1-\rho^2,\quad E_{k_z} = \frac{\hbar^2k_z^2}{2M}$$

and took into account that $\Gamma \ll \tilde{E}$ (this is a criterion that determines the applicability limits for perturbation theory and, hence, for the theory described above), as well as the fact that the probability $W = 1-\rho^2$ of transition through the boundary, depends rather weakly on k_z. As a result, we have found that, in the case of a single half-space, the final result does not contain a generalized distribution function. Consequently, the thermionic current depends on the impurity concentration only via the variation of the reflection coefficient ρ.

Now, let us show that formula (32) for thermionic current reduces to the Richardson–Dashmen law (1) in the classical limit. Neglecting the effects of above-barrier reflection, we can write a simplified expression for the reflection coefficient as

$$\rho = \begin{cases}\dfrac{\hbar^2k_z^2}{2M} < \Phi_0,\\ \dfrac{\hbar^2k_z^2}{2M} > \Phi_0.\end{cases} \tag{34}$$

Then, formula (32) for the current intensity simplifies to

$$j_z^{\text{therm}} = -\frac{MT^2}{2\pi^2}e\exp\left(-\frac{E_F-\Phi_0}{T}\right) = -\frac{MT^2}{2\pi^2}e\exp\left(-\frac{W}{T}\right)$$

in the classical limit.

Thus, in this paper, we have attempted to carry out a detailed quantum-mechanical analysis of the processes in real metals and alloys at low temperatures. We have considered two systems: the first is a composite system consisting of a metal layer of finite thickness and a metal half-space, and the second is a simpler system consisting of a metal half-space without metal layer. In both cases, our aim was to derive a finite expression for the thermionic current emitted by the system into vacuum. We assumed that both the metal half-space and the layer may contain impurities. Therefore, we placed special emphasis on the effect of impurities on the distribution function of electrons in metal. We have demonstrated that impurities may substantially change the distribution function (recall that, in the model of free electrons without impurities, this is the Fermi distribution for an electron gas with effective mass). In this case, the energy and momentum of electrons become independent variables because of the violation of the ideal lattice field in the metal and, hence, the homogeneity of the space, by the impurity field.

We have established that two new parameters, Γ and Δ, appear in the theory of a doped metal. These parameters describe the interaction between the electron gas and the impurity ions. The parameter Γ determines the reflection of electron waves by impurity sites and characterizes the indeterminacy in the electron momentum for a given total energy. The parameter Δ can be interpreted as a shift in the electron energy under the influence of the impurity field. Note that Δ and Γ are not independent; they are related by a formula similar to the Kramers–Kronig equation (see [6]).

We have shown that, in the case of a single heated metal half-space without a layer, the expression for the current does not contain a generalized distribution function of electrons and, hence, does not explicitly contain the width Γ. The dependence on the impurity

concentration results from the reflection coefficient only and is extremely weak. As a consequence, the result obtained is in good agreement with the classical Richardson–Dashmen formula (1) in a wide range of temperatures, in spite of the fact that the distribution function of electrons in metal is different from the Fermi distribution.

The situation changes drastically in the case of a more complex system consisting of a plane-parallel metal layer and a metal half-space. As we have shown, the thermionic current depends explicitly on the generalized distribution function that takes into account the interaction between electrons and impurities through the earlier introduced parameters Γ and Δ. We have found that thermionic current from such a system is highly sensitive to the impurity concentration both in the layer and in the metal half-space. Numerical analysis has shown that this dependence has an oscillating character. Moreover, we have found that the dependence of thermionic current on the layer thickness also exhibits oscillating behavior. Thus, the metal layer can play the role of a resonator for the electron waves emitted by the metal half-space, which either enhances or reduces (depending on its thickness) the contribution of the metal half-space (20) to the total current (22).

We should point out possible applications of the results obtained. Owing to the onrush of the electronics industry, there is increasing demand for compact and reliable electron-beam sources of prescribed intensity (new light sources, ultrathin high-resolution monitors, etc.). The results obtained in this paper can be generalized to the case when a metal layer of finite thickness represents a plasma produced by irradiating the surface of a metal by a short-wavelength (picosecond) laser. Then, one can directly affect the thickness of the metal layer and the temperature of the surface, thus changing the current emitted by this composite system. The relations obtained can answer the question of how the thermionic current is changed under the variation of the intensity of the laser beam. It also becomes possible to measure the temperature of the plasma produced by irradiating the metal by a laser beam. We can point out a method for improving the theory proposed. For example, one can consider a layer and a metal half-space at different temperatures and analyze the effect of electron kinetics on the boundary between two metals on the results presented here (which, strictly speaking, are obtained in the equilibrium limit).

ACKNOWLEDGMENTS

We are grateful to N.V. Stepanova for useful suggestions and advice.

This work was supported in part by the program of the President of the Russian Federation for supporting leading scientific schools, project no. NSh-1257.2003.2.

REFERENCES

1. C. Herring and M. H. Nichols, Rev. Mod. Phys. **21**, 185 (1949).
2. G. Baym and L. Kadanoff, *Quantum Statistical Mechanics. Green's Function Methods in Equilibrium and Nonequilibrium Problems* (Benjamin, New York, 1962).
3. L. V. Keldysh, Zh. Éksp. Teor. Fiz. **47**, 1515 (1964) [Sov. Phys. JETP **20**, 1018 (1964)].
4. E. M. Lifshitz and L. P. Pitaevskiĭ, *Physical Kinetics* (Nauka, Moscow, 1979; Pergamon Press, Oxford, 1981).
5. A. N. Starostin, A. B. Mironov, *et al.*, Physica A (Amsterdam) **305**, 287 (2002).
6. A. A. Abrikosov, L. P. Gor'kov, and I. E. Dzyaloshinskiĭ, *Methods of Quantum Field Theory in Statistical Physics* (Fizmatgiz, Moscow, 1962; Prentice Hall, Englewood Cliffs, N.J., 1963).
7. A. I. Ansel'm, *Introduction to the Theory of Semiconductors* (Fizmatlit, Moscow, 1962).
8. A. N. Starostin, N. L. Alexandrov, A. B. Mironov, and M. V. Schipka, Contrib. Plasma Phys. **41**, 299 (2001).
9. V. M. Galitskiĭ and V. V. Yakimets, Zh. Éksp. Teor. Fiz. **51**, 957 (1966) [Sov. Phys. JETP **24**, 637 (1967)].
10. *Physical Quantities. Handbook*, Ed. by I. S. Grigor'ev and E. Z. Meĭlikhov (Énergoatomizdat, Moscow, 1991).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1231–1239.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1380–1390.

SOLIDS
Electronic Properties

Magnetic Phase Transitions in $Nd_{1-x}Ca_xMnO_3$ Manganites

I. O. Troyanchuk[a], V. A. Khomchenko[a], H. Szymczak[b], and M. Baran[b]

[a]*Institute of Solid State and Semiconductor Physics, National Academy of Sciences of Belarus, Minsk, 220072 Belarus*

[b]*Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland*

e-mail: troyan@ifttp.bas-net.by

Received April 7, 2002

Abstract—The magnetic properties of manganites of the $Nd_{1-x}Ca_xMnO_3$ system with $x \le 0.15$ have been studied. It is shown that, in the $0.06 \le x \le 0.1$ interval, the results can be interpreted using a model according to which the concentrational transition from a weakly ferromagnetic (WFM) state ($x = 0$) to a ferromagnetic (FM) state ($x > 0.15$) proceeds via a mixture of the exchange-coupled FM and WFM phases. In the vicinity of $T = 9$ K, samples with $0.06 \le x \le 0.1$ exhibit a spontaneous magnetic phase transition involving reorientation of the magnetization vectors of the WFM and the exchange-coupled FM phases. In the temperature interval between 5 and 20 K, a sample with the composition $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ exhibits metamagnetic behavior. Magnetic phase diagrams in the *H–T* and *T–x* coordinates are presented. The appearance of the spin-reorientation transitions is explained in terms of the magnetic analog of the Jahn–Teller effect with allowance for the fact that, according to the neutron diffraction data, the magnetic moments of neodymium ions in the FM phase are parallel to the magnetic moments of manganese ions. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The phenomenon of giant (colossal) magnetoresistance discovered in manganese-containing perovskites of the $La_{1-x}A_xMnO_3$ system, where A is a divalent element (Ca, Sr, Ba, Pb), has attracted much attention in recent years [1–3]. This interest is explained to a considerable extent by good prospects for the practical applications of these materials [4, 5]. At the same time, manganites exhibiting a large variety of properties have become model objects for the investigation of strongly correlated electron systems.

The interrelated magnetic and electrical properties of the above materials were most exhaustively studied for lanthanum-containing manganites, whereas the manganites of rare-earth elements have been studied to a much lesser extent. However, it is known that the ion radius of A-cations in the ABO_3 perovskite lattice significantly influences the magnetic and electron-transport properties of manganites with such structures. For example, stoichiometric $LaMnO_3$ is an orbital-ordered antiferromagnet (weak ferromagnet) of the A-type with a Néel temperature of $T_N = 140$ K [6, 7]. The transition to an orbital-disordered state in this composition takes place at a temperature of about 700 K [8, 9]. As the radius of the rare-earth ion in position A decreases, the temperature of magnetic ordering decreases and that of the orbital order–disorder transition increases. For $NdMnO_3$, the Néel temperature is about 85 K [10], while the temperature of breakage of the antiferrodistortion order of the d_{z^2} orbitals increases up to 1080 K [8].

As is known, substitution of divalent ions (Ca^{2+}, Sr^{2+}, etc.) for La^{3+} in the $La_{1-x}A_xMnO_3$ structure is accompanied by certain changes in the magnetic properties of these compounds. In particular, $La_{1-x}A_xMnO_3$ with $0 \le x < 0.5$ exhibits a transition from a dielectric antiferromagnetic (AFM) state (for $x = 0$) to a metallic ferromagnetic (FM) state ($0.2 \le x < 0.5$). The magnetic state in the intermediate interval of compositions ($0 < x < 0.2$) can be described either in terms of noncollinear magnetism [11] or within the framework of a model of the FM–AFM phase separation [12–14]. In comparison to the case of lanthanum manganites, the mechanism of evolution of the magnetic state of rare-earth manganites in the course of doping can significantly vary. For example, compounds of the $La_{1-x}Ca_xMnO_3$ system (where Ln = Tb, Dy) do not exhibit a concentrational transition to the homogeneous FM state and show spin-glass behavior in a broad range of compositions [15, 16].

In addition to the aforementioned size effect in the A-sublattice, manganites are characterized by the dependence of their magnetic properties on the magnetic state of a lanthanide component. In particular, Lu^{3+} and Y^{3+} ions (like La^{3+}) are diamagnetic ($L = S = 0$), which implies that the behavior of $LuMnO_3$ and $LaMnO_3$ is determined only by the manganese sublattice. On the contrary, the magnetic properties of $LnMnO_3$ (with Ln = Pr–Yb) depend not only on the *d–d* interactions of manganese ions, but on the intersublattice (*f–d*) and intrasublattice (*f–f*) interactions of rare-earth ions as well; at low temperatures, the contribution of these interactions to the magnetic state can be com-

parable to an analogous contribution from the B-sublattice.

Stoichiometric $NdMnO_3$ (like the manganites of lanthanum, praseodymium, samarium, and europium) is a weak ferromagnet [17]. The weak FM moment of this compound is due to a small noncollinearity of the magnetic moments of manganese that is related to the Dzyaloshinski–Morija interaction [18]. Similar to lanthanum manganite, $NdMnO_3$ can be nonstoichiometric with respect to oxygen, which (similarly to the case of $LaMnO_{3+\delta}$) must lead to an increase in the FM component. The results of neutron diffraction for $NdMnO_{3+\delta}$ containing excess oxygen (the sample was obtained by low-temperature synthesis in air) were interpreted within the framework of a model according to which the magnetic moments of neodymium ions (1.2 μ_B at $T = 2$ K) are ordered below 13 K, whereas the manganese sublattice exhibits both the predominant AFM component below $T = 80$ K (3.2 μ_B at $T = 2$ K) and an FM component below 70 K (about 1.4 μ_B at $T = 2$ K) [19]. At temperatures below 13 K, the magnetic moments of Nd^{3+} ions are parallel to the magnetic moments of manganese ions in the FM component. Although the available data were insufficient to unambiguously decide whether a two-phase state or a homogeneous noncollinear state is realized, the results were interpreted based on a single-phase noncollinear model. However, the magnetic properties of nonstoichiometric $NdMnO_{3.04}$ indicate that the magnetic moments of neodymium ions are aligned oppositely to the moment of the manganese sublattice in the vicinity of the Néel temperature [10]. The combined data of nuclear magnetic resonance and neutron diffraction obtained for slightly doped lanthanum manganites in a magnetic field were indicative of the formation of a two-phase state [6, 14].

Previous investigation of the properties of perovskites of the $Nd(Mn_{0.9}Me_{0.1})O_3$ type (Me = Al, Fe, Cr, Zn) showed that partial replacement of manganese ions leads to a low-temperature magnetic phase transition, the nature of which is unclear [20]. In order to reveal the features of interaction between the magnetic sublattices of neodymium and manganese and to elucidate the mechanism of the concentrational transition from AFM to FM state, we studied the magnetic properties of samples of the $Nd_{1-x}Ca_xMnO_3$ system with $x \le 0.15$. It was found that neodymium ions significantly modify the properties of manganites, so that a concentrational transition to an FM state proceeds via the formation of an intermediate inhomogeneous state subject to spin-reorientation magnetic phase transitions.

2. EXPERIMENTAL

Polycrystalline samples of $Nd_{1-x}Ca_xMnO_3$ with $0.06 \le x \le 0.15$ were synthesized by standard ceramic technology. The initial components were Nd_2O_3 (preliminarily annealed for 1 h at 1000°C to remove adsorbed water), Mn_2O_3, and $CaCO_3$. A mixture of these reagents with a stoichiometric ratio of cations was annealed for 2 h at 1000°C. The product was ground, pressed into disks, and sintered in air for 5 h at 1500°C, followed by cooling at a rate of 100°C/h. The content of oxygen in the synthesized samples was determined by thermogravimetry. Calculated from the weight loss upon the reduction to simple oxides, the error of determination did not exceed 0.4% of the total oxygen content. Superstoichiometric oxygen was removed by annealing the samples in evacuated quartz ampules. Some of the samples were reduced to an oxygen content below the stoichiometric level using metallic tantalum as a reducing agent.

The X-ray diffraction measurements were performed on a DRON-3 diffractometer using CrK_α radiation. It was established that the synthesized samples of $Nd_{1-x}Ca_xMnO_3$ with $x \le 0.15$ are single-phase and possess a perovskite structure with *O*'-orthorhombic ($c/\sqrt{2} < a < b$) distortions, space group *Pbnm*. As the content of calcium increases, the unit cell volume decreases as a result of the growth in the content of Mn^{4+} ions (in the octahedral oxygen environment, the effective ion radii of Mn^{3+} and Mn^{4+} are 0.645 and 0.530 Å, respectively [21]).

The magnetization measurements were performed on a commercial vibrating-sample magnetometer of the OI-3001 type in a temperature interval from 4.2 to 150 K. The measurements in the heating mode were performed for samples cooled in a nonzero magnetic field (field cooling, FC) and in the absence of field (zero field cooling, ZFC). For some of the samples, the magnetization was studied as a function of the applied magnetic field using a SQUID magnetometer.

3. RESULTS

Figure 1 shows the temperature dependence of magnetization for a series of $Nd_{1-x}Ca_xMnO_3$ samples with $0.06 \le x \le 0.15$. The anomalous behavior of magnetization indicative of the appearance of a magnetic order in $Nd_{0.94}Ca_{0.06}MnO_3$ was observed at 73 K (Fig. 1a). In the region of $T = 68$ K, the ZFC sample magnetization exhibits a peak below which the curves of the ZFC and FC samples sharply differ. The FC sample magnetization exhibits a sharp drop at $T_{eff} \approx 9$ K, whereby the orientation of the total magnetic moment becomes opposite to the field direction. At this temperature, the curve of the ZFC sample exhibits a small peak. The temperature hysteresis observed in this region is indicative of the first-order phase transition.

For a compound with the composition $Nd_{0.92}Ca_{0.08}MnO_3$, the temperature of the transition to a magnetically ordered state increases to 84 K (Fig. 1b). Additional reduction of this sample led to a change in the magnetic properties: the compound $Nd_{0.92}Ca_{0.08}MnO_{2.98}$

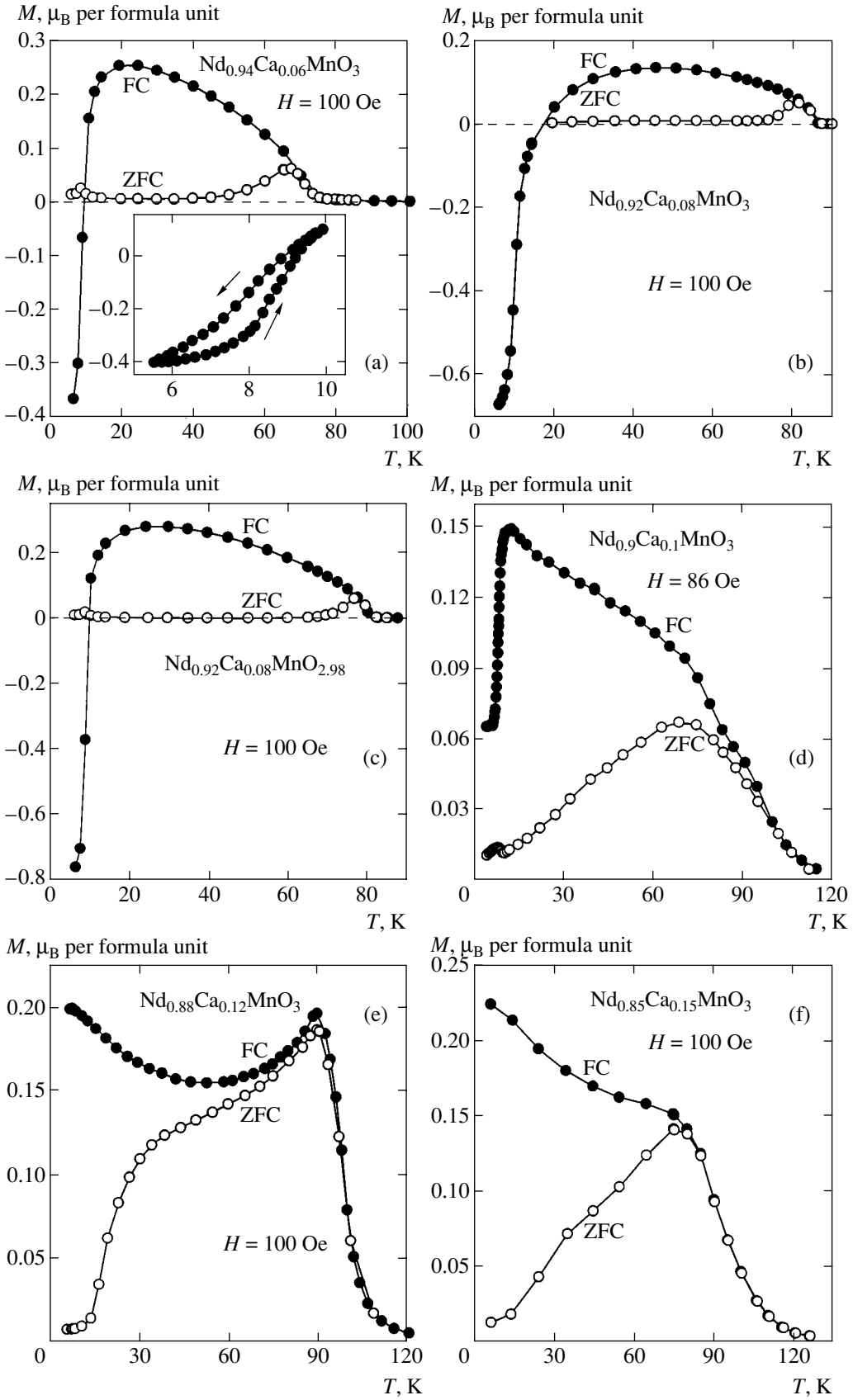

Fig. 1. The temperature dependences of magnetization for samples of the $Nd_{1-x}Ca_xMnO_3$ system with $0.06 \le x \le 0.15$. The inset shows the magnetization of the FC sample in the region of phase transition.

showed some decrease in the temperature of transition to the paramagnetic state and a more pronounced low-temperature phase transition (Fig. 1c). Similar to the case of $Nd_{0.94}Ca_{0.06}MnO_3$, both the stoichiometric (with respect to oxygen) and the reduced samples with $x = 0.08$ exhibited a sharp drop in magnetization in the FC state and a peak in the ZFC state at $T_{eff} \approx 9$ K.

Subsequent increase in the content of Ca^{2+} ions leads to further increase in the Curie temperature, up to 104 K for $Nd_{0.9}Ca_{0.1}MnO_3$ (Fig. 1d). The low-temperature magnetic phase transition takes place near the same temperature of 9 K but, in contrast to the samples with $x = 0.06$ and 0.08, the magnetization of the sample with $x = 0.1$ remains positive in the entire temperature range.

However, further increase in the content of calcium is not accompanied by a significant change in T_c. The temperature of the transition from para- to ferromagnetism for the samples with $x = 0.12$ and 0.15 is 107 K (Figs. 1e and 1f). In these samples, the anomalous behavior of magnetization in the region of 9 K is no longer observed. In contrast to the compositions with $0.06 \le x \le 0.08$, the FC samples with $x = 0.12$ and 0.15 exhibit an increase in magnetization in the low-temperature region.

Figure 2 presents the results of measurements of the field dependence of magnetization for a sample with the composition $Nd_{0.92}Ca_{0.08}MnO_{2.98}$, in which the anomalous low-temperature magnetization behavior was most pronounced. These measurements were performed on the ZFC sample in a temperature interval from 5 to 30 K, using an external field with a strength of up to 50 kOe. It should be noted that low-doped neodymium manganites are strongly anisotropic materials. According to the neutron diffraction data, the spontaneous magnetic moment of $NdMnO_{3+\lambda}$ has to be 2.6 μ_B per formula unit (pfu), while measurements in a field of 30 kOe gave only 2 μ_B [19]. Therefore, it is very difficult to correctly evaluate the spontaneous magnetic moment of such materials by measurements in the fields below 50 kOe.

At a temperature of 25 K, the sample was characterized by a coercive force of 4.6 kOe. This value indicates that $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ is a magnetically hard material. The sample magnetization exhibited no saturation in the fields of up to 50 kOe, which also confirms a high magnetic anisotropy (Fig. 2a). No significant changes in the character of the $M(H)$ curve was observed for analogous measurements performed above 25 K.

As the sample temperature is decreased below 25 K, the field dependence of magnetization acquires qualitatively new features. At $T = 20$ K, there appears a clearly pronounced hysteresis in the range of fields from 19 to 50 kOe (Fig. 2b). The presence of hysteresis is evidence of the first-order magnetic phase transition. Further decrease in the temperature leads to a shift of the hysteresis loop toward lower fields. At a temperature of 20 K and below, the $M(H)$ curves measured in the field increase mode exhibit jumps in magnetization: the lower the sample temperature, the more pronounced the jumps.

In contrast to the case of the field dependences measured at 20 K and above, a decrease in the applied field strength at $T = 17$ K leads to a significant drop in magnetization, this drop being much greater than the growth observed with increasing field. The residual magnetization is characterized by 0.05 μ_B (pfu) (Fig. 2c). An unusual phenomenon is observed in the temperature interval between 9 and 15 K: as the applied field decreases, the magnetization sharply drops and acquires negative values in the positive field (Figs. 2d–2f). This effect is most pronounced at $T = 12.5$ K, whereby the sample magnetization in a zero field is –0.22 μ_B (pfu) (Fig. 2e). With a subsequent decrease of the sample temperature, at $T = 8.5$, the magnetization at $H = 0$ is zero (Fig. 2g).

In the field dependences measured below 8.5 K, the sample magnetization remains positive when the magnetic field strength is decreased to zero. At $T = 5$ K, the sample exhibits a residual magnetization of 1 μ_B (pfu) and a coercive field of 4 kOe (Fig. 2h). However, an increase in the level of magnetization observed in the fields above 13 kOe is evidence of the phase transition. It should also be noted that the level of magnetization in a field of 50 kOe decreases with decreasing temperature.

4. DISCUSSION

Rare-earth magnets are classical objects for the investigation of phase transitions. In manganites, the main attention was devoted to the study of magnetic phase transitions of the order–disorder type (characterized by the Curie and Néel temperatures), whereas transitions of the order–order type involving a change in the magnetic structure have been studied to a lower extent. The magnetic phase transitions involving spin reorientation (orientational transitions) may take place in the course of variation of either the sample temperature (spontaneous transitions) or the external magnetic field (field-induced transitions). Such transitions are most clearly manifested in rare-earth orthoferrites and garnet ferrites [22–24]. To our knowledge, no systematic investigations of the orientational magnetic phase transitions in manganites have been reported so far.

Our investigation of the temperature dependence of magnetization in $Nd_{1-x}Ca_xMnO_3$ with $0.06 \le x \le 0.15$ revealed anomalous shapes of magnetization curves for both ZFC and FC samples in the region of $T = 9$ K (Figs. 1a–1d). Such a behavior of the spontaneous magnetization cannot be related to a scenario involving sharp ordering of the neodymium sublattice. Indeed, according to the results of measurements of the temperature dependence of magnetization for $NdMnO_{3.04}$ in various regimes [10], the neodymium sublattice becomes ordered in the vicinity of the Néel temperature.

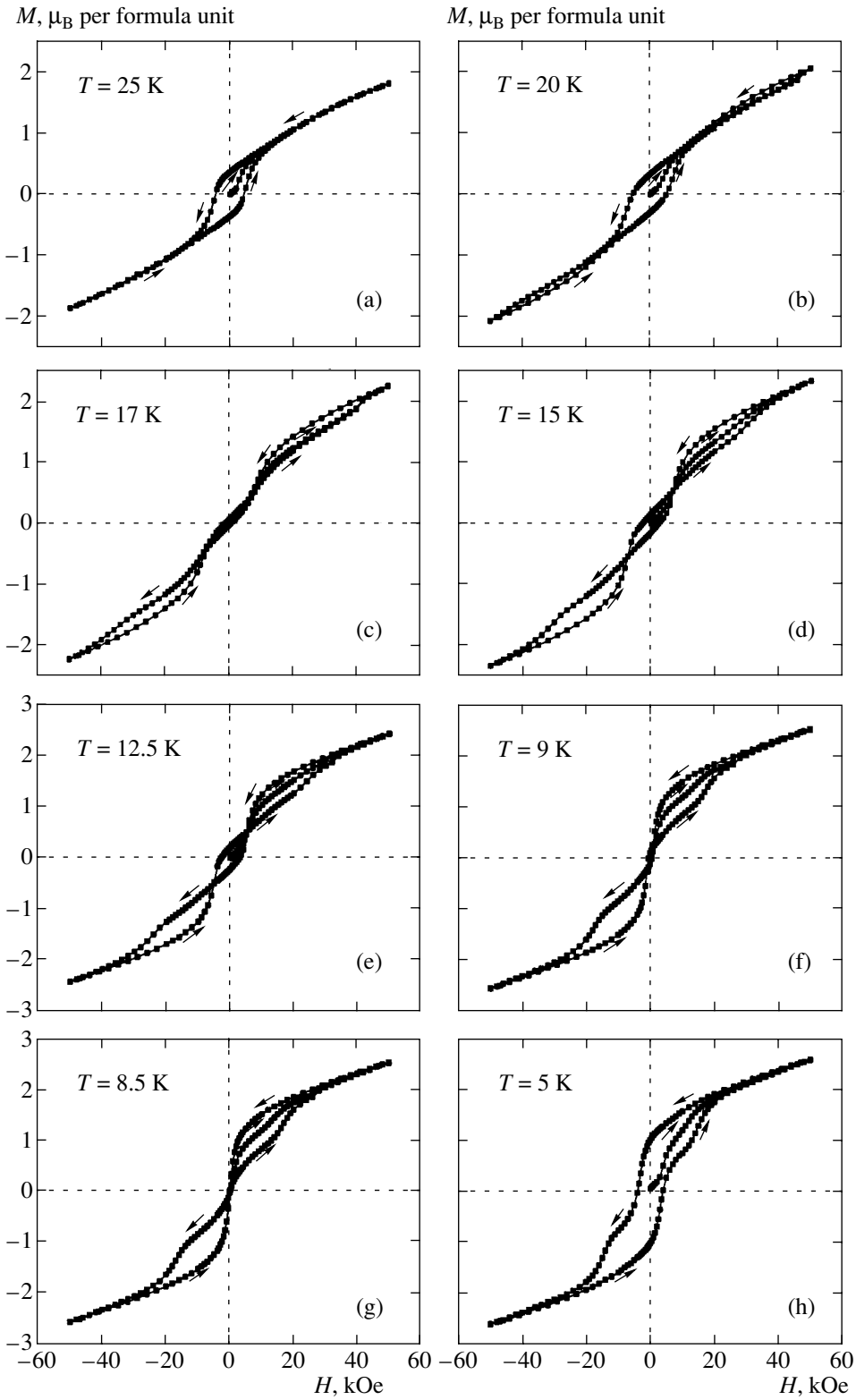

Fig. 2. The field dependences of magnetization in a $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ manganite sample at various temperatures.

Solid solutions of the $Nd_{1-x}Ca_xMnO_3$ system contain magnetic sublattices of neodymium and manganese, so that the magnetic state of our manganite samples is determined by the *f–f*, *f–d*, and *d–d* exchange interactions involving the magnetically active Nd^{3+} and Mn^{3+} ions. Magnetic interactions in the rare-earth sublattice are much lower than the *d–d* exchange between manganese ions. In addition, the replacement of neody-

mium ions by nonmagnetic calcium ions must additionally decrease the *f–f* exchange. In connection with this, we may assume that ordering of the neodymium sublattice is related to the *f–d* exchange between neodymium and manganese sublattices. In all probability, variation of the content of Mn^{4+} ions may lead to a change in sign of the *f–d* exchange. This is evidenced by the following experimental facts.

(i) In an almost ferromagnetic sample of $Nd_{0.88}Ca_{0.12}MnO_3$, the magnetic moments of neodymium ions are aligned parallel to those of the manganese sublattice (the *f–d* exchange is positive) [25].

(ii) The temperature dependence of magnetization in the vicinity of the Néel temperature in $NdMnO_{3.04}$ shows that the *f–d* exchange is negative [10].

(iii) The neutron diffraction data for $NdMnO_{3+\lambda}$ can be interpreted based on a model according to which the *f–d* exchange below 13 K is positive, while above 13 K the neodymium sublattice is disordered [19].

Note that the latter two statements are mutually exclusive. The discrepancy can be eliminated if we adopt a two-phase model of slightly doped manganites. In this model, the *f–d* exchange in the FM phase must be positive and that in the weak ferromagnetic (WFM) phase above 13 K, negative. Since the fraction of the FM phase in the samples studied in [19] was not less than 40%, the contributions from the two phases to the magnetization of the neodymium sublattice above 13 K must almost compensate each other. Below 13 K, the magnetic moments of neodymium ions in the WFM phase exhibit reorientation and become aligned parallel to the FM component, which leads to the observation of the magnetic moment of neodymium (1.2 μ_B) parallel to that of the FM component (the magnetic moment of manganese is equal to 1.4 μ_B) [19].

The question naturally arises as to why do the magnetic moments of neodymium ions exhibit this reorientation. In order to answer this question, we have to take into account that the two phases occur in an exchange-coupled state similar to that of thin magnetic layers in multilayer film structures. Apparently, the molecular field of the FM phase is, in a certain sense, analogous to an external field oriented opposite to the internal exchange field of the WFM phase. At a certain temperature, these fields may compensate each other. We suggest that the ground state of Nd^{3+} ions in the vicinity of this temperature exhibits degeneracy (crossover). Theoretically, this degenerate state cannot be stable and the magnetic structure is subject to transformation [24]. The magnetic phase transition removes the degeneracy. The crossover accounts for the spin reorientation manifested as the first-order phase transition. The presence of closely spaced energy levels and electron transitions between these levels in Nd^{3+} ions are confirmed by spectroscopic data [26].

Note the following peculiarity of the magnetic phase transition in the vicinity of 9 K. The magnetic moment of the neodymium sublattice in the FM phase of manganese must be about 0.6 μ_B (pfu). This estimate takes into account that the content of the FM phase is about 50% (we assuming that the samples with $x \geq 0.06$ are doped with Mn^{4+} stronger than $NdMnO_{3+\lambda}$ studied in [19]) and that the magnetic moment of neodymium is 1.2 μ_B in both the FM and WFM phases [19, 25]. However, a change in the magnetic moment related to the phase transition is significantly greater than 1 μ_B even for the FC sample in an applied field of 100 Oe (Fig. 1c). Taking into account that the sample represents a strongly anisotropic polycrystalline material, the change in the magnetization must be even more pronounced.

In connection with this, we may assume that the magnetic moments of manganese ions must be also involved into the process of spin reorientation. We suggest that the phase transition in a WFM phase leads to reorientation of the magnetic moment of a less anisotropic FM phase in the direction opposite to that of the applied field (100 Oe). This is related to a strong exchange coupling between the two phases. The magnetic moments of neodymium in the WFM phase do not change their orientation, while the magnetic moments of manganese ions in the WFM phase change their direction. Thus, the sample below 9 K occurs in a state with the magnetic moments of all four sublattices of the two phases oriented opposite to an external field of 100 Oe. The proposed mechanism is confirmed by the results of measurements in strong magnetic fields.

We believe that the jumps in the field dependence of magnetization observed for a sample of $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ in an external magnetic field (Fig. 2) are analogous to the behavior observed on cooling the sample in a field of 100 Oe. As is known, the splitting of levels for a rare-earth ion in a magnetically ordered crystal is determined by the combined action of the crystal field, exchange coupling, and external field [23]. If the *f–d* exchange coupling leads to an increase in the separation of sublevels, the external field must bring these levels closer to each other because the two fields have opposite directions (while the *f–d* exchange field renders the magnetic moments of the neodymium and manganese sublattices in the WFM phase antiparallel, the external field tends to orient these moments in the same direction). A sufficiently strong magnetic field may lead to crossover and degeneracy of the sublevels.

At $T = 20$ K, the jump in magnetization related to reorientation of the neodymium sublattice is rather small. This is probably due to the fact that the external magnetic field is insufficient for completing the transition and also due to the strong temperature dependence of the magnetic moment of neodymium. Beginning with 17 K, the jump observed with decreasing field significantly exceeds that in the field increase mode; in the temperature interval from 9 to 15 K, magnetization in the positive field becomes negative (Fig. 2e). This behavior indicates that the field switching on and off is

accompanied by different processes. We believe that increasing field leads to reorientation of the magnetic moments of neodymium in the WFM phase along the field. A decrease in the field strength to zero is accompanied by reorientation of the magnetic moments of manganese ions in the WFM and FM phases, while the magnetic moments of neodymium in the WFM phase remain oriented along the field. For this reason, the negative magnetic moment in the positive field is smaller than that observed in the case of cooling in a field of 100 Oe (Fig. 1c).

The pattern somewhat changes when the field dependences are measured at temperatures below T_{eff} (Fig. 2h). A negative residual magnetization no longer takes place, but reorientation of the magnetic moments of neodymium can be observed in a field applied in the direction opposite to that of the magnetic moments. Thus, a state with the antiparallel orientation of the magnetic moments of neodymium and manganese ions in the WFM phase can exist in a broad range of field strengths.

Based on the above data, we have constructed a magnetic phase diagram of $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ in the *H–T* coordinates (Fig. 3). This phase diagram shows that, as the temperature increases, the range of field strengths in which the components with parallel or antiparallel orientations of magnetic moments in the neodymium and manganese sublattices occur (depending ion the prehistory) in the WFM phase shifts toward higher fields. This type of magnetic phase diagram corresponds to interpretation of the phase transitions in terms of crossover. It should be noted that, since the measurements were performed on a polycrystalline sample, the range of fields featuring metamagnetic phase transitions is rather wide. In single crystals, these transitions will proceed in a jumplike manner in a narrow range of field strengths.

Using data on the temperature dependence of magnetization for manganites of the $Nd_{1-x}Ca_xMnO_3$ system with $x \le 0.15$, we have also constructed a hypothetical magnetic phase diagram in the *T–x* coordinates (Fig. 4). Data for $NdMnO_3$ were taken from our previous study [10].

Let us consider the possible mechanisms of concentrational magnetic phase transition between AFM and FM states in the $Nd_{1-x}Ca_xMnO_3$ system. As was mentioned above, there are several models describing the magnetic state of manganites in the range of intermediate compositions. A double exchange model developed by Zener [27] for explaining the ferromagnetism of compounds with high electric conductivity is based on the concept of real transitions of *d* electrons between Mn^{4+} and Mn^{3+} ions. Within the framework of this model, it is assumed that the AFM–FM transition proceeds via the formation of a noncollinear magnetic structure [11].

An alternative mechanism of the formation of an FM state in manganites was proposed by Wollan and Koehler [6]. According to this model, an inhomogeneous (two-phase) state with collinear magnetic moments in these phases is energetically more favorable than a homogeneous (over the whole sample) mag-

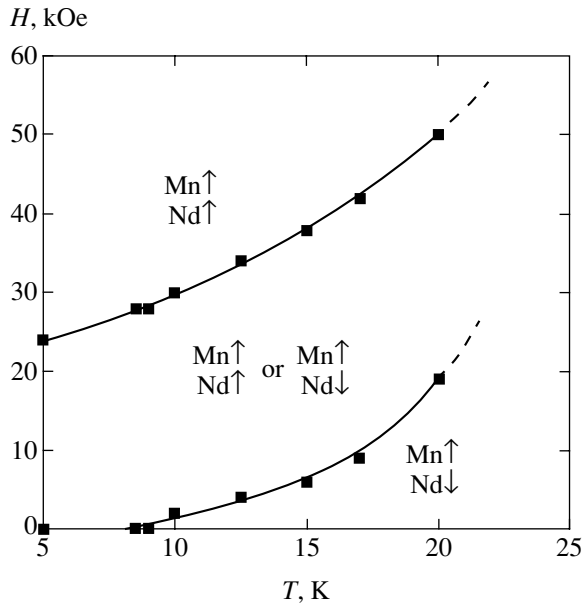

Fig. 3. A magnetic phase diagram of $Nd_{0.92}Ca_{0.08}MnO_{2.98}$ manganite in the *H–T* coordinates. Arrows indicate the magnetic moment orientations in the neodymium and manganese sublattices of the WFM phase (in the FM phase, the magnetic moments of both sublattices are always parallel).

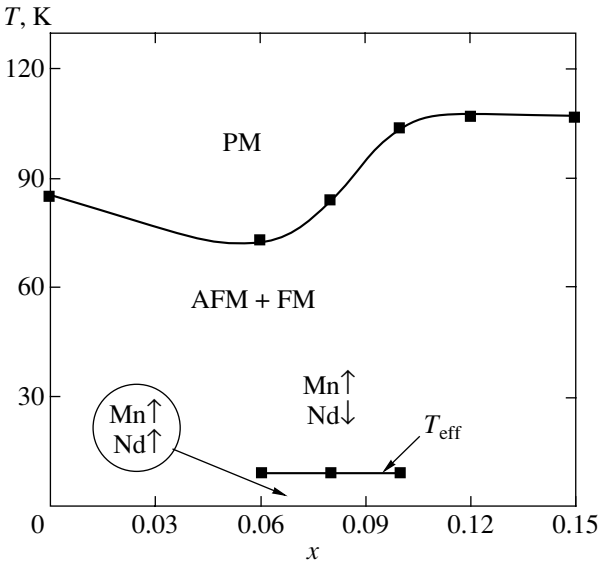

Fig. 4. A magnetic phase diagram of the $Nd_{1-x}Ca_xMnO_3$ system ($x \le 0.15$): (AFM) antiferromagnetic phase; (FM) ferromagnetic phase; (PM) paramagnetic phase; (T_{eff}) the effective temperature at which the anomalous jump of magnetization of the FC samples, related to the spin-reorientation phase transition, is most clearly pronounced. Arrows indicate the magnetic moment orientation in the neodymium and manganese sublattices of the WFM phase.

netic structure with noncollinear ordering of the magnetic moments. The separation of magnetic phases can be related to the fact that the FM order of magnetic moments favors the motion of charge carriers. In this case, the minimum energy is achieved at the expense of charge carrier concentration in certain parts of the crystal. As a result, the crystal separates into highly conducting FM regions and dielectric AFM regions. As the charge carrier density increases, the FM phase volume grows accordingly. Beginning with a certain carrier density corresponding to a percolation threshold, the FM droplets come into contact with each other. This corresponds to the dielectric–metal transition, whereby the crystal passes to a ferromagnetic state [12].

Difficulties in judging between the above models are related to the fact that, in experiment, a two-phase magnetic state can be manifested in the same manner as a noncollinear magnetic structure. For example, the results of neutron diffraction measurements can be interpreted both within the noncollinear ordering model [28] and assuming a two-phase state representing a mixture of FM and AFM regions [6]. Both models explain the giant magnetoresistance and the FM ordering in conducting materials, but neither of the two can explain the FM behavior of nonmetallic manganites with a sufficiently high dopant content.

Taking into account the effect of orbital ordering, Goodenough *et al.* [29, 30] suggested that the FM of manganites is related, besides a double exchange, to a special character of the superexchange interactions in Mn^{3+}–O–Mn^{3+} and Mn^{3+}–O–Mn^{4+} systems containing Jahn–Teller ions. In this approach, the sign of the 180° Mn^{3+}–O–Mn^{3+} superexchange is determined by the orbital state of Mn^{3+} ions. Removal of the static Jahn–Teller distortions gives rise to an isotropic FM interaction resulting from a relation between the electron configuration and atomic oscillations. According to this model, a two-phase state is related to separation of a crystal into regions featuring different orbital dynamics.

Since the FM state in solid solutions of the $Nd_{1-x}Ca_xMnO_3$ system is not highly conducting [10], advantages of the description of concentrational transitions from AFM to FM state in this system within the framework of the double exchange model or the scenario of electron-induced phase separation are not clearly manifested. More likely, the $Nd_{1-x}Ca_xMnO_3$ manganite system features a mixed magnetic state related to the mechanism of the orbital-induced phase separation. According to the approach of Goodenough *et al.* [29, 30], the orbital-ordered phase is antiferromagnetic, while the orbital-disordered phase is ferromagnetic. Both phases possess close chemical compositions and equal carrier densities, while differing in the local crystal structure distortions and the orbital dynamics.

5. CONCLUSIONS

The results of our investigation of the magnetic properties of manganites of the $Nd_{1-x}Ca_xMnO_3$ system lead to the following conclusions.

(i) The behavior of magnetization in weakly substituted manganites of the $Nd_{1-x}Ca_xMnO_3$ system can be explained using a model according to which the samples consist of exchange-coupled FM and WFM phases. In the FM phase, the magnetic moments of neodymium ions are oriented parallel to the moments of manganese ions. In the WFM phase at $T > T_{eff}$, the magnetic moments of neodymium ions are antiparallel with the vector of weak ferromagnetism.

(ii) The low-temperature phase transition at $T_{eff} \approx 9$ K in $Nd_{1-x}Ca_xMnO_3$ with $0.06 \le x \le 0.10$ is related to reorientation of the vector of weak ferromagnetism along the magnetic moments of neodymium ions in the WFM phase. This is accompanied by reorientation of the magnetic moment of the exchange-coupled FM phase.

(iii) The jumps in magnetization observed in increasing external magnetic field are related to reorientation of the magnetic moments of neodymium ions in the WFM phase along the vector of weak ferromagnetism and the magnetic moment of the FM phase. As the applied field strength decreases, the sample features reorientation of the vector of weak ferromagnetism and the magnetic moments of the FM phase, while the magnetic moments of neodymium ions in the WFM phase remain oriented along the field.

(iv) The spin-reorientation phase transitions are explained within the framework of the magnetic analog of the Jahn–Teller theorem.

ACKNOWLEDGMENTS

This study was supported by the Foundation for Basic Research of the Republic of Belarus, project no. F02R-122.

REFERENCES

1. S. Jin, T. H. Tiefel, M. McCormack, *et al.*, Science **264**, 413 (1994).
2. S. Jin, M. McCormack, T. H. Tiefel, and R. Ramesh, J. Appl. Phys. **76**, 6929 (1994).
3. R. von Helmolt, J. Wecker, K. Samwer, *et al.*, J. Appl. Phys. **76**, 6925 (1994).
4. K. Chabara, T. Ohno, M. Kasai, and Y. Kozono, Appl. Phys. Lett. **63**, 1990 (1993).
5. C. N. R. Rao, A. K. Cheetham, and R. Mahesh, Chem. Mater. **8**, 2421 (1996).
6. E. O. Wollan and W. C. Koehler, Phys. Rev. **100**, 545 (1955).
7. C. Ritter, M. R. Ibarra, J. M. De Teresa, *et al.*, Phys. Rev. B **56**, 8902 (1997).
8. N. V. Kasper and I. O. Troyanchuk, J. Phys. Chem. Solids **57**, 1601 (1996).

9. J. Rodriguez Carvajal, M. Hennion, F. Moussa, *et al.*, Phys. Rev. B **57**, R3189 (1998).
10. I. O. Troyanchuk, D. A. Efimov, N. V. Samsonenko, *et al.*, J. Phys.: Condens. Matter **10**, 7957 (1998).
11. P. G. De Gennes, Phys. Rev. **118**, 141 (1960).
12. E. L. Nagaev, Phys. Rep. **346**, 387 (2001).
13. E. Dagotto, T. Hotta, and A. Moreo, Phys. Rep. **344**, 1 (2001).
14. G. Allodi, R. De Renzi, G. Guidi, *et al.*, Phys. Rev. B **56**, 6036 (1997).
15. J. M. De Teresa, M. R. Ibarra, J. Garcia, *et al.*, Phys. Rev. Lett. **76**, 3392 (1996).
16. A. Sundaresan, A. Maignan, and B. Raveau, Phys. Rev. B **55**, 5596 (1997).
17. S. Quezel-Ambrunas, Bull. Soc. Mineral. Crystallogr. **91**, 339 (1968).
18. I. E. Dzyaloshinski, J. Solid State Chem. **4**, 241 (1958).
19. A. Munoz, J. A. Alonso, M. J. Martinez-Lope, *et al.*, J. Phys.: Condens. Matter **12**, 1361 (2000).
20. I. O. Troyanchuk, J. Magn. Magn. Mater. **231**, 53 (2001).
21. R. D. Shannon, Acta Crystallogr. A **32**, 751 (1976).
22. K. P. Belov, A. K. Zvezdin, A. M. Kadomtseva, and R. Z. Levitin, *Reorientational Transitions in Rare-Earth Magnets* (Nauka, Moscow, 1979).
23. K. P. Belov, *Rare-Earth Magnets and Their Application* (Nauka, Moscow, 1980).
24. A. K. Zvezdin, V. M. Matveev, A. A. Mukhin, and A. I. Popov, *Rare-Earth Ions in Magnetic-Ordered Crystals* (Nauka, Moscow, 1985).
25. H. Gamari-Seale, H. Szymczak, I. O. Troyanchuk, and A. Hoser, Physica B (Amsterdam) **276**–**278**, 668 (2000).
26. A. A. Mukhin, V. Yu. Ivanov, V. D. Travkin, and A. M. Balbashov, J. Magn. Magn. Mater. **226**–**230**, 1139 (2001).
27. C. Zener, Phys. Rev. **82**, 403 (1951).
28. Z. Jirak, S. Vratislav, and J. Zajicek, Phys. Status Solidi A **52**, K39 (1979).
29. J. B. Goodenough, A. Wold, R. J. Arnott, and N. Menyuk, Phys. Rev. **124**, 373 (1961).
30. J.-S. Zhou, H. Q. Yin, and J. B. Goodenough, Phys. Rev. B **63**, 184423 (2001).

Translated by P. Pozdeev

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1240–1245.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1391–1397.

SOLIDS
Electronic Properties

Statistical Analysis of Low-Voltage Electron Emission from Nanocarbon Cathodes

Al. A. Zakhidov, A. N. Obraztsov*, A. P. Volkov, and D. A. Lyashenko

Moscow State University, Vorob'evy gory, Moscow, 119992 Russia
e-mail: obraz@acryst.phys.msu.ru

Received April 21, 2003

Abstract—The current–voltage (*I–V*) characteristics of low-voltage electron emission from nanocarbon (nC) film cathodes consisting of carbon nanotubes and/or nanosized graphite crystallites is analyzed. It is shown that an adequate qualitative description of the *I–V* characteristics can be obtained within the classical Fowler–Nordheim (FN) theory with regard to the normal statistical distribution of the parameters of emission sites situated on the cathode surface. However, the application of this classical theory to obtain quantitative estimates leads to a considerable discrepancy between the results obtained and experimental data. A quantitative agreement between experimental data and theoretical results can be achieved under the assumption that the effective areas of emission sites increase at the expense of the lateral surfaces of nC structures. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Field emission attracts considerable interest from the viewpoint of abstract science, because this phenomenon is based on quantum-mechanical effects occurring on the surface and interfaces of solids, as well as from the viewpoint of applied investigations involving electron beams in various devices. In the latter case, of special importance is the voltage used to generate field-emission electrons and is determined by the electric field required to induce electron tunneling through the potential barrier on the cathode surface [1, 2]. In many recent publications, it has been shown that various carbon nanotubes can emit electrons under anomalously low (compared with conventional metal field-emission cathodes) voltages (see, for example, the review [3]). Similar properties are also exhibited by other nanosized carbon structures; this fact suggests that there exists a general mechanism of low-voltage field emission attributed to nanometric sizes of emitters [4, 5]. Within the classical Fowler–Nordheim (FN) theory, such a low-voltage emission may be associated with two factors: a decrease in the work function of electrons in the cathode material and the enhancement of the electric field due to the geometric shape of the emitter, which, for example, may have the form of a thin tip or edge [1, 2]. However, for nanosized emitters, the quantitative characteristics and the physical sense of both the work function and the geometric shape of the equipotential surface, which is responsible for the electric field strength, may be different from those of macroscopic cathodes. This fact requires that one should carry out an additional analysis on the applicability of the classical FN theory to nanoemitters. Moreover, since the absolute values of currents emitted by separate nanosized emitters are relatively small, low-voltage emission is experimentally observed, as a rule, on cathodes containing a large number of emission sites whose parameters are characterized by a certain statistical distribution over a certain range of values. In the present paper, the analysis of these questions is based on the investigation of the current-voltage (*I–V*) characteristics of nanocarbon (nC) film field-emission cathodes and takes into account the statistical distribution of the parameters of emission sites.

2. FEATURES OF EXPERIMENT

Nanocarbon film field-emission cathodes for our experiment have been obtained by chemical vapor deposition (CVD) in hydrogen–methane gas mixture activated by a dc discharge by the method described in our earlier papers (see, for example, [4, 6]). As a substrate, we used standard polished silicon wafers. The surface morphology of the samples of field-emission cathodes is determined by carbon nanotubes and platelike graphite crystallites that constitute the film; these nanotubes and crystallites are predominantly oriented along the normal to the substrate [4]. For lengths ranging from one to several micrometers, the diameter of the nanotubes and the thickness of the crystallites range from 10 to 20 nm. Figure 1 illustrates the typical surface morphology of an nC cathode obtained by a scanning electron microscope (SEM). In this image, several carbon nanotubes that can be distinguished for a given magnification are indicated by arrows. The Raman spectroscopy and the X-ray photoelectron spectroscopy displayed a high degree of crystallographic ordering in nC film materials and the absence of any substantial amount of noncarbon impurities, as was pointed out earlier in [4].

1063-7761/03/9706-1240$24.00

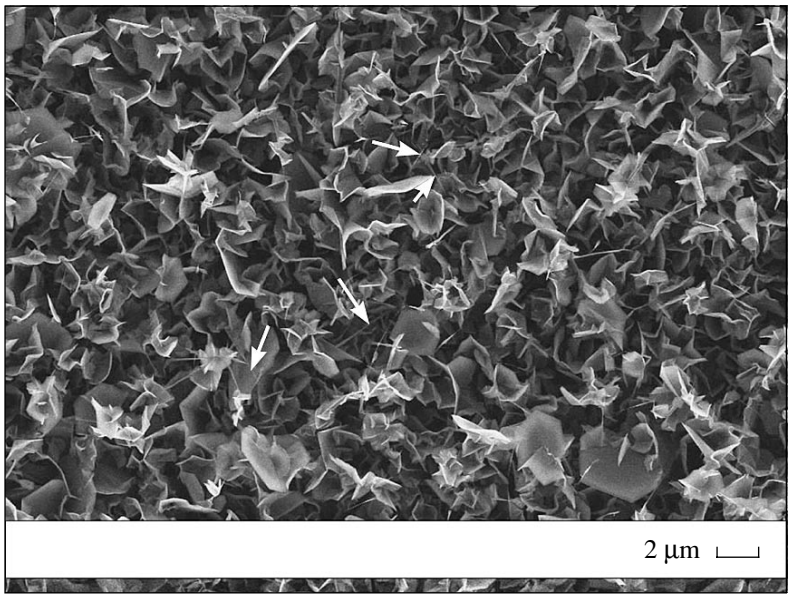

Fig. 1. Typical SEM image of the surface morphology of an nC film. The arrows indicate separate nanotubes.

The *I–V* characteristics of film nC cathodes were measured in the configuration of a vacuum diode with flat electrodes. The anode was a glass plate coated with a transparent conducting layer of indium and tin oxides (ITO). The conducting ITO film was coated by a layer of phosphor that emitted light under the action of electrons emitted from the cathode. In contrast to configurations with spherical or tip-shaped anodes [3, 7], this configuration allows one to determine easily and adequately the macroscopic field and the density of emission sites. The image obtained on the anode plate corresponds to the distribution of emission sites on the cathode, and the intensity of the macroscopic electric field on the cathode is given by the simple relation

$$F = \frac{V}{d},$$

where V is the voltage between cathode and anode and d is the anode–cathode spacing. In this case, the density of the field-emission current is given by the ratio of the total current I of the cathode to its area S:

$$J = \frac{I}{S}.$$

As the threshold value of the electric field strength F, we took a field value corresponding to a current density of $J = 10^{-9}$ A/cm^2. For typical nC cathodes, this threshold field was lower than 1.5 V/μm. The *I–V* characteristics were measured in a vacuum of at least 10^{-6} Torr at room temperature.

3. RESULTS AND DISCUSSION

Typical images of luminescence of phosphor on the anode plate under applied voltages of 300, 400, and 1000 V and an anode–cathode spacing of 200 μm are presented in Fig. 2. These images show that emission sites are uniformly distributed over the cathode surface. The number of emission sites increases as the applied voltage increases. Under a voltage of higher than 800 V, individual emission sites cannot be resolved because the phosphor grain size is finite and amounts to about 5 μm; the density of emission sites under this voltage is estimated to be 10^6 cm^{-2} in order of magnitude.

Figure 3 presents a typical experimental diagram of the current density J of a vacuum-tube diode with an nC field-emission cathode as a function of the intensity of the macroscopic electric field F (dots). The *I–V* curve presented in this figure in FN coordinates has two essentially different regions. Under relatively high electric fields, the dependence is linear, similar to that predicted by the classical FN theory. However, in the region of weak fields (low-voltage emission), which can be called a region of switching on of emission sites, the curve is essentially nonlinear. It should be noted that the *I–V* characteristics of the nC field-emission cathodes considered here are independent of temperature [8], as can occur in the case of emission from semiconductor materials. This fact suggests that the deviation of the *I–V* characteristic in the low-voltage region from the classical linear characteristic is associated with a gradual increase in the number of emission sites as voltage increases. Since the material of the nC film is sufficiently homogeneous and there are hardly any significant variations in the work functions of different emission sites, the variation in the number of these sites can only be attributed to the difference in the geometric characteristics of these sites.

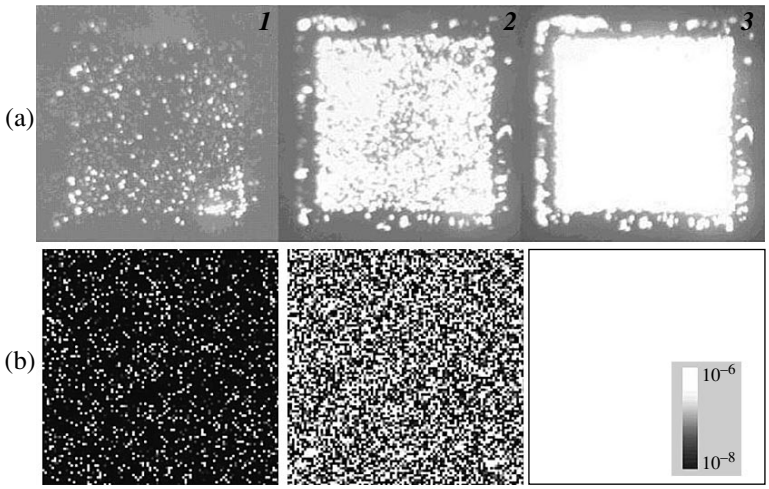

Fig. 2. (a) The image of a luminescent screen (anode) under excitation by electrons emitted from a 10 × 10-mm nC film cathode under an applied voltage of (*1*) 300, (*2*) 400, and (*3*) 1000 V for an anode–cathode spacing of 200 μm. The dots along the cathode perimeter are due to the emission from the sample boundaries. Under maximum voltage, the images of separate emission sites are not distinguished because of the finite grain size of the phosphor. (b) The results of numerical simulation of the distribution of field-emission current density from the cathode surface with regard to the statistical distribution of geometric characteristics of individual emission sites for the same values of the electric-field strength. The scale of currents is shown in the rightmost image, where separate emission sites are indistinguishable for a given magnification factor because of the high density of these sites.

According to the classical FN theory [1, 2], the density of the field-emission current can be expressed as

$$J = \frac{A}{\varphi t(y)} E^2 \exp\left[-B\frac{\varphi^{3/2}}{E}\Theta(y)\right], \quad (1)$$

where E is a local electric field near the emitting surface, φ is the work function of the emitter material, $\Theta(y)$ is the tabulated Nordheim function,

$$A = \frac{e^3}{8\pi h}, \quad B = \frac{8\pi\sqrt{2m}}{3he}, \quad y = \frac{e\sqrt{eF}}{\varphi},$$

and

$$t(y) = \Theta(y) - \frac{2y}{3}\frac{d\Theta(y)}{dy}$$

(e and m are the charge and mass of electron, and h is the Planck constant). Taking into account that, for $E < 10^4$ V/μm, the Nordheim function can be expressed as

$$\Theta(y) = 0.95 - 1.03y^2$$

to within one percent, the expression for the current density is rewritten as

$$J = \frac{A}{\varphi}E^2 \exp(1.03Be^3\varphi^{-1/2}) \times \exp(-0.95B\varphi^{3/2}E^{-1}). \quad (2)$$

Using the dimensions φ [eV], E [V/cm], and J [A/cm^2] for the coefficients in formula (2), we obtain

$$A = 1.5414 \times 10^{-6}\ \text{A eV V}^{-2},$$

$$B = 6.8309 \times 10^7\ \text{eV}^{-3/2}\ \text{V cm}^{-1},$$

$$1.03Be^3 = 10.1\ \text{eV}^{1/2}.$$

The local electric-field strength E, in contrast to its macroscopic value F, depends not only on the applied voltage and the anode–cathode spacing, but also on the

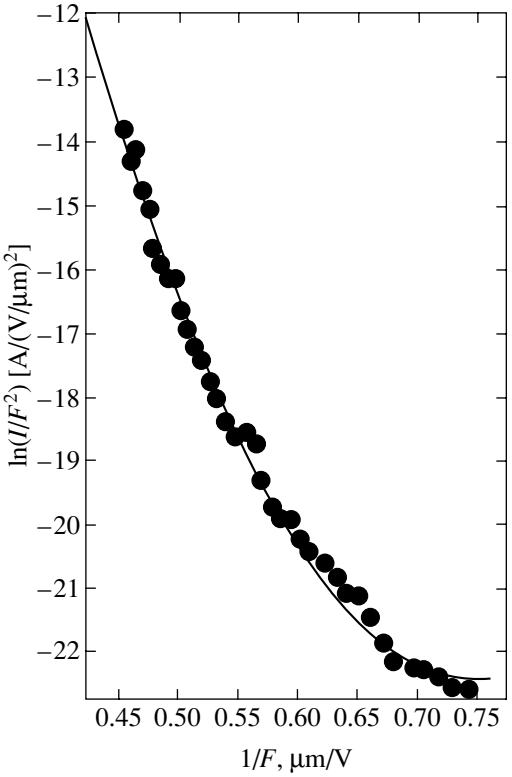

Fig. 3. Typical *I–V* curve of an nC cathode in the FN coordinates. The dots represent experimental data. The solid curve represents the theoretical *I–V* curve calculated by formula (3) with $\sigma = 0.1r_0$.

morphology of the cathode surface. In the most general form, this relation can be expressed as

$$E = \beta F = \frac{\beta V}{d},$$

where β is the local field enhancement factor, which is determined by the geometric characteristics of emitters. In the simplest case of an isolated emitter in the form of a cylindrical tip of length L with a hemispherical apex of radius r, the field enhancement factor can be expressed as

$$\beta = \frac{L}{r}$$

to a good approximation (see [1, 2]). Formula (2) shows that, when all emission sites are identical, the *I–V* diagram in the FN coordinates (i.e., in the coordinates where the ordinate is $\ln(I/F^2)$ and the abscissa is $1/F$), represents a straight line with a negative slope ratio.

As in [9], we assume that the statistics of the geometric parameters of individual emission sites follow the normal distribution. For simplicity, we assume that all emission sites have the same length L but differ in the curvature radii r of their apexes. Then,

$$n(r) = \frac{N}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(r-r_0)^2}{2\sigma^2}\right],$$

where r_0 is the mathematical expectation and σ is the rms deviation of the radii of emission sites. The total number of emission sites is

$$N = \int_{r_{\min}}^{r_{\max}} n(r)dr.$$

It is obvious from physical considerations that $r_{\min} > 0$, $r_{\max} \ll L$, and $\sigma \ll r_0$. The last condition is due to the requirement that the integral should be much less than N outside admissible boundaries. Taking into account that the area of each emission site is $s = 2\pi r^2$, we can represent the total current from the whole surface of the cathode as

$$I = \int JdS = \int_{r_{\min}}^{r^{\max}} J(r)2\pi r^2 n(r)dr.$$

Substituting the expression for J from (2) into the latter formula, we finally obtain

$$I = CNF^2\exp\left[-\frac{Dr}{F} + \frac{\sigma^2}{2}\left(\frac{D}{F}\right)^2\right], \qquad (3)$$

where

$$C = \frac{2A\pi L^2}{\varphi}\exp\left(\frac{10.1}{\sqrt{\varphi}}\right), \quad D = 0.95B\frac{\varphi^{3/2}}{L}.$$

In contrast to the classical FN formula (1), expression (3) provides a satisfactory approximation to the experimental *I–V* curves, including the low-voltage region (see Fig. 3). The application of the least squares method allows one to obtain a value of $\sigma \approx 0.1r_0$ for the rms deviation of the radii of emission sites. Taking into account that $r_0 \approx 5$–10 nm according to the electron-microscopy data, we have $\sigma \approx 5$–10 Å; i.e., the difference in the sizes of nC emitters is no greater than several atomic layers with a graphite-like crystal structure.

The nonlinear behavior of the *I–V* curve in the FN coordinates is determined by the second term in square brackets in (3), which may be comparable to the first term for small values of the field F. The range of currents and fields corresponding to the nonlinear region of the *I–V* curve is sufficiently small precisely due to the smallness of σ. This nonlinear region corresponds to the gradual switching on of an increasing number of emission sites with smaller enhancement factors as the applied voltage increases. Under sufficiently strong fields, the quadratic term in (3) becomes considerably smaller than the linear term. This result corresponds to the situation when the majority of emission sites have already been switched on and make the dominant contribution to the current, while the role of the remaining sites with minimal geometric field enhancement factors is insignificant, in spite of the fact that the number of such sites may be sufficiently large.

The results of the computer simulation of the switching on of different emission sites are illustrated in Fig. 2b. In this simulation, we assumed that all the emission sites represent thin cylindrical tips with hemispherical apexes that have different field enhancement factors. These tips were randomly distributed over a given surface, and the statistical distribution of β was specified with parameters similar to those given above. Taking into account that the experimental data were obtained for phosphors with essentially nonlinear dependence of the luminance on the current density, we can assume that the agreement between simulated and experimental images in Fig. 2 is quite satisfactory.

The comparison of the experimental *I–V* curve with that given by formula (3) leads to the following relation between the work function and the field enhancement factor, which appear in (3) as parameters:

$$\frac{\varphi^{3/2}}{\beta} = 7.4\times10^{-3}\ \text{eV}^{3/2}.$$

This relation is represented graphically in Fig. 4. It is obvious that the leftmost and rightmost parts of this graph represent physically meaningless values. Indeed, for $\varphi < 1$ eV, one should expect intense field emission

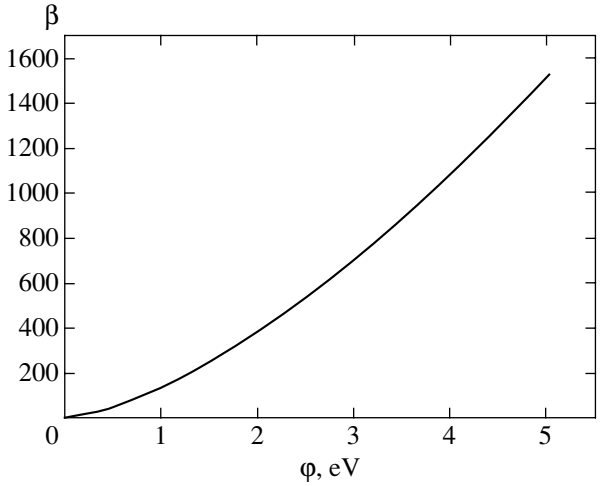

Fig. 4. Empirical relation between the work function φ and the field amplification coefficient β, $\varphi^{3/2}/\beta = 7.4 \times 10^{-3}$ eV$^{3/2}$, obtained by approximating the experimental *I–V* characteristic by the theoretical *I–V* characteristic calculated by formula (3).

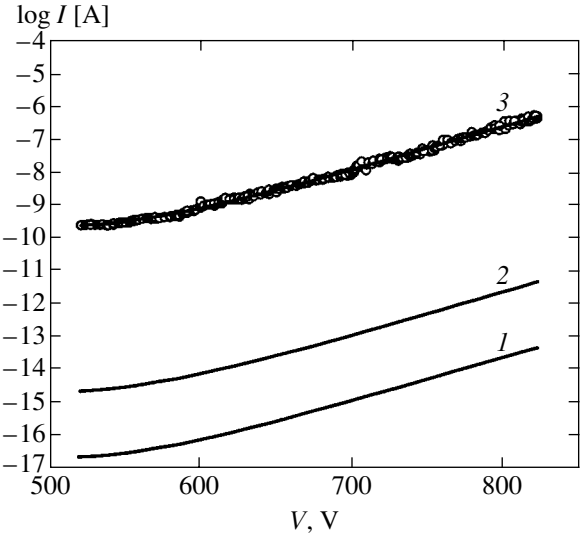

Fig. 5. *I–V* curve of an nC cathode in semilogarithmic coordinates. The dots represent experimental data. The solid curves represent the theoretical *I–V* curve calculated by formula (3) for various densities *N* of emission sites; (*1*) $N = 10^6$, (*2*) 10^8, and (*3*) 10^{13} cm^{-2}.

even at room temperature, whereas the value of $\beta > 1000$, corresponding to $\varphi = 4$–5 eV, does not agree with the electron-microscopy data (see [3, 4, 6–8] and Fig. 1). According to the data obtained, values of *L* ranging from 1 to 5 μm and r_0 ranging from 5 to 10 nm are characteristic of nC emitters. As a rule, the largest nC structures have the largest characteristic sizes of their apexes. Thus, according to our estimates, the geometric factor β ranges from 50 to 500 [4, 8].

Formula (3) contains the total number *N* of emission sites as a parameter. Figure 5 presents several curves obtained by substituting the values of $N = 10^6$, 10^8, and 10^{13} cm^{-2} into formula (3) for a sample size of 1 cm^2. This figure shows that the best approximation is obtained for $N = 10^{13}$ cm^{-2}, which is substantially different from the estimate made earlier by examining images on the cathodoluminescent screen. The graphs in Fig. 5 show that such a high density of emission sites ($N = 10^{13}$ cm^{-2}) is attributed to the experimentally observed anomalously high density of the field-emission current. However, for such density of emission sites, the distance between them must be about 3 nm, which is less than the characteristic size of the sites. This contradiction can be resolved under the assumption that the emitting area *s* of a separate emission site is much greater than $2\pi r^2$. Such an assumption is possible when the electron emission occurs not only from the tip of an nC emitter (for example, a carbon nanotube or nanographite crystallite) but also from its lateral surface. For example, for a cylindrical nanoemitter, we then have

$$s = 2\pi r^2 + \pi r(kL),$$

where *k* ranges from 0 to 1 and indicates from what part of the lateral surface electrons are emitted.

The fact that the emission of electrons from the lateral surface of nanotubes or other graphite-like nanosized structures occurs at a considerably lower strength of electric field than that predicted by the FN theory can be attributed to lowering the potential barrier for electrons tunneling into vacuum. The possibility of lowering the barrier follows immediately from the formal quantum-mechanical analysis of electron tunneling through a thin layer of a dielectric (or a large-gap semiconductor) on the surface of a conductor (see, for example, [1, 5, 10]). The practical implementation of this mechanism in nC emitters is facilitated by two unique circumstances: the presence of atomically thin clusters with the properties of a large-gap diamond-like material on the surface of these emitters, and the presence of a sharp interface between the dielectric (diamond-like) and conducting (graphite-like) phases of a carbon nanomaterial [4, 8, 11, 12]. Due to the presence of these diamond-like clusters, which, according to this model, are emission sites, the surfaces of nC emitters are not equipotential. This fact was experimentally verified in *in situ* observations of the electric-field distribution near the surface of carbon nanotubes during electron emission (see, for example, [13]). In this case, the relative contribution of such clusters to the redistribution of the surface potential of nC emitters is appreciable only under a relatively low strength of the electric field, which corresponds to a low-voltage field emission. When the applied voltage increases, these *in situ* images of the field distribution become similar to those obtained in the classical case of a metal tip [14].

4. CONCLUSIONS

In this work, we have demonstrated that an adequate qualitative description of *I–V* curves of vacuum-tube diodes with nC field-emission cathodes is possible within the classical Fowler–Nordheim theory with regard to the statistical distribution of geometric parameters of individual emission sites. The quantitative discrepancy observed between the experimental data and the results of the classical theory are attributed to the specific features of nC emitters consisting of a conducting graphite-like material with diamond-like clusters on its surface. Such a heterogeneous structure of nC emitters leads to a nonuniform redistribution of the potential on the emitter surface, which manifests itself, for example, in the experimentally observed anomalously low threshold voltage of field emission.

ACKNOWLEDGMENTS

This work was supported in part by INTAS, grant no. 01-0254.

REFERENCES

1. L. N. Dobretsov and M. V. Gomoyunova, *Emission Electronics* (Nauka, Moscow, 1966; Israel Program for Sci. Transl., Jerusalem, 1971).
2. R. Gomer, *Field Emission and Field Ionization* (AIP, New York, 1993).
3. A. V. Eletskiĭ, Usp. Fiz. Nauk **172**, 401 (2002) [Phys. Usp. **45**, 369 (2002)].
4. A. N. Obraztsov, A. P. Volkov, A. I. Boronin, and S. V. Koshcheev, Zh. Éksp. Teor. Fiz. **120**, 970 (2001) [JETP **93**, 846 (2001)].
5. V. D. Frolov, A. V. Karabutov, S. M. Pimenov, *et al.*, Diamond Relat. Mater. **10**, 1719 (2001).
6. A. N. Obraztsov, A. A. Zolotukhin, A. O. Ustinov, *et al.*, Carbon **41**, 836 (2003).
7. J.-M. Bonard, H. Kind, Th. Stöckli, and L.-O. Nilsson, Solid-State Electron. **45**, 893 (2001).
8. A. N. Obraztsov, I. Yu. Pavlovsky, and A. P. Volkov, J. Vac. Sci. Technol. B **17**, 674 (1999).
9. J. D. Levine, J. Vac. Sci. Technol. B **13**, 553 (1995).
10. V. T. Binh and Ch. Adessi, Phys. Rev. Lett. **85**, 864 (2000).
11. A. N. Obraztsov, A. P. Volkov, and I. Yu. Pavlovskiĭ, Pis'ma Zh. Éksp. Teor. Fiz. **68**, 56 (1998) [JETP Lett. **68**, 59 (1998)].
12. A. N. Obraztsov, A. P. Volkov, I. Yu. Pavlovskiĭ, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **69**, 381 (1999) [JETP Lett. **69**, 411 (1999)].
13. Z. L. Wang, R. P. Gao, W. A. de Heer, and P. Poncharal, Appl. Phys. Lett. **80**, 856 (2002).
14. J. Cumings, A. Zettl, M. R. McCaetney, and J. C. H. Spence, Phys. Rev. Lett. **88**, 056804 (2002).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 97, No. 6, 2003, pp. 1246–1249.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 124, No. 6, 2003, pp. 1398–1402.

MISCELLANEOUS

Optical Excitations in Hexagonal Nanonetwork Materials¶

K. Harigaya

Nanotechnology Research Institute, AIST 305-8568, Tsukuba, Japan
Interactive Research Center of Science, Tokyo Institute of Technology 152-8551, Tokyo, Japan
e-mail: k.harigaya@aist.go.jp

Received March 1, 2003

Abstract—Optical excitations in hexagonal nanonetwork materials, for example, Boron–Nitride (BN) sheets and nanotubes, are investigated theoretically. Exciton dipoles directed from the B site to the N site are considered along the BN bond. When the exciton hopping integral is restricted to the nearest neighbors, two flat bands of excitons appear. The symmetry of these exciton bands is optically forbidden. Possible relations to experiment are discussed. *© 2003 MAIK "Nauka/Interperiodica".*

Hexagonal nanonetwork materials composed of atoms with ionic characters, for example, Boron–Nitride (BN) sheets and nanotubes [1, 2], have been intensely investigated. They are intrinsic insulators with an energy gap of about 4 eV, as the preceding band calculations have indicated [3, 4]. The possible photogalvanic effects depending on the chiralities of BN nanotubes have been proposed by the model calculation [5]. Although not many optical measurements on the BN systems have been reported, it is quite interesting to predict condensed matter properties of hexagonal nanonetwork materials.

In this paper, we investigate optical excitation properties in BN systems. The bonding is positively polarized at the B site and is negatively polarized at the N site. There is a permanent electric dipole moment along the BN bond directed from the B site to the N site: When we assume the one-orbital model [5], as shown in Fig. 1, the energy of the highest occupied atomic orbital of N is larger than that of B and the energy of the lowest unoccupied orbital of B is smaller than that of N. There is a band gap

$$\Delta \equiv \varepsilon_B - \varepsilon_N \sim 4 \text{ eV}$$

(see [3, 4]). Low-energy optical excitations are the excitations of electron–hole pairs between the higher occupied states of N and the lower unoccupied states of B atoms. The presence of the dipole moments gives rise to strong excitonic properties, illustrated in Fig. 2.

In what follows, we discuss optical excitations in hexagonal nanonetwork materials, BN sheets and nanotubes. We show that two flat bands of excitons, which are optically forbidden, appear in the energy dispersions. Possible relations to experiments are discussed.

The interactions between the electric dipole moment along the BN bond has the strongest interaction strengths when the exciton hopping integral is restricted to the nearest neighbor dipoles. In Fig. 3a, the B and N atoms are represented by full and open circles,

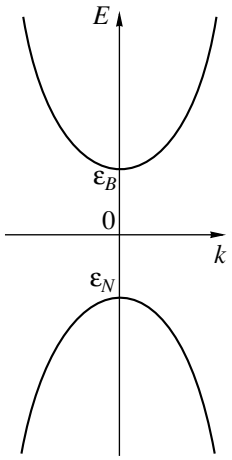

Fig. 1. One-orbital model for low-energy dispersions of the BN network.

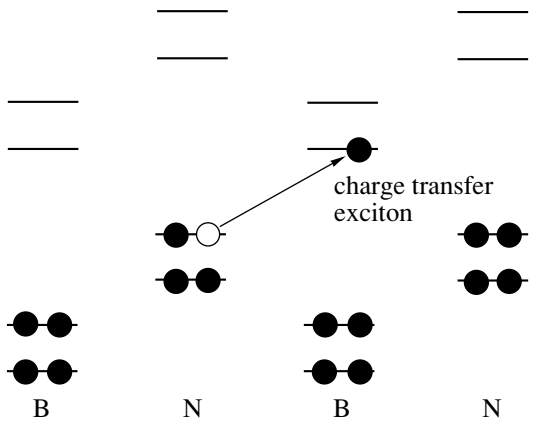

Fig. 2. Optical excitations along the BN alternations.

¶This article was submitted by the author in English.

respectively. We assume the one-orbital Hubbard model with the hopping integral of electrons t, the on-site repulsion U, and the energy difference Δ between the B and N sites. After second-order perturbations, we obtain the nearest-neighbor interactions

$$J_1 = \frac{t^2}{-\Delta + U}$$

for the conserved excited spin (type-1 interaction) and

$$J_2 = \frac{t^2}{\Delta} + \frac{t^2}{-\Delta + U}$$

in the case where the spin of the excited electron flips (type-2 interaction). The meaning of the interactions J_1 and J_2 is illustrated in Figs. 4 and 5, respectively. When the condition $U > \Delta$ applies, J_1 and J_2 become positive. We discuss this case first, then comment on the case where J_1 and J_2 are negative afterwards. The interactions are present along the thin lines in Fig. 3a. The several arrows show the directions of dipole moments. After the extraction of interactions J_1 and J_2, there remains the two-dimensional Kagomé lattice, shown in Fig. 3b. As described in [6], the Kagomé lattice is obtained as a line graph of the hexagonal lattice. Therefore, the optical excitation Hamiltonian becomes

$$H = \sum_{\langle i,j\rangle} \sum_{\sigma = \alpha,\beta} J_1(|i,\sigma\rangle\langle j,\sigma| + \text{H.c.}) + \sum_{\langle i,j\rangle} J_2(|i,\alpha\rangle\langle j,\beta| + |i,\beta\rangle\langle j,\alpha| + \text{H.c.}), \quad (1)$$

where indices i and j denote the vertex points of the Kagomé lattice and the sum is taken over the nearest neighbor pairs $\langle i, j\rangle$ and the excited spin σ. The unit cell has three lattice points 1, 2, and 3, as shown in Fig. 3b.

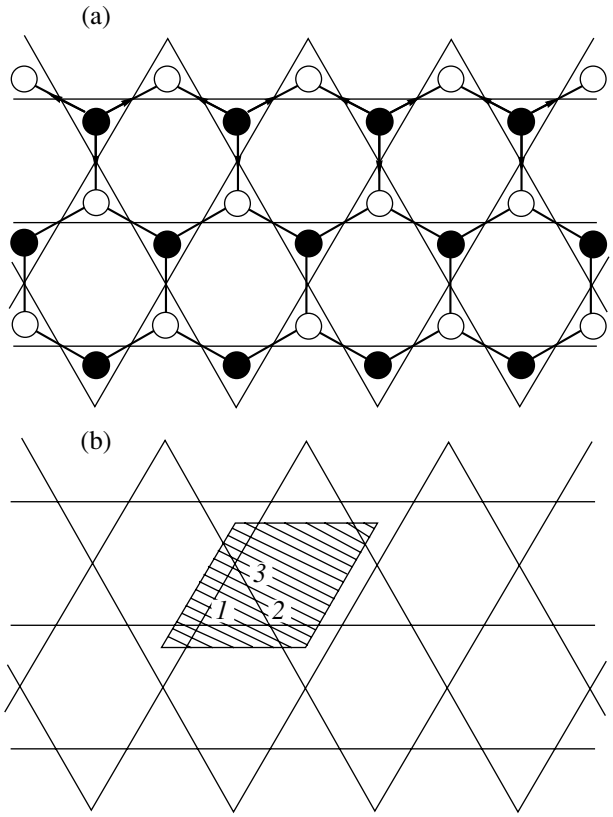

Fig. 3. (a) hexagonal nanonetwork of boron (solid circles) and nitrogen sites (empty circles). Several arrows indicate the directions of dipole moments, and the thin lines represent the conjugate Kagomé lattice network. (b) Kagomé lattice extracted from Fig. 3a. The shaded area is the unit cell, which has three lattice points indicated by numbers.

The model has six eigenenergies that are expressed in terms of wavenumbers $\mathbf{k} = (k_x, k_y)$ as

$$E = \begin{cases} -2(J_1 + J_2), \\ (J_1 + J_2)\left\{1 \pm \sqrt{1 + 4\cos(k_x b/2)[\cos(k_x b/2) + \cos(\sqrt{3}k_y b/2)]}\right\}, \\ 2(-J_1 + J_2), \\ (J_1 - J_2)\left\{1 \pm \sqrt{1 + 4\cos(k_x b/2)[\cos(k_x b/2) + \cos(\sqrt{3}k_y b/2)]}\right\}, \end{cases} \quad (2)$$

where the two-dimensional x and y axes are defined as usual in Fig. 3, $b = \sqrt{3}\,a$ is the unit cell length of the Kagomé lattice in Fig. 3b, and a is the bond length in Fig. 3a. The dispersion relations are shown in Fig. 6 for the representative dimensionless parameters $J_1 = 1$ and $J_2 = 2$ that correspond to the case where $U = 2t$ and $\Delta = t$ in the second-order perturbation relations with $t = 1$. There appears a dispersionless band (triplet state) with

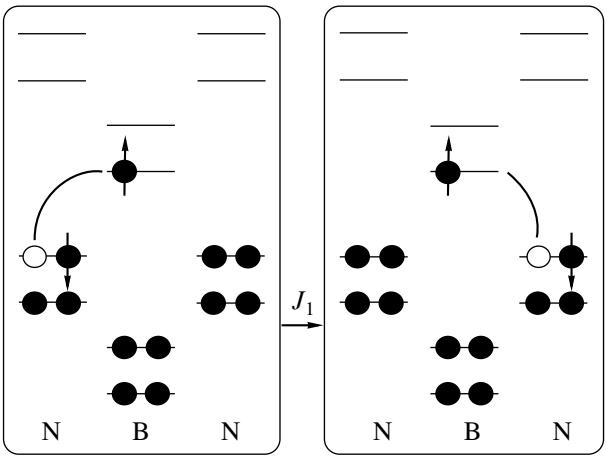

Fig. 4. Type-1 interaction J_1 that preserves an excited spin.

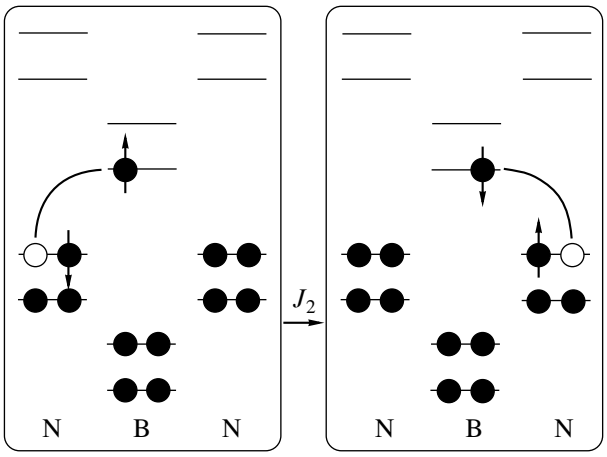

Fig. 5. Type-2 interaction J_2 where an excited spin flips.

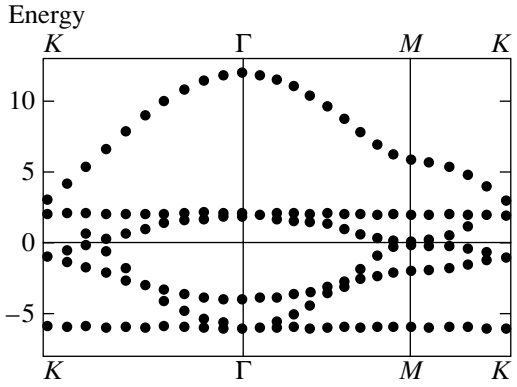

Fig. 6. Energy dispersions for the dimensionless parameters $J_1 = 1$ and $J_2 = 2$.

the lowest energy $-2(J_1 + J_2)$ when J_1 and J_2 are positive. There is another dispersionless band (singlet state) at the higher energy $2(-J_1 + J_2)$. The other four bands have dispersions that are similar to those of the two-dimensional network of electrons on graphite [7]. In fact, with the electron hopping integral of graphite denoted as t, the dispersion is

$$E = \pm t \times \sqrt{1 + 4\cos(k_x b/2)[\cos(k_x b/2) + \cos(\sqrt{3}k_y b/2)]}. \quad (3)$$

We note that the x and y axes are interchanged compared with the definitions used in [7].

Such an appearance of the flat band, for example, in the Kagomé lattice, has been discussed in the literature [6, 8] from the standpoint of possible ferromagnetism. In the present case, the lowest optical excitation band becomes flat in the honeycomb BN plane when the interactions J_1 and J_2 are positive. When the BN plane is rolled up into nanotubes, the flat band is also dispersionless. The interesting properties of excitons on the Kagomé lattice have been investigated recently [9].

To discuss how the excitons appear in optical experiments, we must consider symmetries of the wave functions. The most interesting part is the wave function of the lowest excitons with the energy $-2(J_1 + J_2)$ when J_1 and J_2 are positive.

Solving the eigenvalue problem at the wavenumber $\mathbf{k} = (0, 0)$,

$$\begin{pmatrix} 0 & 0 & 2J_1 & 2J_2 & 2J_1 & 2J_2 \\ 0 & 0 & 2J_2 & 2J_1 & 2J_2 & 2J_1 \\ 2J_1 & 2J_2 & 0 & 0 & 2J_1 & 2J_2 \\ 2J_2 & 2J_1 & 0 & 0 & 2J_2 & 2J_1 \\ 2J_1 & 2J_2 & 2J_1 & 2J_2 & 0 & 0 \\ 2J_2 & 2J_1 & 2J_2 & 2J_1 & 0 & 0 \end{pmatrix} \Psi = E\Psi, \quad (4)$$

we obtain the twofold degenerate solutions for the energy $E = -2(J_1 + J_2)$,

$$\Psi^\dagger = \frac{1}{2}(1, 1, -1, -1, 0, 0), \quad (5)$$

and

$$\Psi^\dagger = \frac{1}{2\sqrt{3}}(1, 1, 1, 1, -2, -2). \quad (6)$$

It follows that the spin of these states is triplet. Similarly, we obtain the solutions for the energy $E = 2(-J_1 + J_2)$,

$$\Psi^\dagger = \frac{1}{2}(1, -1, -1, 1, 0, 0), \quad (7)$$

and

$$\Psi^\dagger = \frac{1}{2\sqrt{3}}(1, -1, 1, -1, -2, 2). \quad (8)$$

The spin alignment of these states is singlet.

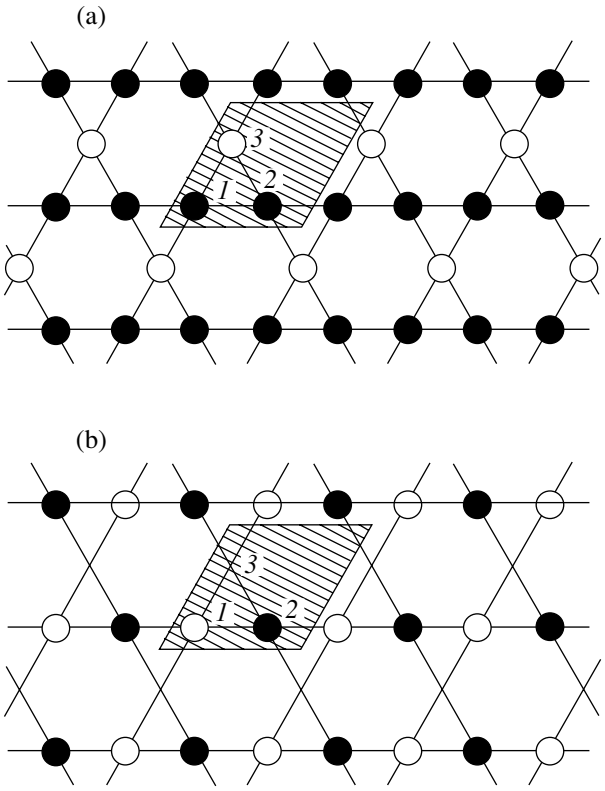

Fig. 7. Symmetries of two wavefunctions at $E = -2(J_1 + J_2)$ for the spin up or down sector. The solid and empty circles indicate positive and negative values at the lattice point, respectively. If the value at the lattice point is zero, nothing is shown there. The numbers *1*, *2*, and *3* in the unit cell correspond to the first, third, and fifth (second, fourth, and sixth) elements for the up (down) spin sector of the wavefunction Ψ in Eq. (4), respectively.

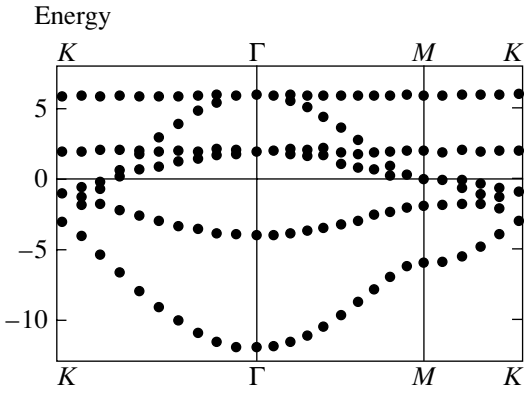

Fig. 8. Energy dispersions for the dimensionless parameters $J_1 = -2$ and $J_2 = -1$.

The symmetry of solution (5) of the up or down spin sector is shown in Fig. 7a, and that of solution (6) is displayed in Fig. 7b. We find that both wave functions are symmetric under spatial inversion, and therefore they have the gerade symmetry. The transition to the flat band exciton is optically forbidden. Such properties might result in interesting optical measurements in hexagonal nanonetwork materials. Similarly, the symmetry of states (7) and (8) is gerade with respect to the spatial inversion.

Finally, the dispersions of excitons in the case where $J_1 < 0$ and $J_2 < 0$ are of interest. Representative dispersions are shown in Fig. 8 for the dimensionless parameters $J_1 = -2$ and $J_2 = -1$. These correspond to $U = 0.5t$ and $\Delta = t$ with $t = 1$ in the formula of the second-order perturbations J_1 and J_2. The energy band structures are almost interchanged between top and bottom compared with those in Fig. 6. There is a flat band at the top of the excitonic bands. The lowest exciton has a finite dispersion, but this state is still an optically forbidden triplet.

The optically forbidden transition is known in C_{60} molecules [10]. The luminescence from the lowest exciton has a long lifetime due to the forbidden transition nature [11, 12]. We have analyzed possible phonon couplings in the luminescence spectra [13]. A formalism similar to the one in this paper could be applied to systems with honeycomb or Kagomé network materials, where neighboring interactions with dipoles are effective.

In summary, optical excitations in BN sheets and nanotubes have been investigated theoretically. We have shown that two flat bands of excitons, which are optically forbidden, appear in the energy dispersions. Possible relations to experiments were discussed.

REFERENCES

1. D. Golberg, Y. Bando, K. Kurashima, and T. Sato, Solid State Commun. **116**, 1 (2000).
2. D. Golberg, Y. Bando, L. Bourgeois, *et al.*, Appl. Phys. Lett. **77**, 1979 (2000).
3. A. Rubio, J. L. Corkill, and M. L. Cohen, Phys. Rev. B **49**, 5081 (1994).
4. X. Blase, A. Rubio, S. G. Louie, and M. L. Cohen, Europhys. Lett. **28**, 335 (1994).
5. P. Král, E. J. Mele, and D. Tománek, Phys. Rev. Lett. **85**, 1512 (2000).
6. A. Mielke, J. Phys. A **24**, 3311 (1991).
7. R. Saito, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, Phys. Rev. B **46**, 1804 (1992).
8. A. Mielke, J. Phys. A **25**, 4335 (1992).
9. H. Ishii, T. Nakamura, and J. Inoue, Surf. Sci. **514**, 206 (2002); cond-mat/0110360.
10. K. Harigaya and S. Abe, Phys. Rev. B **49**, 16746 (1994).
11. M. Matus, H. Kuzmany, and E. Sohmen, Phys. Rev. Lett. **68**, 2822 (1992).
12. D. Dick, X. Wei, S. Jeglinski, *et al.*, Phys. Rev. Lett. **73**, 2760 (1994).
13. B. Friedman and K. Harigaya, Phys. Rev. B **47**, 3975 (1993).