Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 371–391.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 435–458.

GRAVITATION, ASTROPHYSICS

An Exact General-Relativity Solution for the Motion and Intersections of Self-Gravitating Shells in the Field of a Massive Black Hole

M. V. Barkov[a, *], V. A. Belinski[b, **], and G. S. Bisnovatyi-Kogan[c, *]**

[a]*Space Research Institute, Russian Academy of Sciences, ul. Profsoyuznaya 84/32, Moscow, 117810 Russia*
**e-mail: barmv@sai.msu.ru*
[b]*National Institute of Nuclear Research (INFN) and International Center of Relativistic Astrophysics (ICRA), Rome, Italy*
***e-mail: belinski@icra.it*
[c]*Institute des Hautes Études Scientifiques (IHES), Bures-sur-Yvette, France*
****e-mail: gkogan@mx.iki.rssi.ru*
Received March 13, 2002

Abstract—The motion with intersections of relativistic gravitating shells in the Schwarzschild gravitational field of a central body is considered. Formulas are derived for calculating parameters of the shells after intersection via their parameters before intersection. Such special cases as the Newtonian approximation, intersections of light shells, and intersections of a test shell with a gravitating shell are also considered. The ejection of one of the shells to infinity in the relativistic region is described. The equations of motion for the shells are analyzed numerically. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

An interesting application of the theory of gravitating, spherically symmetric shells is related to the possibility of modeling some properties of star clusters by the evolution of such shells, each moving in the field produced by all the remaining shells and the central body [1–6]. Let there be steady motion of a large number of gravitating particles (stars) around a nonrotating central body at rest. Because of the large number of particles, each of them may be assumed to move in an average stationary spherically symmetric field (which is not a Schwarzschild field inside the cluster but transforms into a Schwarzschild field outside the cluster). Each particle moves along a trajectory that is a generalization of elliptical orbits in Newtonian mechanics; the total energy and total angular momentum of each particle are conserved. Let us single out a sphere of arbitrary radius r_0 inside the cluster at some arbitrary time t_0 and single out the particles on this sphere with equal radial velocities and with the same absolute value of the total angular momentum per unit rest mass, whence it follows that the total energies per unit rest mass are also equal. One can see from the first integrals of geodesics in a stationary spherically symmetric field that the radial motion of all the singled-out particles obeys the same equations. This implies that the radial motion of all these particles is the same; i.e., they all form what is called a shell. Whereas the shell as a whole moves only radially, its constituent particles also move tangentially inside the shell. When the number of particles is large, the tangential particle motions may be treated as chaotic motions with the parameters distributed uniformly and isotropically on the two-dimensional shell surface. Consequently, these motions can be taken into account hydrodynamically as the presence of a tangential pressure. The radial behavior of a cluster is modeled by the motion of a discrete set of infinitely thin shells, each moving in a vacuum. The radial motion is accurately taken into account by solving the gravitational equations. Of course, not all cluster properties can be described in terms of this model but several aspects of its dynamics are well modeled. These include, for example, the violent relaxation of a star cluster [1, 5] and the ejection of some of the cluster layers during its collapse after the loss of stability.

Clearly, the model described above has the following two main determining elements: first, the motion of one shell in the field of a central mass and, second, the motion of two shells in the field of a central mass, including the possible intersections of the shells with one another. The first problem was solved by Chase [7] more than thirty years ago. This author did not raise the question of applications to analysis of the cluster behavior but derived the equation of motion for one shell with a tangential pressure (in general, also with an electric charge) and studied its stability against collapse. He derived the equation of motion for a shell in a general form, i.e., without making any special assumptions about the equation of state for the shell matter.

Up until now, there has been no complete general-relativity solution to the second problem. We hope that this paper will bridge this gap to some extent. We emphasize that, unless the shells intersect, their equations of motion are a trivial generalization of the equa-

tion of motion for one shell. This is because a spherically symmetric shell produces no gravitational field within itself and, therefore, does not affect anything inside it. At the same time, the effect of the inner shell on the outer shell can be easily taken into account: only the mass parameters of the Schwarzschild metrics in outer (relative to the inner shell) regions change. Thus, the main problem here is related precisely to the intersection of shells and involves allowance for the effects of a large number of such intersections. In this case, we treat the intersection as a purely ballistic process; i.e., each shell is affected only by its surrounding gravitational field, and there is no direct nongravitational interaction between the shells at the time of their intersection. For this condition to be satisfied, the joining must be such that, in the limit of test shells (whose gravitation may be ignored), they would intersect without responding in any way to this process; i.e., all parameters of the motion of each shell in this limit would remain continuous when passing through the point of intersection.

Note that the above problem was considered in terms of Newtonian gravitation in [1–6, 8]. An interesting result (first obtained in [4]) is the ejection of one of the shells to spatial infinity through energy transfer between the shells as they intersect. In [8], we also described this effect but with some new features. In addition, we numerically solved the equations of motion for two shells by taking into account a large number of their intersections with one another. We found that, after a sufficiently large number of intersections, the motion of the shells became chaotic with the strong sensitivity to small variations in initial parameters typical of chaos.

Noteworthy are also [9, 10], in which the motion and intersection of shells was considered in a strong gravitational field. As our comparison with the exact solution presented here shows, the approximate procedure used in these papers to describe the intersection has a high accuracy only when one (or both) of the velocities of the intersecting shells is low and when the ratios of the effective shell masses to the central-body mass are small.

Finally, we emphasize that our study is intended for specific astrophysical applications and, of course, it does not cover all aspects of the theory for gravitating shells. Therefore, our list of references is quite limited and reflects the development of those applied issues that are dealt with in this paper. Among the most recent papers that have a direct bearing on our study, the paper by Berezin and Okhrimenko [11] is particularly noteworthy. These authors also investigated the behavior of a gravitating shell composed of particles moving in the field of a central mass and considered both astrophysical and quantum aspects of the problem. However, the authors did not touch on the issue of the intersection of two shells. Neronov [12] considered the problem of inelastic collisions between an arbitrary number of n-dimensional shells in $(n + 2)$-dimensional space-time and pointed out the conservation law that related a certain combination of shell parameters before collision with a similar combination of shell parameters after collision. The author applied this result to the collision of three-dimensional shells in five-dimensional anti–de Sitter space-time to study the properties of the so-called ekpyrotyc cosmological model [13, 14]. For completeness, note also [15], which mainly repeats Neronov's results. The interested reader will find in the cited papers [11–15] quite an extensive review of the literature on various aspects of the theory for thin gravitating shells.

2. A GRAVITATING SHELL WITH A TANGENTIAL PRESSURE

Chase [7] obtained his result by the geometrical method that was first used in [16], where the equation of motion for a spherically symmetric dust shell was derived by this method. Naturally, the same result can be obtained in a more habitual (for the physicist) way by making up the energy–momentum tensor with an appropriate δ-shaped source and by directly integrating the Einstein equations with this right-hand side. We used precisely this derivation method, and our result closely matched Chase's result. Of course, we do not detail our calculations here but only note the main points, because this is of methodological interest and, in addition, provides a convenient means of introducing all the necessary concepts and notation.

Let there be a central body of mass m_{in} and let a spherically symmetric shell move outside this body. Even before solving any equations, it is clear that the metric inside and outside the shell is the Schwarzschild metric but with different mass parameters. Using the coordinates $x^0 = ct$ and r, which are continuous when passing through the shell, we can write the intervals[1] inside, outside, and on the shell as

$$-(ds^2)_{\text{in}} = -e^{T(t)} f_{\text{in}}(r)c^2dt^2 + f_{\text{in}}^{-1}(r)dr^2 + r^2d\Omega^2, \quad (1)$$

$$-(ds^2)_{\text{out}} = -f_{\text{out}}(r)c^2dt^2 + f_{\text{out}}^{-1}(r)dr^2 + r^2d\Omega^2, \quad (2)$$

$$-(ds^2)_{\text{on}} = -c^2d\tau^2 + r_0^2(\tau)d\Omega^2, \quad (3)$$

where we denoted

$$d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$$

and

$$f_{\text{in}} = 1 - \frac{2km_{\text{in}}}{c^2r}, \quad f_{\text{out}} = 1 - \frac{2km_{\text{out}}}{c^2r}. \quad (4)$$

In the interval (3), τ is the proper time of the shell. The factor e^T in (1) is required to ensure that the time coordinate t be continuous when passing through the shell. The parameter $m_{\text{out}}c^2$ is the total energy of the system.

[1] The interval is written as $-ds^2 = g_{ik}dx^idx^k$ and the metric signature is (–, +, +, +), i.e., $g_{00} < 0$. The Roman indices take on 0, 1, 2, and 3. The standard notation for spherical coordinates is $(x^0, x^1, x^2, x^3) = (ct, r, \theta, \phi)$. The Newtonian gravitational constant is denoted by k.

If the equation of motion for the shell is

$$r = R_0(t),$$

then joining the angular parts of all three intervals (1)–(3) yields

$$r_0(\tau) = R_0[t(\tau)], \tag{5}$$

where the function $t(\tau)$ describes the relationship between the global time and the proper time of the shell. Joining the radial-time parts of the intervals (1)–(3) on the shell requires that the following relations hold:

$$f_{\text{in}}(r_0)\left(\frac{dt}{d\tau}\right)^2 e^{T(t)} - f_{\text{in}}^{-1}(r_0)\left(\frac{dr_0}{cd\tau}\right)^2 = 1, \tag{6}$$

$$f_{\text{out}}(r_0)\left(\frac{dt}{d\tau}\right)^2 - f_{\text{out}}^{-1}(r_0)\left(\frac{dr_0}{cd\tau}\right)^2 = 1. \tag{7}$$

If the equation of motion for the shell [i.e., the function $r_0(\tau)$] is known, then the function $t(\tau)$ follows from (7) and we can then derive $T(t)$ from (6). Thus, the problem consists only in determining $r_0(\tau)$, which can be done by directly integrating the Einstein equations for the metric

$$-(ds^2) = g_{00}(t, r)c^2dt^2 + g_{11}(t, r)dr^2 + r^2d\Omega^2 \tag{8}$$

and the energy–momentum tensor

$$T_i^k = \varepsilon u_i u^k + (\delta_i^2\delta_2^k + \delta_i^3\delta_3^k)p. \tag{9}$$

Here, $u_2 = u_3 = 0$ and ε, p, u_0, and u_1 depend on the coordinates t and r alone, with

$$u^0u_0 + u^1u_1 = -1.$$

The energy density ε is

$$\varepsilon = \frac{M(t)c^2\delta[r - R_0(t)]}{4\pi r^2u^0\sqrt{-g_{00}g_{11}}}, \tag{10}$$

where δ is the standard δ function. In the absence of tangential pressure p, the quantity M in (10) would be constant and the energy density ε would be the sum of the rest energies of the shell particles per unit volume of the radially comoving frame. In this case, M would be the total rest mass of the shell or its bare or baryonic mass. In the presence of pressure, Mc^2 includes the rest energy along with the energy (in the radially comoving frame) of the tangential motions of the shell particles that produce this pressure. In this case, Mc^2 can no longer be constant but depends on the degree of compression of the shell, i.e., on its radius $R_0(t)$ or simply on time t, which is explicitly specified in (10).

As in any spherically symmetric problem, we can choose the hydrodynamic equations $T_{i;\,k}^k = 0$ and the $\binom{0}{0}$, $\binom{1}{1}$, and $\binom{1}{0}$ components of the Einstein equations taken in the form

$$R_i^k - \frac{1}{2}\delta_i^k R = \frac{8\pi k}{c^4}T_i^k$$

(all the remaining components of these equations are identically satisfied either in view of the Bianchi identities or because of the symmetry of the problem) as the complete system of equations. As we see from (9), these components of the Einstein equations contain no pressure. It is easy to show that they actually lead[2] to the solution (1)–(4) with arbitrary constants m_{in} and m_{out} and, in addition, to the following first-order equation for the function $r_0(\tau)$:

$$\sqrt{f_{\text{in}}(r_0) + \left(\frac{dr_0}{cd\tau}\right)^2} + \sqrt{f_{\text{out}}(r_0) + \left(\frac{dr_0}{cd\tau}\right)^2} = \frac{2(m_{\text{out}} - m_{\text{in}})}{\mu(\tau)}, \tag{11}$$

where we denoted

$$\mu(\tau) = M[t(\tau)]. \tag{12}$$

Two of the four equations $T_{i;\,k}^k = 0$ are identically satisfied because of the symmetry of the problem. The other two equations can be reduced to such a form that one of them will serve simply to determine the pressure,

$$p = -\frac{dM}{dt}\frac{c\delta[r - R_0(t)]}{8\pi r u^1\sqrt{-g_{00}g_{11}}}, \tag{13}$$

and the other is identically satisfied after substituting this expression for p in it (it is a differential equation of the second order in τ for the function $r_0(\tau)$, which is nothing else but the result of differentiating Eq. (11) with respect to τ). It should be noted, however, that the latter holds only in the general unsteady-state case where $R_0 \neq \text{const}$.

The case of a steady-state shell, $R_0 = \text{const}$, requires a special analysis. In this special case, all quantities depend only on one variable r and $u^1 = 0$. Three of the four equations $T_{i;\,k}^k = 0$ are identically satisfied and the

[2] When integrating the equations of the problem under consideration, we need only two standard rules to work with symbolic functions: $d\theta(x)/dx = \delta(x)$ (where θ is the Heaviside step function) and $F(x)\delta(x) = (1/2)[F(-0) + F(+0)]\delta(x)$.

fourth equation again serves to determine the pressure, which can be written as

$$p = \frac{Mc^2}{32\pi R_0^2} \times \left[\frac{1-f_{\text{in}}(R_0)}{\sqrt{f_{\text{in}}(R_0)}} + \frac{1-f_{\text{out}}(R_0)}{\sqrt{f_{\text{out}}(R_0)}}\right]\delta(r-R_0). \quad (14)$$

Expression (10) for the energy density now takes the form

$$\varepsilon = \frac{Mc^2}{8\pi R_0^2}[\sqrt{f_{\text{in}}(R_0)} + \sqrt{f_{\text{out}}(R_0)}]\delta(r-R_0). \quad (15)$$

The shell radius R_0 itself must be determined from Eq. (11), in which we should now set $dr_0/d\tau = 0$ and $\mu = M = \text{const}$. It is easy to establish that a positive solution for R_0 at positive parameters m_{in} and m_{out} exists only when the inequality

$$(m_{\text{out}} - m_{\text{in}})^2 < M^2 \quad (16)$$

is satisfied; this solution is

$$R_0 = \frac{kM^2}{2c^2}\frac{m_{\text{out}} + m_{\text{in}} + \sqrt{4m_{\text{out}}m_{\text{in}} + M^2}}{M^2 - (m_{\text{out}} - m_{\text{in}})^2}. \quad (17)$$

Expressions (14)–(17) give the solution to the problem for a steady-state shell.

Let us return to the general unsteady-state case. The equation of motion for the shell (11) can be written in several equivalent forms. Below, we give the following two forms:

$$\sqrt{f_{\text{in}}(r_0) + \left(\frac{dr_0}{cd\tau}\right)^2} = \frac{m_{\text{out}} - m_{\text{in}}}{\mu(\tau)} + \frac{k\mu(\tau)}{2c^2 r_0}, \quad (18)$$

$$\sqrt{f_{\text{out}}(r_0) + \left(\frac{dr_0}{cd\tau}\right)^2} = \frac{m_{\text{out}} - m_{\text{in}}}{\mu(\tau)} - \frac{k\mu(\tau)}{2c^2 r_0}. \quad (19)$$

Given expressions (4) for f_{in} and f_{out}, it is easy to verify that both (18) and (19) directly follow from Eq. (11). At the same time, Eq. (11) is the sum of Eqs. (18) and (19). Yet another equivalent form of the equations of motion for the shell can be derived by squaring each of Eqs. (18) and (19) and then adding them term by term. As a result, we obtain

$$1 + \left(\frac{dr_0}{cd\tau}\right)^2 = \frac{(m_{\text{out}} - m_{\text{in}})^2}{\mu^2(\tau)} + \frac{k(m_{\text{out}} + m_{\text{in}})}{c^2 r_0} + \frac{k^2\mu^2(\tau)}{4c^4 r_0^2}. \quad (20)$$

The equation of motion for the shell is given in [7] precisely in this form. We emphasize that, for actual astrophysical applications, all the radicals encountered in our paper should be taken to be positive.

To proceed further, we must specify the function $\mu(\tau)$, which, as we see from (10) and (13), is equivalent to specifying an equation of state. Here, of course, there is a wide range of possibilities to choose from, but we restrict our analysis to the shell model described in the Introduction, i.e., a shell composed of particles that move in the field of a central mass. The quantity Mc^2 in (10) in the absence of pressure is the total rest energy of the shell, i.e.,

$$Mc^2 = \sum_a m_a c^2,$$

where the sum is taken over all particles and m_a is the rest mass of the individual particle. The shell is meant to be in the state of rest with respect to its radial motion, and this state takes place in the radially comoving frame. Since the tangential motions are normal to the radial motion, their role reduces only to producing an effective rest mass with respect to the radial motion of the entire shell; i.e., in their presence, we have the following expression for Mc^2:

$$Mc^2 = \sum_a \sqrt{m_a^2c^4 + p_a^2c^2} = \sum_a \left(m_a c^2 \sqrt{1 + \frac{p_a^2}{m_a^2c^2}}\right), \quad (21)$$

where p_a is the tangential momentum of each particle. By the definition of the shell (see the Introduction), all its particles are at the same distance R_0 from the center and they all have the same $|l_a|/m_a$ ratio (where l_a is the total angular momentum of particle a). Since

$$\frac{p_a^2}{m_a^2} = \frac{l_a^2}{m_a^2 R_0^2} = \frac{\text{const}}{R_0}, \quad (22)$$

the square root in (21) does not depend on the index a and this root can be taken outside the sum and the sum

$$\sum_a m_a c^2 = mc^2,$$

where the constant m is the total rest mass of the entire shell. Because the $|l_a|/m_a$ ratio is independent of a, we have

$$\left(\sum_a m_a c^2\right)^2 \frac{p_a^2}{m_a^2c^2} = \frac{c^2}{R_0^2}\left(\frac{|l_a|}{m_a}\sum_a m_a\right)^2 = \frac{c^2}{R_0^2}\left(\sum_a |l_a|\right)^2. \quad (23)$$

As a result, formula (21) [given the designation (12)] can be written as

$$\mu(\tau) = \sqrt{m^2 + \frac{L^2}{c^2 r_0^2(\tau)}}, \quad (24)$$

where

$$L = \sum_a |l_a|$$

is the sum of the absolute values of the total angular momenta for the shell particles. Substituting (24) in Eq. (11) [or in one of the equivalent equations (18)–(20)] yields the final equation. We can determine the function $r_0(\tau)$ from this equation if the initial shell radius and the four arbitrary constants m_{in}, m_{out}, m, and L are specified.

According to (10), (12), (13), and (24), the equation of state that relates the shell energy density ε to the tangential pressure p is

$$p = \frac{\varepsilon}{2}\frac{L^2}{m^2c^2R_0^2}\left(1 + \frac{L^2}{m^2c^2R_0^2}\right)^{-1}.$$

The relation

$$\frac{L^2}{m^2c^2R_0^2} = \frac{p_a^2}{m_a^2c^2}$$

(recall that p_a^2/m_a^2 do not depend on particle number a) follows from the definition of the constants L and m and from formula (23). Hence, we see the limiting forms of the equation of state: we have a dust state, $p \ll \varepsilon$, for nonrelativistic tangential velocities ($p_a^2 \ll m_a^2c^2$) and obtain $p = (1/2)\varepsilon$ for ultrarelativistic tangential particle velocities ($p_a^2 \gg m_a^2c^2$), as should be the case for a two-dimensional ultrarelativistic gas. As the shell expands to infinity ($R_0 \longrightarrow \infty$), the equation of state always tends to the dust one, because the contribution of the tangential particle motions becomes negligible.

3. THE INTERSECTION OF SHELLS

Let us now consider the next (in complexity) case of two shells that move in the field of a central mass. When the shells do not intersect, the equations of their motion can be immediately written without difficulty by using the previously derived equations of motion for one shell and the fact that a spherically symmetric shell has no effect on anything inside it. When there is an intersection, this information is no longer enough, because additional joining conditions that are not contained in the theory of motion for one shell are required at the point of intersection. As previously, we cover the part of the physical space-time of interest by a global coordinate system $(r, x^0) = (r, ct)$ with continuous r and t. Let shell 1 be inside shell 2 at the initial time and in its vicinity and then let these shells intersect at some point of space-time (r^*, t^*), so shell 2 turns out to be inside shell 1 after t^* and these relative positions of the shells are maintained for some time after t^*. The intersection process is shown in Fig. 1. If $r = R_1(t)$ and $r = R_2(t)$ are the equations of motion for the first and second shells, then we have four space-time regions:

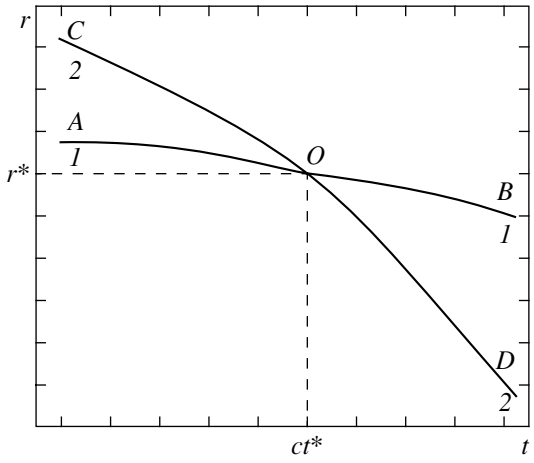


Fig. 1. A schematic view of the intersection of two gravitating shells. The coordinates of point O are ct^* and r^*.

$$COB \quad (r > R_1, r > R_2),$$

$$COA \quad (R_1 < r < R_2),$$

$$AOD \quad (r < R_1, r < R_2),$$

$$BOD \quad (R_2 < r < R_1).$$

In each of these regions, the metric is the Schwarzschild metric but supplemented with the dilaton factor $e^{T(t)}$, which is required if we wish to cover all four regions by a global continuous time coordinate t. In this case, it remains possible to choose the metric coefficient g_{00} in a purely Schwarzschild form without the dilaton factor in any of these four regions (but only in one of them), thereby fixing the choice of global time t. It would be natural to introduce a purely Schwarzschild metric outside the shells, i.e., in region COB adjacent to spatial infinity. Thus, the metric in these four regions has the form (8) but with different metric coefficients g_{00} and g_{11}:

$$g_{00}^{(COB)} = -f_{out}(r), \quad g_{11}^{(COB)} = f_{out}^{-1}(r), \tag{25}$$

$$g_{00}^{(COA)} = -e^{T_1(t)}f_{12}(r), \quad g_{11}^{(COA)} = f_{12}^{-1}(r), \tag{26}$$

$$g_{00}^{(AOD)} = -e^{T_0(t)}f_{in}(r), \quad g_{11}^{(AOD)} = f_{in}^{-1}(r), \tag{27}$$

$$g_{00}^{(BOD)} = -e^{T_2(t)}f_{21}(r), \quad g_{11}^{(BOD)} = f_{21}^{-1}(r). \tag{28}$$

Here, f_{in} and f_{out} are the same as those in (4) and f_{12} and f_{21} are given by similar expressions:

$$f_{12} = 1 - \frac{2km_{12}}{c^2r}, \quad f_{21} = 1 - \frac{2km_{21}}{c^2r}. \tag{29}$$

The mass parameters m_{out}, m_{in}, and m_{12} are assumed to be specified by the initial conditions. The mass param-

eter m_{21} formed in the region between the shells after their intersection is to be determined from the dynamics of the process and from the joining conditions at point (r^*, t^*).

Let us first write out the equations of motion for the shells before their intersection at t^*. To do this, it will suffice to turn to the equations of motion for one shell, for example, in the form (18), (19) and to take into account the fact that the mass parameters in regions *AOD*, *COA*, and *COB* are m_{in}, m_{12}, and m_{out}, respectively. In this case, it is convenient to use the equations in the forms (19) and (18) for (inner) shell 1 and (outer) shell 2, respectively:

$$\sqrt{f_{12}(r_1)+\left(\frac{dr_1}{cd\tau_1}\right)^2} = \frac{m_{12}-m_{\text{in}}}{M_1(r_1)} - \frac{kM_1(r_1)}{2c^2r_1}, \quad (30)$$

$$\sqrt{f_{12}(r_2)+\left(\frac{dr_2}{cd\tau_2}\right)^2} = \frac{m_{\text{out}}-m_{12}}{M_2(r_2)} + \frac{kM_2(r_2)}{2c^2r_2}, \quad (31)$$

$$M_1 = \sqrt{m_1^2+\frac{L_1^2}{c^2r_1^2}}, \quad M_2 = \sqrt{m_2^2+\frac{L_2^2}{c^2r_2^2}}. \quad (32)$$

Here, τ_1 and τ_2 are the proper times of the first and second shells and $r_1(\tau_1) = R_1[t(\tau_1)]$ and $r_2(\tau_2) = R_2[t(\tau_2)]$. These equations must be supplemented with the joining conditions for the intervals on both shells. Joining on the first shell (on curve *AO*) yields

$$e^{T_1(t)}f_{12}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 - f_{12}^{-1}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = 1, \quad (33)$$

$$e^{T_0(t)}f_{\text{in}}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 - f_{\text{in}}^{-1}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = 1. \quad (34)$$

Joining on the second shell (on curve *CO*) yields

$$f_{\text{out}}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 - f_{\text{out}}^{-1}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = 1, \quad (35)$$

$$e^{T_1(t)}f_{12}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 - f_{12}^{-1}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = 1. \quad (36)$$

If all free parameters (m_{in}, m_{out}, m_{12}, m_1, m_2, L_1, L_2) and initial data to Eqs. (30)–(32) were specified and if the functions $r_1(\tau_1)$ and $r_2(\tau_2)$ were derived, then their substitution into (33)–(36) gives the functions $\tau_1(t)$, $\tau_2(t)$ and $T_1(t)$, $T_0(t)$, i.e., all that is required to determine the motion of the shells before their intersection. The intersection time t^* corresponds to some proper times $\tau_1(t^*)$ and $\tau_2(t^*)$. Therefore, the coordinates of the point of intersection r^* and t^* can be found from two relations,

$$r^* = r_1(\tau_1(t^*)), \quad r^* = r_2(\tau_2(t^*)). \quad (37)$$

Of course, we assume that Eqs. (37) have a solution (the cases where there is no solution correspond to the motion without intersections but we are not interested in such possibilities here).

The equations of motion for the shells after the intersection time t^* can be easily set up in just the same way again by turning to Eqs. (18) and (19) and by taking into account the fact that a new parameter m_{21} emerges in region *BOD*. We use Eq. (18) for (now outer) shell 1 and Eq. (19) for (now inner) shell 2:

$$\sqrt{f_{21}(r_1)+\left(\frac{dr_1}{cd\tau_1}\right)^2} = \frac{m_{\text{out}}-m_{21}}{M_1(r_1)} + \frac{kM_1(r_1)}{2c^2r_1}, \quad (38)$$

$$\sqrt{f_{21}(r_2)+\left(\frac{dr_2}{cd\tau_2}\right)^2} = \frac{m_{21}-m_{\text{in}}}{M_2(r_2)} - \frac{kM_2(r_2)}{2c^2r_2}. \quad (39)$$

Here, $M_1(r_1)$ and $M_2(r_2)$ are given by the same expressions (32) but, naturally, with the values of the functions $r_1(\tau_1)$ and $r_2(\tau_2)$ (and the times τ_1 and τ_2) after the intersection.

Joining the intervals on the first shell (on curve *OB*) yields

$$f_{\text{out}}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 - f_{\text{out}}^{-1}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = 1, \quad (40)$$

$$e^{T_2(t)}f_{21}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 - f_{21}^{-1}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = 1. \quad (41)$$

Joining on the second shell (on curve *OD*) requires that the following relations hold:

$$e^{T_2(t)}f_{21}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 - f_{21}^{-1}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = 1, \quad (42)$$

$$e^{T_0(t)}f_{\text{in}}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 - f_{\text{in}}^{-1}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = 1. \quad (43)$$

We see from Eqs. (38)–(43) that, if the parameter m_{21} were known, then the evolution of the shells after their intersection (for $t > t^*$) would be completely determined by their evolution before the intersection, because the initial data to Eqs. (38) and (39) have already been specified (it is known that the shell positions at time t^* are $r_1 = r_2 = r^*$ and the coordinates of the point of intersection r^* and t^* have already been found from the previous evolution). Thus, we must have an additional physical condition from which we could determine m_{21}.

Actually, we have not yet used the continuity condition for the relative velocity of the shells when passing through the point of intersection. We can make sure that this condition is necessary as follows. It is well known that the motion of the shells can be described in coordinates in which all metric coefficients are everywhere continuous and only their first derivatives together with

the Γ symbols undergo discontinuities (like finite jumps). This implies that the shell accelerations in these coordinates can only have discontinuities like finite jumps and, hence, the velocities must be everywhere continuous. It thus follows that the invariants expressed in terms of the velocities and metric coefficients alone will be everywhere continuous not only in these coordinates but also in any other coordinates, in particular, in our coordinates r and t. If there is only one shell, then only one such invariant exists, the norm of the 4-velocity vector for the shell. In the coordinates in which the metric is everywhere continuous, this norm is defined relative to the metric at the points of the shell trajectory themselves and is assumed to be a unit norm. In any other coordinates, the limiting values of the norm as the shell is approached on both of its sides remain unit ones. This circumstance is expressed by relations (6) and (7), which may be called the joining conditions.

In the presence of two shells, the aforesaid remains valid for each of them, which is expressed by the joining conditions (33)–(36) and (40)–(43). However, we now have one more invariant that can be made up only from the velocities and metric coefficients, namely, the scalar product of the unit 4-velocity vectors for the first and second shells. Before the intersection (trajectories *CO* and *AO*), this quantity can be determined everywhere in the region between the shells (sector *COA*) by making an appropriate parallel transport of the 4-velocity vectors for both shells to the points of this region from the shell sides adjacent to sector *COA*. We emphasize that this scalar product cannot be determined in such a way in sectors *COB* and *AOD*, because, in this case, we would have to transport the 4-velocity vector for one of the shells through the trajectory of the other shell, i.e., through the point of discontinuities in the Γ symbols, and this parallel transport is not uniquely determined. After the intersection, the scalar product of the unit 4-velocity vectors for the shells is uniquely determined for the same reason only in the region between the shells, i.e., in sector *BOD*. Actually, there is no need to make all these parallel transports, because both vectors are at the same point of spacetime at the intersection time, and we are interested only in this point. The previous analysis serves only to show that the limit of the scalar product of the unit 4-velocity vectors for the shells at the point of intersection *O* should be calculated relative to the metric in sector *COA* when it is approached from the side "before the intersection" (i.e., at $t = t^* - 0$) and relative to the metric in sector *BOD* when point *O* is approached on the side "after the intersection" (i.e., at $t = t^* + 0$). The continuity of this scalar product in the coordinate system where this metric is continuous and its invariance require that these limits also be equal in our coordinates. This is precisely the additional condition from which the parameter m_{21} can be determined. Let us now derive this condition in an explicit form. The unit tangent vector to trajectory *AO* is

$$u^i_{AO} = (u^0_{AO}, u^1_{AO}, u^2_{AO}, u^3_{AO}) = \left(\frac{dt}{d\tau_1}, \frac{dr_1}{cd\tau_1}, 0, 0\right)_{t \le t^*}, \quad (44)$$

and the unit tangent vector to trajectory *CO* is

$$u^i_{CO} = (u^0_{CO}, u^1_{CO}, u^2_{CO}, u^3_{CO}) = \left(\frac{dt}{d\tau_2}, \frac{dr_2}{cd\tau_2}, 0, 0\right)_{t \le t^*}. \quad (45)$$

The fact that these are actually unit vectors follows from the joining equations (33) and (36).

The components of vector (44) can be easily expressed from Eqs. (30) and (33) as

$$\left(\frac{dt}{d\tau_1}\right)_{t \le t^*} = \frac{1}{f_{12}(r_1)} e^{-T_1(t)/2}\left[\frac{m_{12} - m_{\text{in}}}{M_1(r_1)} - \frac{kM_1(r_1)}{2c^2 r_1}\right], \quad (46)$$

$$\left(\frac{dr_1}{cd\tau_1}\right)_{t \le t^*} = \delta_1 \sqrt{\left[\frac{m_{12} - m_{\text{in}}}{M_1(r_1)} - \frac{kM_1(r_1)}{2c^2 r_1}\right]^2 - f_{12}(r_1)}, \quad (47)$$

where

$$\delta_1 = \operatorname{sgn}\left(\frac{dr_1}{cd\tau_1}\right)_{t \le t^*}. \quad (48)$$

For the components of vector (45), we obtain the following expressions from Eqs. (31) and (36):

$$\left(\frac{dt}{d\tau_2}\right)_{t \le t^*} = \frac{1}{f_{12}(r_2)} e^{-T_1(t)/2}\left[\frac{m_{\text{out}} - m_{12}}{M_2(r_2)} + \frac{kM_2(r_2)}{2c^2 r_2}\right], \quad (49)$$

$$\left(\frac{dr_2}{cd\tau_2}\right)_{t \le t^*} = \delta_2 \sqrt{\left[\frac{m_{\text{out}} - m_{12}}{M_2(r_2)} + \frac{kM_2(r_2)}{2c^2 r_2}\right]^2 - f_{12}(r_2)}, \quad (50)$$

$$\delta_2 = \operatorname{sgn}\left(\frac{dr_2}{cd\tau_2}\right)_{t \le t^*}. \quad (51)$$

Using the metric (26) in region *COA*, we now need to calculate the quantity

$$Q = \{g^{(COA)}_{00} u^0_{AO} u^0_{CO} + g^{(COA)}_{11} u^1_{AO} u^1_{CO}\}_{t = t^*, r = r_1 = r_2 = r^*}. \quad (52)$$

From the preceding results, we obtain

$$Q = \frac{1}{f_{12}(r^*)}\left\{-\left[\frac{m_{12}-m_{\text{in}}}{M_1(r^*)}-\frac{kM_1(r^*)}{2c^2r^*}\right]\times\left[\frac{m_{\text{out}}-m_{12}}{M_2(r^*)}+\frac{kM_2(r^*)}{2c^2r^*}\right]+\delta_1\delta_2\sqrt{\left[\frac{m_{12}-m_{\text{in}}}{M_1(r^*)}-\frac{kM_1(r^*)}{2c^2r^*}\right]^2-f_{12}(r^*)}\times\sqrt{\left[\frac{m_{\text{out}}-m_{12}}{M_2(r^*)}+\frac{kM_2(r^*)}{2c^2r^*}\right]^2-f_{12}(r^*)}\right\}. \quad (53)$$

Let us now turn to the region after the intersection time. For the unit tangent vectors to trajectories *OB* and *OD*, we have

$$u^i_{OB} = (u^0_{OB}, u^1_{OB}, u^2_{OB}, u^3_{OB}) = \left(\frac{dt}{d\tau_1}, \frac{dr_1}{cd\tau_1}, 0, 0\right)_{t\geq t^*}, \quad (54)$$

$$u^i_{OD} = (u^0_{OD}, u^1_{OD}, u^2_{OD}, u^3_{OD}) = \left(\frac{dt}{d\tau_2}, \frac{dr_2}{cd\tau_2}, 0, 0\right)_{t\geq t^*}; \quad (55)$$

we see from the joining conditions (41) and (42) that these are actually unit vectors. For the components of vector (54), we obtain expressions from Eqs. (38) and (41) similar to (46) and (47) with the substitutions $f_{21}(r_1)$ for $f_{12}(r_1)$, m_{21} for m_{12}, m_{out} for m_{in}, $T_2(t)$ for $T_1(t)$, δ_1' for δ_1, $[-M_1(r_1)]$ for $M_1(r_1)$, and $t \geq t^*$ for $t \leq t^*$. The components of vector (55) follow from Eqs. (39) and (42); the expressions for them can be derived from (49) and (50) by substituting $f_{21}(r_2)$ for $f_{12}(r_2)$, m_{21} for m_{12}, m_{in} for m_{out}, $T_2(t)$ for $T_1(t)$, δ_2' for δ_2, $[-M_2(r_2)]$ for $M_2(r_2)$, and $t \geq t^*$ for $t \leq t^*$.

Accordingly, δ_1' and δ_2' are defined as in (48) and (51) but for $t \geq t^*$. Using the metric (28) in region *BOD*, let us calculate the scalar product

$$Q' = \{g^{(BOD)}_{00}u^0_{OB}u^0_{OD} + g^{(BOD)}_{11}u^1_{OB}u^1_{OD}\}_{t=t^*,\, r=r_1=r_2=r^*}. \quad (56)$$

We easily obtain

$$Q' = \frac{1}{f_{21}(r^*)}\left\{-\left[\frac{m_{\text{out}}-m_{21}}{M_1(r^*)}+\frac{kM_1(r^*)}{2c^2r^*}\right]\times\left[\frac{m_{21}-m_{\text{in}}}{M_2(r^*)}-\frac{kM_2(r^*)}{2c^2r^*}\right]+\delta_1'\delta_2'\sqrt{\left[\frac{m_{21}-m_{\text{in}}}{M_2(r^*)}-\frac{kM_2(r^*)}{2c^2r^*}\right]^2-f_{21}(r^*)}\times\sqrt{\left[\frac{m_{\text{out}}-m_{21}}{M_1(r^*)}+\frac{kM_1(r^*)}{2c^2r^*}\right]^2-f_{21}(r^*)}\right\}. \quad (57)$$

The necessary continuity condition is now the requirement that

$$Q = Q', \quad (58)$$

where Q and Q' are given by (53) and (57), respectively. Since the intersection coordinate r^* is assumed to be known, condition (58) is the equation for determining the parameter m_{21}.

We began the derivation of Eq. (58) by calling it the continuity condition for the relative velocity of the shells at point (r^*, t^*) in advance. Let us explain this in more detail. We define the "physical" shell velocities v_1 and v_2 in region *COA* as

$$\frac{v_1^2}{c^2} = \frac{g^{(COA)}_{11}(r_1)dr_1^2}{-g^{(COA)}_{00}(r_1)c^2dt^2}, \qquad \frac{v_2^2}{c^2} = \frac{g^{(COA)}_{11}(r_2)dr_2^2}{-g^{(COA)}_{00}(r_2)c^2dt^2}. \quad (59)$$

Using the joining equations (33) and (36), we then obtain

$$g^{(COA)}_{11}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = \frac{v_1^2/c^2}{1-v_1^2/c^2}, \qquad g^{(COA)}_{00}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 = -\frac{1}{1-v_1^2/c^2}, \quad (60)$$

$$g^{(COA)}_{11}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = \frac{v_2^2/c^2}{1-v_2^2/c^2}, \qquad g^{(COA)}_{00}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 = -\frac{1}{1-v_2^2/c^2}. \quad (61)$$

It now follows from (44), (45), and (52) that

$$Q = \left\{\frac{v_1v_2/c^2-1}{\sqrt{1-v_1^2/c^2}\sqrt{1-v_2^2/c^2}}\right\}_{t=t^*,\, r_1=r_2=r^*}. \quad (62)$$

The "physical" velocities of the shells after their intersection, v_1' and v_2', are defined similarly:

$$\frac{v_1'^2}{c^2} = \frac{g^{(BOD)}_{11}(r_1)dr_1^2}{-g^{(BOD)}_{00}(r_1)c^2dt^2}, \qquad \frac{v_2'^2}{c^2} = \frac{g^{(BOD)}_{11}(r_2)dr_2^2}{-g^{(BOD)}_{00}(r_2)c^2dt^2}. \quad (63)$$

We then find from the joining equations (41) and (42) that

$$g_{11}^{BOD}(r_1)\left(\frac{dr_1}{cd\tau_1}\right)^2 = \frac{v_1'^2/c^2}{1-v_1'^2/c^2},$$
$$g_{00}^{BOD}(r_1)\left(\frac{dt}{d\tau_1}\right)^2 = -\frac{1}{1-v_1'^2/c^2}, \quad (64)$$

$$g_{11}^{BOD}(r_2)\left(\frac{dr_2}{cd\tau_2}\right)^2 = \frac{v_2'^2/c^2}{1-v_2'^2/c^2},$$
$$g_{00}^{BOD}(r_2)\left(\frac{dt}{d\tau_2}\right)^2 = -\frac{1}{1-v_2'^2/c^2}. \quad (65)$$

It follows from (54)–(56) that

$$Q' = \left\{\frac{v_1'v_2'/c^2-1}{\sqrt{1-v_1'^2/c^2}\sqrt{1-v_2'^2/c^2}}\right\}_{t=t^*,\, r_1=r_2=r^*}. \quad (66)$$

Naturally, the continuity of Q when passing through the point of intersection (r^*, t^*) also implies the continuity of any function of Q, in particular, of

$$\frac{\sqrt{Q^2-1}}{Q} = \frac{v_1/c - v_2/c}{1-v_1v_2/c^2}. \quad (67)$$

The latter is nothing else but the relative velocity of two "particles," as it is defined in relativistic mechanics. This gives grounds to call condition (58) the continuity condition for the relative velocity of the shells.

To work with Eq. (58), it is convenient first to simplify the form of expressions (53) and (57) for Q and Q' by denoting

$$\sigma_1 = \left[\frac{kM_1(r^*)}{c^2r^*}\right]^2, \quad \sigma_2 = \left[\frac{kM_2(r^*)}{c^2r^*}\right]^2 \quad (68)$$

and by expressing the differences between the mass parameters in Q and Q' using formulas (4) and (29) in terms of the differences between the functions $f(r)$ taken at point $r = r^*$ in accordance with the relation

$$m_a - m_b = \frac{c^2r^*}{2k}[f_b(r^*) - f_a(r^*)], \quad (69)$$

where a and b mean the indices "in," "out," 12, and 21. Now, Q and Q' can be represented as

$$Q = -(4\sqrt{\sigma_1\sigma_2}f_{12})^{-1} \times [(f_{\text{in}} - f_{12} - \sigma_1)(f_{12} - f_{\text{out}} + \sigma_2) - \delta_1\delta_2\sqrt{(f_{\text{in}} - f_{12} - \sigma_1)^2 - 4\sigma_1 f_{12}} \times \sqrt{(f_{12} - f_{\text{out}} + \sigma_2)^2 - 4\sigma_2 f_{12}}], \quad (70)$$

$$Q' = -(4\sqrt{\sigma_1\sigma_2}f_{21})^{-1} \times [(f_{21} - f_{\text{out}} + \sigma_1)(f_{\text{in}} - f_{21} - \sigma_2) - \delta_1'\delta_2'\sqrt{(f_{21} - f_{\text{out}} + \sigma_1)^2 - 4\sigma_1 f_{21}} \times \sqrt{(f_{\text{in}} - f_{21} - \sigma_2)^2 - 4\sigma_2 f_{21}}], \quad (71)$$

where we omitted the argument r^* in the functions f_a to save space. In all the subsequent formulas given below, the quantities f_a without any argument will mean their values at $r = r^*$.

If the equation $Q = Q'$ is written using expressions (70) and (71) and if the first term from Q' is carried over to the left-hand side of this equation, then squaring the resulting relation yields a quadratic equation for the unknown f_{21}, i.e., for the parameter m_{21}. This procedure is somewhat cumbersome, and, in addition, it does not answer the question of which of the two solutions should be chosen. However, there is a method not only to easily find a solution to the equation $Q = Q'$ in a simple form but also to find this solution unambiguously and directly from the continuity equations without resorting to any additional physical considerations. The point is that, apart from Eq. (58), there are other similar continuity requirements when passing through the point of intersection of the shells; their use makes it easier to find the solution and singles it out unequivocally. As we saw, Eq. (58) was derived by requiring that the scalar product of the unit 4-velocity vectors for the shells be continuous when passing through the point of their intersection from sector *COA* (before intersection) into sector *BOD* (after intersection). It is easy to see, however, that all the reasoning that accompanied the derivation of Eq. (58) is also equally applicable to the passage through the point of intersection from sector *AOD* into sector *COB*; i.e., we may also assert that the limit of the scalar product of the unit tangent vectors to trajectories *AO* and *OD* when point *O* is approached from sector *AOD* must be equal to the limit of the scalar product of the unit tangent vectors to trajectories *OB* and *CO* when point *O* is approached from sector *COB*. This continuity condition can be easily obtained in an explicit form by repeating a procedure similar to the procedure that led to Eq. (58). We first calculate the scalar product

$$P = \{g_{00}^{(AOD)}u_{AO}^0u_{OD}^0 + g_{11}^{(AOD)}u_{AO}^1u_{OD}^1\}_{t=t^*,\, r=r_1=r_2=r^*} \quad (72)$$

using the equations of motion for shells *AO* and *OD* in the form (18) and the joining conditions (34) and (43) and then calculate the scalar product

$$P' = \{g_{00}^{(COB)}u_{CO}^0u_{OB}^0 + g_{11}^{(COB)}u_{CO}^1u_{OB}^1\}_{t=t^*,\, r=r_1=r_2=r^*} \quad (73)$$

using the equations of motion for shells *CO* and *OB* in the form (19) and the joining conditions (35) and (40). Subsequently, we equate the results:

$$P' = P. \tag{74}$$

In the same notation that we used to derive expressions (70) and (71), the quantities P and P' are

$$\begin{aligned} P &= -(4\sqrt{\sigma_1\sigma_2}f_{\text{in}})^{-1} \\ &\times[(f_{\text{in}}-f_{12}+\sigma_1)(f_{\text{in}}-f_{21}+\sigma_2) \\ &-\delta_1\delta_2'\sqrt{(f_{\text{in}}-f_{12}+\sigma_1)^2-4\sigma_1 f_{\text{in}}} \\ &\times\sqrt{(f_{\text{in}}-f_{21}+\sigma_2)^2-4\sigma_2 f_{\text{in}}}], \end{aligned} \tag{75}$$

$$\begin{aligned} P' &= -(4\sqrt{\sigma_1\sigma_2}f_{\text{out}})^{-1} \\ &\times[(f_{21}-f_{\text{out}}-\sigma_1)(f_{12}-f_{\text{out}}-\sigma_2) \\ &-\delta_1'\delta_2\sqrt{(f_{21}-f_{\text{out}}-\sigma_1)^2-4\sigma_1 f_{\text{out}}} \\ &\times\sqrt{(f_{12}-f_{\text{out}}-\sigma_2)^2-4\sigma_2 f_{\text{out}}}]. \end{aligned} \tag{76}$$

In addition, it is clear that all that we have said above about the continuity conditions at the point of intersection is not restricted only to the passages from *COA* into *BOD* and from *AOD* into *COB* but is equally applicable to the passages in general from any sector into any other sector. This implies that all four quantities Q, Q', P, and P' must actually be identical and, in addition to conditions (58) and (74), we must require that one more equation, for example, $P = Q$, be satisfied. Thus, the complete set of continuity conditions at the point of intersection can be written as

$$Q = Q', \quad Q = P, \quad Q = P'. \tag{77}$$

It turns out that these three quadratic equations for one unknown f_{21} actually have one single common root. This root can be easily determined[3] from the first two equations from (77) and is

$$\begin{aligned} f_{21} &= f_{\text{out}}+f_{\text{in}}-f_{12}-(2f_{12})^{-1} \\ &\times[(f_{\text{in}}-f_{12}-\sigma_1)(f_{12}-f_{\text{out}}+\sigma_2) \\ &-\delta_1\delta_2\sqrt{(f_{\text{in}}-f_{12}-\sigma_1)^2-4\sigma_1 f_{12}} \\ &\times\sqrt{(f_{12}-f_{\text{out}}+\sigma_2)^2-4\sigma_2 f_{12}}]. \end{aligned} \tag{78}$$

[3] This requires expressing the components of the unit 4-velocity vectors in terms of hyperbolic functions. Thus, for example, if we introduce the angles α_1 and α_2 according to $\delta_1\sqrt{(f_{\text{in}}-f_{12}-\sigma_1)^2-4\sigma_1 f_{12}} = \sqrt{4\sigma_1 f_{12}}\sinh\alpha_1$ and $\delta_2\sqrt{(f_{12}-f_{\text{out}}+\sigma_2)^2-4\sigma_2 f_{12}} = \sqrt{4\sigma_2 f_{12}}\sinh\alpha_2$, then $Q = -\cosh(\alpha_1-\alpha_2)$. Similar notation should be introduced for the quantities from which the scalar products P, P', and Q' are made up. Equation (77) can then be easily resolved for the "angles" that contain the unknown f_{21}.

The third equation from (77) after substituting the expression for f_{21} in it is satisfied identically.

Formula (78) solves the problem of determining the mass parameter m_{21} from the quantities specified at the evolutionary stage before intersection. The energy transfer between the shells as they intersect is determined along with it. Indeed, the total conserved energies of the shells before their intersection are defined as

$$E_1 = (m_{12}-m_{\text{in}})c^2, \quad E_2 = (m_{\text{out}}-m_{12})c^2, \tag{79}$$

where E_1 is the energy of (inner) shell 1 and E_2 is the energy of (outer) shell 2. During the intersection, the energy of each shell undergoes a discontinuity; subsequently, their energies become equal to E_1' and E_2', respectively:

$$E_1' = (m_{\text{out}}-m_{21})c^2, \quad E_2' = (m_{21}-m_{\text{in}})c^2. \tag{80}$$

The conservation of total energy of the shells when passing through the point of intersection automatically follows from these expressions:

$$E_1+E_2 = E_1'+E_2'. \tag{81}$$

Using this relation, we can write the energies of the shells after their intersection as

$$E_1' = E_1-\Delta E, \quad E_2' = E_2+\Delta E. \tag{82}$$

It follows from (80), (81), and (69) that

$$\Delta E = \frac{c^4 r^*}{2k}(f_{\text{in}}+f_{\text{out}}-f_{12}-f_{21}). \tag{83}$$

Substituting the solution (78) for f_{21} in this expression yields

$$\begin{aligned} \Delta E &= \frac{c^4 r^*}{4kf_{12}}[(f_{\text{in}}-f_{12}-\sigma_1)(f_{12}-f_{\text{out}}+\sigma_2) \\ &-\delta_1\delta_2\sqrt{(f_{\text{in}}-f_{12}-\sigma_1)^2-4\sigma_1 f_{12}} \\ &\times\sqrt{(f_{12}-f_{\text{out}}+\sigma_2)^2-4\sigma_2 f_{12}}]. \end{aligned} \tag{84}$$

Note that the square bracket multiplied by the factor f_{12}^{-1} in formulas (78) and (84) can be easily expressed in terms of the scalar product Q [see formula (70)], which, in turn, can be easily expressed in terms of the shell velocities v_1 and v_2 at the point of intersection [relation (80)]. This representation of ΔE is

$$\begin{aligned} \Delta E &= -\frac{kM_1(r^*)M_2(r^*)}{r^*}Q \\ &= \left(\frac{kM_1M_2}{r}\frac{1-v_1v_2/c^2}{\sqrt{1-v_1^2/c^2}\sqrt{1-v_2^2/c^2}}\right)_{r=r^*}, \end{aligned} \tag{85}$$

and it is convenient in some cases (in particular, for obtaining the nonrelativistic approximation). The cor-

responding expression for the metric coefficient f_{21} can be written as

$$f_{21} = f_{\text{out}} + f_{\text{in}} - f_{12} - \frac{2k}{c^4 r^*}\Delta E. \quad (86)$$

4. THE NONRELATIVISTIC (NEWTONIAN) APPROXIMATION

In the nonrelativistic approximation, the total energies of the shells (79) and (80) can be expanded as

$$E = mc^2 + \mathscr{E} + o(1/c^2),$$

where m and $\mathscr{E}$ do not depend on c. This means that the differences between the mass parameters, to within $1/c^2$ inclusive, are

$$m_{12} - m_{\text{in}} = m_1 + \frac{\mathscr{E}_1}{c^2}, \quad m_{\text{out}} - m_{12} = m_2 + \frac{\mathscr{E}_2}{c^2}, \quad (87)$$

$$m_{\text{out}} - m_{21} = m_1 + \frac{\mathscr{E}'_1}{c^2}, \quad m_{21} - m_{\text{in}} = m_2 + \frac{\mathscr{E}'_2}{c^2}. \quad (88)$$

Here, m_1 and m_2 are the rest masses of the shells, those that appear in formulas (32) for the effective masses M_1 and M_2. The quantities $\mathscr{E}$ in (87) and (88) are the total nonrelativistic energies of the shells; the law of their conservation follows from (87) and (88):

$$\mathscr{E}_1 + \mathscr{E}_2 = \mathscr{E}'_1 + \mathscr{E}'_2. \quad (89)$$

Relation (82) now takes the form

$$\mathscr{E}'_1 = \mathscr{E}_1 - \Delta\mathscr{E}, \quad \mathscr{E}'_2 = \mathscr{E}_2 + \Delta\mathscr{E}, \quad (90)$$

where $\Delta\mathscr{E} = (\Delta E)_{c=\infty}$ follows from expression (85). Since $M_1 = m_1$ and $M_2 = m_2$ in the first nonvanishing order in $1/c^2$, we obtain the following nonrelativistic formulas for energy transfer during the intersection from (85):

$$\Delta\mathscr{E} = \frac{km_1m_2}{r^*}. \quad (91)$$

The nonrelativistic equations of motion for the shells before their intersection can be easily derived from the exact equations (30)–(32). The proper times τ_1 and τ_2 in the principal order are equal to the global time t, so $(dr_1/cd\tau_1)^2$ and $(dr_2/cd\tau_2)^2$ in these equations are nothing else but v_1^2/c^2 and v_2^2/c^2, respectively, where

$$v_1 = \frac{dr_1}{dt}, \quad v_2 = \frac{dr_2}{dt}. \quad (92)$$

Expanding Eqs. (30)–(32) up to the order $1/c^2$ inclusive yields

$$\mathscr{E}_1 = \frac{m_1v_1^2}{2} - \frac{km_1(m_{\text{in}} + m_1/2)}{r_1} + \frac{L_1^2}{2r_1^2m_1}, \quad (93)$$

$$\mathscr{E}_2 = \frac{m_2v_2^2}{2} - \frac{km_2(m_{\text{in}} + m_2/2 + m_1)}{r_2} + \frac{L_2^2}{2r_2^2m_2}. \quad (94)$$

A similar operation with Eqs. (38) and (39) leads to the following equations of motion for the shells after their intersection:

$$\mathscr{E}'_1 = \frac{m_1v_1^2}{2} - \frac{km_1(m_{\text{in}} + m_1/2 + m_2)}{r_1} + \frac{L_1^2}{2r_1^2m_1}, \quad (95)$$

$$\mathscr{E}'_2 = \frac{m_2v_2^2}{2} - \frac{km_2(m_{\text{in}} + m_2/2)}{r_2} + \frac{L_2^2}{2r_2^2m_2}. \quad (96)$$

Together with relations (90) and (91) and the initial data $r_1 = r_2 = r^*$ at $t = t^*$, Eqs. (95) and (96) completely determine the evolution of the shells immediately after t^* (i.e., before the next possible intersection).

It follows from Eqs. (93)–(96) applied to the point $r_1 = r_2 = r^*$ that

$$\begin{aligned}\mathscr{E}_1 - \mathscr{E}'_1 &= \frac{m_1}{2}[v_1^2(r^*) - v_1'^2(r^*)] + \frac{km_1m_2}{r^*},\\ \mathscr{E}_2 - \mathscr{E}'_2 &= \frac{m_2}{2}[v_2^2(r^*) - v_2'^2(r^*)] - \frac{km_1m_2}{r^*}.\end{aligned} \quad (97)$$

Here, we denoted the limits of the velocities $v_{1,2}$ at point r^* from the side "after intersection" by $v'_{1,2}(r^*)$, while $v_{1,2}(r^*)$ are the limits of these velocities at point r^* from the side "before intersection." Relations (97) and the law of change in energy (90) and (91) show that the shell velocities are continuous at the point of intersection in the nonrelativistic approximation:

$$v'_1(r^*) = v_1(r^*), \quad v'_2(r^*) = v_2(r^*) \quad (98)$$

(the cases where the velocities change sign are excluded from the additional physical requirements for joining the evolutions discussed in the Introduction). Of course, condition (98) is obvious in advance and requires no discussion. We deduced it here only to show the consistency of the entire procedure. Note that, when studying the Newtonian approximation, we can begin directly from condition (98) as the main postulate (as we did in [8]).

The exact formula (84) for ΔE written in terms of the velocities in the form (85) also allows the post-Newtonian correction to the energy transfer to be easily

calculated. Expanding the right-hand side of Eq. (85) to within order $1/c^2$ inclusive yields

$$\Delta E = \frac{km_1m_2}{r^*} + \frac{1}{2c^2}\left\{\frac{km_1m_2}{r^*}[v_1(r^*) - v_2(r^*)]^2 + \frac{km_2L_1^2}{m_1r^{*3}} + \frac{km_1L_2^2}{m_2r^{*3}}\right\}. \quad (99)$$

The velocities $v_1(r^*)$ and $v_2(r^*)$ in this formula should be determined from Eqs. (93) and (94) applied to the point $r_1 = r_2 = r^*$; i.e., these are the standard velocities of the Newtonian approximation. Of course, if we take into consideration the post-Newtonian correction to ΔE, then we must take into account such corrections to all energies E; i.e., we must write out the next terms of the expansion in $1/c^2$ in energies after the Newtonian ones (93)–(96). We performed the corresponding calculations but do not present their results here, because they are relatively cumbersome and are not needed at the current stage of our studies.

5. SHELLS WITH ZERO EFFECTIVE REST MASSES

If the shell particles move only radially and have a zero rest mass, then $m_1 = m_2 = 0$ and $L_1 = L_2 = 0$. In this case, both shells have a zero effective rest mass, i.e., $M_1 = M_2 = 0$. As previously, the shell intersection corresponds to Fig. 1, with the only difference that trajectories *CO*, *AO*, *OB*, and *OD* are now isotropic. In all four sectors, the metric is given by the same formulas (25)–(28), in which the functions f have the same form (4), (29). As previously, the problem consists in determining the radius r^* of the point of intersection of the shells and then the mass parameter m_{21} or, equivalently, $f_{21}(r^*)$. Naturally, the latter must follow from formula (78), in which we should set $\sigma_1 = \sigma_2 = 0$ and assume that $\delta_1\delta_2 = -1$, because such *s*-wave lightlike shells can intersect only if they initially move toward each other. Given also that the mass parameters in the physical region are arranged in the order

$$m_{\text{out}} > m_{12} > m_{\text{in}}, \quad m_{\text{out}} > m_{21} > m_{\text{in}}, \quad (100)$$

or, equivalently,

$$f_{\text{out}} < f_{12} < f_{\text{in}}, \quad f_{\text{out}} < f_{21} < f_{\text{in}}, \quad (101)$$

we obtain the following result from (78):

$$f_{21}(r^*)f_{12}(r^*) = f_{\text{in}}(r^*)f_{\text{out}}(r^*). \quad (102)$$

This is nothing else but the relation derived by Dray and 't Hooft [17]; these authors considered the intersection of two light spherically symmetric shells with a purely radial motion of their constituent "photons."

To completely describe the behavior of light shells, we must also set up the equations of their motion. Now, the proper time τ cannot be used in these equations, because it does not exist. If there is only one shell, then expressions (1) and (2) for the intervals remain the same, while expression (3) changes to

$$-(ds^2)_{\text{on}} = R_0^2(t)d\Omega^2. \quad (103)$$

Thus, the joining conditions now imply that the radial-time part of the interval must become zero on both sides of the shells; i.e., instead of (6) and (7), we obtain

$$f_{\text{in}}(R_0)e^{T(t)} - f_{\text{in}}^{-1}(R_0)\left(\frac{dR_0}{cdt}\right)^2 = 0, \quad (104)$$

$$f_{\text{out}}(R_0) - f_{\text{out}}^{-1}(R_0)\left(\frac{dR_0}{cdt}\right)^2 = 0. \quad (105)$$

The equation of motion for the shell $r = R_0(t)$ follows from (105),

$$\left(\frac{dR_0}{cdt}\right)^2 = f_{\text{out}}^2(R_0), \quad (106)$$

and Eq. (104) then gives the dilaton factor $e^{T(t)}$:

$$e^{T(t)} = \frac{f_{\text{out}}^2(R_0)}{f_{\text{in}}^2(R_0)}. \quad (107)$$

It is easy to see that the equation of motion in the forms (18)–(20) is now no longer required, because it simply copies Eq. (106). This can be easily shown first (at $\mu \neq 0$) by changing to the global time t instead of τ in the equations and then passing to the limit $\mu \longrightarrow 0$.

In the case of two shells, the equations of motion for the second (outer) shell before their intersection follow the equality of the radial-time interval to zero on both of its sides:

$$f_{\text{out}}(R_2) - f_{\text{out}}^{-1}(R_2)\left(\frac{dR_2}{cdt}\right)^2 = 0, \quad (108)$$

$$e^{T_1(t)}f_{12}(R_2) - f_{12}^{-1}(R_2)\left(\frac{dR_2}{cdt}\right)^2 = 0, \quad (109)$$

whence it follows that

$$\left(\frac{dR_2}{cdt}\right)^2 = f_{\text{out}}^2(R_2), \quad e^{T_1(t)} = \frac{f_{\text{out}}^2(R_2)}{f_{12}^2(R_2)}. \quad (110)$$

For (inner) shell 1 before the intersection,

$$e^{T_1(t)}f_{12}(R_1) - f_{12}^{-1}(R_1)\left(\frac{dR_1}{cdt}\right)^2 = 0, \quad (111)$$

$$e^{T_0(t)}f_{\text{in}}(R_1) - f_{\text{in}}^{-1}(R_1)\left(\frac{dR_1}{cdt}\right)^2 = 0. \quad (112)$$

Substituting e^{T_1} from (110) in (111) yields

$$\left(\frac{dR_1}{cdt}\right)^2 = \frac{f_{\text{out}}^2(R_2) f_{12}^2(R_1)}{f_{12}^2(R_2)},$$
$$e^{T_0(t)} = \frac{f_{\text{out}}^2(R_2) f_{12}^2(R_1)}{f_{12}^2(R_2) f_{\text{in}}^2(R_1)}. \quad (113)$$

After the intersection, we have for (now outer) shell 1

$$f_{\text{out}}(R_1) - f_{\text{out}}^{-1}(R_1)\left(\frac{dR_1}{cdt}\right)^2 = 0, \quad (114)$$

$$e^{T_2(t)} f_{21}(R_1) - f_{21}^{-1}(R_1)\left(\frac{dR_1}{cdt}\right)^2 = 0, \quad (115)$$

whence it follows that

$$\left(\frac{dR_1}{cdt}\right)^2 = f_{\text{out}}^2(R_1), \quad e^{T_2(t)} = \frac{f_{\text{out}}^2(R_1)}{f_{21}^2(R_1)}. \quad (116)$$

For (inner) shell 2 after the intersection,

$$e^{T_2(t)} f_{21}(R_2) - f_{21}^{-1}(R_2)\left(\frac{dR_2}{cdt}\right)^2 = 0, \quad (117)$$

$$e^{T_0(t)} f_{\text{in}}(R_2) - f_{\text{in}}^{-1}(R_2)\left(\frac{dR_2}{cdt}\right)^2 = 0. \quad (118)$$

Substituting e^{T_2} from (116) into (117) yields

$$\left(\frac{dR_2}{cdt}\right)^2 = \frac{f_{\text{out}}^2(R_1) f_{21}^2(R_2)}{f_{21}^2(R_1)},$$
$$e^{T_0(t)} = \frac{f_{\text{out}}^2(R_1) f_{21}^2(R_2)}{f_{21}^2(R_1) f_{\text{in}}^2(R_2)}. \quad (119)$$

If the initial data to the first of Eqs. (110) and to the first of Eqs. (113) are specified and if the mass parameters m_{12}, m_{in}, and m_{out} are also specified, then the trajectories $r = R_1(t)$ and $r = R_2(t)$ of the shells before their intersection are completely determined together with the point of their intersection $r^* = R_1(t^*) = R_2(t^*)$. The mass parameter m_{21} is then derived from (102), and the trajectories of the shells after their intersection are completely determined from the first Eqs. (116) and (119).

The energy transfer can be easily calculated from (83) by substituting $f_{21}(r^*)$ expressed from relation (102):

$$\Delta E = \frac{2k(m_{12} - m_{\text{in}})(m_{\text{out}} - m_{12})}{r^* f_{12}(r^*)}. \quad (120)$$

Note also that, if we took $\delta_1\delta_2 = 1$, then we would obtain

$$f_{21}(r^*) = f_{\text{in}}(r^*) + f_{\text{out}}(r^*) - f_{12}(r^*), \quad (121)$$

and the energy transfer ΔE would become zero, reflecting the fact that such light shells moving in the same direction cannot intersect. Indeed, a simple examination of Eqs. (110) and (113) shows that no intersection is possible for $\delta_1\delta_2 = 1$, which, of course, is obvious in advance. At the same time, the existence of solution (121) and, hence, region 21 after the intersection stems in this case from the fact that the exact solution for zero effective masses may be treated as the first approximation for massive shells but in the ultrarelativistic regime, when the effective rest masses M_1 and M_2 are insignificant and the solution can be expanded in small parameters σ_1 and σ_2. In this case, the intersection is possible even if the shells move in the same direction and the energy transfer in the first nonvanishing order is a small quantity linear in σ_1 and σ_2, i.e., in M_1^2 and M_2^2. Our calculation yields

$$\Delta E = \frac{kM_1^2(m_{\text{out}} - m_{12})^2 + kM_2^2(m_{12} - m_{\text{in}})^2}{2r^*(m_{12} - m_{\text{in}})(m_{\text{out}} - m_{12})}, \quad \delta_1\delta_2 = 1; \quad (122)$$

the solution (121) for $f_{21}(r^*)$ refers precisely to this situation.

In conclusion, note that, if the photons that constitute the shells move nonradially, then the parameters L_1 and L_2 are nonzero at zero m_1 and m_2. In this case, the effective rest masses M_1 and M_2 are nonzero and the qualitative behavior of such light shells is the same as that of massive shells.

6. THE INTERSECTION OF A TEST SHELL WITH A GRAVITATING SHELL

Clearly, the motion and intersections of two test shells are of no interest. Each of them moves as if no other shell exists at all. A nontrivial situation arises when only one of the shells is a test one, while the gravitational field of the other shell is completely taken into account. Naturally, the solution of this problem must follow from the general case using the corresponding passage to the limit. Let us first consider how the passage to the limit of a test shell is accomplished when there is only one shell. To obtain this limit, we must redesignate the constant parameters as

$$m = \lambda m_p, \quad L = \lambda L_p, \quad m_{\text{out}} = m_{\text{in}} + \lambda\frac{E_p}{c^2}, \quad (123)$$

and assume the constants m_p, L_p, and E_p to be λ-independent. Subsequently, we must pass to the limit $\lambda \longrightarrow 0$.

In the limit $\lambda \longrightarrow 0$, we have $m_{\text{out}} = m_{\text{in}}$ and $e^{T(t)} = 1$ follows from the joining equations (6) and (7). We now see from (1) and (2) that the metric is the same both

inside and outside the shell (as should be the case if it is a test shell):

$$-ds^2 = f_{\text{in}}(r)c^2dt^2 + f_{\text{in}}^{-1}(r)dr^2 + r^2d\Omega^2. \quad (124)$$

Only one relation now remains from the joining conditions:

$$f_{\text{in}}(r_0)\left(\frac{dt}{d\tau}\right)^2 - f_{\text{in}}^{-1}(r_0)\left(\frac{dr_0}{cd\tau}\right)^2 = 1, \quad (125)$$

which no longer makes sense to call the joining condition. This is simply the condition for normalizing the 4-velocity of a test particle to unity.

We write the effective mass $\mu(\tau)$ as

$$\mu(\tau) = \lambda\mu_p(\tau), \quad (126)$$

and relation (24) then defines μ_p:

$$\mu_p(\tau) = \sqrt{m_p^2 + \frac{L_p^2}{c^2r_0^2(\tau)}}. \quad (127)$$

Substituting (126) and (127) into Eq. (18) and passing to the limit $\lambda \longrightarrow 0$ yields the following equation of motion for the test shell in the field of the central mass m_{in}:

$$\frac{E_p}{c^2} = \mu_p(\tau)\sqrt{f_{\text{in}}(r_0) + \left(\frac{dr_0}{cd\tau}\right)^2}. \quad (128)$$

As we see, the zero parameters disappeared from the final equations (124), (125) and (127), (128), and only the finite constants m_p, L_p, m_{in}, and E_p remained; the latter characterize the test shell.

It would be natural to expect that the equation of motion for the test shell must match the equation of a geodesic in the field of a central mass. It turns out that this is actually the case. Using relation (125), we can easily write the equation of motion (128) in the Schwarzschild time t rather than in the proper time:

$$\frac{E_p}{c^2} = \sqrt{m_p^2 + \frac{L_p^2}{c^2R_0^2(t)}}\sqrt{\frac{f_{\text{in}}^3(R_0)}{f_{\text{in}}^2(R_0) - \left(\frac{dR_0}{cdt}\right)^2}}. \quad (129)$$

This expression is now easy to compare with that following from the solution of the geodesic equations in the metric (124). For the test shell, Eq. (129) is the only integral of motion in the sense that there are no motions except radial one in the shell as a whole. Therefore, we deal with the match between (129) and the first integral of the geodesic equations that describes the radial part of the motion of the test particle (although the particle, of course, also has a tangential motion). This integral of the geodesic equations is known to exist. It is easy to show that it exactly matches (129) if we identify the total conserved particle energy with E_p, the particle rest mass with m_p, and the square of the norm of the conserved particle angular momentum with L_p^2.

Let us now consider the situation with two shells shown in Fig. 1 and assume that shell 2 (trajectories *CO* and *OD*) is a test shell. We leave the parameters of shell 1 unchanged and redesignate the parameters of shell 2 in accordance with (123):

$$m_2 = \lambda m_p, \quad L_2 = \lambda L_p, \quad E_2 = \lambda E_p, \quad E_2' = \lambda E_p'.$$

It thus follows that

$$M_2(r) = \lambda M_p(r),$$

where

$$M_p(r) = \sqrt{m_p^2 + \frac{L_p^2}{c^2r^2}}. \quad (130)$$

Relations (79) and (80) now yield

$$m_{12} = m_{\text{out}} - \lambda\frac{E_p}{c^2}, \quad m_{21} = m_{\text{in}} + \lambda\frac{E_p'}{c^2},$$

whence we see that the mass parameters on both sides of shell 2 before and after the intersection are equal in the limit $\lambda = 0$:

$$m_{12} = m_{\text{out}}, \quad m_{21} = m_{\text{in}}. \quad (131)$$

It follows from the joining conditions (35), (36) and (42), (43) in the limit $\lambda = 0$ that

$$e^{T_1} = 1, \quad e^{T_2} = e^{T_0}. \quad (132)$$

Thus, as we see from (25), (26) and (27), (28), the metrics in regions *COB* and *COA* are identical, as are the metrics in regions *AOD* and *BOD*. In other words, the metrics are continuous when passing through trajectories *CO* and *OD*, as should be the case for a test shell.

Substituting the redesignated parameters in Eqs. (30) and (31) and passing to the limit $\lambda = 0$ yields the following equations of motion for the shells before their intersection:

$$\sqrt{f_{\text{out}}(r_1) + \left(\frac{dr_1}{cd\tau_1}\right)^2} = \frac{m_{\text{out}} - m_{\text{in}}}{M_1(r_1)} - \frac{kM_1(r_1)}{2c^2r_1}, \quad (133)$$

$$\sqrt{f_{\text{out}}(r_2) + \left(\frac{dr_2}{cd\tau_2}\right)^2} = \frac{E_p}{c^2M_p(r_2)}. \quad (134)$$

Similarly, we derive the equations of motion after their intersection from (38) and (39):

$$\sqrt{f_{\text{in}}(r_1) + \left(\frac{dr_1}{cd\tau_1}\right)^2} = \frac{m_{\text{out}} - m_{\text{in}}}{M_1(r_1)} + \frac{kM_1(r_1)}{2c^2r_1}, \quad (135)$$

$$\sqrt{f_{\text{in}}(r_2) + \left(\frac{dr_2}{cd\tau_2}\right)^2} = \frac{E_p'}{c^2M_p(r_2)}. \quad (136)$$

The equations of motion (133) and (135) for the gravitating shell 1 are actually the same equation, only the former is written in the form (19), while the latter is written in the form (18); i.e., the entire evolution of shell 1 can be described only by one of these equations extended to both stages of motion, before and after the intersection. Thus, the gravitating shell moves without being affected by the test shell, as should be the case. The situation with the test shell 2 is different. The equations of its motion (134) and (136) are distinctly different. In addition, we must also define its energy E_p' after the intersection via the parameters specified before the intersection. The latter is accomplished by using Eq. (78), in which we must substitute the redesignated parameters and then pass to the limit $\lambda = 0$. This operation leads to the following:

$$E_p' = E_p + \frac{1}{2f_{\text{out}}}[(f_{\text{in}} - f_{\text{out}} - \sigma_1)E_p - \delta_1\delta_2\sqrt{(f_{\text{in}} - f_{\text{out}} - \sigma_1)^2 - 4\sigma_1 f_{\text{out}}} \times \sqrt{E_p^2 - f_{\text{out}}M_p^2c^4}], \quad (137)$$

where f_{in}, f_{out}, and M_p are also the values of these functions at the point of intersection $r = r^*$.

Equation (137) gives a jump in the energy of the test shell. If necessary, we can also determine the jump in its velocity. To avoid misunderstandings, we first note the following. For the gravitating shell 1, because the equations of its motion before and after the intersection are identical, the derivative $dr_1/d\tau_1$ is continuous at the point of intersection. In contrast, the velocity of this shell defined by formulas (60) and (64) is discontinuous, because

$$g_{11}^{COA}(r^*) \neq g_{11}^{BOD}(r^*).$$

Of course, this discontinuity is not physical and results only from different definitions of the velocity before and after the intersection: the velocity of the gravitating shell is defined with respect to the metric outside it (sector *COB*) before the intersection and with respect to the metric inside it (sector *BOD*) after the intersection. It is easy to verify that, if we continued to define the velocity of shell 1 after the intersection with respect to the metric outside it (i.e., in sector *COB*), then this velocity would be continuous at the point of intersection. The velocity of shell 1 defined everywhere with respect to the metric inside it would also be continuous. Since the gravitating shell does not feel the presence of the test shell, its intersection with the latter is not distinguished in any way and the change in the definition of any quantities at an undistinguished time of evolution, naturally, bears no relation to the physics of the process. Of course, there may be reasons why this strange definition of the velocities is, nevertheless, convenient, but this is another thing altogether. In our study, this was required only to relate the joining conditions at the point of intersection to the continuity of the relative velocity, which may not have been done. Calculating the scalar products Q, Q', P, and P' and their continuity conditions by no means require introducing any velocities.

The situation with the test shell with the same metric on both of its sides is different. The "physical" velocity of this shell with respect to this metric is unambiguously defined everywhere and cannot have fictitious discontinuities of the type described above. Therefore, the discontinuity in this velocity at the point of intersection is actually connected with physics of the process. Before the intersection, the velocity of the test shell is defined by formulas (61) and (26), in which we should also pass to the limit $\lambda = 0$. Using (134), we then obtain

$$\frac{M_pc^2}{\sqrt{1 - v_2^2/c^2}} = \frac{E_p}{\sqrt{f_{\text{out}}}}. \quad (138)$$

The same operations with (65), (28), and (136) yield

$$\frac{M_pc^2}{\sqrt{1 - v_2'^2/c^2}} = \frac{E_p'}{\sqrt{f_{\text{in}}}}. \quad (139)$$

In these formulas, as previously, all functions of r are taken at the point of intersection $r = r^*$. Since the jump in energy is known [relation (137)], the jump in the velocity of the test shell can be determined from (138) and (139).

7. MASS EJECTION FROM A STAR CLUSTER

The dynamical processes near supermassive black holes (SBHs), quasars, blazars, and active galactic nuclei are characterized by violent events that give rise to jets and ejections. The formation of jets is commonly associated with processes that take place in magnetized accretion disks [18, 19]. The formation of quasi-spherical ejections, which are possibly observed in broad absorption lines, may prove to be related to other ejection mechanisms. In this section, based on the ballistic interaction between gravitating shells described in the preceding sections, we point out the possibility of shell ejection from the neighborhood of a SBH surrounded by a dense massive star cluster.

Numerical calculations for the collapse of a star cluster in the shell approximation [4, 9, 10] showed that, even if all shells were initially bound, after several intersections, some of the shells acquire enough energy to become unbound and to fly away to infinity. The remnant can be a stationary star cluster in the Newtonian approximation and a SBH in general relativity.

Ejections can be produced by the interaction between shells moving near a SBH. In a homogeneous star cluster with or without a SBH, stars evaporate through pair collisions with modest kinetic energy transfer. The formation of rapidly escaping stars is approximately a factor of 100 less probable, because

collisions with weak momentum transfer prevail [20]. If the cluster is denser and contains several compact parts, then the collisions between these parts will be completely different; significant momentum transfer during a collision becomes possible. In this case, the gravitational interaction between compact parts can lead to high-velocity ejections, and if such an intersection takes place near a SBH, then the shell escape velocity from the cluster can account for an appreciable fraction of the speed of light c. Such a situation can arise through the collision of galaxies during a close encounter of their nuclei. In that case, one nucleus can pull part of the matter from the other nucleus in the form of collapsing shells. The interaction of such shells with cluster stars can lead not only to collapse onto the SBH but also to the reverse phenomenon: shell ejection with a velocity much higher than the fall velocity at a given radius; the shell will not fall to the SBH because of the large angular momentum of its stars.

The ejection mechanism manifests itself even for the interaction between two shells in the Newtonian approximation considered in Section 4. If two gravitationally bound shells with energies $\mathscr{E}_1 < 0$ and $\mathscr{E}_2 < 0$ that obey the equations of motion (93) and (94) intersect at point $r = r_1^*$, then their next intersection can occur at point $r = r_2^*$ farther from the center, i.e., at $r_2^* > r_1^*$. According to (90) and (91), the shell energies will be

$$\mathscr{E}_1' = \mathscr{E}_1 - \frac{km_1m_2}{r_1^*}, \quad \mathscr{E}_2' = \mathscr{E}_2 + \frac{km_1m_2}{r_1^*} \tag{140}$$

after the first intersection and

$$\begin{aligned}\mathscr{E}_1'' &= \mathscr{E}_1' + \frac{km_1m_2}{r_2^*} = \mathscr{E}_1 - km_1m_2\left(\frac{1}{r_1^*} - \frac{1}{r_2^*}\right),\\ \mathscr{E}_2'' &= \mathscr{E}_2' - \frac{km_1m_2}{r_2^*} = \mathscr{E}_2 + km_1m_2\left(\frac{1}{r_1^*} - \frac{1}{r_2^*}\right)\end{aligned} \tag{141}$$

after the second intersection. If the absolute values of $\mathscr{E}_1$ and $\mathscr{E}_2$ are sufficiently small and if r_1^* is moderately large and not too close to r_2^*, then $\mathscr{E}_2'' > 0$ and (outer) shell 2 after the intersection can go to infinity. Clearly, there is a broad class of such solutions, and one specific example (with the highest possible ejection velocity) is given in [8].

Naturally, this effect also remains in the relativistic theory of gravitation. If two gravitationally bound shells with energies $E_1 < m_1c^2$ and $E_2 < m_2c^2$ that move according to Eqs. (30)–(32) intersect at point $r = r_1^*$, then the energy transfer is described by formulas (82) and (84), (85):

$$\begin{aligned}E_1' &= E_1 - \frac{kM_1(r_1^*)M_2(r_1^*)}{r_1^*}(-Q),\\ E_2' &= E_2 + \frac{kM_1(r_1^*)M_2(r_1^*)}{r_1^*}(-Q).\end{aligned} \tag{142}$$

Here, Q is given by expression (70), in which all functions of r are taken at point r_1^*. For simplicity, let us consider only those cases where the second intersection occurs at $r_2^* > r_1^*$ but at such a large r_2^* that the Newtonian approximation can be used for estimates in this region. Thus, the shell energies after the second intersection will be

$$\begin{aligned}E_1'' &= E_1' + \frac{km_1m_2}{r_2^*}\\ &= E_1 - \left[\frac{kM_1(r_1^*)M_2(r_1^*)}{r_1^*}(-Q) - \frac{km_1m_2}{r_2^*}\right],\\ E_2'' &= E_2' - \frac{km_1m_2}{r_2^*}\\ &= E_2 + \left[\frac{kM_1(r_1^*)M_2(r_1^*)}{r_1^*}(-Q) - \frac{km_1m_2}{r_2^*}\right].\end{aligned} \tag{143}$$

Now, an important circumstance is that, whatever the value of r_1^*, the first term in the square brackets in (143) satisfies the inequality

$$\frac{kM_1(r_1^*)M_2(r_1^*)}{r_1^*}(-Q) > \frac{km_1m_2}{r_1^*}. \tag{144}$$

This follows from the fact that $M_1(r) > m_1$ and $M_2(r) > m_2$ at any r and, in addition, the absolute value of Q is always larger than unity (see Footnote 3). Comparison of expressions (143), (144), and (141) indicates that the shell ejection effects in the relativistic region not only remain but can even be more intense.

8. A NUMERICAL REALIZATION OF THE EXACT SOLUTION

Let us now consider a numerical solution of the exact equations of motion for two intersecting shells. To calculate the motions of the shells between their intersections, we used Eqs. (29)–(36), where m_{in}, m_{12}, m_{out}, m_1, m_2, L_1, and L_2 are the free initial parameters of the system. It is also required to specify the initial shell radii; the calculation start time may be taken to be zero. We deduce expressions for the derivatives of the proper times τ_1 and τ_2 with respect to t from (33)–(36). Substituting them into Eqs. (30) and (31) yields the equations of motion for the shells in time relative to an infinitely distant observer.

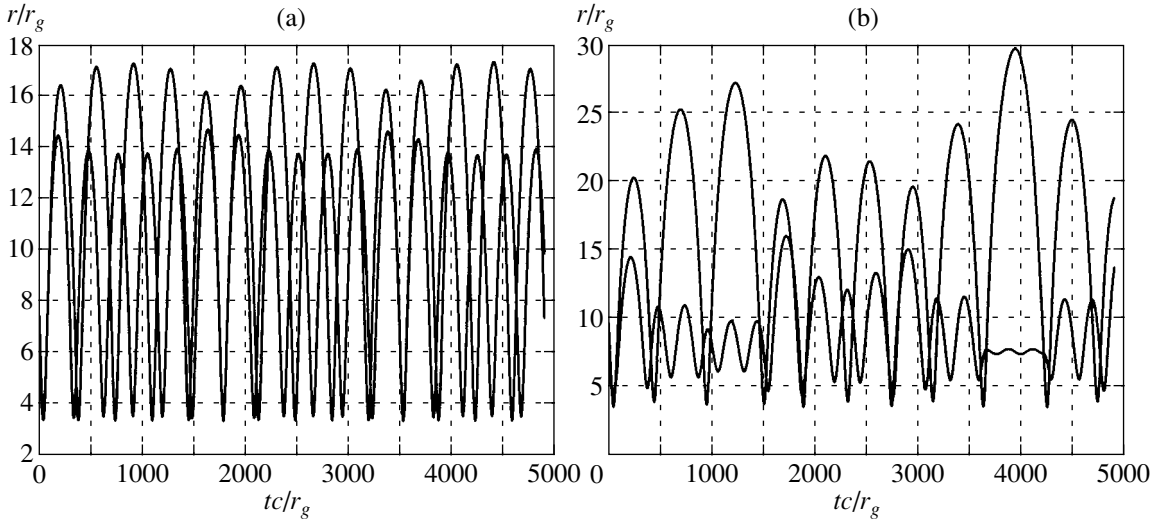


Fig. 2. The motion of two gravitationally bound shells with intersections when the mass of each shell accounts for 1% of the mass of the central body: (a) $m_1 = m_2 = 0.01m_{\text{in}}$ and (b) $0.15m_{\text{in}}$.

The energy of the shells is redistributed during their intersection, which can be taken into account by using Eq. (78). The shells were renamed after their intersection, which allows us to count as many intersections as we wish without being concerned about the shell numbers (shell 2 is always the outer one).

Figure 2 shows the motion of two shells around a central body. The rest masses of the shells were assumed to be identical and equal to 1 and 15% of the mass of the central body, and the initial relative energies E_i/m_ic^2 were 0.975 in both cases. Shell "beating" and energy transfer from one shell to the other, which manifests itself in changes in the radius of maximum shell recession from the center, are clearly seen in Fig. 2a. However, at larger shell masses in Fig. 2b, the shell energies and trajectories even after one intersection change so greatly that the next intersection can occur for quite different shell radii and velocity directions (i.e., the change in phase is comparable to the shell oscillation period itself). We clearly see chaotization of the shell motion from this example. Chaotic motions for various shell parameters in the Newtonian case are illustrated in [8].

To achieve the largest gain in energy of the escaping shell, by analogy with the Newtonian case, we should choose the shell parameters as follows: first, the initial total energies must be close to the rest energies m_1c^2 and m_2c^2; and, second, the first intersection must occur at a point as close to the gravitating center as possible, while the second intersection must occur as far as possible from this center. The characteristic relativistic potential can be used to satisfy these conditions. Near the peak of the potential curve (near $2r_g$), one of the shells can be arbitrarily long; the shell, as it were, sticks to the radius of the potential peak, which gives time for the other shell to fly far away (see Fig. 3).

Figure 4 shows the motion of shells with intersection and with the ejection of one shell after the second intersection. The rest masses of the shells were assumed to be identical and equal to 1% of the mass of the central body, $m_1 = m_2 = 0.01m_{\text{in}}$ (Fig. 4a). The other initial parameters were taken to be $r_1 = 7.5$, $r_2 = 7.75$, $L_1 = 2.013$, $L_2 = 2.0279481$, $m_{\text{out}} - m_{12} = (1 - 10^{-12})m_2$, and $m_{12} - m_{\text{in}} = (1 - 10^{-12})m_1$. Here, r_i and L_i are given in units of r_g and m_ir_gc, respectively, with $r_g = 2km_{\text{in}}/c^2$. The first and second intersections occur at $r_1^* = 2.126104$ and $r_2^* = 43.8996$, respectively. The escaping shell acquires an energy Δmc^2 approximately equal to its kinetic energy $mV^2/2 = \Delta mc^2$, $\Delta m = 4.3604 \times 10^{-5}m_{\text{in}} \approx 4.4 \times 10^{-3}m_1$, which corresponds to the velocity at infin-

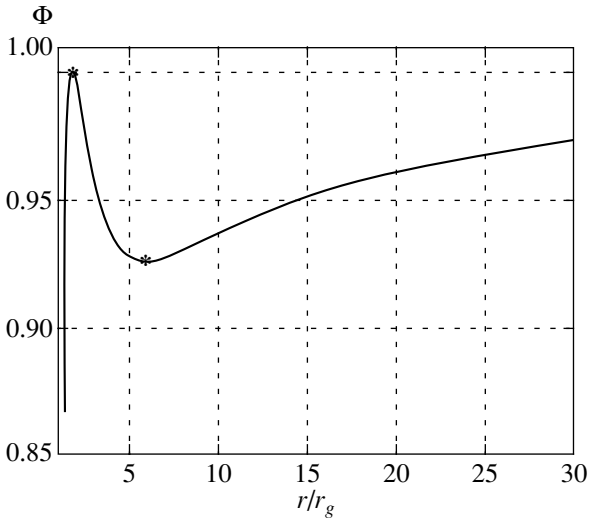


Fig. 3. The total (gravitational plus centrifugal) effective potential Φ for the motion of one shell, $r_g = 2km_{\text{in}}/c^2$.

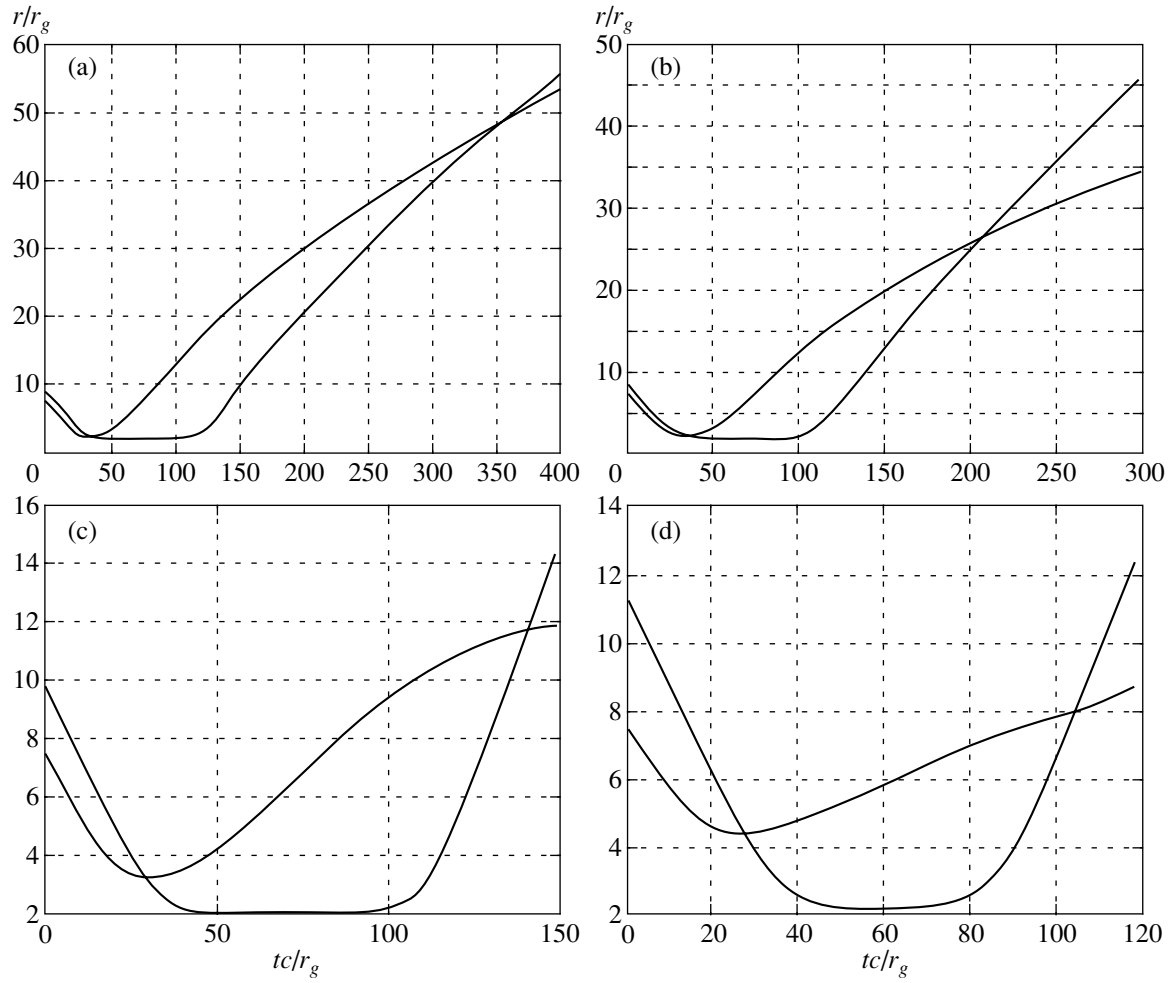


Fig. 4. The ejection of one of the shells after two intersections under the conditions most favorable for the attainment of the maximum ejection velocity: (a) $m_1 = m_2 = 0.01m_{in}$, (b) $0.03m_{in}$, (c) $0.15m_{in}$, and (d) $0.30m_{in}$.

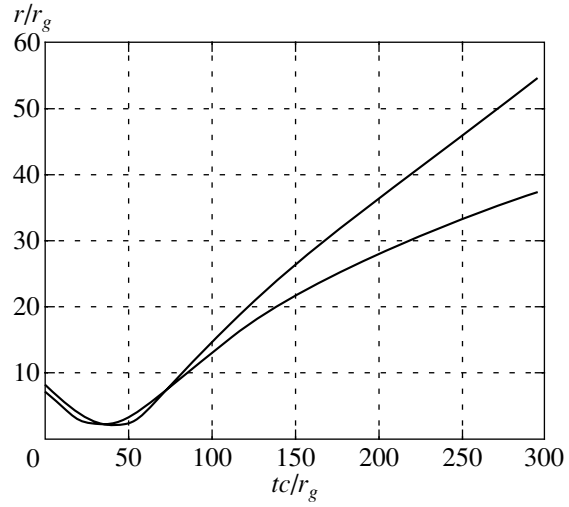


Fig. 5. The ejection of one of the shells after two intersections for initial parameters identical to those in Fig. 4b with the angular momentum of the particles constituting the second shell increased by 0.5%, which violates the most favorable conditions for ejection.

ity $v = 0.0931c$. The initial and resulting parameters for the various calculations shown in Figs. 4a–4d and Fig. 5 are given in the table. Figures 6–8 show plots for variations in the radii of the first and second intersections and in the escape velocity of one of the shells to infinity for the conditions most favorable for shell ejection.

Figure 5 shows the motion of the same shells as those in Fig. 4b but with the angular momentum of shell 2 increased by half a percent. The latter causes the "sticking" phase to disappear; as a result, the second intersection occurs much earlier and the efficiency of the mechanism decreases sharply. The change in energy was found to be $\Delta m = 2.634 \times 10^{-4} m_{in}$, which accounts for about 0.88% of the shell rest mass; i.e., the efficiency of the mechanism decreased by 17% (see the table).

In conclusion, note that, as the shell masses rise, the efficiency of the ballistic ejection mechanism initially increases. However, when the shell rest masses reach about 20% of the mass of the central body, the mini-

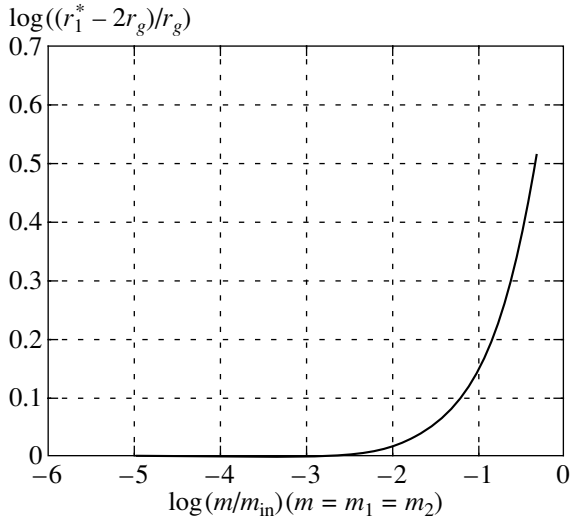


Fig. 6. $\log[(r_1^* - 2r_g)/r_g]$ versus logarithm of the mass ratio, $\log(m_1/m_{in})$; r_1^* is the radius of the first intersection for equal rest masses of the shells, $m_1 = m_2$, under the conditions most favorable for ejection.

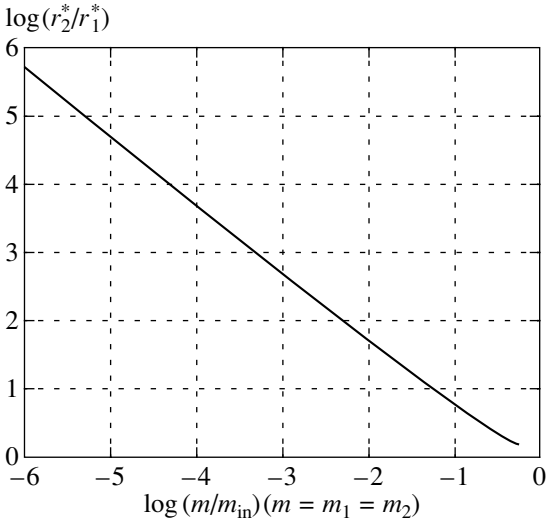


Fig. 7. $\log(r_2^*/r_1^*)$ versus $\log(m/m_{in})$; r_2^* is the radius of the second intersection for the same conditions as in Fig. 6.

mum possible radius of the first intersection increases and the maximum possible radius of the second intersection decreases because of the strong interaction between the shells, which causes the efficiency of this scenario to decrease (see Figs. 6–8). In turn, as was mentioned above, a small deviation of the parameters from the optimal position results in a significant deviation from the limiting shell escape velocity.

The calculations of the shell motion with simplified conditions during intersections from [9, 10] (see the Appendix) proved to be in good agreement with the exact calculations for low shell masses, $m_{1,2}/m_{in} \leq 0.03$. This is because, under the conditions most favorable for escape, the first intersection occurs near the point of minimum radius, where either $v_1 \ll c$ or $v_2 \ll c$ (the necessary condition for the validity of the simplified condition) and the second intersection occurs far from the center in the nonrelativistic region, where $v_1, v_2 \ll c$ as well. Thus, the conditions under which the approximate solution is in good agreement with the exact solution are satisfied.

For low-mass ($m_1, m_2 \ll m_{in}$) shells, we can take the radius of the first intersection to be $r_1^* = 2r_g$ and the

Table

	Fig. 4a	Fig. 4b	Fig. 4c	Fig. 4d	Fig. 5
m_1/m_{in}	0.01	0.03	0.015	0.30	0.03
m_2/m_{in}	0.01	0.03	0.015	0.30	0.03
r_1/r_g	7.5	7.5	7.5	7.5	7.5
r_2/r_g	7.75	8.5	9.8	11.27	8.5
r_1^*/r_g	2.126104	2.3638	3.249	4.3698	2.3765
r_2^*/r_g	43.8996	26.986	11.7076	7.9374	7.9345
$L_1/m_1 c r_g$	2.013	2.05	2.285	2.61	2.05
$L_2/m_2 c r_g$	2.0279481	2.0753315	2.305431	2.5393	2.0857082
$(m_{out}-m_{12})/m_2$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$
$(m_{12}-m_{in})/m_1$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$	$1–10^{-12}$
$\Delta m/m_{in}$	4.3604×10^{-5}	3.1889×10^{-4}	4.22686×10^{-3}	7.52×10^{-3}	2.634×10^{-4}
v/c	0.0931	0.1447	0.2336	0.2198	0.1384
$\Delta m/m_2$, %	0.44	1.06	2.8	2.5	0.88

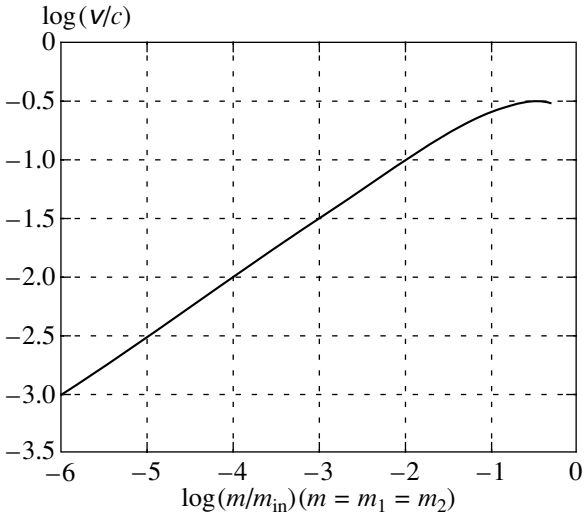


Fig. 8. Logarithm of the ejection velocity to infinity, $\log(v/c)$], versus $\log(m/m_{in})$ for the same conditions as in Fig. 6.

radius of the second intersection to be infinite. As a result, the maximum energy carried away by a low-mass shell through the intersections with a shell of the same rest mass is proportional to the shell mass and the escape velocity is proportional to the square root of the shell mass. Using the first row in the table, we have

$$\frac{\Delta m}{m} \approx 0.044\sqrt{\frac{m}{m_{in}}}, \quad \frac{v}{c} \approx 0.931\sqrt{\frac{m}{m_{in}}}$$

for $m/m_{in} \ll 1$, where $m = m_1 = m_2$.

ACKNOWLEDGMENTS

This study was supported in part by the Russian Foundation for Basic Research (project no. 99-02-18180), the INTAS-ECO (project no. 120), and the INTAS (project no. 00-491).

APPENDIX

Since there was no exact solution to the problem of the intersection of relativistic gravitating shells in the course of the study by Bisnovatyi-Kogan and Yangurazova [9, 10], these authors proposed the following approximate continuity conditions at the point of intersection:

$$\frac{E_1}{\sqrt{1 - r_{g1}/r^*}} = \frac{E_1'}{\sqrt{1 - r_{g1}'/r^*}}, \tag{A.1}$$

$$E_1' + E_2' = E_1 + E_2. \tag{A.2}$$

Here, the total energies of the shells E_1, E_2, E_1', and E_2' are given by formulas (79) and (80); r_{g1} and r_{g1}' are[4]

$$r_{g1} = \frac{2k}{c^2}\left(m_{in} + \frac{E_1}{2c^2}\right),$$
$$r_{g1}' = \frac{2k}{c^2}\left(m_{in} + \frac{E_2'}{c^2} + \frac{E_1'}{2c^2}\right). \tag{A.3}$$

Equation (A.2) is the exact energy equation (81) and requires no discussion. In contrast, as our comparison with the exact theory shows (see below), Eq. (A.1) is approximate. It can be used only for low effective rest masses of the shells M_1 and M_2 (compared to the mass of the central body) together with the condition for the velocity of at least one of the shells being low (compared to the speed of light) at the intersection time. Our exact solution with the most favorable (for escape) conditions corresponds to the case where the first intersection occurs near the turning point (the minimum possible distance from the central body) of one of the shells, i.e., where the velocity of this shell is nearly zero. The energy transfer responsible for the ejection of one of the shells to infinity is determined by this first intersection, because the second intersection occurs far from the center, where the energy transfer is negligible. For these reasons, the results of our numerical calculations using conditions (A.1) and (A.2) for sufficiently low shell masses proved to be similar to those obtained by using the exact theory when describing the cases corresponding to maximum shell escape velocities.

Let us briefly explain the derivation of the approximate condition (A.1) from the exact solution. Consider the intersection of shells described in Section 3 for low (compared to $m_{in}c^2$) energies E_1 and E_2. In addition, we assume that the intersection does not occur too close to the gravitational radius of the central body and that, although the shell velocities at the intersection time can account for a sizeable fraction of the speed of light, they are, nevertheless, not ultrarelativistic. This means that

$$E_1, E_2 \ll m_{in}c^2, \quad f_{in}(r^*) \sim 1,$$
$$\sqrt{1 - v_1^2/c^2}, \sqrt{1 - v_2^2/c^2} \sim 1. \tag{A.4}$$

Below, v_1 and v_2 are the velocities given by relations (59)–(61) but taken at point $r = r^*$. It is easy to show that, under conditions (A.4), the equations of motion (30) and (31), to a first approximation, yield

$$E_1 = \left(\frac{M_1c^2\sqrt{f_{in}}}{\sqrt{1 - v_1^2/c^2}}\right)_{r = r^*},$$
$$E_2 = \left(\frac{M_2c^2\sqrt{f_{in}}}{\sqrt{1 - v_2^2/c^2}}\right)_{r = r^*}, \tag{A.5}$$

[4] Actually, the terms $E_1/2c^2$ and $E_1'/2c^2$ related to shell self-gravitation were taken in [9, 10] without the factor 1/2. The more accurate expressions (A.3) were used later.

whence it also follows that conditions (A.4) mean M_1, $M_2 \ll m_{in}$. In this approximation, formula (85) for energy transfer can be written as

$$\Delta E = \frac{k(1 - v_1 v_2/c^2)}{c^4 r^* f_{in}(r^*)} E_1 E_2. \quad (A.6)$$

If we also add the requirement that the intersection occur near the turning point of one of the shells, i.e., for $v_1 v_2/c^2 \ll 1$, then it follows from (A.6) that

$$\Delta E = \frac{k E_1 E_2}{c^4 r^* f_{in}(r^*)}. \quad (A.7)$$

It is easy to show that the same expression for ΔE also follows from Eqs. (A.1) and (A.2) if we express $E_1 - E_1' = \Delta E$ in them as a function of E_1, E_2, and m_{in} and take the first term of its expansion in small parameters $E_1/m_{in}c^2$ and $E_2/m_{in}c^2$ for $f_{in}(r^*) \sim 1$.

REFERENCES

1. M. Hénon, Ann. Astrophys. **27**, 83 (1964).
2. M. Hénon, Astron. Astrophys. **24**, 229 (1973).
3. S. L. Shapiro and S. A. Teukolsky, Astrophys. J. **298**, 34 (1985).
4. L. R. Yangurazova and G. S. Bisnovatyi-Kogan, Astrophys. Space Sci. **100**, 319 (1984).
5. J. R. Gott, Astrophys. J. **201**, 296 (1975).
6. J. L. Spitzer and H. M. Hart, Astrophys. J. **164**, 399 (1971).
7. J. E. Chase, Nuovo Cimento B **67**, 136 (1970).
8. M. V. Barkov, V. A. Belinski, and G. S. Bisnovatyi-Kogan, astro-ph/0107051; Mon. Not. R. Astron. Soc. **334**, 338 (2002)
9. G. S. Bisnovatyi-Kogan and L. R. Yangurazova, Astrofizika **27**, 79 (1987).
10. G. S. Bisnovatyi-Kogan and L. R. Yangurazova, Astrophys. Space Sci. **147**, 121 (1988).
11. V. Berezin and M. Okhrimenko, Class. Quantum Grav. **18**, 2195 (2001).
12. A. Neronov, hep-th/0109090.
13. J. Khoury, B. Ovrut, P. Stainhardt, and N. Turok, hep-th/0103239.
14. M. Bucher, hep-th/0107148.
15. D. Langlois, K. Maeda, and D. Wands, gr-qc/0111013.
16. W. Israel, Nuovo Cimento B **44**, 1 (1966).
17. T. Dray and G.'t Hooft, Commun. Math. Phys. **99**, 613 (1985).
18. R. V. E. Lovelace, Nature **262**, 649 (1976).
19. G. S. Bisnovatyi-Kogan and S. I. Blinnikov, Pis'ma Astron. Zh. **2**, 489 (1976) [Sov. Astron. Lett. **2**, 191 (1976)].
20. V. A. Ambartsumyan, Uch. Zap. Leningr. Gos. Univ. **22**, 19 (1938).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 392–403.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 459–471.

GRAVITATION, ASTROPHYSICS

Global Monopole in General Relativity¶

K. A. Bronnikov[a, b, *], B. E. Meierovich[c, **], and E. R. Podolyak[c]

[a]*Russian Research Institute for Metrological Service, Moscow, 117313 Russia*
**e-mail: kb@rgs.mccme.ru*
[b]*Institute of Gravitation and Cosmology, Peoples' Friendship University of Russia, Moscow, 117198 Russia*
***e-mail: meierovich@yahoo.com; http://geocities.com/meierovich*
[c]*Kapitza Institute for Physical Problems, Moscow, 117334 Russia*
Received April 15, 2002

Abstract—We consider the gravitational properties of a global monopole on the basis of the simplest Higgs scalar triplet model in general relativity. We begin with establishing some common features of hedgehog-type solutions with a regular center, independent of the choice of the symmetry-breaking potential. There are six types of qualitative behaviors of the solutions; we show, in particular, that the metric can contain at most one simple horizon. For the standard Mexican hat potential, the previously known properties of the solutions are confirmed and some new results are obtained. Thus, we show analytically that solutions with the monotonically growing Higgs field and finite energy in the static region exist only in the interval $1 < \gamma < 3$, where γ is the squared energy of spontaneous symmetry breaking in Planck units. The cosmological properties of these globally regular solutions apparently favor the idea that the standard Big Bang might be replaced with a nonsingular static core and a horizon appearing as a result of some symmetry-breaking phase transition at the Planck energy scale. In addition to the monotonic solutions, we present and analyze a sequence of families of new solutions with the oscillating Higgs field. These families are parametrized by n, the number of knots of the Higgs field, and exist for $\gamma < \gamma_n = 6/[(2n + 1)(n + 2)]$; all such solutions possess a horizon and a singularity beyond it.

1. INTRODUCTION

In accordance with the standard cosmological model [1], the Universe has been expanding and cooling from a split second after the Big Bang to the present moment and has remained uniform and isotropic in doing so. In the process of its evolution, the Universe has experienced a chain of phase transitions with spontaneous symmetry breaking, including the Grand Unification and electroweak phase transitions, formation of neutrons and protons from quarks, recombination, and so forth. Regions with spontaneously broken symmetry that are more than the correlation length apart are statistically independent. At interfaces between these regions, the so-called topological defects necessarily arise. A systematic exposition of the potential role of topological defects in our Universe was given by Vilenkin and Shellard [2]. The particular types of defects—domain walls, strings, monopoles, or textures—are determined by topological properties of the vacuum [3]. If the vacuum manifold is not shrinkable to a point after the breakdown, then the Polyakov–t'Hooft monopole-type solutions [4, 5] appear in quantum field theory.

Spontaneous symmetry breaking plays a fundamental role in modern attempts to construct particle theories. In this context, a symmetry is commonly associated with internal rather than spacetime transformations, e.g., the isotopic, electroweak, and Grand Unification symmetries, and supersymmetry, whose transformations mix bosons and fermions. Topological defects that are caused by spontaneous breaking of internal symmetries and are independent of spacetime coordinates are said to be global.

A fundamental property of a global symmetry violation is the Goldstone degree of freedom. For the monopole, the term related to the Goldstone boson in the energy–momentum tensor decreases rather slowly away from the center. As a result, the total energy of a global monopole grows linearly with the distance, or, in other words, diverges. Without gravity, this divergence is a general property of spontaneously broken global symmetries. In his pioneering paper [4], Polyakov mentioned two possibilities of avoiding this difficulty. The first one was to combine the monopole with the Yang–Mills field. This idea was independently considered by t'Hooft [5]. This, among other reasons, gave rise to numerous papers on gauge (magnetic) monopoles. The second possibility was to consider a bound monopole–antimonopole system, whose total energy would be large (proportional to the distance between the components) but finite.

One more possibility is to take the self-gravity of global monopoles into account; this can in principle remove the above self-energy problem and is also necessary for potential astrophysical applications. Such a study was first performed by Barriola and Vilenkin [6],

¶This article was submitted by the authors in English.

who found that the gravitational field outside a monopole is characterized by a solid angle deficit proportional to the energy scale of the spontaneous symmetry breaking. Harari and Lousto [7] showed that the gravitational mass of a global monopole, calculated using the Tolman integral, is negative. Solutions with a horizon for supermassive global monopoles were found by Liebling [8], who also confirmed the estimate in [9] for the upper value of the symmetry-breaking energy compatible with a static configuration. The existence of de Sitter cores inside global monopoles and other topological defects have led to the idea of topological inflation [10–12].

For global strings in flat space, the energy per unit length (without gravitation) also diverges with growing distance from the axis, but only logarithmically. But in general relativity, integration over the cross section yields a finite result [13, 14]. The gravitational interaction thus leads to self-localization of a global string. Does a similar effect occur for a global monopole? An attempt to answer this question, which does not appear to be answered in the existing papers, was one of the motivations for reconsidering the gravitational properties of a global monopole.

The previous studies have used the boundary condition according to which the symmetry-breaking potential must vanish at spatial infinity. Our approach is different: we do not even assume the existence of a spatial asymptotics, but require regularity at the center and try to observe the properties of the entire set of global monopole solutions. In doing so, among other quantities, we discuss the behavior of the total scalar field energy, which turns out to be finite in static regions of supermassive global monopoles.

In Section 2, we present the complete sets of equations for a static spherically symmetric gravitating global monopole in the two most convenient coordinate systems—those with quasiglobal and harmonic radial coordinates. The general properties of static global monopoles are summarized in Section 3. In Section 4, we analytically and numerically analyze the specific features of a global monopole in the particular case of the Mexican hat potential. Section 5 contains a general discussion of our results, including their possible cosmological interpretation.

2. EQUATIONS AND BOUNDARY CONDITIONS

2.1. General Setting of the Problem

We begin with the most general form of a static spherically symmetric metric, without specifying the radial coordinate $x^1 = u$,

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = e^{2F_0}dt^2 - e^{2F_1}du^2 - e^{2F_\Omega}d\Omega^2. \quad (1)$$

Here, $d\Omega^2 = d\theta^2 + \sin^2\theta d\varphi^2$ is the linear element on a unit sphere and F_0, F_1, and F_Ω are functions of u. The nonzero components of the Ricci tensor are (the prime denotes d/du)

$$\begin{aligned} R_0^0 &= e^{-2F_1}[F_0'' + F_0'(-F_1' + 2F_\Omega' + F_0')], \\ R_1^1 &= e^{-2F_1}[F_0'' + 2F_\Omega'' + 2F_\Omega'^2 + F_0'^2 - F_1'(2F_\Omega' + F_0')], \\ R_2^2 &= R_3^3 = -e^{-2F_\Omega} \\ &\quad + e^{-2F_1}[F_\Omega'' + F_\Omega'(-F_1' + 2F_\Omega' + F_0')]. \end{aligned} \quad (2)$$

We consider the Lagrangian describing a triplet of real scalar fields ϕ^a ($a = 1, 2, 3$) in general relativity,

$$L = \frac{R}{16\pi G} + \frac{1}{2}g^{\mu\nu}\partial_\mu\phi^a\partial_\nu\phi^a - V(\phi), \quad (3)$$

where R is the scalar curvature, $V(\phi)$ is a potential depending on $\phi = \pm\sqrt{\phi^a\phi^a}$, and G is the gravitational constant. We use the natural units such that

$$\hbar = c = 1, \quad (4)$$

and, therefore, $G = m_{Pl}^{-2}$, where $m_{Pl} = 1.22 \times 10^{19}$ GeV is the Planck mass.

To obtain a global monopole with unit topological charge [2], we assume that the metric has form (1) and ϕ^a comprise the following "hedgehog" configuration:

$$\begin{aligned} \phi^a &= \phi(u)n^a, \\ n^a &= \{\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta\}. \end{aligned} \quad (5)$$

The Einstein equations can be written as

$$R_\mu^\nu = -8\pi G\tilde{T}_\mu^\nu = -8\pi G\left(T_\mu^\nu - \frac{1}{2}\delta_\mu^\nu T_\alpha^\alpha\right), \quad (6)$$

where T_μ^ν is the energy–momentum tensor and the nonzero components of $\tilde{T}_\mu^\nu$ are

$$\begin{aligned} \tilde{T}_0^0 &= -V, \quad \tilde{T}_1^1 = -V - e^{-2F_1}\phi'^2, \\ \tilde{T}_2^2 &= \tilde{T}_3^3 = -V - e^{-2F_\Omega}\phi^2. \end{aligned} \quad (7)$$

The conditions for the metric (1) to be regular at the center are that

$$\begin{aligned} e^{F_\Omega} &\longrightarrow 0, \quad F_0 = F_{0c} + O(e^{2F_\Omega}), \\ e^{-F_1 + F_\Omega}&|F_\Omega'| \longrightarrow 1 \end{aligned} \quad (8)$$

at the corresponding value u_c of the coordinate $x^1 = u$. The last condition is necessary for local flatness and ensures the correct ratio of the circumference to the radius for coordinate circles at small $r = e^{F_\Omega}$.

The scalar field energy, defined as the partial time derivative of the scalar field action, $E = -\partial S/\partial t$, is a conserved quantity for our static system,

$$E = \int \sqrt{-g}\,T_0^0 d^3x = 4\pi \int e^{F_0 + F_1 + 2F_\Omega}\left(\frac{1}{2}e^{-2F_1}\phi'^2 + e^{-2F_\Omega}\phi^2 + V\right)du, \quad (9)$$

where g is the determinant of the metric tensor.

In what follows, we make some general inferences without specifying the potential $V(\phi)$ and then perform a more detailed study for the simplest and most frequently used symmetry-breaking potential

$$V(\phi) = \frac{1}{4}\lambda(\phi^a\phi^a - \eta^2)^2 = \frac{1}{4}\eta^4\lambda(f^2 - 1)^2, \quad (10)$$

where $\eta > 0$ characterizes the energy of symmetry breaking, λ is a dimensionless constant, and $f(u) = \phi(u)/\eta$ is the normalized field magnitude playing the role of an order parameter. The model has a global SO(3) symmetry, which can be spontaneously broken to SO(2) by potential wells ($V = 0$) at $f = \pm 1$.

We now explicitly write the Einstein equations and the boundary conditions in the two coordinate frames to be used.

2.2. The Quasiglobal Coordinate ρ

The first choice is the coordinate $u = \rho$ specified by the condition $F_0 + F_1 = 0$. Setting $e^{2F_0} = e^{-2F_1} = A(\rho)$ and $e^{F_\Omega} = r(\rho)$, we obtain the metric

$$ds^2 = A(\rho)dt^2 - \frac{d\rho^2}{A(\rho)} - r^2(\rho)d\Omega^2. \quad (11)$$

The scalar field equation

$$\Box\phi^a + \partial V/\partial\phi^a = 0, \quad (12)$$

where $\Box = \nabla^\alpha\nabla_\alpha$ is the d'Alembert operator, and certain combinations of the Einstein equations are given by

$$(Ar^2\phi')' - 2\phi = r^2 dV/d\phi, \quad (13)$$

$$(A'r^2)' = -16\pi G r^2 V, \quad (14)$$

$$2r''/r = -8\pi G\phi'^2, \quad (15)$$

$$A(r^2)'' - r^2A'' = 2(1 - 8\pi G\phi^2), \quad (16)$$

$$A'rr' + Ar'^2 - 1 = 8\pi G\left(\frac{1}{2}Ar^2\phi'^2 - \phi^2 - r^2V\right), \quad (17)$$

where the prime denotes $d/d\rho$. Only three of these five equations are independent: scalar field equation (13) follows from the Einstein equations and Eq. (17) is a first integral of the others. Given a potential $V(\varphi)$, this is a determined set of equations for the unknowns r, A, and ϕ.

This choice of the coordinates is preferable for considering Killing horizons, which correspond to zeros of the function $A(\rho)$, because such zeros are regular points of Eqs. (13)–(17); moreover, in a close neighborhood of a horizon, the coordinate ρ defined in this manner varies (up to a positive constant factor) as the manifestly well-behaved Kruskal-like coordinates used for analytic continuation of the metric [15, 16]. Therefore, the regions at both sides of a horizon can be simultaneously considered in terms of ρ and the entire range of ρ, can contain several horizons in general. For this reason, the coordinate ρ can be called *quasiglobal*.

The regularity conditions at the center, Eq. (8), are satisfied if

$$A(\rho) = A_c + O((\rho - \rho_c)^2), \quad r(\rho) \approx (\rho - \rho_c)/\sqrt{A_c} \quad (18)$$

near some value ρ_c of the coordinate ρ.

In regions where $A < 0$ (sometimes called *T* regions [1]), whenever they exist, the coordinate ρ is timelike and t is spacelike. Changing the notation as $t \longrightarrow x \in \mathbb{R}$ and introducing the proper time of a comoving observer in the *T* region,

$$\tau = \int d\rho/\sqrt{|A(\rho)|}, \quad (19)$$

we can rewrite the metric as

$$ds^2 = d\tau^2 - |A(\tau)|dx^2 - r^2(\tau)d\Omega^2. \quad (20)$$

The spacetime geometry then corresponds to a homogeneous anisotropic cosmological model of the Kantowski–Sachs type [17, 18], where spatial sections have the topology of $\mathbb{R} \times \mathbb{S}^2$.

2.3. The Harmonic Coordinate u

Another convenient variable that allows considerably simplifying the form of the equations is the harmonic coordinate u specified by the condition [19][1]

$$F_1 = 2F_\Omega + F_0, \quad (21)$$

such that $\Box u = 0$. The field equations can then be written as

$$\phi'' - 2e^{F_0 + F_1}\phi = e^{2F_1}dV/d\phi, \quad (22)$$

$$F_0'' = -8\pi G e^{2F_1}V, \quad (23)$$

[1] A cylindrical version of the harmonic radial coordinate has been used previously in the analysis of gravitational properties of current-conducting filaments [20] and cosmic strings [21, 22].

$$F_1'' - 2F_\Omega'(F_\Omega' + 2F_0') = -8\pi G(\phi'^2 + e^{2F_1}V), \quad (24)$$

$$F_\Omega'' - e^{2(F_0+F_\Omega)} = -8\pi G(\phi^2 e^{2(F_0+F_\Omega)} + e^{2F_1}V), \quad (25)$$

$$\begin{aligned} &-e^{-2F_\Omega} + e^{-2F_1}(F_\Omega'^2 + 2F_\Omega'F_0') \\ &= 8\pi G\left(\frac{1}{2}e^{-2F_1}\phi'^2 - e^{-2F_\Omega}\phi^2 - V\right), \end{aligned} \quad (26)$$

where the prime denotes d/du.

It is straightforward to find that the regularity condition at the center can only correspond to $u \longrightarrow \pm\infty$; we choose $u \longrightarrow -\infty$, where we must have

$$\begin{aligned} e^{F_\Omega} \sim 1/|u|, \quad e^{F_0} = \sqrt{A_c}(1 + O(u^{-2})), \\ e^{F_1} \sim 1/u^2, \end{aligned} \quad (27)$$

and A_c is the same as in (18).

3. GENERAL PROPERTIES OF GLOBAL MONOPOLES

3.1. Monopoles in Minkowski Spacetime

The Minkowski metric written in the usual spherical coordinates,

$$ds^2 = dt^2 - dr^2 - r^2 d\Omega^2, \quad (28)$$

is a special case of (11) with $r \equiv \rho$ and $A \equiv 1$. In flat spacetime, the only unknown is $\phi(r)$ and the only field equation is (13), which becomes

$$(r^2\phi')' - 2\phi = r^2 dV/d\phi. \quad (29)$$

For the particular potential in Eq. (10), we have $dV/d\phi = \lambda\phi(\phi^2 - \eta^2)$ and the scalar field equation can then be written in terms of $f = \phi/\eta$ as

$$r^{-2}(r^2 f')' - 2fr^{-2} + \lambda\eta^2 f(1 - f^2) = 0. \quad (30)$$

The energy integral in Eq. (9) takes the form

$$E = 4\pi\int r^2\left(\frac{1}{2}\phi'^2 + \frac{\phi^2}{r^2} + V\right)dr. \quad (31)$$

In the case where $V(\phi) \geq 0$, its convergence implies that all three terms must vanish as $r \longrightarrow \infty$ sufficiently rapidly:

$$\phi = o(r^{-1/2}), \quad \phi' = o(r^{-3/2}), \quad V = o(r^{-3}). \quad (32)$$

This actually implies that a finite-energy configuration is only possible with $V(0) = 0$, contrary to the symmetry-breaking assumption according to which V has minima in nonsymmetric states, $\phi \neq 0$. In particular, potential (10) does not give rise to global monopoles with a finite energy. A consideration of self-gravity of the field triad ϕ^a is one of the ways to overcome this difficulty.

In flat spacetime, the harmonic coordinate u is related to r as $u - u_0 = \pm 1/r$, where u_0 is an arbitrary constant; choosing the minus sign, we find that u ranges from $-\infty$, which corresponds to the center $r = 0$, to u_0 corresponding to spatial infinity.

3.2. Solutions with Constant ϕ

Under the assumption that $\phi = \phi_0 = \text{const}$, the corresponding value of the potential $V(\phi_0) = V_0$ (times $8\pi G$) plays the role of a cosmological constant, and the Einstein equations can be integrated explicitly.

Indeed, in the region where $\phi = \text{const}$, Eq. (15) reduces to $r'' = 0$, whence $r = \alpha\rho + r_0$, where α, $r_0 = \text{const}$. It remains to find $A(r)$, and this is immediately done by integrating Eq. (16),

$$A(r) = \frac{1 - \Delta}{\alpha^2} - \frac{2GM}{r} + Cr^2, \quad \Delta = 8\pi G\phi_0^2, \quad (33)$$

where M and C are integration constants. Substituting (33) into (14), we find

$$C = -8\pi G V_0/3. \quad (34)$$

Thus, the solution is essentially determined by the values of ϕ_0, V_0, and M. One more constant, α, reflects the freedom in choosing the unit of time. We note that this is not a monopole solution. Even if we set $M = 0$, which is evidently necessary for regularity at $r = 0$, this solution with constant $\phi \neq 0$ is singular at the center: for $A(r)$ given by (33) with $M = 0$, the Kretschmann scalar is $\mathcal{K} = R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta} \approx 4\Delta^2/r^4$ at small r.

Regarding the global monopole, two cases of the solution in Eq. (33) are of interest. The case where $\phi_0 \equiv 0$ describes the symmetric state, and the case where $V_0 = 0$ gives a possible asymptotic behavior at spatial or temporal infinity.

In the case where $\phi_0 = 0$ (the symmetric state), setting $M = 0$ (which is necessary for a regular center), we arrive at the de Sitter metric

$$\begin{aligned} ds^2 = \left(1 - \frac{r^2}{r_h^2}\right)dt^2 - \left(1 - \frac{r^2}{r_h^2}\right)^{-1} dr^2 - r^2 d\Omega^2, \\ r_h^2 = \frac{8\pi G V_0}{3}. \end{aligned} \quad (35)$$

This metric has a horizon at $r = r_h$. At $r > r_h$, outside the horizon, r becomes a timelike coordinate, and t is a spacelike one. Changing the notation as in (19) and (20), we obtain the metric

$$ds^2 = d\tau^2 - \sinh^2(\tau/r_h)dx^2 - r_h^2\cosh^2(\tau/r_h)d\Omega^2. \quad (36)$$

This is the Kantowski–Sachs cosmology with the isotropic inflationary expansion at late times ($\tau \longrightarrow \infty$).

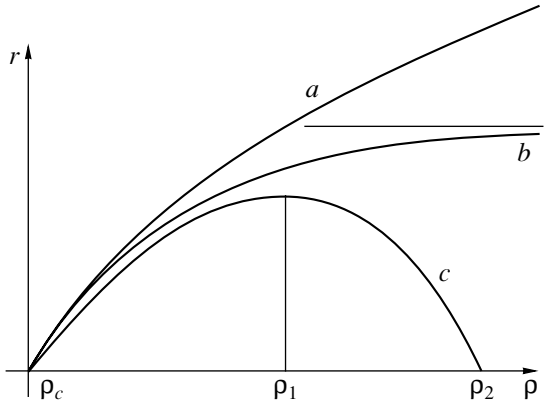


Fig. 1. Possible behavior of $r(\rho)$ in global monopole solutions.

In the other case, $\phi_0 \neq 0$ but $V_0 = 0$ (the case of broken symmetry, such as $\phi = \eta$ in potential (10)), the metric becomes [6]

$$ds^2 = \left(\frac{1-\Delta}{\alpha^2} - \frac{2GM}{r}\right)dt^2 - \left(\frac{1-\Delta}{\alpha^2} - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2, \quad (37)$$

where the constant M has the meaning of mass in the sense that test particles at rest experience the acceleration $-GM/r^2$ in gravitational field (37) at large r.

Furthermore, a nonzero value of ϕ_0 leads to a solid angle deficit Δ defined in (33) in the asymptotic region as $r \longrightarrow \infty$ (see [2] for more detail) and to a linear divergence of integral (9) at large r.

The general case of Eq. (33) describes the large-r asymptotic behavior of any solution to Eqs. (13)–(17), provided that such an asymptotic form exists and ϕ tends to a constant value sufficiently rapidly.

For the monopoles to be studied, metric (37) gives a large-r asymptotic behavior in the case where $\Delta < 1$. We also consider solutions with $\Delta > 1$, for which a static asymptotic regime is absent. Metric (37) then describes cosmological evolution at late times.

3.3. General Properties of Solutions with Varying ϕ

We now consider the general form of Eqs. (13)–(17) with varying ϕ, without specifying the potential $V(\phi)$.

We first note that, because of (15), we have $r'' \leq 0$, which forbids any nonsingular configurations without a center such as wormholes and horns (see Theorem 1 in [16] for further details).

Second, Eq. (16) can be rewritten as

$$(r^4 B')' = -2(1 - 8\pi G\phi^2), \quad B \stackrel{\text{def}}{=} A/r^2, \quad (38)$$

and at a point where $B' = 0$, we have $r^4 B'' = -2(1 - 8\pi G\phi^2)$. Hence, it follows that, as long as $\phi^2 < 1/(8\pi G)$ (i.e., the ϕ field does not reach trans-Planckian values), $B'' < 0$ at possible extrema of the function B. In other words, B cannot have a regular minimum.

Our interest is in systems with a regular center satisfying conditions (18) with $A(\rho) > 0$ and $B(\rho) > 0$ near $\rho = \rho_c$. At a possible horizon $\rho = h$, both A and B vanish, and because this cannot be a minimum of B, $B < 0$ at $\rho > h$ near the horizon. At greater ρ, the function $B(\rho)$, having no minima, can only decrease and never returns to zero; therefore, $A = Br^2 < 0$ at $\rho > h$. We conclude that there can be no more than one horizon, and if it exists, it is simple (corresponds to a simple zero of $A(\rho)$). Because the global causal structure of spacetime is determined (up to possible identifications of isometric hypersurfaces) by the number and disposition of Killing horizons [23–25], we have the following result.

Statement 1. *Under the assumption that* $\phi^2 < 1/(8\pi G)$ *in the entire space, our system with a regular center can have either no horizon or one simple horizon*; *in the latter case, its global structure is the same as that of de Sitter spacetime.*

The above reasoning is essentially the same as in the proof of Theorem 2 in [16] on the disposition of horizons in scalar-vacuum spacetimes. It uses only Eq. (16), which does not involve the potential V. The conclusion is therefore valid for systems with any potentials, positive or negative.

We now return to Eq. (15), according to which $r'' \leq 0$. Because $r' > 0$ at a regular center, this leaves three possibilities for the function $r(p)$ (Fig. 1):

(a) monotonic growth with a decreasing slope, but $r \longrightarrow \infty$ as $\rho \longrightarrow \infty$;

(b) monotonic growth with $r \longrightarrow r_{\max} < \infty$ as $\rho \longrightarrow \infty$; and

(c) growth up to $r_{\max}$ at some $\rho_1 < \infty$ and further decrease, reaching $r = 0$ at some finite $\rho_2 > \rho_1$.

In each case, according to Statement 1, a horizon can occur at some $\rho = h$ within the range of ρ, and we therefore have a T region with the geometry of the Kantowski–Sachs cosmological model at $\rho > h$.

We conclude that there are six classes of qualitative behaviors of the solutions, i.e., (a), (b), and (c), each with or without a horizon, which we indicate with the respective symbols 1 or 0. Thus, all solutions with a spatial asymptotic behavior belong to class (a0). Class (b0) includes spacetimes ending with a "tube" consisting of two-dimensional spheres of equal radii. Solutions in class (c0) contain a second center at $\rho = \rho_2$, and this center can *a priori* be regular or singular. We thus obtain a static analogue of closed cosmologies. Classes (a1), (b1), and (c1) describe different late-time cosmological behaviors in the two directions corresponding to $\mathbb{S}^2$, whereas the fate of the third spatial direction ($\mathbb{R}$) is determined by the function $A(\rho)$. In

particular, the possible de Sitter asymptotic metric in Eq. (36) belongs to class-(a1) solutions, and the expansion is isotropic at late times in this case. On the other hand, class (c1) contains models that at late times behave as the Schwarzschild spacetime inside the horizon, contracting to $r = 0$.

This classification is obtained without any assumptions about $V(\phi)$. Solutions with given $V(\phi)$contain some of these classes, not necessarily all of them.

In the case, where $V \geq 0$, Eq. (14) leads to one more important observation: because $A'r^2 = 0$ at a regular center, we can write (14) in the integral form

$$A'r^2 = -16\pi G\int_0^\rho V(\bar\rho)r^2(\bar\rho)d\bar\rho, \qquad (39)$$

and therefore, $A(\rho)$ is a decreasing function unless $V \equiv 0$. Equation (39) leads to the following conclusions.

Statement 1a. *If $V(\phi) \geq 0$, our system with a regular center can have either no horizon or one simple horizon*; *in the latter case, its global structure is the same as that of de Sitter spacetime*.

Statement 2. *If $V(\phi) \geq 0$, the second center in class-*(c0) *solutions is singular.*

Statement 3. *If $V(\phi) \geq 0$ and the solution is asymptotically flat, the mass M of the global monopole is negative*.

Statement 1a shows that, for nonnegative potentials, the assumption $\phi^2 < 1/(8\pi G)$ in Statement 1 is unnecessary, and the causal structure types are known for any magnitudes of ϕ.

Statement 2 follows from $A'(\rho_2) < 0$, whereas at a regular center, it should be $A' = 0$ (see (18)). The equality $A'(\rho_2) = 0$ could only be possible with $V \equiv 0$, but in this case, the only solution with a regular center is trivial (flat space, $\phi = 0$).

In Statement 3, the asymptotic flatness is understood up to the solid angle deficit, i.e., $r = \rho$ and A is given by (33) with $C = 0$ at large ρ. As $\rho \longrightarrow \infty$, we then obtain $2GM$ on the left-hand side of Eq. (39) and a negative quantity on the right-hand side.

To our knowledge, this simple conclusion, valid for all nonnegative potentials, has so far been obtained only numerically for the particular potential (10) [9]. We note that Statement 3 is an extension to global monopoles of the so-called generalized Rosen theorem [16, 26], previously known for scalar-vacuum configurations.

Therefore, even before studying particular solutions with potential (10), we have a more or less complete knowledge of what can be expected of from such global monopole systems.

4. THE MEXICAN HAT POTENTIAL

4.1. Equations and Boundary Conditions

In what follows, we analyze the particular Mexican hat potential in Eq. (10). For numerical integration, we prefer to use the harmonic coordinate u and to work with Eqs. (22)–(25). This variable enters the equations only via derivatives and is therefore invariant under translations $u \longrightarrow u + \text{const}$.

Introducing the dimensionless quantities

$$\tilde u = u/(\sqrt\lambda\eta), \quad e^{\tilde F_\Omega} = \sqrt\lambda\eta e^{F_\Omega}, \quad e^{\tilde F_1} = \lambda\eta^2 e^{F_1}, \qquad (40)$$

we eliminate the parameter λ from the equations. Indeed, omitting the tildes, we obtain

$$f'' = e^{2(F_0+F_\Omega)}[2 - e^{2F_\Omega}(1-f^2)]f, \qquad (41)$$

$$F_0'' = -\frac{\gamma}{4}e^{2(F_0+2F_\Omega)}(f^2-1)^2, \qquad (42)$$

$$F_\Omega'' = e^{2(F_0+F_\Omega)}\left[1 - \gamma f^2 - \frac{\gamma}{4}e^{2F_\Omega}(1-f^2)^2\right]. \qquad (43)$$

Condition (21) is preserved for the newly defined quantities, but the metric becomes

$$ds^2 = e^{2F_0}dt^2 - \frac{e^{2F_1}du^2 + e^{2F_\Omega}d\Omega^2}{\lambda\eta^2}. \qquad (44)$$

The boundary conditions as $u \longrightarrow -\infty$ are given by

$$f = 0, \quad F_0 = 0, \quad F_0' = 0, \quad F_\Omega = -\ln(-u) + o(1/|u|). \qquad (45)$$

They follow from the regularity requirement at the center and a particular choice of the time unit ($F_0 = 0$) and of the origin of the u coordinate (the fourth condition).

There remains only one dimensionless parameter in Eqs. (41)–(43),

$$\gamma = 8\pi G\eta^2, \qquad (46)$$

which is the squared energy of symmetry breaking in Planck units.

It is easy to find that $\gamma = 1$ is a critical value of this parameter. Indeed, if we assume the existence of a large-r asymptotic behavior such that $f \longrightarrow 1$, i.e., the field tends to the minimum of potential (10), then the asymptotic form of the metric at large r is given by (37) with $\Delta = \gamma$. Consequently, the asymptotics can be static only if $\gamma \leq 1$, whereas, for $\gamma > 1$, the large-r asymptotics can only be cosmological (the Kantowski–Sachs type), and there is a horizon separating such an outer region from the static monopole core.

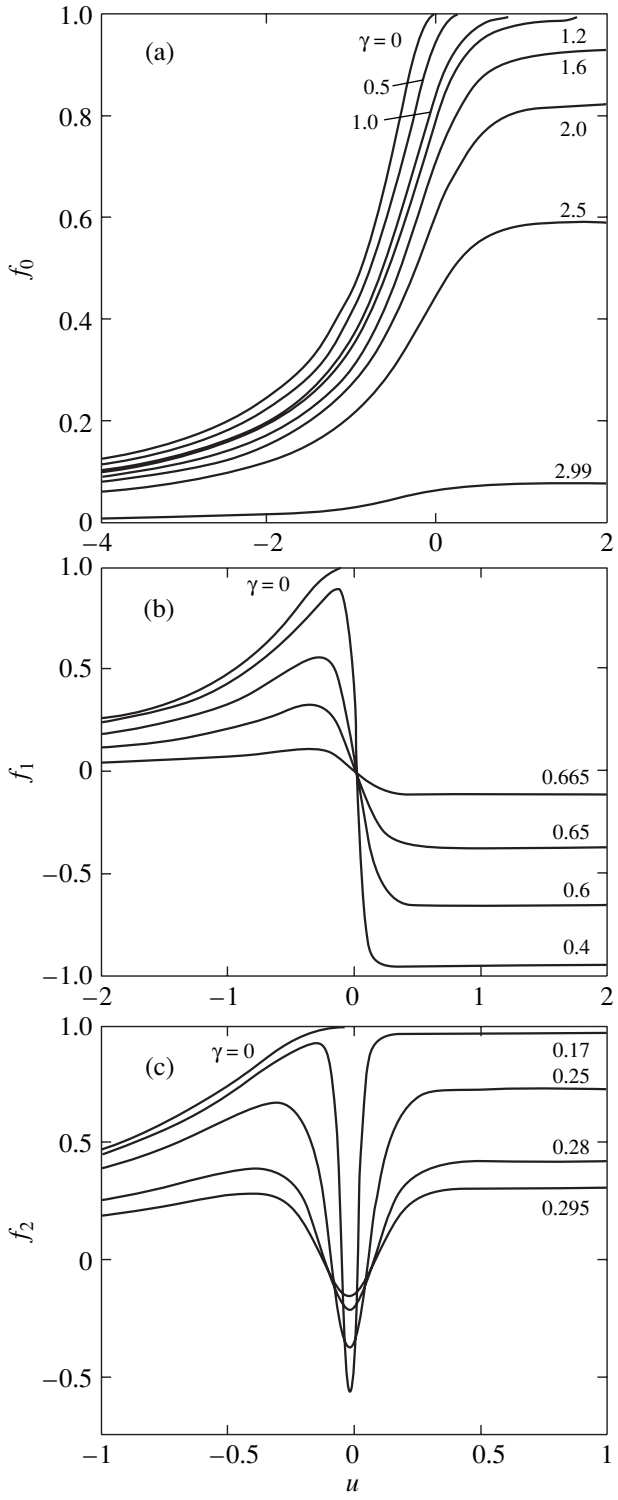


Fig. 2. The field magnitude f as a function of the harmonic coordinate u for different values of γ. Solutions with monotonically growing $f = f_0(u)$ (a) exist for $0 < \gamma < \gamma_0 = 3$. In the region $\gamma < \gamma_1 = 2/3$, there are solutions with $f = f_1(u)$ changing their sign once (b); in the region $\gamma < \gamma_2 = 0.3$, there are solutions with $f = f_2(u)$ changing their sign twice (c). As $\gamma \longrightarrow \gamma_n - 0$, the function $f_n(u)$ vanishes in the entire range $-\infty < u < \infty$ from the center to the horizon.

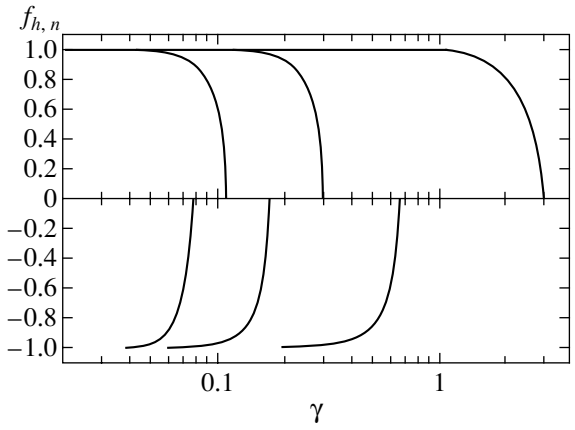


Fig. 3. The γ dependence of the values of $f(u)$ on the horizon, $f_{h,n} = f_n(u)|_{u \to \infty}$.

On the other hand, if a configuration with $\gamma < 1$ possesses a horizon, there is again the Kantowski–Sachs cosmology outside it, but there cannot be a large-r asymptotic form, and in accordance with Section 3, the solutions belong to classes (b1) or (c1).

Now, leaving aside the sufficiently well studied case of solutions with a static asymptotics [2, 6, 7] belonging to class (a0) in accordance with Section 3, we suppose that there is a horizon and return to Eqs. (41)–(43). The horizon corresponds to $u \longrightarrow +\infty$. In such cases, in addition to (45), we impose the boundary condition

$$f(u) \longrightarrow f_h, \quad |f_h| < \infty \ \text{ as } \ u \longrightarrow +\infty. \qquad (47)$$

This condition is necessary for the regularity of a solution on the horizon and is applicable to classes (a1), (b1), and (c1).

For class-(a0) solutions, having a spatial asymptotic behavior and no horizon, condition (47) is meaningless. Moreover, the coordinate u then ranges from $-\infty$ to some $u_0 < \infty$ such that $r(u_0) = \infty$.

For configurations of classes (a0) and (a1), the commonly used boundary condition is

$$f \longrightarrow 1 \ \text{ as } \ r \longrightarrow \infty. \qquad (48)$$

It is of interest that, in case (a1), to which both conditions are applicable, condition (47), being less restrictive, still leads to solutions satisfying (48) because of the properties of the physical system itself.

The set of equations (41)–(43) with boundary conditions (45) and (47) comprise a well-posed nonlinear eigenvalue problem. Its trivial solution, with $f = 0$ and de Sitter metric (35), describes the symmetric state (with unbroken symmetry). Nontrivial solutions describing hedgehog configurations with spontaneously broken symmetry can be found numerically and yield a sequence of eigenvalues γ_n, $n = 0, 1, \ldots$, and the corresponding values of the horizon radius $r_{h,n}$ for each

given value of f_h. Conversely, for a given (admissible) value of γ, we obtain a sequence of values of f_h and r_h.

4.2. The Linear Eigenvalue Problem

Liebling [8] has empirically found the upper critical value $\gamma_0 \approx 3$ for the existence of static solutions.[2] In this section, we find a theoretical ground for this limit.

Actually, we analytically find a sequence of critical values γ_n, $n = 0, 1, \ldots$, such that, for $\gamma < \gamma_n$, there exist static configurations with the field magnitude $f(u)$ changing its sign n times.

Only the analysis for $f(u) > 0$ can be found in the literature. Our numerical integration of Eqs. (41)–(43) shows that, in addition to solutions with monotonically growing $f(u)$ (which exist for $\gamma < \gamma_0 = 3$, Fig. 2a), there exist regular solutions for $\gamma < \gamma_1 = 2/3$ with $f(u)$ changing its sign once (Fig. 2b). Solutions with two zeros of $f(u)$ exist for $\gamma < \gamma_2 = 0.3$ (Fig. 2c), etc. All these solutions have a horizon, and the absolute value of f on the horizon, $|f_{h,n}| = |f_n(\infty)|$, is a decreasing function of γ, vanishing as $\gamma \longrightarrow \gamma_n - 0$ (Fig. 3).

As $\gamma \longrightarrow \gamma_n$, the function $f(u)$ vanishes in the entire range of u, and it is this circumstance that allows us to find the critical values γ_n analytically. In a close neighborhood of γ_n, the field $f(u)$ is small within the horizon, $f^2 \ll 1$, and Eq. (41) therefore reduces to a linear equation with given background functions F_0 and F_Ω, corresponding to the de Sitter metric (35). In terms of the dimensionless spherical radius r, Eq. (41) becomes

$$\frac{d}{dr}\left[r^2\left(1-\frac{r^2}{r_h^2}\right)\frac{df}{dr}\right] - (2-r^2)f = 0, \qquad (49)$$

where $r_h = \sqrt{12/\gamma}$ is the value of r on the horizon. The boundary conditions are

$$f|_{r=0} = 0, \quad |f(r_h)| < \infty. \qquad (50)$$

Nontrivial solutions of (49) with these boundary conditions exist for a sequence of eigenvalues $\gamma = \gamma_n$, $n = 0, 1, 2, \ldots$, and the corresponding eigenfunctions $f_n(r)$, which are regular in the interval $0 \le r \le r_h$, are simple polynomials,

$$f_n(r) = \sum_{k=0}^{n} a_k \left(\frac{r}{r_h}\right)^{2k+1}. \qquad (51)$$

[2] In the notation of [8], $\eta^* \approx \sqrt{3/(8\pi)}$.

Substituting (51) in (49), we find the eigenvalues

$$r_{h,n}^2 = 2(2n+1)(n+2), \quad \gamma_n = \frac{3}{(n+1/2)(n+2)} \qquad (52)$$

and the recurrent relation

$$a_k = a_{k-1}\frac{(2k-1)(2k+2) - r_{h,n}^2}{(2k+1)(2k+2)-2}, \quad k = 1, 2, \ldots, \qquad (53)$$

allowing us to express all a_k, $k = 1, 2, \ldots, n$, in terms of a_0. Because Eq. (49) is linear and homogeneous, a_0 is an arbitrary constant.[3] For fixed n, the coefficients a_k in (51) are

$$a_k = a_0 \prod_{i=1}^{k} \frac{(2i-1)(2i+2) - r_{h,n}^2}{(2i+1)(2i+2)-2}, \qquad (54)$$
$$n > 0, \quad 1 \le k \le n.$$

The case where

$$n = 0, \quad r_{h,0} = 2, \quad f_0(r) = a_0 r / r_{h,0},$$

gives a monotonically growing function $f(u)$ in a close vicinity of $\gamma = \gamma_0 = 3$ (see Fig. 2a). Thus, the upper limit $\gamma_0 = 3$ for the existence of static monopole solutions, previously found by Liebling [8] numerically, is now obtained analytically.

The case where

$$n = 1, \quad r_{h,1} = 3\sqrt{2},$$
$$f_1(r) = a_0 \frac{r}{r_{h,1}}\left[1 - \frac{7}{5}\left(\frac{r}{r_{h,1}}\right)^2\right]$$

describes the function $f(u)$ changing its sign once, at γ close to $\gamma_1 = 2/3$ (see Fig. 2b). The case where $n = 2$, $\gamma_2 = 3/10$, $r_{h,2} = 2\sqrt{10}$ and

$$f_2(r) = a_0 \frac{r}{r_{h,2}}\left[1 - \frac{18}{5}\left(\frac{r}{r_{h,2}}\right)^2 + \frac{99}{35}\left(\frac{r}{r_{h,2}}\right)^4\right]$$

gives the field function $f(u)$ changing its sign twice (Fig. 2c).

For $n \gg 1$, the function $f_n(r)$ rapidly oscillates,

[3] As $\gamma \longrightarrow \gamma_n - 0$, the general equation (41) has the same solution as (49), with $a_0 \ll 1$. To find the dependence $a_0(\gamma)$, we must take the next terms that are nonlinear in f into account.

$$f_n(r) = a_0 \frac{\cos\left[r_{h,n}\arcsin\sqrt{\dfrac{r^2-2}{r_{h,n}^2-2}} - \sqrt{2}\arcsin\left(\dfrac{\sqrt{2}}{r}\sqrt{\dfrac{1-(r/r_{h,n})^2}{1-2/r_{h,n}^2}}\right)\right]}{\sqrt[4]{r^2(r^2-2)[1-(r/r_{h,n})^2]}}.$$

But this semiclassical formula is not valid near the left turning point[4] $r = \sqrt{2}$ (see the dashed curve in Fig. 4). Its applicability range is $1 \ll r < r_{h,n} \approx 2n$, $n \gg 1$.

We have not previosly met regular monopole configurations with the field function $f(u)$ changing its sign. It seems that this is their first presentation.

4.3. Solutions with Monotonically Growing f(u)

As is clear from the aforesaid, the interval $0 < \gamma < 3$ of the existence of nontrivial solutions with monotonically growing $f(u)$ splits into two qualitatively different regions separated by $\gamma = 1$.

In the interval $0 < \gamma < 1$, the solutions have spatial asymptotics (37); according to our general classification, they belong to class (a0). The spherical radius $r(u) = e^{F_\Omega(u)}$ varies from zero to infinity, $f(u)$ grows from zero to unity, and $A(u)$ decreases from unity to its limiting positive value (cf. Eq. (37))

$$A|_{r\to\infty} = \frac{1-\gamma}{\alpha^2},$$
$$\alpha = \left.\frac{dr}{d\rho}\right|_{\rho\to\infty} = 1 - \frac{\gamma}{2}\int_0^\infty f'^2(\rho) r(\rho) d\rho, \qquad (55)$$

and the energy integral (9) diverges.

In the interval $1 < \gamma < 3$, solutions with monotonically growing $f(u)$ belong to class (a1). Instead of a spatial asymptotics, there is a horizon and the Kantowski–Sachs cosmology outside it. The functions A and r inside and outside the horizon are presented in Fig. 5 for $\gamma = 2$. In the presence of a global monopole, the cosmological expansion is slower than the de Sitter one (Eq. (36)). As $\tau \longrightarrow \infty$, the radius $r(\tau)$ grows linearly, while A tends to the negative constant value $-(\gamma-1)/\alpha^2$.

Within the horizon, $f(u)$ monotonically grows from zero at $u = -\infty$ to a value $f_h = f_{h,0}(\gamma)$ on the horizon, $u \longrightarrow \infty$ (see Fig. 2a). As a function of γ, the value $f_{h,0}$ of f on the horizon decreases from unity at $\gamma = 1$ to zero at $\gamma = \gamma_0 = 3$ (see Fig. 3). Integral (9) taken over the static region converges, and we can conclude that, at $1 < \gamma < 3$, the gravitational field is sufficiently strong to suppress the Goldstone divergence and to localize the monopole. At $\gamma > 3$, gravity is probably so strong that it restores the high symmetry of the system.

Outside the horizon, the field f as a function of the proper time τ grows from $f_{h,0}$ on the horizon to unity as $\tau \longrightarrow \infty$. Introducing the proper radial length l inside the horizon by the relation $dl = d\rho/\sqrt{A}$, we can ascertain that the functions $f(l(\rho))$ at $\rho < h$ and $f(\tau(\rho))$ at $\rho > h$ are two parts of a single smooth curve (Fig. 6).

When the parameter γ is close to its critical value $\gamma = 1$ separating the (a0) and (a1) branches of the solution, i.e., when

$$0 < \gamma - 1 \ll 1, \qquad (56)$$

the horizon radius $r_{h,0}$ and the scalar field value on the horizon $f_{h,0}$ can be found analytically under certain additional assumptions on the system behavior that follow from the results of numerical analysis. In particular, there is an "intermediate" region of the u range, $1 \ll u \ll u_0 = \text{const}$, where the first term f'' in the scalar field equation (41) is very small, whereas the function $e^{2(F_0+F_\Omega)}$ is quite large (despite the fact that this function eventually vanishes as $u \longrightarrow \infty$). In this region, the expression in square brackets in (41) must therefore be small, i.e.,

$$e^{2F_\Omega}(1-f^2) \approx 2.$$

This relation can be used for further estimates. The results are

$$\ln r_{h,0} \approx \ln[1/(\gamma-1)] \gg 1,$$
$$f_{h,0} \approx 1 - C(\gamma-1)^2, \qquad (57)$$

where the constant C can be found by comparison with the numerical results; our estimate is $C \approx 0.2$.

The behavior of the solution in the critical regime, $\gamma = 1$, can be characterized as a globally static model with a "horizon at infinity" [8], because $A \longrightarrow 0$ as $r \longrightarrow \infty$.

The fact that monotonic solutions with horizons are absent for $\gamma < 1$ becomes evident from the analysis of the inflection point $u = u_{\inf}$ of the function $F_\Omega(u)$. A horizon, whenever it exists, corresponds to $u \longrightarrow \infty$, where F_Ω remains finite. Because it behaves logarithmically as $u \longrightarrow -\infty$, there is (at least one) inflection point where the second-order derivative is zero, and it follows from Eq. (43) that

$$1 - \gamma f^2 - \frac{\gamma}{4} e^{2F_\Omega}(f^2-1)^2 = 0, \quad u = u_{\inf}.$$

[4] We recall that, in view of the substitution (40), the distances are measured in the units $(\sqrt{\lambda}\eta)^{-1}$.

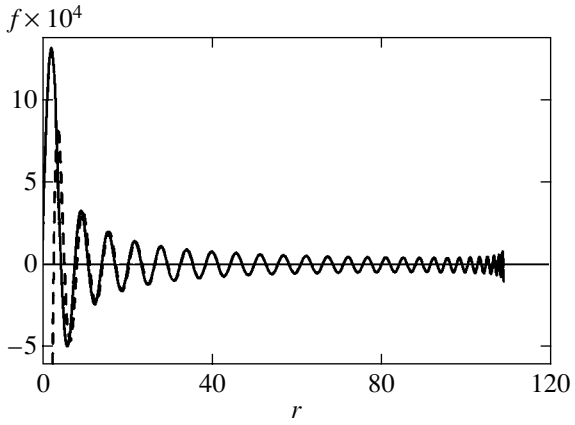


Fig. 4. The field magnitude f as a function of the spherical radius r for $\gamma = 0.001$.

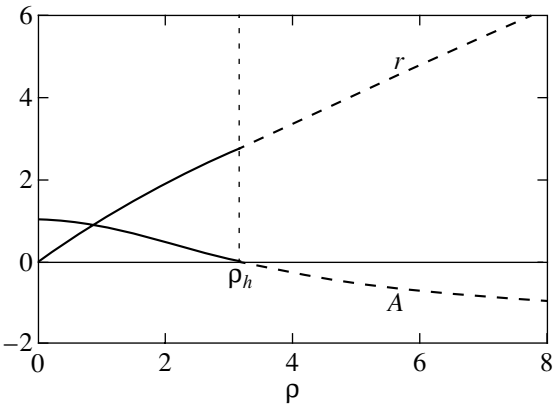


Fig. 5. The functions $A(\rho)$ and $r(\rho)$ form unified smooth curves in the regions inside (solid curves) and outside (dashed) the horizon; $\gamma = 2$.

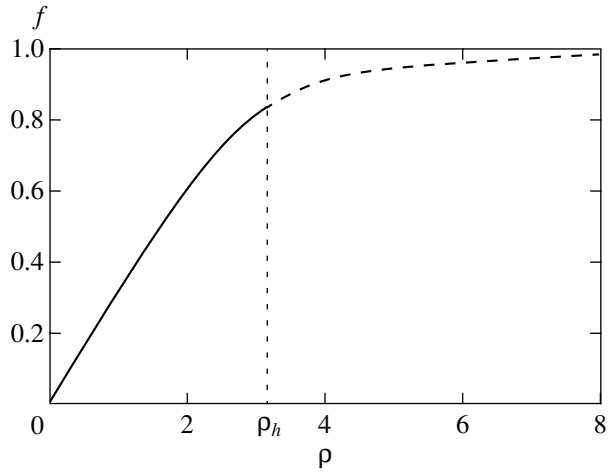


Fig. 6. The function $f(\rho)$ inside (solid curve) and outside (dashed curve) the horizon; $\gamma = 2$.

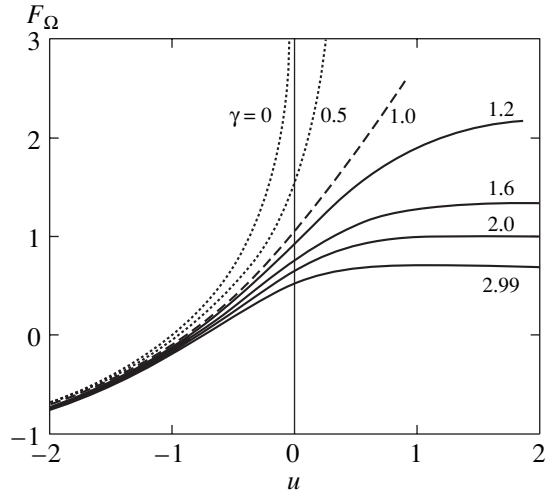


Fig. 7. The function $F_\Omega(u)$ for different values of γ. At $\gamma < 1$, there is a limiting value $u_{\max}$ of u such that $F_\Omega \longrightarrow \infty$ as $u \longrightarrow u_{\max}(\gamma)$ (dotted curves). As $\gamma \longrightarrow 1$, the value $u_{\max}(\gamma) \longrightarrow \infty$ (dashed curve); for $\gamma > 1$, the function $F_\Omega(u)$ tends to a finite constant value as $u \longrightarrow \infty$ (solid curves).

This is a quadratic equation for $1 - f^2$, and hence,

$$1 - f^2 = 2e^{-2F_\Omega}\left(1 \pm \sqrt{1 - \frac{\gamma - 1}{\gamma}e^{2F_\Omega}}\right). \qquad (58)$$

A monotonically growing function $f(u)$ corresponds to greater values of f, i.e., to the "minus" branch of (58) (as is confirmed by numerical results). But the right-hand side of (58) is then negative for $\gamma < 1$, leading to $f^2 > 1$, which cannot occur because $f^2 = 1$ is the maximum attainable value for the solutions under study. Therefore, for $\gamma < 1$, all solutions with monotonically growing $f(u)$ belong to class (a0) and possess a spatial asymptotics with a solid angle deficit and a divergent field energy.

Numerical integration confirms these conclusions. The different behavior of $F_\Omega(u)$ for $\gamma < 1$ and $\gamma > 1$ is shown in Fig. 7.

4.4. Solutions with f(u) Changing Its Sign

For $\gamma < \gamma_1 = 2/3$, there are solutions with the function $f(\phi)$ changing its sign once (see Fig. 2b). For $\gamma < \gamma_2 = 0.3$, there are solutions with $f(\phi)$ changing its sign twice (see Fig. 2c), etc. Unlike the monotonic solutions discussed in Section 4.3, all of them possess a horizon and, in agreement with the general inferences in Section 4.1, belong to class (c1). This implies that, beginning with a regular center, the spherical radius $r(\rho)$ first grows, then passes its maximum $r_{\max}$ at some ρ_1, and then decreases to zero at finite $\rho = \rho_2$, which is a singularity. The horizon occurs at some $\rho = h < \rho_2$, which can be greater or smaller than ρ_1, but in any case, the singularity occurs in a T region and is of cosmological nature. The dependence $r(\rho)$ before and after the horizon is a single smooth curve (Fig. 8a).

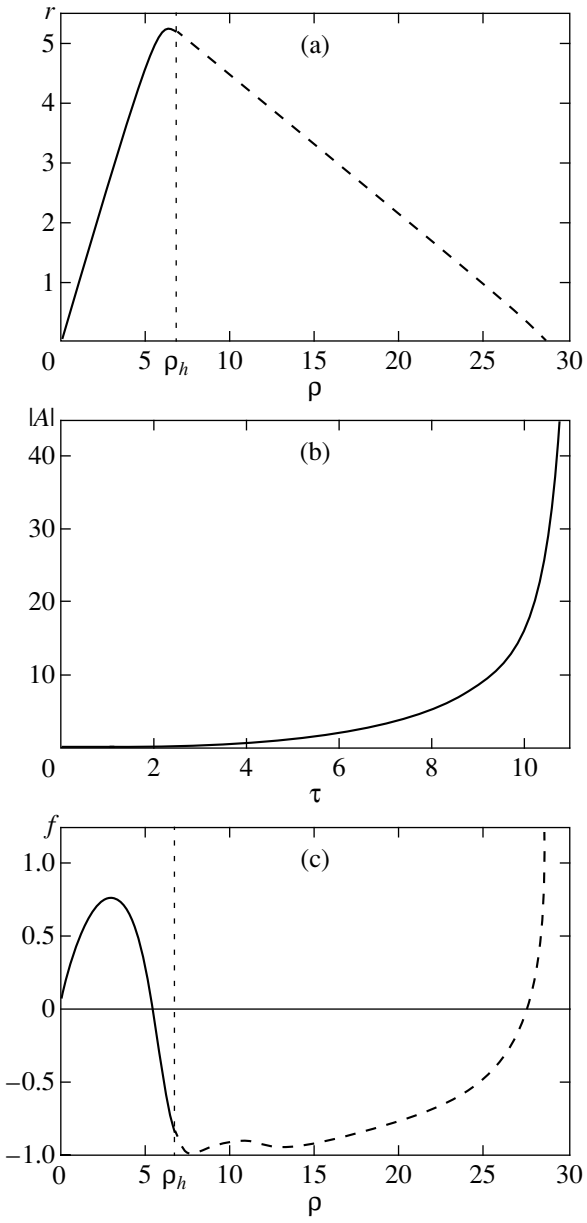


Fig. 8. Solutions with f of alternating sign: the functions (a) $r(\rho)$, (b) $|A(\tau)|$, and (c) $f(\rho)$ for $\gamma = 0.5$ before (solid lines) and after (dashed lines) the horizon.

Beyond the horizon, $|A(\tau)|$ grows from zero at $\tau = 0$ (the horizon) to infinity as $\tau \longrightarrow \tau_s = \tau(\rho_2)$ (the singularity) as a function of the proper time τ of a comoving observer (see Fig. 8b). Beyond the horizon, the scalar field magnitude $|f|$ first grows and then slightly varies around unity. Approaching the singularity, $f(\rho(\tau))$ changes its sign and finally $|f(\rho(\tau))| \longrightarrow \infty$ as $\tau \longrightarrow \tau_s$ (see Fig. 8c).

5. CONCLUSION AND DISCUSSION

We have performed a general study of the properties of static global monopoles in general relativity. We have shown that, independently of the shape of the symmetry-breaking potential, the metric can contain either no horizon or one simple horizon, and in the latter case, the global structure of spacetime is the same as that of the de Sitter spacetime. Outside the horizon, the geometry corresponds to homogeneous anisotropic cosmological models of the Kantowski–Sachs type, where spatial sections have the topology $\mathbb{R} \times \mathbb{S}^2$. In general, all possible solutions can be divided into six classes with different qualitative behaviors. This classification is obtained without any assumptions about $V(\phi)$. Solutions with given $V(\phi)$ contain some of these classes, not necessarily all of them. This qualitative analysis gives a complete picture of what can be expected of global monopole systems with particular symmetry-breaking potentials.

Our analytical and numerical analysis for the Mexican hat potential confirms the previous results of other authors concerning the configurations with the monotonically growing Higgs field magnitude f. Among other things, we have analytically obtained the upper limit $\gamma_0 = 3$ for the existence of static monopole solutions, previously found numerically by Liebling [8]. We have also found and analyzed a new family of solutions with the field function f. Changing its sign, which we have not met in the existing literature.

Of particular interest can be the class-(a1) solutions with a static nonsingular monopole core and the Kantowski–Sachs cosmological model outside the horizon. Its anisotropic evolution is determined by the functions of the proper time $|A(\tau)|$ (the squared scale factor in the $\mathbb{R}$ direction), $r(\tau)$ (the scale factor in the two $\mathbb{S}^2$ directions), and the field magnitude $f(\tau)$. For a comoving observer in the T region, the expansion starts with a rapid growth of $|A(\tau)|$ from zero to finite values, resembling inflation, and ends with $A \longrightarrow$ const as $\tau \longrightarrow \infty$. The expansion in the o $\mathbb{S}^2$ directions described by $r(\tau)$ is comparatively uniform and linear at late times, i.e., much slower than $\cosh^2(\tau/r_h)$ corresponding to de Sitter's spacetime (see (35)). We stress that all such models with the de Sitter-like causal structure (i.e., with a static core and expansion beyond the horizon) drastically differ from the standard Big Bang models in that the expansion starts from a nonsingular surface and cosmological comoving observers can receive information in the form of particles and light quanta from the static region situated in the absolute past with respect to them. Moreover, the static core is nonsingular in our case, and it therefore provides an example of an entirely nonsingular cosmology in the spirit of papers by Gliner and Dymnikova [27–30].

The nonzero symmetry-breaking potential plays the role of a time-dependent cosmological constant, a kind of hidden vacuum matter. Because the field function Λ tends to unity as $\tau \longrightarrow \infty$, the potential vanishes and the "hidden vacuum matter" disappears.

The lack of isotropization at late times does not seem to be a fatal shortcoming of the model for two reasons. First, if the model is used to describe the near-Planck epoch of the Universe evolution, then, at the next stage, the anisotropy can probably be damped by diverse particle creation, and the further stages with lower energy densities may conform to the standard picture (with possible further phase transitions). Second, if we add a comparatively small positive quantity Λ to potential (10) ("slightly raise the Mexican hat"), this must change nothing but the late-time asymptotics, which then becomes the de Sitter one corresponding to the cosmological constant Λ. In our view, these ideas deserve a further study.

Evidently, the present simple model cannot be directly applied to our Universe. It would be too naive to expect that a macroscopic description based on a simple toy model of a global monopole with only one dimensionless parameter γ can explain all the variety of early-Universe phenomena. Nevertheless, it may be considered as an argument in favor of the idea that the standard Big Bang might be replaced with a nonsingular static core and a horizon appearing as a result of some symmetry-breaking phase transition at the Planck energy scale.

ACKNOWLEDGMENTS

The authors are grateful to A.F. Andreev for a useful discussion at the seminar at the Kapitza Institute for Physical Problems.

REFERENCES

1. Ya. B. Zeldovich and I. D. Novikov, *Relativistic Astrophysics* (Nauka, Moscow, 1967; Univ. of Chicago Press, Chicago, 1971).
2. A. Vilenkin and E. P. S. Shellard, *Cosmic Strings and Other Topological Defects* (Cambridge Univ. Press, Cambridge, 1994).
3. T. W. B. Kibble, J. Phys. A **9**, 1387 (1976).
4. A. M. Polyakov, Pis'ma Zh. Éksp. Teor. Fiz. **20**, 430 (1974) [JETP Lett. **20**, 194 (1974)].
5. G.'t Hooft, Nucl. Phys. B **79**, 276 (1974).
6. M. Barriola and A. Vilenkin, Phys. Rev. Lett. **63**, 341 (1989).
7. D. Harari and C. Lousto, Phys. Rev. D **42**, 2626 (1990).
8. S. L. Liebling, Phys. Rev. D **61**, 024030 (1999).
9. N. Sakai, H. Shinkai, T. Tachizawa, and K. Maeda, Phys. Rev. D **53**, 655 (1996).
10. A. Vilenkin, Phys. Rev. Lett. **72**, 3137 (1994).
11. A. Linde, Phys. Lett. B **327**, 208 (1994).
12. R. Basu and A. Vilenkin, Phys. Rev. D **50**, 7150 (1994).
13. B. E. Meierovich, Gen. Relativ. Gravit. **33**, 405 (2001).
14. B. E. Meierovich and E. R. Podolyak, Gravit. Cosmology **7**, 117 (2001).
15. K. A. Bronnikov, G. Clément, C. P. Constantinidis, and J. C. Fabris, Phys. Lett. A **243**, 121 (1998); gr-qc/9801050; Gravit. Cosmology **4**, 128 (1998); gr-qc/9804064.
16. K. A. Bronnikov, Phys. Rev. D **64**, 064013 (2001).
17. A. S. Kompaneets and A. S. Chernov, Zh. Éksp. Teor. Fiz. **47**, 1939 (1964) [Sov. Phys. JETP **20**, 1303 (1965)].
18. R. Kantowski and R. K. Sachs, J. Math. Phys. **7**, 443 (1966).
19. K. A. Bronnikov, Acta Phys. Pol. B **4**, 251 (1973).
20. B. E. Meierovich, Zh. Éksp. Teor. Fiz. **112**, 385 (1997) [JETP **85**, 209 (1997)]; Gravit. Cosmology **3**, 29 (1997); Phys. Rev. D **61**, 024004 (2000).
21. B. E. Meierovich and E. R. Podolyak, Phys. Rev. D **61**, 125 007 (2000).
22. B. E. Meierovich, Usp. Fiz. Nauk **171**, 1003 (2001) [Phys. Usp. **44**, 981 (2001)].
23. K. A. Bronnikov, Izv. Vyssh. Uchebn. Zaved., Fiz., No. 6, 32 (1979).
24. M. O. Katanaev, Nucl. Phys. (Proc. Suppl.) **88**, 233 (2000); gr-qc/9912039; Proc. Steklov Inst. Math. **228**, 158 (2000); gr-qc/9907088.
25. T. Klosch and T. Strobl, Class. Quantum Grav. **13**, 1395 (1996); **14**, 1689 (1997).
26. K. A. Bronnikov and G. N. Shikin, Itogi Nauki Tekh., Ser. Klas. Teor. Polya Gravit. **2**, 4 (1991).
27. E. B. Gliner, Usp. Fiz. Nauk **172**, 221 (2002).
28. E. B. Gliner and I. G. Dymnikova, Pis'ma Astron. Zh. **1** (5), 7 (1975) [Sov. Astron. Lett. **1**, 93 (1975)]; Usp. Fiz. Nauk **172**, 227 (2002).
29. I. G. Dymnikova, Zh. Éksp. Teor. Fiz. **90**, 1900 (1986) [Sov. Phys. JETP **63**, 1111 (1986)].
30. I. Dymnikova and M. Khlopov, Gravit. Cosmol. **4**, 50 (1998); Mod. Phys. Lett. A **15**, 2305 (2000); Eur. Phys. J. C **20**, 139 (2001).

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 404–412.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 472–480.

NUCLEI, PARTICLES, AND THEIR INTERACTION

(Quasi)Elastic Electron–Muon Large-Angle Scattering within the Two-Loop Approximation: Vertex Contributions¶

V. V. Bytev, E. A. Kuraev*, and B. G. Shaikhatdenov**
Joint Institute for Nuclear Research, Dubna, Moscow oblast, 141980 Russia
**e-mail: kuraev@thsun1.jinr.ru*
***On leave of absence from IPT, Almaty-82, Kazakhstan*
Received April 12, 2002

Abstract—We consider quasielastic large-angle electron–muon scattering at high energies with radiative corrections up to the two-loop level. The lowest order radiative corrections arising from the one-loop virtual photon emission and a real soft emission are presented within a power accuracy. Two-loop corrections are supposed to be of three gauge-invariant classes. One of them, the so-called vertex contribution, is given in the logarithmic approximation. The relation to the renormalization group approach is discussed. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Interest in the physics at electron–muon colliders is now increasing. The main attention is paid to the investigation of rare processes, for instance, those violating the lepton number conservation law. Another motivation is a test of the models alternative to the Standard Model [1]. The problems of calibration and precise determination of luminosity will be important. For this, the process of quasi-elastic electron–muon scattering can be used.

The processes of quasi-elastic and inelastic large-angle electron–muon scattering (EMS) play an important role in the luminosity calibration at electron–positron colliders. Indeed, they have a clear signature: the scattered leptons move almost back-to-back (in the center-of-mass reference frame), and the cross section is sufficiently large,

$$\frac{d\sigma_0(\theta)}{d\Omega_e} \approx \frac{200\ \text{nb}}{s[\text{GeV}^2]}, \quad \cos\theta \approx \frac{1}{2}, \tag{1}$$

where s is the total energy squared in center-of-mass reference frame, $d\Omega_e$ is the element of angular phase, and θ is the scattering angle.

The modern experimental requirements to the theoretical accuracy are at the level of per mille or even less and therefore necessitate a detailed knowledge of nonleading terms in the two-loop approximation. Some of these terms were recently calculated in a series of papers [2] devoted to the study of large-angle Bhabha scattering. The contribution of the elastic genuine two-loop virtual correction to the Bhabha amplitude was recently evaluated [3] using the prescription developed in [4] to handle singular terms in QCD at the two-loop level.

In this paper, we consider the EMS process in the two-loop approximation. At this level, we are interested in the contribution to the cross section given by the interference of the Born amplitude and the two-loop virtual corrections. An attempt to solve this problem was made in a series of papers [5], where a direct calculation was performed; unfortunately, their result is incorrect even in the part containing the infrared divergence. Other papers (see, e.g., [6]) were devoted to the calculation of two-loop Feynman amplitudes within the dimensional regularization scheme. Once again, their results cannot be straightforwardly applied to the real amplitudes of large-angle EMS. One of the reasons is the requirement of distinct masses of the interacting particles.

Here, we consider only virtual and real soft photon contributions to the cross section of the EMS. In the third order of the perturbation theory, there exist three sets of contributions, each of which is free of infrared singularities. They include the contribution coming from the one-loop virtual photon emission corrections (see Fig. 1) and the one given by a soft photon emission (see Fig. 3, diagram *a*).

In the fourth order, there are four sets free of infrared singularities. One of them, dubbed the vertex, contains virtual corrections to the lepton vertex function up to the second order of the perturbation theory and relevant inelastic processes with the emission and absorption of real soft photons and lepton pairs by the initial and scattered electrons (and similarly muons). We use here the known expression for the lepton vertex function up to the fourth order of the perturbation theory [7]. Together with the contribution coming from the emission of two real soft photons and a soft charged lepton

¶This article was submitted by the authors in English.

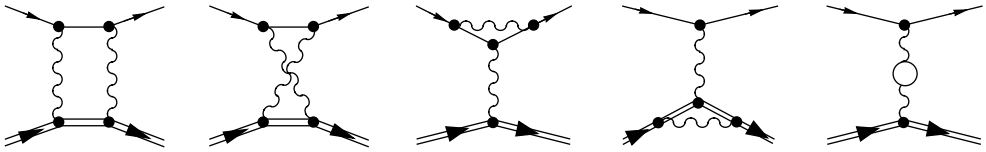


Fig. 1. First-order contributions.

pair (see Fig. 3, diagrams *d–f)*, it is our primary concern in the present paper. We also consider the contribution to the vacuum polarization caused by hadrons and the soft real pion pair production.

Three additional gauge invariant contributions are described by the one-photon exchange containing lepton vertex functions accounting for the vacuum polarization and box-type Feynman diagrams with the self-energy insertion into one of the exchange photon Green's functions. They are left for a separate consideration.

Quasielastic refers to a process with the final particles emitted almost back-to-back in the center-of-mass reference frame. The final particle energies coincide with those of the initial particles up to a small value $\Delta\varepsilon \ll \varepsilon$. This disbalance is due to a possible emission of soft photons and pairs.

We start by giving the results for the Born differential cross section and first-order corrections. The latter contain radiative corrections due to the emission of virtual photons at the one-loop level and the emission of an additional soft photon. These contributions involve infrared divergences that cancel when the two contributions are added.

The result of the calculations agrees with the renormalization group (RG) prediction in the leading logarithmic approximation,

$$\frac{\alpha}{\pi}\rho_t \sim 1, \quad \frac{\alpha}{\pi} \ll 1, \quad \rho_t = \ln\frac{-t}{m_e m_\mu},$$
$$d\sigma = \frac{d\sigma_0}{|1-\Pi(t)|^2}\mathcal{D}_\Delta^4, \tag{2}$$

where $\Pi(t)$is defined below (see Eq. (7)) and $\mathcal{D}_\Delta$ is the Δ part of the nonsinglet lepton structure function [9],

$$\mathcal{D}_\Delta = 1 + \sum_{n=1}^{\infty}\left(\frac{\alpha}{2\pi}\rho_t\right)^n P_{n\Delta}, \quad P_{1\Delta} = 2\ln\Delta + \frac{3}{2},$$
$$P_{2\Delta} = \left(2\ln\Delta + \frac{3}{2}\right)^2 - 4\zeta_2, \tag{3}$$
$$\Delta = \frac{\Delta\varepsilon}{\varepsilon} \ll 1, \quad \zeta_2 = \frac{\pi^2}{6}.$$

Here, ρ_t is the so-called large logarithm, t is the kinematical invariant, and m_e and m_μ are masses of the leptons.

In addition, we give the explicit form of the nonleading terms and present the result of calculating the lowest-order radiative corrections to a power accuracy,

$$1 + O\left(\frac{\alpha}{\pi}\frac{m^2}{s}\rho_t\right). \tag{4}$$

Our calculation of the second-order contribution is performed in the logarithmic approximation. We keep all the logarithmically enhanced terms including those containing logarithms of the mass ratio and omit the terms of the order $O(1)$.

In calculating radiative corrections in the fourth order of perturbation theory, we consider three separate gauge-invariant contributions. We call them the vertex contributions, the decorated boxes, and contributions of the eikonal type. The last two involve amplitudes with the electron–muon exchange enhanced by one or two additional virtual (or real soft) photons and by a virtual (real soft) pair. Their contributions are not considered here.

The first set of Feynman diagrams is of the vertex type with second-order radiative corrections (Fig. 2).

The corresponding contribution involves the fourth power of large logarithms and the infrared divergent terms. Combining this with additional contributions coming from the emission and absorption of one and two soft photons by either of the lepton lines results in the cancellation of the fourth and third powers of large logarithms and of all the infrared-divergent terms. The result is found to be in agreement with the RG predictions.

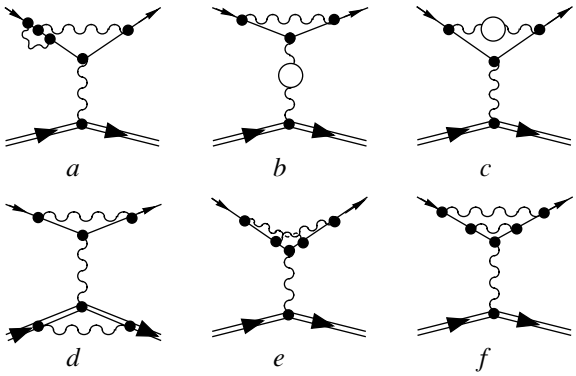


Fig. 2. Some of the second-order *V*-type contributions: *a, d, e, f*—double virtual photon contributions to vertex function; *b, c*—vacuum polarization insertions.

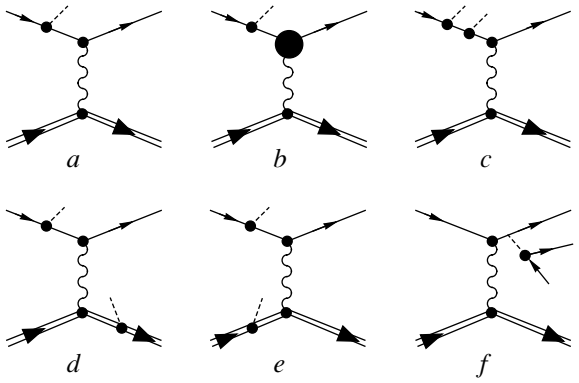


Fig. 3. Some of the soft photon contributions. Diagram *a* corresponds to the first-order radiative corrections; in *b*, the filled circle denotes the vertex one-loop radiative corrections; *c*–*e* represent the emission of two soft photons; and *f*, represents a soft pair production.

Our paper is organized as follows. After some introductory remarks, we discuss the first-order contribution to the cross section of the process in Section 2. In Section 3.1, the radiative corrections coming from the vertex diagrams are considered to the α^4 order of perturbation theory. Section 3.2 is devoted to the study of the vacuum polarization effects including the hadronic contribution to the vertex Feynman diagrams. In Section 3.3, we give the contribution due to the emission of one and two soft photons and a soft pair for the cases of emitted (absorbed) leptons with equal and different masses. In conclusion, we summarize the results obtained.

2. THE BORN CROSS-SECTION AND LOWEST-ORDER RADIATIVE CORRECTIONS

We recall that we consider the large-angle high-energy electron–muon scattering

$$\begin{aligned} &e^-(p_1) + \mu^-(p_2) \longrightarrow e^-(p_1') + \mu^-(p_2'), \\ &p_1^2 = p_1'^2 = m_e^2, \quad p_2^2 = p_2'^2 = m_\mu^2, \end{aligned} \tag{5}$$

with the kinematic invariants s, t, and u much larger than the lepton mass squared,

$$s = (p_1 + p_2)^2, \quad t = (p_1 - p_1')^2 = -\frac{s}{2}(1-c),$$

$$u = (p_1 - p_1')^2 = -\frac{s}{2}(1+c),$$

where $c = \cos(\widehat{\mathbf{p}_1, \mathbf{p}_1'})$ is the cosine of the scatter angle in the center-of-mass reference frame (this reference frame is implied in what follows). The differential cross section in the Born approximation is given by

$$\begin{aligned} d\sigma_0 &= \frac{1}{8s} B d\Gamma, \\ B &= \sum |\mathcal{M}_0|^2 = 8(4\pi\alpha)^2 \frac{s^2+u^2}{t^2}, \end{aligned} \tag{6}$$

$$d\Gamma = \frac{1}{(2\pi)^2}\frac{d^3p_1'}{2\varepsilon_1}\frac{d^3p_2'}{2\varepsilon_2}\delta^4(p_1 + p_2 - p_1' - p_2') = \frac{d\Omega_e}{8(2\pi)^2}.$$

We can then write

$$\frac{d\sigma_0}{d\Omega_e} = \frac{\alpha^2}{2s}\frac{s^2+u^2}{t^2}\left\{1 + O\left(\frac{m_\mu^2}{s}\right)\right\}.$$

The lowest order radiative corrections come from the emission of virtual (one-loop corrections) and real photons. The one-loop radiative corrections are classified into three distinct sets. One of them is related to the vacuum polarization insertion into the propagator of a photon exchanged between leptons. It can be taken into account as

$$\begin{aligned} \left(\frac{d\sigma}{d\Omega_e}\right)^{vp} &= \frac{d\sigma_0}{d\Omega_e}\frac{1}{|1-\Pi(t)|^2}, \\ \Pi(t) &= \frac{\alpha}{3\pi}\left(l_t - \frac{5}{3}\right) + \frac{\alpha}{3\pi}\left(L_t - \frac{5}{3}\right) \\ &\quad + \delta_{\text{had}}(t) + \frac{\alpha^2}{4\pi^2}(l_t + L_t) + \ldots, \\ \delta_{\text{had}}(t) &= \frac{\alpha}{3\pi}\int\limits_{4m_\pi^2}^{\infty}\frac{dM^2}{M^2}\mathcal{R}(M^2)\frac{t}{t-M^2}, \\ l_t &= \ln\frac{-t}{m_e^2} = \rho_t + L, \\ L_t &= \ln\frac{-t}{m_\mu^2} = \rho_t - L, \quad L = \ln\frac{m_\mu}{m_e}, \end{aligned} \tag{7}$$

where M^2 denotes the square of the hadron invariant mass in the process $e\bar{e} \longrightarrow h$ and

$$\mathcal{R}(M^2) = \frac{\sigma_{e\bar{e}\to h}(M^2)}{\sigma_{e\bar{e}\to\mu\bar{\mu}}} \tag{8}$$

is the known ratio of the single-photon annihilation cross sections with hadron and muon pairs produced.

Another set of one-loop radiative corrections contains the vertex function (we recall that only the Dirac

formfactor of the vertex function applies within power accuracy implied in Eq. (4)),

$$\left(\frac{d\sigma}{d\Omega_e}\right)^{v} = \frac{d\sigma_0}{d\Omega_e}[V_e(l_t)V_\mu(L_t)]^2, \tag{9}$$

with the lowest order Dirac formfactors of leptons given by (see [7])

$$\begin{aligned} V_e(l_t) &= 1+\frac{\alpha}{\pi}f_1^{(2)}(l_t)+\frac{\alpha^2}{\pi^2}f_1^{(4)}(l_t)+\ldots,\\ V_\mu(L_t) &= V_e(l_t \longrightarrow L_t),\\ f_1^{(2)}(l_t) &= l_\lambda(1-l_t)-1+\frac{3}{4}l_t-\frac{1}{4}l_t^2+\frac{1}{2}\zeta_2,\\ l_\lambda &= \ln\frac{m_e}{\lambda}. \end{aligned} \tag{10}$$

Here, λ is a fictitious photon mass. It is convenient to represent $f_1^{(4)}(l_t)$ as the sum of two ingredients,

$$f_1^{(4)} = f^{\gamma\gamma}+f^{vp}, \tag{11}$$

where f^{vp} contains QED vacuum polarization effects (to be specified in Section 3.2) and

$$\begin{aligned} f^{\gamma\gamma} &= \frac{1}{32}l_t^4-\frac{3}{16}l_t^3+\left(\frac{17}{32}-\frac{1}{8}\zeta_2\right)l_t^2\\ &+\left(-\frac{21}{32}-\frac{3}{8}\zeta_2+\frac{3}{2}\zeta_3\right)l_t+\frac{1}{2}l_\lambda^2(l_t-1)^2\\ &-l_\lambda(l_t-1)\left(-\frac{1}{4}l_t^2+\frac{3}{4}l_t-1+\frac{1}{2}\zeta_2\right)+O(1),\\ \zeta_3 &\approx 1.2020569. \end{aligned} \tag{12}$$

The contribution of the Pauli formfactor is neglected because it is proportional to the lepton mass squared. The remaining one-loop radiative correction is associated with the interference of the Born amplitude with those containing two virtual photons exchanged between lepton lines.

Depending on the photon energy, the soft region ($\omega < \Delta\varepsilon \ll \varepsilon$) and the hard region ($\omega > \Delta\varepsilon$) of the real photon emission can be distinguished. In the quasireal case, only the soft region is relevant,

$$\begin{aligned} \frac{d\sigma^s}{d\sigma_0} &= -\frac{4\pi\alpha}{(2\pi)^3}\int\frac{d^3k}{2\omega}R^2(k),\\ \omega &= \sqrt{k^2+\lambda^2}<\Delta\varepsilon,\\ R(k) &= Q_k^{p_1p_1'}+Q_k^{p_2p_2'},\quad Q_k^{pp'} = \frac{p'}{p'k}-\frac{p}{pk}. \end{aligned} \tag{13}$$

We now give some useful formulas for the description of a soft photon emission. The center-of-mass reference frame is understood for the initial particles, which implies that the values of 3-momenta of all particles are equal (we consider the elastic EMS).

We first give the expression for a single soft photon emission,

$$\begin{aligned} \delta_{11'}^s &= -\frac{\alpha}{4\pi^2}\int\frac{d^3k}{\omega}\left(\frac{p_1'}{p_1'k}-\frac{p_1}{p_1k}\right)^2\Bigg|_{\omega<\Delta\varepsilon} = \frac{2\alpha}{\pi}\\ &\times\left[(l_t-1)(\ln\Delta+l_\lambda)+\frac{1}{4}l_t^2-\zeta_2+\frac{1}{2}\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right], \end{aligned} \tag{14}$$

with the dilogarithm function

$$\mathrm{Li}_2(z) = -\int_0^z\frac{dx}{x}\ln(1-x).$$

By properly squaring this formula, it is easy to derive the quantity $\delta_{11'}^{SS}$ (see Eqs. (21) and (22)).

The contributions to the lowest order radiative corrections that are free of infrared singularities and originate in two sets containing Dirac formfactors of the leptons, and the relevant contribution coming from a soft photon emission can be cast into the form

$$\begin{aligned} \frac{d\sigma^{(1)}}{d\sigma_0} &= 1+\frac{\alpha}{\pi}[\delta_{11'}^s+2f_1^{(2)}(l_t)+\delta_{22'}^s+2f_1^{(2)}(L_t)] = 1\\ &+\frac{\alpha}{\pi}\left[2(\rho_t-1)\left(2\ln\Delta+\frac{3}{2}\right)-2\zeta_2-1+2\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right], \end{aligned} \tag{15}$$

which agrees with the structure function approach.

After accounting for the soft photon contributions $\delta_{12'}^s$ and $\delta_{21'}^s$ and the interference of the Born and box-type Feynman diagrams, we obtain

$$\begin{aligned} d\sigma &= d\sigma_0\left\{1+\frac{\alpha}{\pi}\frac{1}{|1-\Pi(t)|^2}\right.\\ &\times\left.\left[\rho_t(4\ln\Delta+3)-4\ln\Delta-4-2\zeta_2+2\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right]+K\right\}, \end{aligned}$$

$$\begin{aligned} K &= \frac{\alpha}{\pi}\left\{L_{us}(4\ln\Delta+L_{st}+L_{ut})+2\mathrm{Li}_2\left(\frac{1-c}{2}\right)\right.\\ &+\left.\frac{t^2}{s^2+u^2}\left[\frac{u}{t}L_{st}-\frac{s}{t}L_{ut}+\frac{s-u}{2t}(6\zeta_2+L_{st}^2+L_{ut}^2)\right]\right\}, \end{aligned} \tag{16}$$

$$L_{st} = \ln\frac{s}{-t},\quad L_{ut} = \ln\frac{u}{t},\quad L_{us} = \ln\frac{-u}{s}$$

(details of the lowest order box Feynman diagram contribution can be found in [2]). The factor K represents the sum of the elastic Born and box-type amplitudes and the corresponding inelastic contributions. The

expression for the cross section given above is in agreement with predictions expected from the RG considerations.

The expression for the EMS cross-section in the leading logarithmic approximation can be brought to the form of a Drell–Yan-like process [9] written in terms of structure functions,

$$d\sigma = \frac{d\sigma_0}{|1-\Pi(t)|^2}\left[\mathcal{D}_\Delta\left(\frac{\alpha(t)}{2\pi}l_t\right)\right]^2\left[\mathcal{D}_\Delta\left(\frac{\alpha(t)}{2\pi}L_t\right)\right]^2, \quad (17)$$

with the nonsinglet structure function

$$\mathcal{D}_\Delta(z) = 1 + zP_{1\Delta} + \frac{1}{2}z^2P_{2\Delta} + \ldots + \frac{1}{n!}z^nP_n\Delta + \ldots. \quad (18)$$

Here, $P_{n\Delta}$ is the nth iteration of the Δ-part of the kernel of evolution equations,

$$\begin{aligned} P_n(y) &= \lim_{\Delta\to 0}[P_{n\Delta}\delta(1-y) + \Theta(1-y-\Delta)P_{n\theta}] \\ &= \int_y^1 \frac{dx}{x}P_1(x)P_{n-1}\left(\frac{y}{x}\right), \quad n \geq 2, \\ P_{1\theta} &= \frac{1+y^2}{1-y}, \end{aligned} \quad (19)$$

$$P_{2\theta} = \frac{1+y^2}{1-y}\left[\ln\frac{(1-y)^2}{y} + \frac{3}{2}\right] + \frac{1}{2}(1+y)\ln y - (1-y).$$

Explicit expressions for $P_{1\Delta}$ and $P_{2\Delta}$ are given in (3). The parameter Δ ($\Delta \ll 1$) can be interpreted as the energy fraction carried by soft real photons and pairs escaping detectors. The running QED coupling constant is

$$\alpha(t) = \frac{\alpha}{1-(\alpha/3\pi)t}.$$

3. SECOND-ORDER RADIATIVE CORRECTIONS

Second-order radiative corrections can be represented as the sum of several sets, each of which depends on the choice of the gauge with respect to virtual and real photons. We consider Feynman diagrams describing elastic scattering with the vacuum polarization effects included and with the soft pair production taken into account. They are related to the one-photon exchange Feynman diagrams for both elastic and quasi-elastic processes and can be specified by the emission of two more (either virtual or real) photons from the same lepton lines.

A keystone to this classification is the soft photon radiator cross section. In the case of only one soft photon emitted, it takes the form

$$d\sigma^s = \frac{1}{8s}\frac{d\Omega_e}{8(2\pi)^5}(\delta\sum|\mathcal{M}|^2)_s\frac{d^3k}{2\omega}\bigg|_{2\omega/\sqrt{s}<\Delta}, \quad (20)$$

$$(\delta\sum|\mathcal{M}|^2)_s = 2\mathrm{Re}\sum\mathcal{M}_0^*\mathcal{M}^{(1)}(-4\pi\alpha)R^2(k).$$

For the emission of two soft photons (e.g., by the electron block), we have

$$\begin{aligned} d\sigma^{ss} &= d\sigma_0\frac{1}{2!}(-4\pi\alpha)^2\frac{d^3k_1}{2\omega_1(2\pi)^3} \\ &\times\frac{d^3k_2}{2\omega_2(2\pi)^3}(Q_{k_1}^{p_1p_1'})^2(Q_{k_2}^{p_2p_2'})^2\bigg|_{2\omega_1/\sqrt{s}+2\omega_2/\sqrt{s}<\Delta}. \end{aligned} \quad (21)$$

For the emission of two soft photons such that their total energy does not exceed $\Delta\varepsilon \ll \varepsilon$, we have

$$\begin{aligned} &\left[\int\frac{d^3k_1}{\omega_1}\frac{p_ip_j}{p_ik_1p_jk_1}\int\frac{d^3k_2}{\omega_2}\frac{p_lp_m}{p_lk_2p_mk_2}\right]\bigg|_{\omega_1+\omega_2<\Delta\varepsilon} \\ &= (a_1\ln\Delta + b_1)(a_2\ln\Delta + b_2) - a_1a_2\zeta_2, \end{aligned} \quad (22)$$

where

$$\left[\int\frac{d^3k_1}{\omega_1}\frac{p_ip_j}{p_ik_1p_jk_1}\right]\bigg|_{\omega_1<\Delta\varepsilon} = a_1\ln\Delta + b_1,$$

$$\left[\int\frac{d^3k_2}{\omega_2}\frac{p_lp_m}{p_lk_2p_mk_2}\right]\bigg|_{\omega_2<\Delta\varepsilon} = a_2\ln\Delta + b_2.$$

The general structure of all the above contributions to the differential cross-section reveals the presence of large logarithms up to the fourth power. But the sum involves only their second powers. Such a cancellation is characteristic of each gauge-invariant set of corrections.

3.1. Vertex Graphs

Three gauge-invariant groups of Feynman diagrams containing one photon exchange contribute

$$\frac{d\sigma^v}{d\sigma_0} = \frac{\alpha^2}{\pi^2}[a_1 + \tilde{a}_1 + a_2]. \quad (23)$$

The quantity $\tilde{a}_1$ is related to the emission of two (virtual and real) photons out of a muon line,

$$\tilde{a}_1 = a_1(l_t \longrightarrow L_t).$$

Using the results given in Eq. (12) for the electron Dirac formfactor up to the fourth order of perturbation theory,[1] we can construct the contributions to the squared matrix element of one-photon exchange amplitudes that are free of infrared singularities

$$\begin{aligned} a_1 &= (f_1^{(2)})^2 + 2f^{\gamma\gamma} + 2f_1^{(2)}\delta_{11'}^s + \delta_{11'}^{ss}, \\ a_2 &= 4f_1^{(2)}\tilde{f}_1^{(2)} + 2[f_1^{(2)}\delta_{22'}^s + \tilde{f}_1^{(2)}\delta_{11'}^s] + \delta_{11'}^s\delta_{22'}^s, \end{aligned} \quad (24)$$

where $\tilde{f}_1^{(2)}$ corresponds to the muon formfactor, which is identical to the electron one with the electron mass

[1] Here, we omit the contribution of the vacuum polarization; it is taken into account in what follows.

replaced by that of the muon. The quantities δ_{ij}^{s} and δ_{ij}^{ss} correspond to the emission of one and two soft real photons (their energies are restricted by the condition $\Delta\omega_1 + \Delta\omega_2 < \epsilon$) from the fermion lines i, j. The corresponding expression is given in Eq. (14). We note the factor 1/2! in front of the latter quantities, which is due to the identity of the soft photons emitted.

The relevant contribution to the differential cross section in the logarithmic approximation is then given by

$$\begin{aligned} a_1 + \tilde{a}_1 &= \rho_t^2 P_{2\Delta} \\ &+ \rho_t\left[-\frac{45}{8} + Y + 2\zeta_2 + 6\zeta_3\right] + O(1), \\ a_2 &= \rho_t^2 P_{2\Delta} + \rho_t[-6 + Y + 5\zeta_2] + O(1), \\ Y &= 2P_{1\Delta}\mathrm{Li}_2\left(\frac{1+c}{2}\right) - (4\zeta_2 + 14)\ln\Delta - 8\ln^2\Delta. \end{aligned} \tag{25}$$

This result is in agreement with the radiative corrections form of the large-angle cross section.

3.2. Hadronic Vacuum Polarization

We study the vacuum polarization effects occurring in considering vertex Feynman diagrams (see Fig. 2, diagrams *b*, *c)*. For this, we use the known expression for the hadronic vacuum contribution to the photon Green's function by making the substitution

$$\frac{1}{k^2} \longrightarrow \frac{\alpha}{3\pi}\int\limits_{4m_\pi^2}^{\infty}\frac{dM^2}{M^2}\frac{\mathcal{R}(M^2)}{k^2 - M^2}, \tag{26}$$

where k is the 4-momentum of the virtual photon, M^2 is the hadron invariant mass squared, and the ratio $\mathcal{R}(M^2)$ is given in Eq. (8).

In the next order of perturbation theory, we must consider the three gauge-invariant classes of Feynman diagrams for elastic and quasi-elastic processes with a soft photon and a soft pion pair production.

We first examine the vertex class. The cross sections can be written as

$$\begin{aligned} \frac{d\sigma}{d\sigma_0} &= 1 + (\delta^s + \delta^v)_{\mathrm{had}}, \\ \delta_{\mathrm{had}}^{v} &= \frac{\alpha^2}{6\pi^2}[F(m_e^2, t) + F(m_\mu^2, t)], \\ F(m^2, t) &= \int\limits_{4m_\pi^2}^{\infty}\frac{dM^2}{M^2}\mathcal{R}(M^2)F_1(t, m^2, M^2), \end{aligned} \tag{27}$$

where $\delta_{\mathrm{had}}^{v}$ corresponds to the soft hadron emission of soft pion pairs and F_1 (with the hadronic vacuum polarization of the virtual photon) is the vertex contribution to the Dirac formfactor of a lepton with the mass m. The contribution of the Pauli formfactor F_2 is suppressed by the factor $|m^2/t|$.

The standard calculation with the regularization at $t = 0$ leads to

$$F_1(t, m^2, M^2) = 2\int\limits_0^1 dx\int\limits_0^1 y\,dy\left[\ln\frac{d_0}{d} + \frac{a}{d} - \frac{a_0}{d_0}\right] \tag{28}$$

with

$$\begin{aligned} a &= a_0 + t[1 - y + x(1-x)y^2], \\ d &= d_0 - y^2 x(1-x)t, \\ a_0 &= -m^2(2 - y^2), \quad d_0 = y^2 m^2 + (1-y)M^2 \end{aligned} \tag{29}$$

(for details, see Appendix A). It can be seen that the condition $F_1|_{t=0} = 0$ is satisfied. We now consider two limiting cases for F_1. In the case of a large hadron invariant mass squared compared to $-t$, we find

$$\begin{aligned} F_1(t, m^2, M^2) &= \frac{t}{M^2}\left[\frac{2}{3}\ln\frac{M^2}{-t} + \frac{11}{9}\right], \\ M^2 &\gg -t, \end{aligned} \tag{30}$$

and in the case of a small invariant mass squared,

$$\begin{aligned} F_1(t, m^2, M^2) &= -\ln^2\frac{-t}{m^2} - 2\ln\frac{M^2}{m^2}\ln\frac{-t}{m^2} \\ &- 5\ln\frac{-t}{m^2} + \frac{\pi^2}{3} - \frac{1}{2}, \quad -t \gg M^2 \gg m_\mu^2. \end{aligned} \tag{31}$$

Taking the emission of soft pairs into account (see Appendix B), we obtain the hadronic contribution to the radiative correction

$$\begin{aligned} (\delta^v + \delta^s)_{\mathrm{had}} &= \frac{\alpha^2}{6\pi^2}\int\limits_{4m_\pi^2}^{-t}\frac{dM^2}{M^2}R(M^2) \\ &\times\left[-\ln\frac{-t}{M^2}\left[8\ln\frac{M^2}{m_e m_\mu} - 2\ln\Delta + 10\right]\right. \\ &\left. - 6\ln^2\frac{M^2}{m_e m_\mu} - 10\ln\frac{M^2}{m_e m_\mu} - 6\ln^2\frac{m_\mu}{m_e} + \frac{2}{3}\pi^2 - 1\right]. \end{aligned} \tag{32}$$

3.3. Leptonic Vacuum Polarization and Soft Lepton Pairs

We next study the contribution to the lepton vertex function of the vacuum polarization type. Obviously, there are two possibilities for a vacuum polarization blob to be inserted into the lepton vertex function. The

contribution to the elastic cross section can then be written as

$$\left(\frac{d\sigma^{vp}}{d\sigma_0}\right)_e = 2\frac{\alpha^2}{\pi^2}[Z_1(m_e, m_e) + Z_2(m_e, m_\mu)], \quad (33)$$

where

$$Z_1(m_e, m_e) = -\frac{1}{36}\rho_t^3 + \frac{1}{12}\left(\frac{19}{6} - L\right)\rho_t^2$$
$$-\frac{1}{36}\left(6\zeta_2 + \frac{265}{6} + 3L^2 - 19L\right)\rho_t \equiv f^{vp}$$

is the contribution of the electron blob inserted into the electron vertex function (see (11) for the definition of f^{vp}) and

$$Z_2(m_e, m_\mu) = -\frac{1}{36}\rho_t^3 + \frac{1}{12}\left(\frac{19}{6} + L\right)\rho_t^2$$
$$-\frac{1}{36}\left(6\zeta_2 + \frac{265}{6} + 3L^2 + 63L\right)\rho_r$$

is a muon blob contribution to the electron vertex.

A similar expression is valid for the muon vertex function (the electron blob contribution to the muon vertex),

$$Z_3(m_\mu, m_e) = -\frac{1}{36}\rho_t^3 + \frac{1}{12}\left(\frac{19}{6} - L\right)\rho_t^2$$
$$-\frac{1}{36}\left(6\zeta_2 + \frac{265}{6} + 3L^2 - 25L\right)\rho_t.$$

We now turn to the inelastic process of a lepton–antilepton pair production (of the mass μ, obeying $2\mu \ll \Delta\varepsilon \ll \varepsilon$). For the differential cross section, we obtain

$$\frac{d\sigma^{sp}}{d\sigma_0} = \frac{\alpha^2}{6\pi^2}\left[\frac{1}{3}\mathbf{L}^3 + \mathbf{L}^2\left(2\ln\Delta - \frac{5}{3}\right)\right.$$
$$\left. + \mathbf{L}\left(4\ln^2\Delta - \frac{20}{3}\ln\Delta + \frac{56}{9} - 4\zeta_2 + 2\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right)\right] \quad (34)$$

with

$$\mathbf{L} = \ln(-t/\mu^2).$$

We assume a muon or an electron to be a scattered lepton, and consequently, the quantity μ stands for the corresponding mass.

The sum of contributions (33) and (34) does not contain cubic powers of large logarithms; for the "electron line corrections," it is found to be (see [11])

$$\left(\frac{d\sigma^{sv,\,vp}}{d\sigma_0}\right)_e = \left(\frac{\alpha}{\pi}\right)^2\left\{\left(\frac{2}{3}\ln\Delta + \frac{1}{2}\right)\rho_t^2 + 2\rho_t\left[-\frac{17}{12} - \frac{11}{9}L\right.\right.$$
$$\left.\left. + \frac{2}{3}\ln^2\Delta - \frac{10}{9}\ln\Delta - \zeta_2 + \frac{1}{3}\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right]\right\}. \quad (35)$$

For a muon, it is given by

$$\left(\frac{d\sigma^{sv,\,vp}}{d\sigma_0}\right)_\mu = \left(\frac{\alpha}{\pi}\right)^2\left\{\left(\frac{2}{3}\ln\Delta + \frac{1}{2}\right)\rho_t^2 + 2\rho_t\left[-\frac{17}{12} + \frac{11}{6}L\right.\right.$$
$$\left.\left. + \frac{2}{3}\ln^2\Delta - \frac{10}{9}\ln\Delta - \zeta_2 + \frac{1}{3}\mathrm{Li}_2\left(\frac{1+c}{2}\right)\right]\right\}. \quad (36)$$

It can be seen that the leading terms are in agreement with the RG predictions.

4. SUMMARY

We have evaluated the Born cross section and the first-order radiative correction to it of the EMS process in the quasi-elastic kinematical situation. The relevant formulas are given in (2) and (3) in the leading logarithmic approximation and in (16) with power accuracy.

Among second-order contributions, we have considered gauge-invariant contributions from Feynman diagrams with radiative corrections to the vertex function of either lepton. We have also included soft photon and pair emission with energies less than $\Delta\varepsilon$.

In the leading logarithmic approximation, the results are in agreement with the RG.

The explicit results for virtual and soft real photon emission are given in Eqs. (23) and (25). For the emission of virtual and soft real lepton pairs, the relevant formulas are given in Eqs. (34) and (35).

In Section 3.2, we determined the contributions coming from the hadronic vacuum polarization, where the radiative correction is expressed in terms of an explicit integral of the experimentally measured quantity $R(M^2)$. We also consider a soft pion pair production (see Appendix B). We calculate the hadronic vacuum polarization contribution to the vertex functions of the electron or muon explicitly. The relevant formulas for radiative corrections are given in (32).

The evaluation of contributions of the other gauge-invariant types, the eikonal and the decorated box Feynman diagrams, requires additional investigation.

APPENDIX A

Details of the Hadronic Vacuum Polarization

We consider here the details of the vertex hadron function calculation. For the vertex function, we can write

$$V_\mu = \Gamma_1\gamma_\mu + \Gamma_2(\hat{q}\gamma_\mu - \gamma_\mu\hat{q}), \tag{37}$$

where $q = p_2 - p_2'$ and Γ_1 and Γ_2 are the Dirac and Pauli formfactors respectively. We write the vertex function as

$$\begin{aligned} V_\mu &= \gamma_\mu[\Gamma_1 + 4m\Gamma_2] - 2(p_2 + p_2')_\mu\Gamma_2 \\ &= \gamma_\mu A + (p_2 + p_2')_\mu B, \end{aligned} \tag{38}$$

where

$$\begin{aligned} A &= \int ydy\int dx\left[\frac{2y2p_2p_2'}{-d} + \frac{4y^2}{d}(m^2 + p_2p_2')\right. \\ &\left. - 2y\ln\frac{d}{m^2} + \frac{-2m^2y^3x^2}{d} + \frac{2y^3x(1-x)}{d}(-2p_2p_2')\right], \\ B &= \int ydy\int dx\left[\frac{y^2}{d}(-2m)\right. \\ &\left. + \frac{2y^3x^2}{d}2m + \frac{2y^3x(1-x)}{d}2m\right]. \end{aligned} \tag{39}$$

The quantities d and d_0 are defined in Eq. (29). With the regularization at $t = 0$, we have

$$\begin{aligned} F_1(t, m^2, M^2) &= \Gamma_1 - \Gamma_1|_{t=0} \\ &= 2\int_0^1 dx\int_0^1 ydy\left[\ln\frac{d_0}{d} + \frac{a}{d} - \frac{a_0}{d_0}\right]. \end{aligned} \tag{40}$$

The contribution of the Pauli formfactor Γ_2 is proportional to B and is therefore suppressed by the factor $|m^2/t|$.

APPENDIX B

Soft Pion Pair Production

The general expression for the soft pion pair production is

$$\begin{aligned} \left|\frac{M}{M_0}\right|^2 d\Gamma_\pm &= \left(\frac{4\pi\alpha}{q^2}\right)^2 d^4q\int\frac{d^3q_+}{2\varepsilon_+}\int\frac{d^3q_-}{2\varepsilon_-}(2\pi)^{-6} \\ &\times\delta^4(q_+ + q_- - q)(q_+ - q_-)_\mu(q_+ - q_-)_\nu J_\mu J_\nu, \\ J_\mu &= (Q_q^{p_1p_1'})_\mu, \end{aligned}$$

where

$$m_\mu \ll \sqrt{q^2} \ll \Delta\varepsilon \ll \varepsilon, \quad q_0^2 \gg q^2.$$

Here, $q_\pm$ is the 4-momentum, and $\epsilon_\pm$ is the energy of $\pi^\pm$; q is the 4-momentum, and q_0 is the energy of the soft pair.

Rewriting

$$\int d^4q = \frac{4\pi}{2}\int dq^2\frac{d\Omega_q}{4\pi}\int_{\sqrt{q^2}}^{\Delta\varepsilon} dq_0\sqrt{q_0^2 - q^2},$$

$$\begin{aligned} &\int\frac{d^3q_+}{2\varepsilon_+}\int\frac{d^3q_-}{2\varepsilon_-}(2\pi)^{-6}\delta^4(q_+ + q_- - q)(q_+ - q_-)_\mu(q_+ - q_-)_\nu \\ &= \frac{1}{3}\left(g_{\mu\nu} - \frac{q_\mu q_\nu}{q^2}\right)\times 2^{-7}\pi^{-5}(4m^2 - q^2)\sqrt{1 - \frac{4m^2}{q^2}}, \end{aligned}$$

we obtain

$$\left|\frac{M}{M_0}\right|^2 d\Gamma_\pm = -\frac{\alpha^2}{4\pi^3}\frac{(q^2 - 4m^2)^{3/2}}{(q^2)^3}\int\frac{d\Omega_q}{4\pi}\int dq_0\sqrt{q_0^2 - q^2}J^2,$$

where

$$J^2 = J_\mu J_\nu\left(g_{\mu\nu} - \frac{q_\mu q_\nu}{q^2}\right).$$

Separate contributions are given by

$$\int\frac{d\Omega_q}{4\pi}\frac{m^2}{(p_1q)^2} = O\left(\Delta^2\frac{m^2}{q^2}\right),$$

$$\int\frac{d\Omega_q}{4\pi}\frac{p_1p_2}{p_1qp_2q} = \int\frac{d\Omega_q}{4\pi}\frac{p_1'p_2'}{p_1'qp_2'q} = \frac{1}{2}\ln^2\left(\frac{2\Delta\varepsilon}{\sqrt{q^2}}\right) - \ln 2$$

and the master integral is

$$\begin{aligned} &\int_{\sqrt{q^2}}^{\Delta\varepsilon} dq_0\sqrt{q_0^2 - q^2}\int\frac{d\Omega_q}{4\pi}\frac{p_1p_1'}{p_1qp_1'q} \\ &= \ln^2\left(\frac{2\Delta\varepsilon}{\sqrt{q^2}}\right) + \ln\left(\frac{2\Delta\varepsilon}{\sqrt{q^2}}\right)\ln\left(\frac{1-c}{2}\right) - \zeta_2. \end{aligned}$$

The soft pion pair production contribution to the invariant mass distribution (from the emission of both electron and muon blocks) is given by

$$\frac{M^2}{\sigma_0}\frac{d\sigma}{dM^2} = \frac{\alpha^2}{3\pi^2}\left[\ln^2\frac{-t}{M^2} + \ln\frac{-t}{M^2}\ln\frac{\Delta\varepsilon}{\varepsilon} + O(1)\right].$$

ACKNOWLEDGMENTS

This work was supported in part by the Russian Foundation for Basic Research (project no. 01-02-17437).

REFERENCES

1. G. Boyarkina and O. Boyarkin, Yad. Fiz. **60**, 683 (1997) [Phys. At. Nucl. **60**, 601 (1997)]; V. Barger, S. Pakwasa, and X. Tata, Phys. Lett. B **415**, 200 (1997).
2. A. Arbuzov, E. Kuraev, and B. Shaikhatdenov, Mod. Phys. Lett. A **13**, 2305 (1998); hep-ph/9806215.
3. E. Glover, J. Tausk, and J. van der Bij, Phys. Lett. B **516**, 33 (2001).
4. S. Catani, Phys. Lett. B **427**, 161 (1998).
5. G. Faldt and P. Osland, Nucl. Phys. B **413**, 16 (1994); **413**, 64 (1994); Erratum: **413**, 404 (1994).
6. V. A. Smirnov and O. L. Veretin, Nucl. Phys. B **566**, 469 (2000).
7. R. Barbieri, J. A. Mignaco, and E. Remiddi, Nuovo Cimento A **11**, 824 (1972).
8. J. Schwinger, Phys. Rev. **76**, 790 (1949).
9. E. A. Kuraev and V. S. Fadin, Yad. Fiz. **45**, 782 (1987) [Sov. J. Nucl. Phys. **45**, 486 (1987)].
10. A. B. Arbuzov *et al.*, Yad. Fiz. **60**, 673 (1997) [Phys. At. Nucl. **60**, 591 (1997)].
11. A. B. Arbuzov and E. A. Kuraev, Fiz. Élem. Chastits At. Yadra **27**, 1247 (1996) [Phys. Part. Nucl. **27**, 510 (1996)].

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 413–420.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 481–489.

ATOMS, SPECTRA, RADIATION

A Uniform Quasi-Classical Description of Radiative Transitions for Asymmetric Rare-Gas Atom–Atom Collisions

A. Z. Devdariani[a], A. L. Zagrebin[a], F. Rebentrost[b], S. I. Tserkovnyi[†c], and E. A. Tchesnokov[a, *]

[a]*Institute of Physics, St. Petersburg State University, Peterhof, St. Petersburg, 198904 Russia*

**e-mail: tchesn@ec8174.spb.edu*

[b]*Max-Planck-Institut für Quantenoptik, Garching bei München, 85748 Germany*

[c]*Baltic State Technical University, St. Petersburg, 198905 Russia*

Received December 25, 2001

Abstract—The uniform quasi-classical approximation [14] is used to describe the optical spectra formed during asymmetric collisions between atoms of rare gases in which one of the atoms is in a metastable state. We consider the reactions $He(2^1S) + Ne \longrightarrow He(1^1S) + Ne + \hbar\omega$ and $Ar(^3P_2) + He \longrightarrow Ar(^1S) + He + \hbar\omega$, in which the optical transition mechanisms are typical of most rare gases. Quasi-molecular terms of excited states and radiative widths calculated in a unified semiempirical approach are used. Spectral characteristics are calculated for thermal collision energies in the entire frequency range, including the center and both wings of the forbidden line. For the blue wing, our results are consistent with the widely used Condon approximation at collision energies $E \geq 200$ cm^{-1}. At lower collision energies and in the region of the red wing and center of the forbidden line, the spectral distributions that cannot be described in the Condon approximation are reproduced in the uniform quasi-classical approximation. Comparison with quantum-mechanical calculations by the strong-coupling method confirms the high accuracy of the uniform quasi-classical approximation in the entire range of radiation frequencies. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

This study[1] is devoted to optical transitions (in absorption or emission) that are forbidden in isolated atoms but are permitted in quasi-molecules produced by two colliding atoms. A characteristic example of such transitions is the radiative decay of metastable atoms during collisions with atoms in normal states. Such transitions can now be investigated experimentally [1, 2]. Moreover, these are even easier to investigate than forbidden atomic transitions, because the lifetime of metastable atomic states is limited by the corresponding quasi-molecular optical or nonadiabatic transition. Whereas calculating forbidden atomic transitions requires invoking high orders of perturbation theory, the radiative widths of quasi-molecular transitions can be calculated even in the first order of perturbation theory (see, e.g., [1]). This is because they are comparable to the radiative widths (probabilities) of resonance atomic transitions at mean internuclear distances.

The main difficulty in theoretically analyzing quasi-molecular transitions stems not only from the variety of term structures and dependences of the radiative widths on internuclear distances in various quasi-molecules. The collisionally broadened line of a permitted transition at frequency ω_0 is known [3] to have a Lorentz profile near ω_0 with far wings whose intensity decreases as a power law, $1/(\omega - \omega_0)^s$, $s > 1$. For forbidden transitions, there is no such view of the general spectral patterns.

Below, we develop a quasi-classical theory of quasi-molecular optical transitions by using asymmetric collisions between rare-gas atoms for which the metastable states belong to the $1s2s$ and $np^5(n+1)s$ configurations (the $1s_3$ and $1s_5$ states in the Paschen notation) as an example. Since the transition mechanisms are different even within one group of elements, we consider two reactions,

$$\begin{aligned} &He(1s2s, {}^1S_0) + Ne({}^1S_0) \\ &\longleftrightarrow He(1s^2, {}^1S_0) + Ne({}^1S_0) + \hbar\omega, \end{aligned} \quad (1)$$

[†]Deceased.

[1]Some of its results were reported at the First International Workshop on the physics of electron and atomic collisions (March 2001, Klyaz'ma). We are grateful to the organizers of the workshop for the invitation and to its participants for a helpful discussion.

and

$$\mathrm{Ar}(3p^5 4s,\ ^3P_2) + \mathrm{He}(^1S_0) \longleftrightarrow \mathrm{Ar}(3p^6,\ ^1S_0) + \mathrm{He}(^1S_0) + \hbar\omega. \quad (2)$$

These reactions illustrate the main cases and peculiarities of optical transitions in asymmetric rare-gas quasi-molecules. Note that our results can also be extended to collisions between second-group atoms with the excited *nsnp* configuration and rare-gas ground-state atoms. As any problem of atom–atom collisions at low energies, analysis of the above reactions includes two problems. The first (static) problem consists in determining the quasi-molecular terms and radiative widths (Section 2) and the second (dynamic) problem consists in calculating the spectral shape (Section 3), including the spectra averaged over the collision parameters (Section 4).

2. THE MECHANISMS OF QUASI-MOLECULAR OPTICAL TRANSITIONS

The quasi-molecular terms of interest in the reaction (1) were extensively studied in connection with the first gas laser; the results are summarized in [4]. The experimental data from [5] are most reliable for the ground-state term. These data can be fitted by

$$U_f(R) = 2594\exp(-3.439R).$$ [2]

For the excited-state term produced by the He($1s2s$, $^1S_0$) atomic configuration, we can use the data from [6], which can also be well fitted by an exponential function,

$$U_i(R) = 0.404\exp(-0.917R).$$

For $R_0 \approx 6$, $U_{i0} \equiv U_i(R_0) \approx 260$ cm^{-1} (the term energy is measured from its asymptotic value), this term of 0^+ symmetry intersects the quasi-molecular term of the same symmetry produced by the He($1s^2$)–Ne($2p^5 5s$, $^1P_1$) configuration. This intersection governs the laser level population [7]. At collision energies $E < U_{i0}$, the excitation transfer channel is closed and the quasi-molecular optical transition in the vacuum ultraviolet remains the only collisional quenching channel for the He(2^1S) metastable states.

The radiative width of the 0^+–0^+ transition is determined by the interaction between atoms, which results in the mixing of the $|\mathrm{He}(1s2s,\ ^1S)\rangle|\mathrm{Ne}(2p^6)\rangle$ and $|\mathrm{He}(1snp,\ ^1P)\rangle|\mathrm{Ne}(2p^6)\rangle$ diabatic quasi-molecular states of the same 0^+ symmetry. Here, $|\mathrm{He}\rangle$ and $|\mathrm{Ne}\rangle$ are the corresponding atomic wave functions. The radiative width calculated in terms of the perturbation theory by taking into account the short- and long-range interactions is given in [8, 9]. These data can be fitted by

$$\Gamma(R) = 4.84\times10^{-5}\exp(-1.84R).$$

Naturally, the interaction with the nearest $n = 2$ configuration mainly contributes to the width. Thus, at collision energies lower than 260 cm^{-1}, calculating the spectral characteristics reduces to the problem of optical transitions between two repulsive terms one of which has the radiative width that exponentially depends on distance. Given the 0^+ symmetry of both terms, the formulated problem, probably, represents the simplest example of a quasi-molecular optical transition.

Let us consider the reaction (2). The ground-state term was determined in [10], and it can be fitted by the exponential function $U_f(R) = 0.01\exp(-0.7R)$. The results of calculations for the excited-state term and the corresponding radiative width are summarized in [11]. Of the five excited states (0^-, ±1, ±2) that correspond to the 0^-, 1, and 2 terms produced by the Ar($4s$, $^3P_2$) metastable state, only two states (+1 and –1) corresponding to the 1 term are coupled with the ground state of 0^+ symmetry by an optical transition. Since the dipole matrix elements for the transitions between the states $+1$–0^+ and -1–0^+ are identical, when calculating the spectral distributions, we can take into account the degeneracy of the initial term by introducing the statistical weight of this term $g = 2/5$ in the expression for the radiative transition cross section. In this case, by the probability P we mean the transition probability between two fixed quasi-molecular states.

If we ignore the interaction with other configurations, then there are two more terms of 1-symmetry within the initial excited Ar($3p^5 4s$)–He($1s^2$) configuration. The dependence of the adiabatic term energies on internuclear distance can be found by diagonalizing the Hamiltonian H_{ik} constructed in the basis of diabatic functions $|\mathrm{Ar}(4s,\ ^{1,3}P_J)\rangle|\mathrm{He}(1s^2)\rangle$ made up from the products of *LS*-coupling atomic wave functions with a unit component of the total angular momentum along the internuclear axis:

[2] Unless stated otherwise, we use the atomic system of units.

$$H_{ik} = \begin{array}{|c|c|c|c|}\hline & |^1P_1\rangle & |^3P_1\rangle & |^3P_2\rangle \\ \hline |^1P_1\rangle & U_\sigma + V_\Pi + \frac{2}{3}G^1 + \frac{\zeta}{2} & -\frac{\zeta}{2} & 0 \\ \hline |^3P_1\rangle & -\frac{\zeta}{2} & U_\sigma + \frac{V_\Pi}{2} + \frac{V_\Sigma}{2} + \zeta & \frac{\Delta V}{2} \\ \hline |^3P_2\rangle & 0 & \frac{\Delta V}{2} & U_\sigma + \frac{V_\Pi}{2} + \frac{V_\Sigma}{2} \\ \hline \end{array} \quad (3)$$

Here, $\Delta V = V_\Pi - V_\Sigma$; $V_{\Pi,\Sigma}$ are the ion–atom interaction potentials in the Π and Σ states without spin–orbit coupling; U_σ is the matrix element of the part of the interaction operator that includes the interaction of an excited s electron with the He atom polarized by the Ar^+ field; G^1 and ζ are the Slater exchange integral and the spin–orbit coupling constant for the $Ar(3p^5 4s)$ configuration. For the last two quantities, we used their semiempirical values from [8]. The dependences $V_{\Pi,\Sigma}$ were restored from the ion potentials [12]. The matrix element U_σ was calculated by the pseudopotential method [13] using formulas from [8]. Although the required adiabatic term has a well D ~ 2 cm^{-1} in depth, the term may be assumed to be repulsive and to be fitted by the exponential function $U_i(R) = 0.17\exp(-0.7R)$ for collisions with energies of the order of 100 cm^{-1}. This fit is justified by the fact that the spectrum is formed by transitions in the range R ~ 7–9 (Section 3), where the exchange interaction between atoms gives a dominant contribution to the matrix element U_σ and, hence, to the adiabatic potential.

The adiabatic wave function $|\Omega = 1, {}^3P_2\rangle$ is determined simultaneously with the diagonalization of (3). Since only the $|^1P_1\rangle$ state is coupled with the ground state by an optical transition in the diabatic basis used, the radiative width of the adiabatic state is

$$\Gamma(R) = \frac{\Gamma_0}{a^2}\left|\langle 1^3P_2|^1P_1\rangle\right|^2 = \Gamma_0\frac{b^2}{4}|\Delta V(R)|^2\left(\frac{1}{\varepsilon_1} - \frac{1}{\varepsilon_2}\right)^2. \quad (4)$$

Here, Γ_0 is the width of the $^1P_1$–$^1S_0$ resonance transition in an isolated Ar atom, and $a = -0.892$ and $b = 0.456$ are the amplitudes for the decomposition of the intermediate-coupling atomic wave eigenfunctions in the Ar atom into LS-coupling functions [8]. Since $|\Delta V| \ll \zeta, G^1$ in the range of distances under consideration, we can take into account the spin–orbit coupling exactly and the ion–atom interaction in the first order of perturbation theory. The second part of formula (4) was derived in this way; $\varepsilon_1(\varepsilon_2)$ are the splittings between the s_2 and s_5 (s_4 and s_5) atomic levels. In contrast to reaction (1), a radiative width arises here because of the interaction within one configuration. Since the width in the range of distances under consideration is determined by the exchange Ar^+–He ion–atom interaction, the exponential fit $\Gamma(R) = 0.486\exp(-4R)$ is again justified for it.

Thus, for the two reactions in the range of distances responsible for the spectrum formation, the quasi-molecular terms correspond to repulsion between atoms and they can be fitted by exponential functions, as the radiative widths. This circumstance is important, because it allows us to use the uniform quasi-classical approximation [14] to calculate the spectral characteristics. The quasi-classical approximation is preferred to the semiclassical approximation, which is based on the use of a single classical trajectory, for the following reason. Note that the characteristic scale for the energy of quasi-molecular terms $U_{i,f}$ and energy defect $\Delta\omega$ is 10^{-3}; i.e., this is precisely the order of the spectrum extent. However, the collision energy is of the same order of magnitude. Therefore, the legitimacy of introducing a single classical trajectory is called into question. In contrast, using the quasi-classical approach allows us to avoid inaccuracies related to the improper introduction of a trajectory.

3. THE SPECTRAL RADIATIVE-TRANSITION PROBABILITY DISTRIBUTIONS

Calculating the spectral distributions $dP^l/d\omega = |W|^2$ for the emission probability of a photon with frequency ω in the first order of perturbation theory, when analysis can be restricted to the interaction between only two states, reduces to calculating the integral

$$W = 2\pi\int_0^\infty \Psi_i^l(R)V(R)\Psi_f^l(R)dR. \quad (5)$$

Here, $\Psi_{i,f}^l(R)$ are the real regular (at zero) solutions to the radial Schrödinger equations in the initial and final channels normalized to the δ function of energy, and $V(R) = \sqrt{\Gamma(R)/2\pi}$. When the wave functions $\Psi_{i,f}^l(R)$ are mainly quasi-classical, the potentials $U_{i,f}(R)$ are monotonically repulsive and the interaction is $V(R) = V_0(R)\exp(-\gamma R)$, where $V_0(R)$ is a smooth function of R; the overlap integral (5) was analyzed in detail in [14,

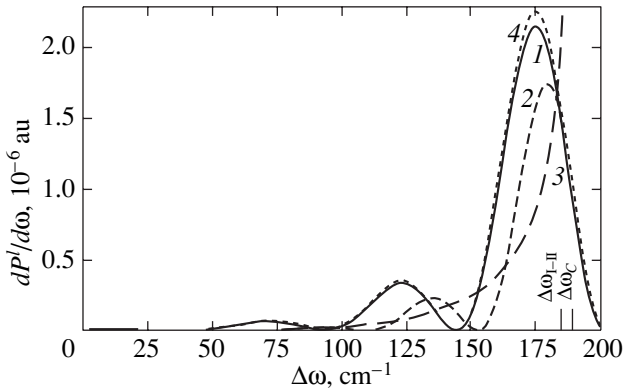


Fig. 1. Transition probability density $dP^l/d\omega$ versus frequency detuning $\Delta\omega$ for the ArHe quasi-molecule at E = 200 cm^{-1} and l = 0. (*1*) A uniform fit in the exponential-interaction approximation (7); (*2*) a uniform fit in the constant-interaction approximation (14); (*3*) the Landau approximation (16); (*4*) the exact calculation by the distorted-wave method (5); $\Delta\omega_{\text{I–II}}$ is the boundary between domains I and II, and $\Delta\omega_C$ is the boundary between the subbarrier (right) and suprabarrier (left) transitions in the constant-interaction approximation.

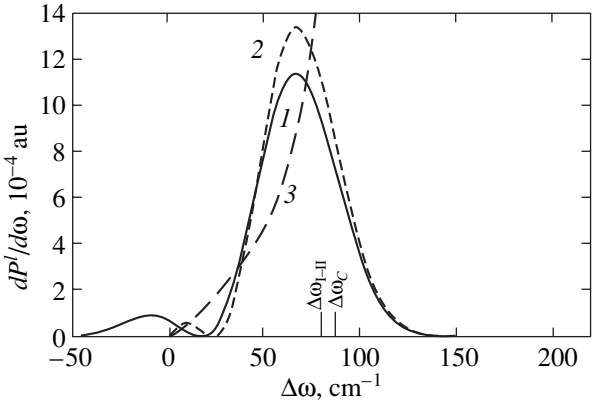


Fig. 2. Transition probability density $dP^l/d\omega$ versus frequency detuning $\Delta\omega$ for the HeNe quasi-molecule at E = 220 cm^{-1} and l = 20. The notation is the same as in Fig. 1.

15]. We established that, in this case, integral (5) is determined by the contribution from two saddle points, R_+ ($\mathrm{Im}(R_+) \geq 0$) and R_- ($\mathrm{Im}(R_-) \leq 0$), which are the roots of the equations

$$\begin{aligned}(k_i - k_f)(R_+) &= i\gamma,\\ (k_i - k_f)(R_-) &= -i\gamma\end{aligned} \tag{6}$$

closest to the real axis. Here, $k_{i,f}$ are the classical momentum functions for motion in the effective potentials $U_{i,f}(R) + J^2/2\mu R^2$ with energies $E_{i,f}$, $J = l + 1/2$. Based on the results from [14, 15], we obtain the following quasi-classical expression for the probability density of optical transitions:

$$\frac{dP^J}{d\omega} = \left|F^+X^+ + F^-X^-\right|^2 \exp(-2\mathrm{Im}A), \tag{7}$$

$$X^\pm = B^{1/4}\mathrm{Ai}(-B) \pm iB^{-1/4}\mathrm{Ai}'(-B). \tag{8}$$

Here, Ai and Ai' are the Airy function and its derivative [16],

$$F^\pm = 2^{1/2}\pi\mu V_0(R_\pm)[k_i k_f (k_i' - k_f')(R_\pm)]^{-1/2}, \tag{9}$$

$$\begin{aligned}2A = {} & i\gamma(R_+ + R_-) + S_f(R_+)\\ & - S_f(R_-) - S_i(R_+) + S_i(R_-),\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{4}{3}B^{3/2} = {} & i\gamma(R_+ - R_-) + S_f(R_+)\\ & + S_f(R_-) - S_i(R_+) - S_i(R_-),\end{aligned} \tag{11}$$

while A and B are purely imaginary and purely real quantities, respectively. The derivatives with respect to internuclear distance R are denoted by a prime in formula (9). The symbols $S_{i,f}(R)$ were introduced to denote the classical action functions that correspond to the momenta $k_{i,f}(R)$ and are accumulated in the interval from the turning points $R_{i,f}$ to R. The characteristic spectral distributions calculated using formulas (6)–(11) for reactions (1) and (2) are shown in Figs. 1 and 2. Figure 1 also shows the exact result for the transition probability with l = 0 obtained in the distorted-wave approximation, i.e., the result of an exact calculation of the integral (5) with the wave functions that are the exact solutions to the Schrödinger equations.

Let us consider our results by drawing analogies, where possible, with the Landau–Zener, Demkov, and Nikitin [17] models widely used in the physics of atomic collisions. It should be emphasized that a merit of our approach is that it is not directly related to these models and includes them as limiting cases in the weak-coupling approximation. An additional advantage is the possibility of analyzing the dependence of the spectra on orbital angular momentum, which is generally also outside the scope of the models mentioned above. Still, the model approach is convenient for discussing the results. Therefore, Figs. 1 and 2 also show the results of our calculations in the constant-interaction approximation, $V(R) = V(R_C) \equiv V_C$, where the position of $R_\pm$ coincides with the position of the Condon point R_C, which can be determined from the equation

$$\Delta\omega = \Delta U(R_C). \tag{12}$$

Here, $\Delta\omega$ is the detuning of the optical transition frequency ω from the frequency of the forbidden atomic

transition $\omega_0 = 60.1$ nm for He($1s2s$, ^1S) and 107.4 nm for Ar($3p^5 4s$, $^3P_2$),

$$\Delta\omega = \omega - \omega_0 = \frac{k_{0i}^2 - k_{0f}^2}{2\mu}, \quad (13)$$

and $k_{0i,f}$ are the momenta before and after collision. The expression for the transition probability density in the case of constant interaction is

$$\frac{dP^J}{d\omega} = 2\pi \frac{dP_{LZ}}{d\omega} B^{1/2} \mathrm{Ai}^2(-B), \quad (14)$$

$$\frac{2}{3} B^{3/2} = S_f(R_C) - S_i(R_C), \quad (15)$$

and it matches Miller's formula [18], which was proven, for example, in [19]. For a linear fit to the terms near the point of intersection, the approximation under consideration was analyzed in detail by Nikitin and coworkers (see, e.g., [17]). In formula (14),

$$\frac{dP_{LZ}}{d\omega} = \frac{4\pi\mu V_C^2}{k_C \Delta F} \quad (16)$$

is the Landau probability [20] calculated in the semiclassical approximation,

$$\Delta F = (U_f' - U_i')(R_C), \quad k_C = k_i(R_C) = k_f(R_C).$$

For comparison, Figs. 1 and 2 also show our calculation using formula (16). It may be called a calculation of the probability density in the quasi-static approximation, because, after integration over the collision parameters and averaging over the velocities, the probability (16) leads to the standard formula of the quasi-static approximation [3].

Let us discuss the results of our calculations in terms of the positions of the saddle points. As was shown in [14], the existence of an imaginary part for the saddle points $R_\pm$ and the choice of branches for the functions of complex variable $k_i(R_+)$ and $k_f(R_+)$ determine four distinct domains of parameters. In Fig. 3, the real parts of the saddle points are plotted against $\Delta\omega$ for the reaction (2) (plots of the same type are shown for both reactions only when the functional dependences for them differ significantly). In domain I (see also Fig. 4), the two complex-conjugate saddle points $R_\pm$ can be determined directly from Eqs. (6). Their contributions to the probability density are of the same order, and interference leads to an oscillatory behavior of the probability in this domain. The probability density in this case can be calculated using formulas (7)–(11). As the kinetic energy of the atoms in the transition region decreases, which is achieved by an increase in $\Delta\omega$ and centrifugal energy, the saddle points merge together (the so-called fold-type catastrophe [21]), whereupon they become real (in domains II–IV). The saddle points are close to each other when the parameters are near the boundary between domains I and II. This leads to a rainbow pattern in the behavior of the transition probability density and allows the boundary between domains I and II, $\Delta\omega_{\mathrm{I-II}}$, specified by the condition $R_+ = R_- = R_s$, to be treated as the boundary between the suprabarrier and subbarrier transitions for an exponential interaction between states. For constant interaction, the location of this boundary $\Delta\omega_C(J)$ is specified by the condition $R_C = R_i = R_f$.

With a further increase in $\Delta\omega$, as one recedes from domain I, the contribution from the point R_- to the transition probability density becomes exponentially small

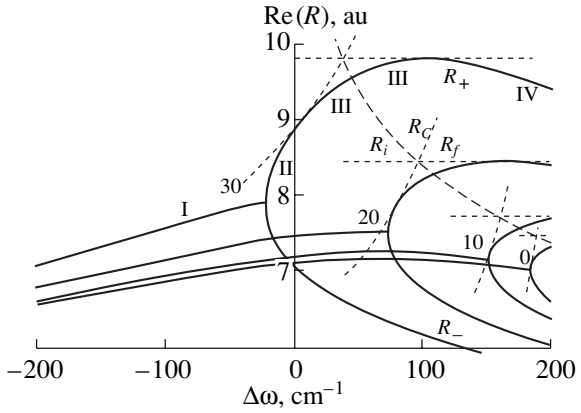


Fig. 3. Real parts of the saddle points Re(R_+) and Re(R_-) versus frequency detuning $\Delta\omega$ (solid curves). The calculation was performed for the ArHe quasi-molecule at E = 200 cm^{-1}. The numbers alongside the curves give angular momenta l. The dashed and dotted lines specify the positions of the Condon point R_C and the turning points R_i and R_f, respectively. The Roman numerals I–IV highlight the domains of $\Delta\omega$ according to the classification from [14].

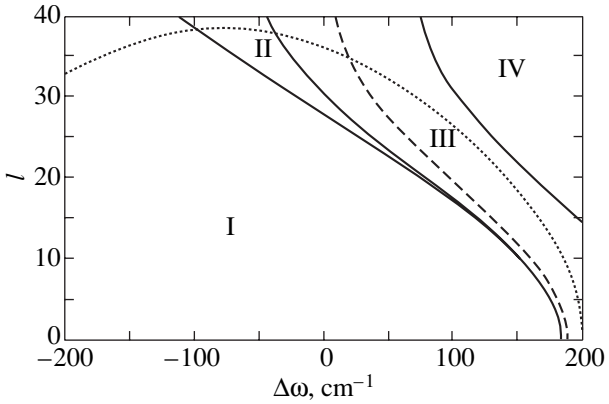


Fig. 4. The locations of domains I–IV in the ($\Delta\omega$, l) plane for E = 200 cm^{-1} for the ArHe quasi-molecule. The dashed line indicates the boundary between the subbarrier and suprabarrier transitions in the constant-interaction approximation (12); the dotted line indicates an approximate upper boundary for the domain of angular momenta l of importance in calculating the cross section in the quasi-classical approximation.

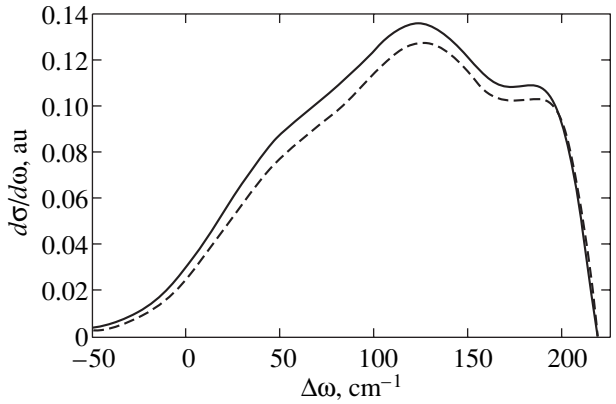


Fig. 5. The spectral density of the emission cross section for the HeNe quasi-molecule at a collision energy of 220 cm^{-1}. The solid and dashed curves represent the close-coupling approximation [22] and the uniform quasi-classical approximation (7), (17), respectively.

against the background of the contribution from the point R_+ (which, in turn, also decreases exponentially). As a result, the intensity monotonically decreases exponentially with increasing $\Delta\omega$. In domains II–IV, the positions of the saddle points, which are now real, are specified by the corresponding analytic continuations of Eq. (6) and the matrix element W is determined by the analytic continuations of Eqs. (7)–(11). The required formulas are given in [14]. Note the following misprint in [14]: there must be a minus before the action $\sigma_f(R_+)$ in formula (34). In general, the optical transitions that form the spectral probability density distribution near its high-frequency boundary $\Delta\omega = E$ may be said to occur when atoms move in the classically forbidden region (for the f channel) of internuclear distances. This causes an exponential decrease in the probability as the high-frequency boundary is approached.

Let us return to the discussion of domain I ($\Delta\omega < \Delta\omega_{I-II}$). As we see from Figs. 1 and 2, the uniform approximation (7) is in good agreement with the exact quantum-mechanical result in the entire frequency range, while the constant-interaction approximation (14), which fits the transition probabilities in the frequency range adjacent to the upper boundary of the spectrum with a small relative error, leads to a significant error at lower frequencies. This error increases with decreasing $\Delta\omega$ as the domain $\Delta\omega \le 0$ is approached; it is attributable to the violation of the Franck–Condon principle. In this case, according to (7) and (10), the oscillation amplitude exponentially decreases with decreasing $\Delta\omega$. For the frequency range adjacent to the upper boundary of the spectrum $\Delta\omega = E$, the probability rapidly decreases with increasing orbital angular momentum l, which is attributable to the subbarrier nature of the wave function Ψ_f^l in the transition region. In contrast, the classically permitted motion of atoms in the transition region is characteristic of low frequencies $\Delta\omega \le 0$. In this case, the probability decreases with increasing l due to a gradual displacement of the transition region toward larger internuclear distances, where the radiative width is smaller, more slowly. Thus, a larger number of partial waves are involved in the formation of the low-frequency domain (Fig. 4). In the constant-interaction approximation, the number of partial waves required to calculate the cross sections increases to infinity as $\Delta\omega \longrightarrow 0$.

At first glance, the mechanism of the transitions that form the domain $\Delta\omega < 0$ could be described, at least qualitatively, by Demkov's model based on parallel, exponentially interacting terms. However, this is not possible, because we consider transitions in the weak-coupling limit; i.e., the strong-coupling region in which $\Delta\omega \approx V$ and which is responsible for the transitions in Demkov's model lies at much smaller internuclear distances than the actual transition region.

4. THE SPECTRAL DISTRIBUTIONS FOR THE RADIATIVE-TRANSITION CROSS SECTIONS

The spectral distributions for the emission or absorption cross sections can be determined from the spectral transition probability densities by the summation over partial waves. For the emission to a nondegenerate term, we have

$$\frac{d\sigma}{d\omega} = g\frac{\pi}{k_{0i}^2}\sum_{l=0}^{\infty}(2l+1)\frac{dP^l}{d\omega}, \qquad (17)$$

where g is the statistical weight of the excited term.

Figure 5 shows the spectral distribution of the emission cross section for the HeNe quasi-molecule calculated both from formulas (7) and (17) and in the exact quantum-mechanical approach [22, 23]. In the latter case, the channel close-coupling method was used for scattering in the potentials U_i and $U_f + \hbar\omega$. The results of these calculations are in good agreement at all frequencies, including $\Delta\omega \le 0$. It should be noted that the change in the angular momentum of a system of colliding atoms due to their interaction with the emitted photon taken into account in the quantum-mechanical theory is completely described in terms of the dipole coupling between states. For the $|0^+, 2^1S_0\rangle \longrightarrow |0^+, 1^1S_0\rangle$ emission under consideration, the matrix elements of the $l \longrightarrow l \pm 1$ transitions should be taken into account. Good quantitative agreement between the quantum-mechanical result and the quasi-classical result, for which the change of l by one was disregarded, shows that, for the spectrum summed over the angular momentum (17), allowance for the change of l is unimportant for the collision energy under consideration. Nevertheless, the change of l should be taken into account when considering partial probabilities, which leads to the characteristic decrease in oscillation ampli-

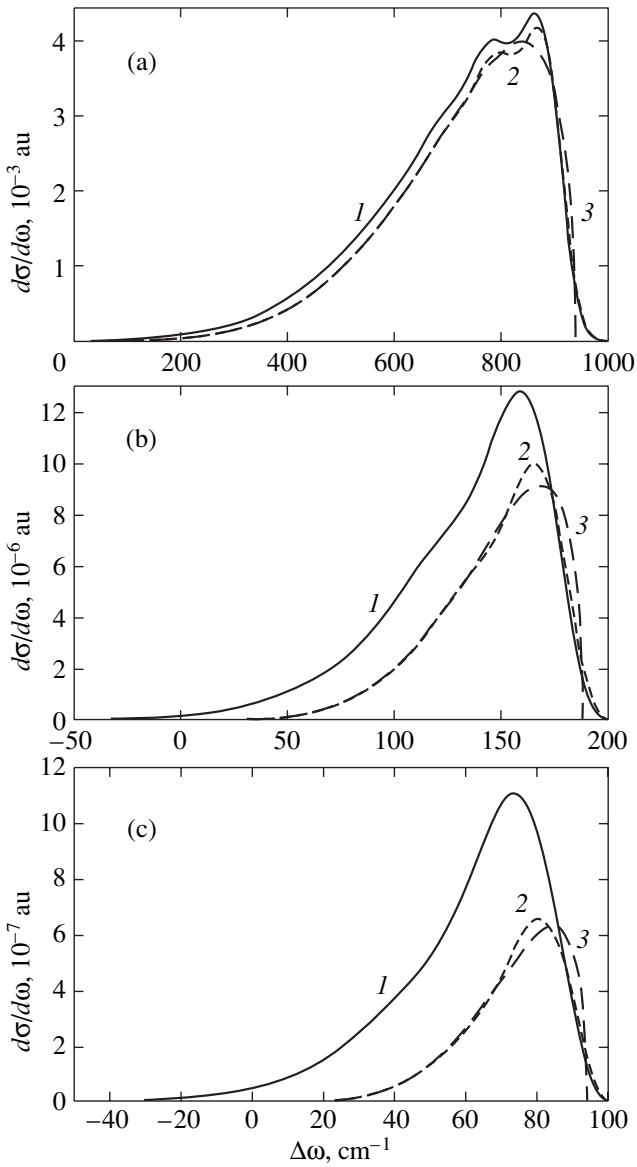


Fig. 6. The spectral density of the emission cross section for the ArHe quasi-molecule at collision energies (a) 1000, (b) 200, and (c) 100 cm^{-1}. The notation is the same as in Fig. 1.

tude in both quasi-classical and quantum-mechanical calculations. Such a more rigorous treatment is needed when analyzing low-temperature collisions.

Our calculations of the emission cross section for the ArHe quasi-molecule (Fig. 6) show the influence of collision energy on the quality of approximations (14) and (16). As we see from the plots, the differences between the approximations used that we pointed out when analyzing the transition probabilities are also retained for the cross sections. As the collision energy decreases, an increasingly large part of the spectrum is shifted to $\Delta\omega < 0$ and to the center of the forbidden line, with the constant-interaction approximation becoming increasingly inaccurate.

Note also that the uniform approximations allow us to trace the oscillation remnants retained after the summation over partial waves; they are clearly seen in Fig. 6a.

5. CONCLUSION

We have been able to trace the formation dynamics of optical quasi-molecular spectra forbidden in the limit of separated atoms by using the uniform quasi-classical approximation. The central point in the analytical description is analysis of the positions of the roots for Eq. (6) that generalize the equation for the Condon point (12) to the radiative width that exponentially depends on distance.

The peak in the distribution is formed by transitions with small values of l near the Condon point. The frequency range near the upper boundary of the spectrum is formed by optical transitions when atoms move in the classically forbidden (for the f channel) region of internuclear distances. A significant number of partial waves contribute to the domain $\Delta\omega \le 0$; optical transitions occur with the violation of the Franck–Condon principle.

The results of our calculations for the HeNe and ArHe quasi-molecules, typical of the metastable states for rare-gas and group-II atoms with an excited outer $nsnp$ shell, are in good agreement with the exact quantum-mechanical calculations. A merit of the uniform quasi-classical approach is that it does not use any standard models of atomic collisions with a distinctive distance dependence of the quasi-molecular terms and explicitly includes the dependence on orbital angular momentum. The latter circumstance is important, because, even for intersecting terms, an increase in orbital momentum qualitatively changes the pattern of spectrum formation. For low l, such transitions can be roughly described in terms of the model of a constant radiative width near the Condon point. For high l, one should use the model of almost parallel terms [14] with one real saddle point located near the turning points.

ACKNOWLEDGMENTS

This study was supported by the INTAS (grant no. 99-00039). A.Z. Devdariani, A.L. Zagrebin, and E.A. Chesnokov are also grateful to the NATO (grant no. CLG 977379) and the Russian Foundation for Basic Research (project no. 99-03-33168a) for partial support.

REFERENCES

1. T. Kurosawa, K. Ohmori, H. Chiba, *et al.*, J. Chem. Phys. **108**, 8101 (1998).
2. E. Bichoutskaia, A. Devdariani, K. Ohmori, *et al.*, J. Phys. B **34**, 2301 (2001).
3. I. I. Sobel'man, *Atomic Spectra and Radiative Transitions* (Fizmatgiz, Moscow, 1963; Springer-Verlag, Berlin, 1979).

4. A. Z. Devdariani, A. L. Zagrebin, and K. B. Blagoev, Ann. Phys. (Paris) **17**, 365 (1992).
5. B. Keil, L. J. Danielson, U. Buck, *et al.*, J. Chem. Phys. **89**, 2866 (1988).
6. H. Haberland, W. Konz, and P. Oesterlin, J. Phys. B **15**, 2969 (1982).
7. A. Z. Devdariani and A. L. Zagrebin, Zh. Éksp. Teor. Fiz. **86**, 1969 (1984) [Sov. Phys. JETP **59**, 1145 (1984)].
8. A. Z. Devdariani, A. L. Zagrebin, and K. B. Blagoev, Ann. Phys. (Paris) **14**, 467 (1989).
9. A. L. Zagrebin and S. I. Tserkovnyĭ, Opt. Spektrosk. **79**, 556 (1995) [Opt. Spectrosc. **79**, 511 (1995)].
10. K. M. Smith, A. M. Rulis, and G. Scoles, J. Chem. Phys. **67**, 152 (1977).
11. A. L. Zagrebin and N. A. Pavlovskaya, Opt. Spektrosk. **66**, 996 (1989) [Opt. Spectrosc. **66**, 582 (1989)].
12. L. Brunetti, F. Vecchiocattivi, and A. Aquilar-Navarro, Chem. Phys. Lett. **126**, 245 (1986).
13. G. K. Ivanov, Teor. Éksp. Khim. **14**, 610 (1978).
14. A. Z. Devdariani, Zh. Éksp. Teor. Fiz. **96**, 472 (1989) [Sov. Phys. JETP **69**, 266 (1989)].
15. A. Z. Devdariani and E. A. Chesnokov, Khim. Fiz. **17**, 57 (1998).
16. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and I. A. Stegun (National Bureau of Standards, Washington, 1964; Nauka, Moscow, 1979).
17. E. E. Nikitin and S. Ya. Umanskiĭ, *Theory of Slow Atomic Collisions* (Atomizdat, Moscow, 1979; Springer-Verlag, New York, 1984).
18. W. H. Miller, J. Chem. Phys. **48**, 464 (1968).
19. J. N. L. Connor, J. Chem. Phys. **74**, 1047 (1981).
20. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 3: *Quantum Mechanics: Non-Relativistic Theory* (Pergamon, New York, 1977, 3rd ed.; Nauka, Moscow, 1989, 4th ed.).
21. T. Poston and I. Stewart, *Catastrophe Theory and Its Applications* (Pitman, London, 1978; Mir, Moscow, 1980).
22. F. Rebentrost, S. Klose, and J. Grosser, Eur. Phys. J. D **1**, 277 (1998).
23. K. C. Kulander and F. Rebentrost, J. Chem. Phys. **80**, 5623 (1984).

Translated by V. Astakhov

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 421–428.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 490–499.

ATOMS, SPECTRA, RADIATION

Spectral Transitions from the Rydberg Autoionization States of a Li-Like Mg X Ion

I. Yu. Skobelev[a, *], A. Ya. Faenov[a], T. A. Pikuz[a], A. I. Magunov[a], F. Flora[b], S. Bollanti[b], P. DiLazzaro[b], D. Murra[b], A. Reale[c], L. Reale[c], G. Tomassetti[c], A. Ritucci[c], G. Petrocelli[d], S. Martellucci[d], N. Lisi[e], and F. B. Rosmej[f]

[a]*Center of Data on the Spectra of Multiply Charged Ions, State Scientific Center, All-Russia Research Institute of Physicotechnical and Radio Engineering Measurements, Mendeleevo, Moscow oblast, 141570 Russia*

**e-mail: skobelev@orc.ru*

[b]*ENEA, Dipartimento Innovazione, Settore Fisica Applicata, 00044 Roma, Italy*

[c]*Fis. Dept., Università de L. Aquila, gc LNGS of INFN, INFN, 67010 L'Aquila, Italy*

[d]*INFM-Dipartimento di Scienze e tecnologie Fisiche ed Energetiche, Università di Roma Tor Vergata, 00133 Roma, Italy*

[e]*ENEA, Unita Nuovi Materiali, C. R. Casaccia, 00060 Roma, Italy*

[f]*Technische Universität Darmastadt, Institut für Kernphysik, Abt. Strahlen- und Kernphysik, D-64289 Darmstadt, Germany*

Received March 29, 2002

Abstract—Satellite lines caused by radiative transitions from the Rydberg autoionization states of a Li-like Mg X ion in a plasma heated by radiation from a XeCl and a Nd laser are identified for the first time, and their wavelengths are measured precisely. Comparison of the experimental data with the atomic structures calculated by the method of relativistic perturbation theory shows that the accuracy of calculations of the energy of autoionization states is rather high even without the use of any semiempirical corrections and is of the order of 0.06%. The experimentally measured wavelengths can be used for a semiempirical estimate of the value of the leading order of perturbation theory among the orders that were neglected in calculations. It is shown that the simulation of the population kinetics of Rydberg autoionization states of Li-like ions in a dense plasma should take into account all possible channels of dielectronic capture, in particular, from the excited states of a He-like ion. The precision experimental wavelengths obtained for satellites of the He_β and He_γ lines of the Mg XI ion make possible to use these satellites as reference lines in studies of complicated spectra of multielectron ions.

1. INTRODUCTION

The radiative decay of autoionization levels of multiply charged ions results in emission of the so-called dielectronic satellites of their resonance-series lines. Such spectral transitions have been extensively studied over the past thirty years in both astrophysical and laboratory (mainly laser-produced) plasma (see, for example, [1–7]). As a result, at present the satellites of the resonance Ly_α and He_α lines of H- and He-like ions are investigated in great detail. They are caused by transitions from the autoionization levels $n_1l_1n_2l_2$ and $1sn_1l_1n_2l_2$ of two- and three-electron multiply charged ions with principal quantum numbers $n_1 = n_2 = 2$. The radiative decays of three-electron configurations with $n_1 = 2$ and $n_2 = 3$, resulting in the appearance of satellites of not only a resonance line itself but also of the second term He_β of the resonance series, have been studied to a lesser degree, but also quite thoroughly [8–13]. Transitions from highly excited (Rydberg) autoionization states with $n_1 + n_2 > 5$ have not been adequately investigated so far. However, as shown in [14, 15], it is these transitions that may dominate in emission spectra of a superdense and not too hot plasma. Such a situation can appear, for example, in a plasma produced by high-contrast subpicosecond laser pulses [16, 17] or in a plasma heated by short-wavelength laser pulses [14, 15]. In these cases, information on Rydberg satellites is very important for a correct identification of emission spectra of the plasma and diagnostics of its parameters.

The point is that the radiative decay of the $1sn_1l_1n_2l_2$–$1s^2n_2l_2$ Rydberg autoionization states results in emission of two groups of satellite lines $1sn_1l_1n_2l_2$–$1s^2n_2l_2$ and $1sn_1l_1n_2l_2$–$1s^2n_1l_1$. If we assume for definiteness that $n_1 < n_2$, then the first group of satellites will be located close to the corresponding resonance transition, i.e., on the wings of the $1sn_1l_1$–$1s^2$ lines, while the second group (for not too large values of n_1) can represent an isolated group of spectral lines located well away from the resonance $1sn_2l_2$–$1s^2$ transition. This means

Rydberg satellites of the resonance series of the Mg XI ion (λ is the wavelength; A is the radiative transition probability; Γ is the autoionization probability; Q_d is the factor determining the intensity of a satellite line under coronal conditions)

Line number	λ_{exp}, Å	λ_{thr}, Å	A, s^{-1}	Γ, s^{-1}	Q_d, s^{-1}	Transition
1	7.5013(6)	7.5023	3.53×10^{11}	4.33×10^{12}	1.94×10^{12}	$1s^22p^2P_{3/2}$–$1sp(^3P)6p^2D_{5/2}$
		7.5028	3.26×10^{11}	2.20×10^{12}	1.09×10^{12}	$1s^22p^2P_{1/2}$–$1sp(^3P)6p^2D_{3/2}$
2	7.5432(6)	7.5408	2.35×10^{11}	9.01×10^{10}	4.09×10^{09}	$1s^22p^2P_{3/2}$–$1s2p(^1P)5p^2P_{3/2}$
		7.5412	2.21×10^{11}	4.62×10^{12}	2.48×10^{11}	$1s^22p^2P_{3/2}$–$1s2p(^1P)5p^2D_{5/2}$
3	7.5468(5)	7.5458	7.88×10^{11}	1.27×10^{12}	9.16×10^{11}	$1s^22s^2S_{1/2}$–$1s2s(^3S)5p^2P_{1/2}$
		7.5458	7.87×10^{11}	1.28×10^{12}	1.84×10^{12}	$1s^22s^2S_{1/2}$–$1s2s(^3S)5p^2P_{3/2}$
4	7.5610(10)	7.5587	8.86×10^{11}	3.23×10^{11}	1.75×10^{11}	$1s^23d^2D_{5/2}$–$1s3d(^1D)4p^2F_{7/2}$
		7.5588	8.06×10^{11}	2.86×10^{11}	1.01×10^{11}	$1s^23d^2D_{3/2}$–$1s3d(^1D)4p^2F_{5/2}$
5	7.5630(6)	7.5606	8.17×10^{11}	1.93×10^{10}	3.15×10^{09}	$1s^23s^2S_{1/2}$–$1s3s(^3S)4p^2P_{3/2}$
6	7.5660(7)	7.5642	6.58×10^{11}	2.09×10^{07}	1.10×10^{07}	$1s^23d^2D_{5/2}$–$1s3p(^1P)4d^2D_{5/2}$
		7.5644	1.38×10^{11}	2.51×10^{12}	1.08×10^{10}	$1s^23p^2P_{1/2}$–$1s3d(^1D)4s^2D_{3/2}$
		7.5651	1.68×10^{11}	2.32×10^{12}	1.96×10^{10}	$1s^23p^2P_{3/2}$–$1s3d(^1D)4s^2D_{5/2}$
		7.5674	5.66×10^{11}	4.07×10^{09}	9.75×10^{08}	$1s^23p^2P_{3/2}$–$1s3s(^1S)4d^2D_{5/2}$
7	7.5760(8)	7.5741	2.17×10^{11}	6.61×10^{11}	1.08×10^{10}	$1s^23d^2D_{5/2}$–$1s3p(^1P)4s^2P_{3/2}$
8	7.5778(8)	7.5773	2.44×10^{11}	5.54×10^{11}	1.10×10^{10}	$1s^23p^2P_{1/2}$–$1s3s(^3S)4d^2D_{3/2}$
		7.5780	2.65×10^{11}	5.40×10^{11}	1.76×10^{10}	$1s^23p^2P_{3/2}$–$1s3s(^3S)4d^2D_{5/2}$
9	7.5840(5)	7.5802	4.76×10^{11}	5.76×10^{12}	2.62×10^{12}	$1s^22p^2P_{3/2}$–$1s2p(^3P)5p^2D_{5/2}$
		7.5803	1.42×10^{11}	1.77×10^{12}	7.77×10^{11}	$1s^22p^2P_{3/2}$–$1s2p(^3P)5f^2F_{5/2}$
10	7.6390(7)	7.6344	3.30×10^{11}	1.41×10^{13}	1.13×10^{12}	$1s^22s^2S_{1/2}$–$1s2s(^1S)4p^2P_{3/2}$
		7.6353	3.97×10^{11}	1.32×10^{13}	6.53×10^{11}	$1s^22s^2S_{1/2}$–$1s2s(^1S)4p^2P_{1/2}$
11	7.6450(6)	7.6423	2.49×10^{11}	1.76×10^{12}	6.04×10^{11}	$1s^22s^2S_{1/2}$–$1s2p(^3P)4s^2P_{3/2}$
		7.6439	1.98×10^{11}	2.91×10^{12}	3.20×10^{11}	$1s^22s^2S_{1/2}$–$1s2p(^3P)4s^2P_{1/2}$
12	7.6915(7)	7.6882	2.25×10^{11}	1.11×10^{13}	3.23×10^{11}	$1s^22p^2P_{1/2}$–$1s2p(^1P)4p^2D_{3/2}$
		7.6896	5.46×10^{11}	2.15×10^{11}	2.24×10^{10}	$1s^22p^2P_{3/2}$–$1s2p(^1P)4p^2P_{3/2}$
		7.6904	3.80×10^{11}	1.03×10^{13}	7.79×10^{11}	$1s^22p^2P_{3/2}$–$1s2p(^1P)4p^2D_{5/2}$
		7.6906	1.12×10^{11}	1.11×10^{13}	1.61×10^{11}	$1s^22p^2P_{3/2}$–$1s2p(^1P)4p^2D_{3/2}$
13	7.6978(5)	7.6962	1.46×10^{12}	1.53×10^{12}	1.39×10^{12}	$1s^22s^2S_{1/2}$–$1s2s(^3S)4p^2P_{1/2}$
		7.6962	1.46×10^{12}	1.57×10^{12}	2.82×10^{12}	$1s^22s^2S_{1/2}$–$1s2s(^3S)4p^2P_{3/2}$
14	7.7242(9)	7.7215	1.24×10^{12}	4.81×10^{12}	1.69×10^{12}	$1s^22p^2P_{3/2}$–$1s2p(^3P)4p^2S_{1/2}$
15	7.7304(5)	7.7279	1.31×10^{12}	1.45×10^{13}	7.13×10^{12}	$1s^22p^2P_{3/2}$–$1s2p(^3P)4p^2D_{5/2}$
		7.7296	3.95×10^{11}	1.54×10^{13}	1.41×10^{12}	$1s^22p^2P_{3/2}$–$1s2p(^3P)4p^2D_{3/2}$
16	7.7350(6)	7.7329	1.84×10^{11}	1.19×10^{11}	9.72×10^{10}	$1s^22p^2P_{1/2}$–$1s2p(^3P)4p^2P_{3/2}$
17	7.7379(7)	7.7346	9.74×10^{11}	2.59×10^{10}	3.14×10^{10}	$1s^22p^2P_{1/2}$–$1s2p(^3P)4p^2P_{1/2}$
18	7.8746(7)	7.8780	7.76×10^{11}	1.65×10^{10}	7.54×10^{09}	$1s^24f^2F_{7/2}$–$1s3d(^1D)4d^2G_{9/2}$
19	7.8759(7)	7.8784	7.10×10^{11}	1.71×10^{10}	5.09×10^{09}	$1s^24f^2F_{5/2}$–$1s3d(^1D)4d^2G_{7/2}$
20	7.8850(8)	7.8875	1.76×10^{12}	1.87×10^{12}	2.97×10^{10}	$1s^23d^2D_{3/2}$–$1s3p(^3P)3d^2F_{5/2}$
		7.8877	1.89×10^{12}	1.86×10^{12}	4.22×10^{10}	$1s^23d^2D_{5/2}$–$1s3p(^3P)3d^2F_{7/2}$
21	7.8972(6)	7.9008	3.08×10^{12}	4.54×10^{10}	3.23×10^{09}	$1s^23d^2D_{3/2}$–$1s3p(^1P)3d^2P_{1/2}$
		7.9012	2.63×10^{12}	5.00×10^{10}	6.07×10^{09}	$1s^23d^2D_{5/2}$–$1s3p(^1P)3d^2P_{3/2}$
22	7.9002(6)	7.9051	4.18×10^{12}	7.41×10^{08}	1.04×10^{08}	$1s^23d^2D_{5/2}$–$1s3p(^1P)3d^2D_{5/2}$
23	7.9010(6)	7.9053	3.98×10^{12}	5.36×10^{08}	4.78×10^{07}	$1s^23d^2D_{3/2}$–$1s3p(^1P)3d^2D_{3/2}$
24	7.9103(5)	7.9157	4.54×10^{12}	2.68×10^{11}	2.45×10^{10}	$1s^23s^2S_{1/2}$–$1s3s(^1S)3p^2P_{3/2}$
		7.9161	4.45×10^{12}	4.07×10^{11}	1.95×10^{10}	$1s^23s^2S_{1/2}$–$1s3s(^1S)3p^2P_{1/2}$
25	7.9113(6)	7.9177	4.34×10^{12}	1.16×10^{08}	3.51×10^{06}	$1s^23p^2P_{1/2}$–$1s3p^2(^3P)^2P_{1/2}$
		7.9178	5.36×10^{12}	9.61×10^{08}	7.27×10^{07}	$1s^23p^2P_{3/2}$–$1s3p^2(^3P)^2P_{3/2}$

Table. (Contd.)

Line number	λ_{exp}, Å	λ_{thr}, Å	A, s^{-1}	Γ, s^{-1}	Q_d, s^{-1}	Transition
26	7.9158(5)	7.9211	3.16×10^{12}	5.74×10^{11}	7.68×10^{11}	$1s^2 3d^2 D_{5/2}$–$1s3p(^1P)3d^2F_{7/2}$
		7.9216	2.87×10^{12}	5.62×10^{11}	4.89×10^{11}	$1s^2 3d^2 D_{3/2}$–$1s3p(^1P)3d^2F_{5/2}$
27	7.9350(8)	7.9410	1.19×10^{12}	4.01×10^{12}	1.00×10^{11}	$1s^2 3d^2 P_{3/2}$–$1s3s(^3S)3d^2D_{5/2}$
28	7.9934(5)	7.9964	1.99×10^{12}	2.71×10^{10}	5.64×10^{10}	$1s^2 2s^2 S_{1/2}$–$1s2s(^1S)3p^2P_{3/2}$
		7.9973	2.00×10^{12}	1.22×10^{11}	1.35×10^{11}	$1s^2 2s^2 S_{1/2}$–$1s2s(^1S)3p^2P_{1/2}$
29	8.0320(6)	8.0350	1.73×10^{12}	1.79×10^{12}	5.44×10^{11}	$1s^2 2p^2 P_{3/2}$–$1s2p(^1P)3p^2P_{3/2}$
		8.0361	8.86×10^{11}	2.68×10^{08}	2.24×10^{07}	$1s^2 2p^2 P_{3/2}$–$1s2p(^1P)3p^2P_{1/2}$
30	8.0341(5)	8.0368	3.07×10^{11}	3.47×10^{13}	1.26×10^{12}	$1s^2 2p^2 P_{3/2}$–$1s2p(^1P)3p^2D_{5/2}$
31	8.0504(5)	8.0522	1.53×10^{12}	3.07×10^{12}	8.27×10^{11}	$1s^2 2p^2 P_{1/2}$–$1s2p(^3P)3p^2S_{1/2}$
		8.0548	2.88×10^{12}	3.07×10^{12}	1.56×10^{12}	$1s^2 2p^2 P_{3/2}$–$1s2p(^3P)3p^2S_{1/2}$
32	8.0544(7)	8.0561	2.61×10^{12}	9.54×10^{11}	1.14×10^{12}	$1s^2 2s^2 S_{1/2}$–$1s2s(^3S)3p^2P_{1/2}$
		8.0562	2.61×10^{12}	1.01×10^{12}	2.37×10^{12}	$1s^2 2s^2 S_{1/2}$–$1s2s(^3S)3p^2P_{3/2}$
33	8.0670(5)	8.0679	8.02×10^{11}	1.46×10^{13}	2.78×10^{12}	$1s^2 2p^2 P_{1/2}$–$1s2s(^1S)3d^2D_{3/2}$
		8.0694	2.26×10^{12}	2.31×10^{13}	1.22×10^{13}	$1s^2 2p^2 P_{3/2}$–$1s2p(^3P)3p^2D_{5/2}$
34	8.0711(7)	8.0715	2.62×10^{12}	9.88×10^{12}	5.64×10^{12}	$1s^2 2p^2 P_{1/2}$–$1s2p(^3P)3p^2D_{3/2}$
		8.0721	1.86×10^{12}	3.30×10^{12}	3.28×10^{12}	$1s^2 2p^2 P_{3/2}$–$1s2s(^1S)3d^2D_{5/2}$
		7.0740	5.82×10^{11}	9.88×10^{12}	1.26×10^{12}	$1s^2 2p^2 P_{3/2}$–$1s2p(^3P)3p^2D_{3/2}$
35	8.0909(10)	8.0954	4.97×10^{11}	1.48×10^{10}	7.50×10^{09}	$1s^2 2p^2 P_{1/2}$–$1s2p(^3P)3p^2P_{3/2}$
		8.0956	1.76×10^{12}	3.79×10^{09}	3.48×10^{09}	$1s^2 2p^2 P_{1/2}$–$1s2p(^3P)3p^2P_{1/2}$
36	8.0924(6)	8.0980	1.92×10^{12}	1.48×10^{10}	2.90×10^{10}	$1s^2 2p^2 P_{3/2}$–$1s2p(^3P)3p^2P_{3/2}$
37	8.1259(6)	8.1301	2.45×10^{11}	5.08×10^{12}	1.24×10^{12}	$1s^2 2p^2 P_{3/2}$–$1s2s(^3S)3d^2D_{5/2}$

that the contribution of the first satellite group to the intensity of resonance lines and their shape should be taken into account in the X-ray spectral plasma diagnostics (where resonance lines are studied, as a rule). On the other hand, the isolated second satellite group provides additional diagnostic possibilities, the more so as the optical thickness of the plasma for such transitions will be much smaller than for resonance lines themselves or satellites with small quantum numbers n_1 and n_2.

Therefore, the study of Rydberg states is an urgent problem of the development of methods for diagnostics of dense high-temperature plasmas. In addition, such studies yield experimental information for comparison with calculations of highly excited autoionization configurations (including configurations corresponding to the so-called hollow ions) performed by various methods by modern atomic theory. The first step in such studies should be the identification of Rydberg satellites in experimental spectra, precision measurements of their wavelengths, and comparison of the experimental data with atomic and kinetic calculations. In this paper, we solved this problem for Rydberg satellites caused by transitions from the autoionization $1sn_1l_1n_2l_2$ levels of the Li-like Mg X ion.

2. EXPERIMENTAL SETUPS AND CALCULATION METHODS

We used two laser setups in our experiments.

In the first setup (Hercules, Fraskatti, Italy) [14, 15], a plasma was produced by a 0.308-μm XeCl excimer laser with an active volume of $9 \times 4 \times 100$ cm^3. The energy of a 12-ns laser pulse was 1.0–1.5 J, and the pulse repetition rate was 0.5 Hz. Laser radiation was focused on a solid target to a spot of diameter 50–70 μm, providing the power density q_{las} on the target of about $(1–4) \times 10^{12}$ W/cm^2.

The second series of experiments was performed at Tor Vergata University of Rome, Italy [18]. A plasma was heated by a Quantel Nd laser emitting 20-J, 12- to 15-ns pulses. The laser setup consisted of two Nd:YAG amplifiers and two Nd:glass amplifiers. The pulse repetition rate was 1/60 Hz to minimize the thermal lens effect. A Faraday cell with an aperture of 1 inch was placed behind the second amplifier to block radiation reflected by the plasma. Laser radiation was focused with a two-component objective with a focal length of 20 cm. The laser beam diameter in the focal plane was ~200 μm. The pulse energy was 4 J, corresponding to the power density on the target equal to 7×10^{11} W/cm^2.

Soft X-rays emitted by the plasma were detected with two spectrographs with spherically bent quartz

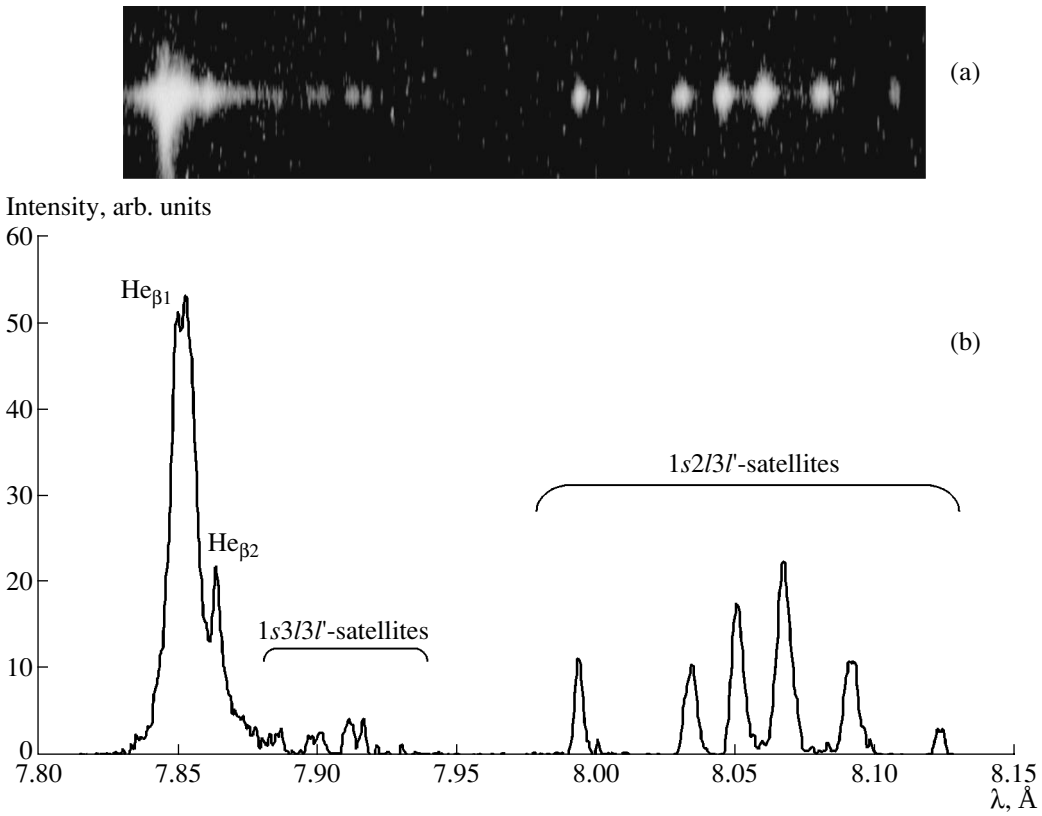


Fig. 1. (a) Emission spectrum (in the 7.80–8.15 Å region) of the magnesium plasma heated by a pulse from the XeCl excimer laser and (b) the corresponding densitogram.

and mica crystals, the radii of spherical surfaces being 150 and 100 mm, respectively. The crystals, plasma, and a photographic film were arranged in the FSPR-1M mounting [19–22]. The X-ray spectra were detected simultaneously with a high spectral resolution ($\lambda/\Delta\lambda \approx$ 3000–10 000) and a high spatial resolution ($\Delta x \approx 20$ μm) along the laser-plasma expansion direction. Because each spectrograph covered a spectral range of 0.3–0.7 Å, the entire spectral range under study from 7.1 to 8.3 Å was covered by using several spectrographs simultaneously. The dispersion curve of the spectrographs was measured by the position of the resonance doublet of the H-like Mg XI ion and the resonance series of the He-like Mg XII ion (see table), which served as reference lines. The wavelengths of the reference lines [23] are know with an accuracy of 0.1 mÅ. This value is much smaller than the total accuracy of measurements, which amounts to 0.3–1.5 mÅ and is determined by such factors as the width of the reflection curve of the crystal, the grain size of the X-ray film, the width and shape of X-ray lines, their intensity, overlap, and position relative to the reference lines.

Mg foils of thickness 90 μm were used as targets.

Atomic structures were calculated with the help of the relativistic perturbation theory taking into account the main quantum-electrodynamic corrections using the MZ code described in papers [24–28].

The theoretical emission spectra of the plasma were plotted using the MARIA kinetics code [29]. The system of stationary kinetic equations was solved for the ground states of magnesium ions of all multiplicities and the excited states nl (n = 2–7) of the Mg XII ion, 1snl (n = 2–7) of the Mg XI ion, and 1s^2nl and 1$snln'l$ (n = 2–7, n' = 2–4) of the Mg X ion, taking into account all possible transitions caused by the electron-impact excitation and relaxation, collision ionization, three-body and radiative recombination, autoionization, dielectronic capture, and radiative decay. The self-absorption effect was considered in the Biberman–Holstein escape factor approximation. The spectra were constructed assuming Gaussian profiles of the spectral lines of the same width.

3. RESULTS

First of all, we studied the 7.8–8.15 Å spectral range containing the $He_{\beta 1}$ ($1s3p^1P_1$–$1s^2{}^1S_0$) and $He_{\beta 2}$ ($1s3p^3P_1$–$1s^2{}^1S_0$) spectral lines and satellites caused by transitions from the 1$s2l3l'$ levels (Fig. 1). Such satellites were identified earlier in the emission spectra of many ions [6–11]; however, the precision measure-

ments of their wavelengths were performed only for Si [11] and Ar [10] ions. The high spectral resolution of our experiments allowed us to measure the wavelengths of the $1s2l3l'$ satellites in the spectrum of the Mg ion with accuracy permitting the use of these lines as reference lines. In addition, we managed to resolve for the first time the closely spaced spectral transitions. Our experimental data are presented in the table (lines 28–37), where the results of our calculations are also reported. One can see from the table that the MZ calculations reproduce the experimental wavelengths with a relative error no worse than 0.06%.

In the region between 7.8 and 7.95 Å (Fig. 1), another group of spectral lines is located. These lines are rather intense only within a very narrow spatial region of size about 50 μm (in the direction perpendicular to the target surface); i.e., the lines are emitted only from the densest plasma region. Because of this, upon detection of the spectra without a spatial resolution or with the resolution exceeding 100 μm, which is typical of slit spectrographs used earlier, the signal-to-noise ratio for this group of lines reduces down to a value that is inadequate for quantitative measurements.

Our calculations showed that this group of lines can be assigned to transitions from the $1s3l3l'$ levels of the Li-like Mg X ion. Figure 2a shows the experimental emission spectrum of the plasma produced by a Nd laser, and Fig. 2b presents the emission spectra calculated for the plasma with the electron density $N_e = 10^{21}$ cm^{-3}, the electron temperature $T_e = 140$ eV, and size of 500 μm (the plasma size affects the efficiency of self-absorption of spectral lines). Two variants of calculations correspond to the consideration (spectrum 2) and neglect (curve *1*) of radiative transitions from the $1s3l4l'$ levels. Comparison of these spectra shows that, while the $1s3l3l'$–$1s^23l'$ transitions produce a well-resolved satellite structure, the $1s3l4l'$–$1s^24l'$ transitions are mainly responsible for the deformation of the profile of the He_β line, increasing the intensity of its long-wavelength wing. The assignment of these transitions is presented in the table (lines 18–27).

Note that the intensity of the $1s3l3l'$–$1s^23l'$ satellites can be comparable with that of the $1s2l3l'$–$1s^22l$ satellites only in a high-density plasma. Indeed, the population of autoionization levels in a low-density plasma is described by the coronal model, and the intensity of satellites is proportional to the factors

$$Q_d = \frac{g_u}{g_l}\frac{A_{ul}\Gamma_u}{\sum_l A_{ul} + \sum \Gamma_u}$$

presented in the table. Here, g_u and g_l are the statistical weights of the initial and final levels of the $u \longrightarrow l$ radiative transition, A_{ul} is the transition probability, Γ_u is the autoionization probability, and $\sum \Gamma_u$ is taken over all possible autoionization channels of the u level. One can see from the table that the values of Q_d for the $1s2l3l'$–$1s^22l$ satellites substantially (by two to three orders of magnitude) exceed Q_d for the $1s2l3l'$–$1s^23l$ transitions. In the opposite case of local thermodynamic equilibrium, the line intensities are proportional to g_uA_{ul}. Unlike the factors Q_d, the products g_uA_{ul} and, hence, the line intensities in the case of local thermodynamic equilibrium prove to be of the same order for both types of satellite lines. As follows from our calculations, the populations of autoionization levels for the Mg X ion at the electron density $N_e \geq 10^{21}$ cm^{-3} already greatly differ from the coronal population (although the local thermodynamic equilibrium is not yet achieved), which allows the observation of these lines emitted from the regions of the laser plasma of critical density.

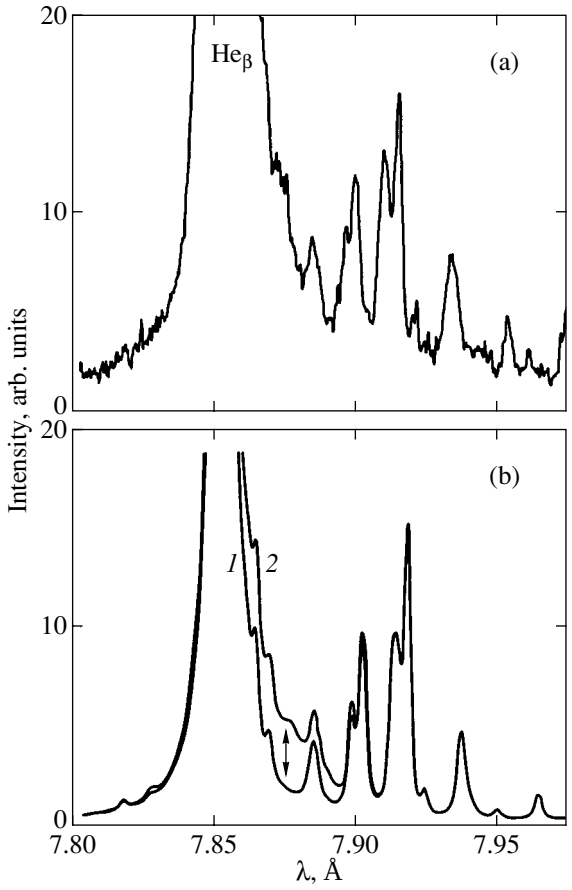


Fig. 2. (a) Emission spectrum (in the 7.80–7.95 Å region) of the magnesium plasma heated by a pulse from the Nd laser and (b) emission spectra calculated neglecting (curve *1*) and taking into account (curve *2*) the $1s3l4l'$ levels.

In the spectral range from 7.4 to 7.8 Å, we observed satellite transitions caused by the radiative decay of even higher excited autoionization states. Figure 3a shows the emission spectrum of the magnesium plasma heated by the Nd laser, and Fig. 3b presents the emission spectrum calculated using the plasma parameters specified above.

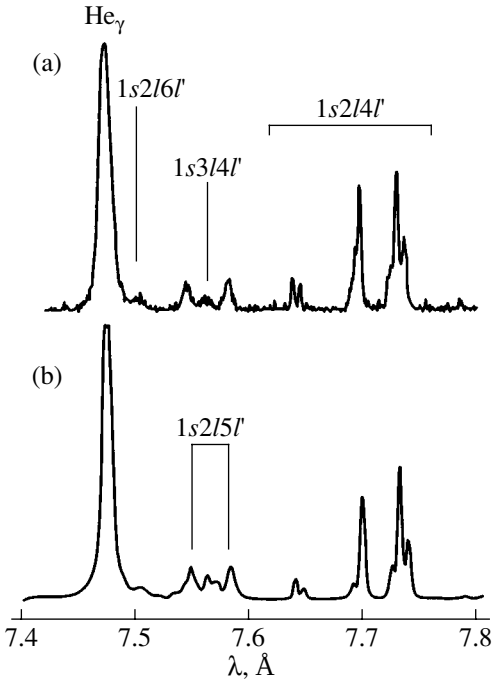


Fig. 3. (a) Emission spectrum (in the 7.40–7.80 Å region) of the magnesium plasma heated by a pulse from the Nd laser and (b) the calculated emission spectrum.

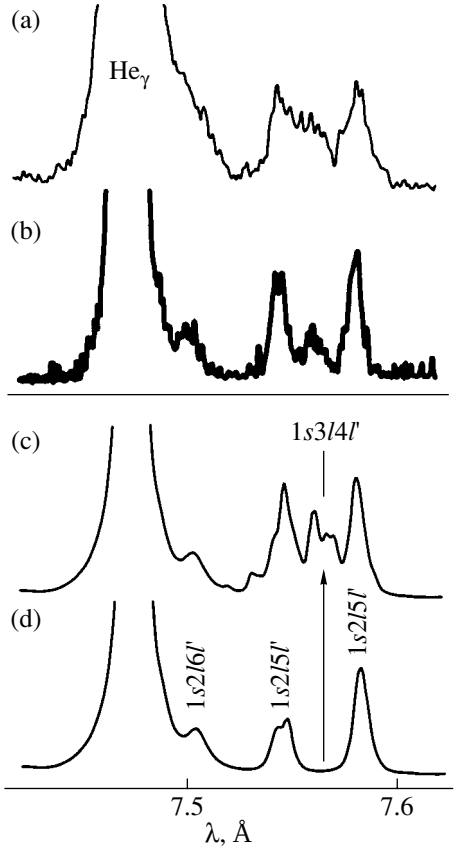


Fig. 4. (a) Emission spectrum (in the 7.45–7.60 Å region) of the magnesium plasma heated by pulses from (a) the XeCl and (b) Nd lasers and emission spectra calculated (c) neglecting and (d) taking into account the 1*s*3*l*4*l*' levels.

One can clearly see from Fig. 3 that the 1*s*2*l*4*l*'–1$s^2$2*l* transitions are the strongest in this spectral range. The lines corresponding to these transitions are the satellites of the He$_\gamma$ line. By comparing the model and experimental spectra observed upon heating the plasma by the XeCl and Nd lasers, we identified eight spectral lines corresponding to the transitions of this type and measured their wavelengths with an accuracy of 0.5–0.9 mÅ. The results are presented in the table (lines 9–17).

The less intense satellite transitions caused by the radiative decay of the autoionization levels 1*s*2*l*5*l*' and 1*s*2*l*6*l*' lie in the same spectral range. They are in fact satellites of the He$_\delta$ and He$_\varepsilon$ lines, but because their shift relative to the corresponding resonance transitions 1*s*5*p*–1s^2 and 1*s*6*p*–1s^2 exceeds the separation between the He$_\gamma$ and He$_{\delta,\,\varepsilon}$ lines, they are located to the right of the He$_\gamma$ line (Fig. 3). The 1*s*2*l*5*l*' configuration is represented in the emission spectrum by two groups of lines, one of which exhibits a fine structure (lines 2, 3, and 9 in the table), whereas transitions from the 1*s*2*l*6*l*' states produce one rather broad line (line 1).

Figure 3 also shows the group of lines located between two groups of the 1*s*2*l*5*l*'–1$s^2$2*l* transitions. Our calculations showed that these lines are related to the 1*s*3*l*4*l*'–1$s^2$3*l* transitions. Figure 4 presents the emission spectra calculated taking these transitions into account (spectrum c) and neglecting them (spectrum d). Also, the emission spectra of the plasma produced by the XeCl laser (spectrum a) and by the Nd laser (spectrum b) are shown. Our calculations showed that the 1*s*3*lnl*' states in a plasma with density $N_e \geq 10^{21}$ cm^{-3} are mainly populated due to the dielectronic capture to the excited 1*s*2*l* levels of the He-like ion, i.e., due to the 1*s*2*l* + $e \longrightarrow$ 1*s*3*lnl*' process (see also [12]). Therefore, the intensity of the 1*s*3*l*4*l*'–1$s^2$3*l* transitions can be sufficient for their observation only in a dense plasma. By using wide-aperture, high-resolution X-ray spectrographs and a dense plasma as an emission source, we have managed for the first time to identify these spectral lines and measured their wavelengths (see lines 4–8 in the table).

4. CONCLUSIONS

We have identified for the first time the satellite lines caused by radiative transitions from the Rydberg autoionization states of the Li-like Mg X ion and performed precision measurements of their wavelengths.

The comparison of our experimental data with the calculation of atomic structures by the method of relativistic perturbation theory has shown that the accuracy of calculations of the energy of autoionization states is very high, being of the order of 0.06%, even without the

use of any semiempirical corrections. The wavelengths measured in the paper can be used for a semiempirical estimate of the value of the leading order of perturbation theory among the orders that were neglected in calculations. These wavelengths, together with earlier wavelength measurements performed for the Si XII ion, can permit the estimate of the value of main corrections both in the expansion of the nonrelativistic part of the energy in powers of the parameter $1/Z$ and in the expansion of the relativistic part of the energy in powers of αZ. This, in turn, makes it possible to perform much more accurate calculations of the energy diagram of autoionization levels for other Li-like ions with nuclear charges from 10 to 20.

Comparison of the experimental intensities of Rydberg satellites with kinetic calculations shows that the simulation of the population kinetics of the Rydberg autoionization levels of Li-like ions in a dense plasma should take into account all possible channels of dielectronic capture. This is necessary because the energy of the $1snln'l'$ states with two sufficiently strongly excited electrons with $n, n' > 2$ can exceed the energy of singly excited $1s(n-1)l$ states of the He-like ion. As a result, the additional $1snln'l' \longrightarrow 1s(n-1)l + e$ autoionization channel appears, as well as the corresponding channel of dielectronic capture. Under coronal conditions (in a low-density plasma), the populations of excited levels of the He-like ion are low, and this channel of dielectronic capture in fact makes no contribution to the population of satellite states. However, as follows from our calculations, this channel can be quite efficient for ions with $Z \sim 10$ in a plasma with $N_e \sim 10^{21}$ cm^{-3}.

One can see from the above figures that the calculated spectra of Rydberg satellites adequately describe their experimental spectra. This means that such satellites, as usual satellites of the He_α line, can be employed for diagnostics of a high-temperature plasma. Since their sensitivity to the electron temperature of the plasma is almost the same as that for usual satellites, their employment can be preferable in some cases because the question about the optical thickness of the plasma both for satellites and for the He_β and He_γ resonance lines can be avoided. Note that the X-ray spectral diagnostics of a plasma by the shape of resonance-series lines of the He-like ions should take into account the contribution from Rydberg satellites to the intensity of the long-wavelength wings of these lines.

Another possible channel of dielectronic capture (which, however, does not lead to the appearance of satellites) is the condensation of highly excited Rydberg atoms described by Manykin *et al.* [30].

Note also that the precision experimental wavelengths of satellites of the He_β and He_γ lines of the Mg XI ion measured in the paper allow one to employ these satellites (at least, the most intense of them) as reference lines in studies of the complicated spectra of multielectron ions. For example, such reference lines will be quite useful is studies of transitions in Ne-like copper and zinc ions.

ACKNOWLEDGMENTS

The work of A.Ya.F. was partly supported by a grant from the Ministry of Foreign Affairs of Italia within the framework of the Landau Network Centro Volta Competition.

This study was partially supported by the ISTC (grant no. 1785) and the Russian–Italian program of scientific and technical cooperation.

REFERENCES

1. A. H. Gabriel, Mon. Not. R. Astron. Soc. **160**, 99 (1972).
2. C. P. Bhalla, A. H. Gabriel, and L. P. Presnyakov, Mon. Not. R. Astron. Soc. **172**, 359 (1975).
3. E. V. Aglitskiĭ, V. A. Boĭko, S. M. Zakharov, *et al.*, Kvantovaya Élektron. (Moscow) **1**, 908 (1974).
4. V. A. Boiko, A. Ya. Faenov, and S. A. Pikuz, J. Quant. Spectrosc. Radiat. Transf. **51**, 11 (1978).
5. V. A. Boiko, A. V. Vinogradov, A. Ya. Faenov, *et al.*, J. Sov. Laser Res. **6**, 85 (1985).
6. E. V. Aglitskiĭ and U. I. Safronova, *Spectroscopy of Autoionization States of Atomic Systems* (Énergoatomizdat, Moscow, 1985).
7. F. B. Rosmej, U. N. Funk, M. Geissel, *et al.*, J. Quant. Spectrosc. Radiat. Transf. **65**, 477 (2000).
8. V. A. Boiko, S. A. Pikuz, U. I. Safronova, and A. Ya. Faenov, Mon. Not. R. Astron. Soc. **185**, 789 (1978).
9. U. I. Safronova, M. S. Safronova, R. Bruch, and L. A. Vainshtein, Phys. Scr. **51**, 471 (1995).
10. I. Yu. Skobelev, A. Ya. Faenov, V. M. Dyakin, *et al.*, Phys. Rev. E **55**, 3773 (1997).
11. I. Yu. Skobelev, A. Bartnik, E. Behar, *et al.*, Kvantovaya Élektron. (Moscow) **25**, 697 (1998).
12. F. B. Rosmej, A. Ya. Faenov, T. A. Pikuz, *et al.*, J. Phys. B **31**, L921 (1998).
13. F. B. Rosmej, D. H. H. Hoffmann, M. Geißel, *et al.*, Phys. Rev. A **63**, 063409 (2001).
14. F. Rosmej, A. Ya. Faenov, T. A. Pikuz, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **65**, 679 (1997) [JETP Lett. **65**, 708 (1997)].
15. F. B. Rosmej, A. Ya. Faenov, T. A. Pikuz, *et al.*, J. Quant. Spectrosc. Radiat. Transf. **58**, 859 (1997).
16. A. Ya. Faenov, J. Abdallah, Jr., R. E. H. Clark, *et al.*, Proc. SPIE **3157**, 10 (1997).
17. A. M. Urnov, J. Dubau, A. Ya. Faenov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **67**, 467 (1998) [JETP Lett. **67**, 489 (1998)].
18. K. B. Fournier, A. Ya. Faenov, T. A. Pikuz, *et al.*, submitted to Phys. Rev. A (2002).
19. A. Ya. Faenov, S. A. Pikuz, A. I. Erko, *et al.*, Phys. Scr. **50**, 333 (1994).
20. T. A. Pikuz, A. Ya. Faenov, S. A. Pikuz, *et al.*, J. X-ray Sci. Technol. **5**, 323 (1995).

21. I. Yu. Skobelev, A. Ya. Faenov, B. A. Bryunetkin, *et al.*, Zh. Éksp. Teor. Fiz. **108**, 1263 (1995) [JETP **81**, 692 (1995)].

22. B. K. F. Young, A. L. Osterheld, D. F. Price, *et al.*, Rev. Sci. Instrum. **69**, 4049 (1998).

23. V. A. Boiko, V. G. Pal'chikov, I. Yu. Skobelev, and A. Ya. Faenov, in *Spectroscopic Constants of Atoms and Ions* (CRC Press, Boca Raton, 1995).

24. L. A. Vainshtein and U. I. Safronova, At. Data Nucl. Data Tables **21**, 49 (1978).

25. L. A. Vainshtein and U. I. Safronova, At. Data Nucl. Data Tables **25**, 311 (1980).

26. L. A. Vainshtein and U. I. Safronova, At. Data Nucl. Data Tables **31**, 519 (1985).

27. V. P. Shevelko and L. A. Vainshtein, *Atomic Physics for Hot Plasmas* (Institute of Physics Publ., Bristol, 1993).

28. U. I. Safronova and M. S. Safronova, J. Phys. B **28**, 2803 (1995).

29. F. B. Posmej, J. Phys. B **30**, L819 (1997).

30. É. A. Manykin, M. I. Ozhovan, and P. P. Poluéktov, Zh. Éksp. Teor. Fiz. **102**, 1109 (1992) [Sov. Phys. JETP **75**, 602 (1992)]; Zh. Éksp. Teor. Fiz. **105**, 50 (1994) [JETP **78**, 27 (1994)].

Translated by M. Sapozhnikov

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 429–439.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 500–512.

PLASMA, GASES

Dust–Acoustic Instability in an Inductive Gas-Discharge Plasma

A. V. Zobnin, A. D. Usachev*, O. F. Petrov, and V. E. Fortov

Institute of High Energy Densities (IVTAN), Russian Academy of Sciences, Izhorskaya ul. 13/19, Moscow, 127412 Russia

**e-mail: usachev@ihed.ras.ru*

Received April 1, 2002

Abstract—Spontaneous excitation of a dust-particle density wave is observed in a dust cloud levitating in the region of the diffused edge of an rf inductive low-pressure gas-discharge plasma. The main physical parameters of this wave and of the background plasma are measured. The analytic model proposed for the observed phenomenon is based on the theory of dust sound and successfully correlates with experimental data in a wide range of experimental conditions. The effect of variable charge of dust particles on the evolution of the observed dust-plasma instability is studied analytically. It is shown that the necessary condition for the development of the dust–acoustic instability is the presence of a dc electric field in the dust cloud region. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The instability of the dust component, which is manifested in spontaneous buildup of random or organized motion of an ensemble of dust particles, is a general and fundamental property of a dust plasma like "classical" plasma instabilities in a particle-free plasma. A wide spectrum of dust-plasma instabilities has been observed in laboratory dust plasmas. This spectrum includes

(i) buildup (heating) of the random motion (temperature) of dust particles in a plasma of dc glow discharges [1, 2] and of a capacitive rf discharge [3–5];

(ii) oscillations of dust particles in the double electric layer near an electrode [6];

(iii) rotation of a "needle" in the plasma of an rf capacitive discharge [7];

(iv) instability manifested under microgravitation conditions and known as "heart beats" [8];

(v) dust vortices in a dc glow-discharge plasma [9], nuclear-excited plasma [10], and in the plasma of an rf capacitive discharge under microgravitation condition [8, 11];

(vi) dust–acoustic instability in 3D dust clouds in dc glow-discharge plasmas [12–17].

A distinguishing feature of such instabilities is an extremely long characteristic time of their evolution (up to several seconds) and the existence of a fundamentally new parameter determining the development of instability (variable charge of dust plasma) in most cases [18, 19].

On the one hand, various dust-plasma instabilities reflect the dynamics of collective processes in a complex plasma, which is of fundamental scientific importance. On the other hand, macroscopic parameters of dust instabilities are associated with microscopic parameters of the plasma and, hence, can be used for its diagnostics. In addition, dust-plasma instabilities have become a frequent and undesirable factor like "classical" plasma instabilities. In particular, dust-plasma instabilities destroy the ordering of dust structures (especially 3D structures), thus creating considerable difficulties in the study of dust crystals. All this necessitates theoretical and experimental investigations of various forms of dust-plasma instabilities. A systematic study of dust-plasma instabilities is still at the initial stage and entails considerable difficulties in the interpretation of many aspects of this phenomenon [4].

Among the above-mentioned types of dust instabilities, dust-acoustic instability (DAI) has been studied most comprehensively [12–17]. This instability is manifested in the form of self-excitation of wavelike oscillations of volume concentration of dust particles in a laboratory dust plasma. It is interesting to note that the development of DAI was observed only in 3D dust clouds. The first such observations were apparently made in [20] in the plasma of an rf capacitive discharge and were interpreted in [21] as a dust–acoustic wave. Later, spontaneous formation of traveling waves of dust particle density was observed in [12] in the anode column of a *Q* machine. In this experiment, the chamber was filled with nitrogen under a pressure of 60–80 mtorr. Particles of AlSiO having a size of 1–15 μm were also supplied to the discharge. The observed length of the dust wave was 6 mm, its phase velocity was 9 cm/s, and the frequency was approximately 15 Hz. The axial electric current in the region of the dust cloud was estimated at approximately 1 V/cm. The observed wave propagated along the horizontal from the anode to the cath-

1063-7761/02/9503-0429$22.00

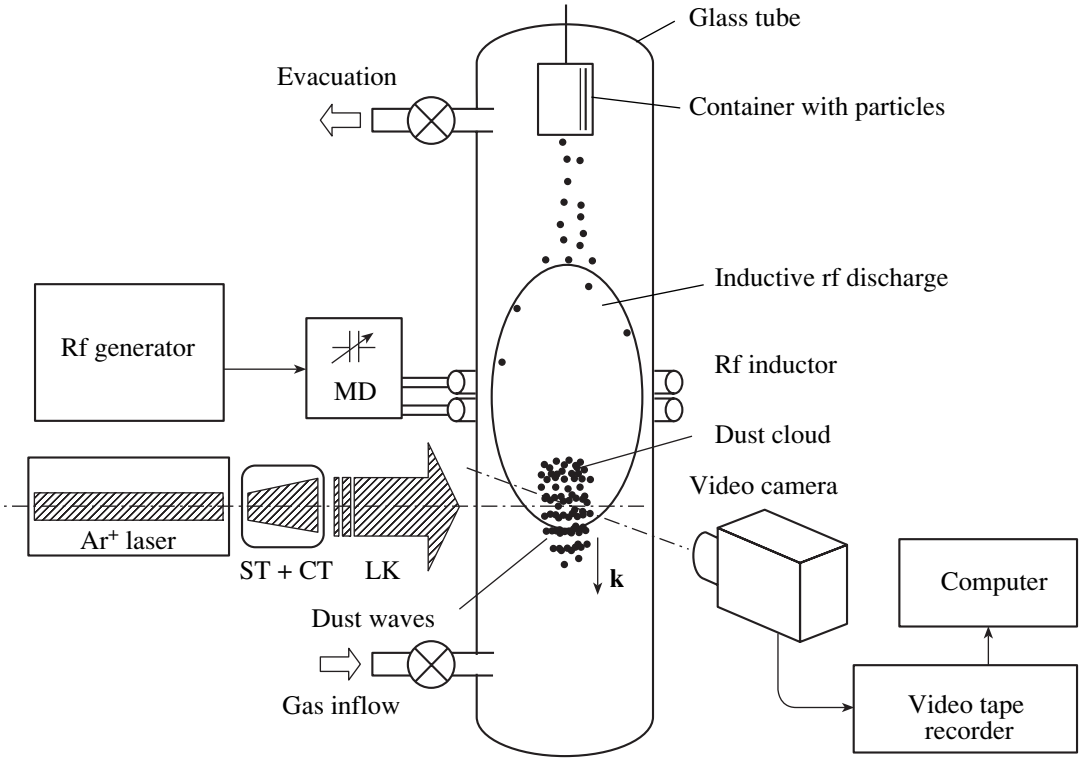


Fig. 1. Diagram of the experimental setup for observing dust–acoustic instability in the plasma of an inductive gas discharge. Notation: MD is a matching device, ST + CT stands for a combination of a spherical and cylindrical telescope, and LK denotes a laser knife.

ode. The parameters of this wave (the frequency and the magnitude of the wave vector) were successfully described by the dispersion relation for a dust–acoustic wave, which led the authors to the conclusion that the observed phenomenon is a dust–acoustic wave. However, the lack of exact information on the size of dust particles (taken at 5 μm) and their charge (assumed to be equal to $40\,000e^-$) implies that such estimates should be made with certain care. As regards the DAI buildup mechanism, the authors of [12], referring to the theoretical publication [22], conclude that this buildup is due to the ion flow through the region of dust cloud.

In [14], DAI was observed in the strata of a dc glow discharge, while the same authors propose in [16] a new mechanism of its buildup in addition to the old mechanism described in [1], which can be presented as the result of variation of the macroparticle charge in the presence of an external electric field. However, the main parameters of the background plasma were not measured in [16]. These data were borrowed from publications by other authors. In addition, the variation of experimental parameters required for a comparison of experimental and theoretical functional dependences was not carried out.

In a recent publication [17], a detailed analysis of the spatial distribution of DAI wave parameters over the inhomogeneous space of a stratum was carried out, but the mechanism of DAI buildup was not investigated. In addition, polydisperse iron particles were used in these experiments, and the question concerning one of the main parameters of dust plasmas, viz., the size of particles in the cloud, remained unanswered.

This work is devoted to the experimental investigation and numerical simulation of the dust–acoustic instability in the plasma of an rf inductive low-pressure gas discharge. The employment of the diffusive edge of an rf inductive discharge for this purpose has the advantage that smoother spatial gradients of the main parameters of the background plasma (such as the electron concentration and temperature) in the macroparticle suspension region as compared to the stratum region of a dc glow discharge are formed. This circumstance, as well as the probe diagnostics of the background plasma parameters, the use of particles with a calibrated size, numerical calculation of the charge of dust particles taking into account the effect of collision processes in the Debye sphere around a dust particle on the charge of this particle [23, 24], and a wide variation of experimental conditions, enabled us to draw the conclusion about the adequacy of the proposed analytic model of the observed dust-plasma instability.

2. EXPERIMENTAL TECHNIQUE AND METHODS OF DIAGNOSTICS

2.1. Experimental Setup

The diagram of the experimental setup intended for studying the dust–acoustic instability in inductively coupled dust plasma is presented in Fig. 1. An rf induc-

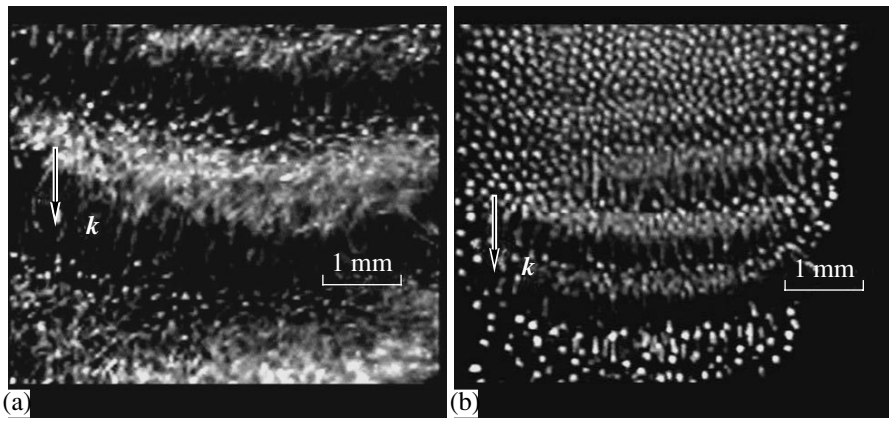


Fig. 2. Video images of dust–acoustic waves in the lower part of the dust cloud under a buffer gas (neon) pressure of (a) 15 and (b) 50 Pa.

tive discharge was excited in a 65-cm-long vertical cylindrical glass tube of diameter 3 cm with the help of a two-turn ring inductor in a buffer gas (neon). The tube was supplied with lower and upper ports for introduction and removal of the working gas, which allowed us to carry out DAI investigations in the stationary gas as well as in the gas flow and provided an additional opportunity to modify the gas cloud. In addition, a slight pumping of the working gas was required for prolonged probe measurements in order to maintain the constant chemical composition of the plasma medium. The frequency of the voltage supplied to the inductor was 100 MHz, and the power introduced to the discharge was about one watt. The voltage across the inductor was controlled with the help of a high-resistance voltage divider. The inductive discharge was a glow cloud in the form of an ellipsoid of revolution. The length of this formation during the experiment was equal to approximately 12 cm, but it could be varied over a wide range by changing the rf power supplied to the discharge. The pressure of the buffer gas (neon) varied from 1 to 120 Pa. The dust cloud was created in the discharge by shaking the container filled with monodisperse particles of melamine formaldehyde of diameter 1.87 ± 0.04 μm, prepared at Microparticles GmbH. The particles were sieved through the bottom of the container with holes, fell down, and were suspended in an electrostatic trap at the lower part of the inductive discharge. Such a trap is formed due to the combination of the fields of ambipolar diffusion and of the charged surface of the discharge tube [24]. The characteristic size of the dust cloud was 5×8 mm^2. As soon as the number of particles falling into the cloud exceeded approximately 500, dust–acoustic instability in the form of macroparticle density waves was spontaneously self-excited in the cloud as a rule. The density of dust particles in the cloud increased with the total number of suspended dust particles. It was found that the observed DAI emerges spontaneously under buffer gas pressures from 10 to 60 Pa, the length of the observed waves and their phase velocity depending considerably on the buffer gas (neon) pressure. The waves were generated in the upper part of the dust cloud and propagated downwards with rapidly increasing amplitude. The uppermost part of the dust cloud remained unperturbed. Under pressures lower than 10 Pa, the random motion of particles started dominating over the wave motion, and no waves were observed. Under pressures exceeding 60 Pa, self-oscillations were suppressed by the high viscosity of the buffer gas. We carried out our experiments at the following pressures of the plasma-forming neon: 10, 15, 20, 30, and 50 Pa. Figure 2 illustrates the typical shape of the density waves of dust particles under the neon pressures of 15 and 50 Pa.

2.2. Video Recording and Processing of Video Images

The dynamics of the behavior of monodisperse particles suspended in the dust trap was monitored by illuminating them with a "laser knife" oriented in the vertical plane. The laser knife was formed from an argon laser beam (with a wavelength of 488 and 511.4 nm and a power of 1 W) with the help of a combination of two telescopes: a spherical telescope (×10) and a cylindrical telescope (×1.5). The cross-sectional area of the laser knife in the waist region was 15×0.3 mm^2 and varied insignificantly (by 20%) within the dust cloud width. The image of the dust cloud was recorded with the help of a high-speed video camera (Redlake 500) with a spatial resolution of 120×120 pixels, which ensured video recording with a frequency up to 500 frames per second. The buffer RAM of the video camera made it possible to carry out continuous video recording for 4 s at the maximum recording speed, after which the obtained video information was rerecorded on a videotape recorder. An analog video signal from the videotape recorder was converted into digital information with the help of a computer video plate and then transformed into a sequence of digital black-and-white images having a size of 540×540 pixels and a brightness depth of 8 bits. Each such image corresponded to a frame of the high-speed video camera and had the form of a 2D array of pixel intensities $I(i, k)$, where $i, k = 1, \ldots, 540$

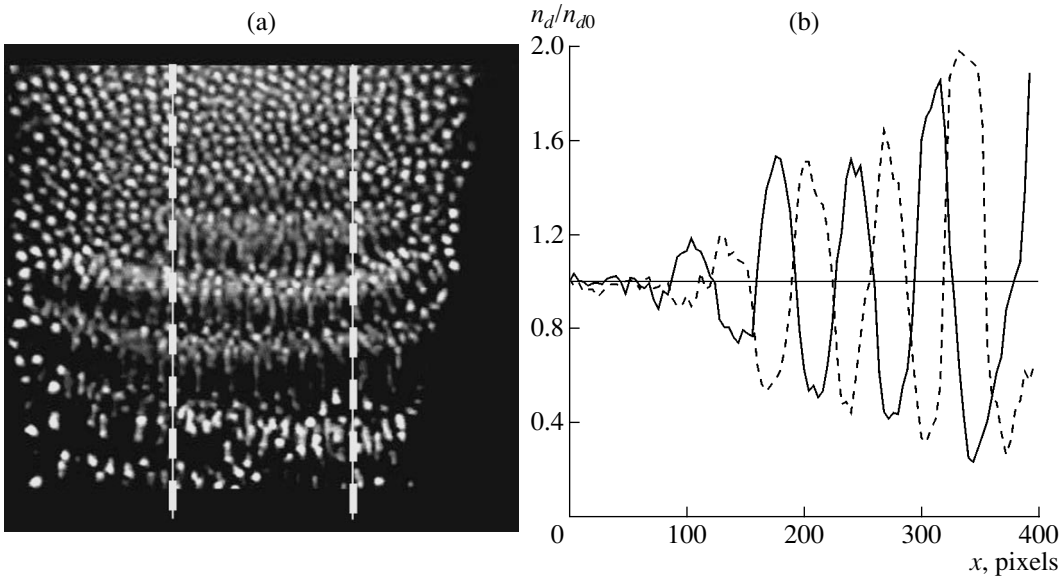


Fig. 3. (a) Video image of a dust–acoustic self-excited wave under a buffer gas pressure of 50 Pa. (b) Relative variation of the video image brightness (proportional to $n_d(x)$) along the vertical x axis in the region between two dashed lines, normalized to the image brightness in the upper part of the unperturbed region of the dust cloud in two consecutive video frames.

are the indices corresponding to the number of a pixel in the horizontal and vertical directions of the CCD matrix, respectively. The obtained digital arrays were processed with the help of standard or specially recorded computer codes. These data were used for determining the countable concentration n_d of dust particles in the upper unperturbed part of the cloud and the distribution of the relative concentration $n_d(x)$ of dust particles in the field of a frame along the x axis of the tube. In order to determine $n_d(x)$, the field of a video frame was divided into discretization elements. Since the observed dust waves were virtually planar, the discretization elements of the video frame field were chosen in the form of horizontal rectangles having a size of 300×6 pixels. The relative concentration of dust particles in the field of a discretization element was assumed to be proportional to the sum of intensities of all pixels within this element minus the background intensity. The inhomogeneity of the laser knife was compensated by corresponding normalization. The background intensity was determined from video frames recorded before the introduction of particles. The sensitivity of the CCD camera was chosen so as to avoid saturation effects in the pixel brightness depth. By varying the size of the discretization element, we attained a compromise between the spatial resolution of the $n_d(x)$ distribution and the statistical noise intensity $n_d/\sqrt{N_d}$, where N_d is the number of dust particles in a discretization cell. The measured $n_d(x)$ distribution was used to determine the wave vector k_d of the wave (having wavelength λ_d) and to estimate the increment in the oscillation amplitude γ_d. The phase velocity v_d of the wave was determined from a comparison of the values of $n_d(x)$ for two consecutive frames. Figure 3a shows a video frame of the dust cloud, in which two vertical lines mark the region of the digital processing of the video field, while Fig. 3b presents the distribution $n_d(x)/n_{d0}$ along the wave vector. The table contains the complete set of experimental data on v_d, λ_d, ω_d, and k_d, obtained under different pressures of the plasma-forming gas. These data correspond to the region of the dust cloud in which the amplitude of oscillations had the minimal value, which allowed us to interpret the experimental data by comparing them with the predictions of the theory of linear dust–acoustic waves [16, 26–30]. Unfortunately, the increment of increase in γ_d was determined by us quite approximately due to a strong statistical noise in the region of small-amplitude waves.

2.3. Probe Measurements

In order to obtain a quantitative description of the physics of DAI development, we need information on electrophysical parameters of the neon background plasma, such as the electron concentration n_e and temperature T_e, and electric field strength E in the region of suspension of dust particles. These parameters were measured by using a solitary Langmuir probe in the entire region of the inductive discharge. The measurements were made with the help of a movable 3-mm long cylindrical probe of diameter 0.05 mm in the form of a molybdenum wire placed in a thin glass tube of diameter 1 mm drawn to a narrow tip, terminating in the glass-covered part of the probe having a diameter of 0.3 mm. The displacement system made it possible to move the probe with the help of permanent magnets in two coordinates: along the axis x of the tube and along

Measured and calculated parameters of dust plasma and dust–acoustic instability waves in seven experiments

No. of experiment	1(*)	2	3(*)	4	5	6	7	Error
Background plasma parameters								
p, Pa	10	15	20	20	30	50	50	±2
η, s^{-1}	37	56	74	74	110	185	185	±10%
n_e, cm^{-3}	2×10^8	2×10^8	3×10^8	3×10^8	3×10^8	4×10^8	4×10^8	±40%
T_e, eV	4.2	4.1	4.0	4.0	3.7	3.5	3.5	±1 eV
$n_i = n_e + Z_d n_d$, cm^{-3}	2.8×10^8	5×10^8	3.3×10^8	5×10^8	4.7×10^8	6.6×10^8	5.7×10^8	±50%
$\tilde{T}_i$, K	1030	680	515	515	340	300	300	±25%
r_D, mm	0.132	0.081	0.086	0.086	0.059	0.047	0.050	±30%
Parameters of dust component								
$n_d\times10^{-4}$, cm^{-3}	2.4	10	1.5	7	7	12	7	±30%
Z_d	3400	3000	2900	2900	2360	2160	2160	–
Parameters of dust wave								
v_{DAI}, cm/s	8.3 ± 1	5.8 ± 2.3	4.8 ± 0.5	4.8 ± 1.6	4.2 ± 1.4	2.9 ± 0.3	2.3 ± 0.3	←
λ_{DAI}, mm	5.2 ± 0.7	1.26 ± 0.5	3.0 ± 0.3	1.05 ± 0.35	0.95 ± 0.3	0.65 ± 0.07	0.67 ± 0.07	←
$\omega_{DAI} = 2\pi v$, s^{-1}	100 ± 16	290 ± 30	100 ± 30	290 ± 40	280 ± 40	285 ± 10	220 ± 40	←
k_{DAI}, cm^{-1}	12 ± 1.5	50 ± 20	21 ± 2	60 ± 20	66 ± 20	97 ± 10	97 ± 10	←
γ_{DAI}, cm^{-1}	–	>10	>10	>10	>10	6 ± 4	5 ± 4	←
ω_{max}, $\chi = 0$, s^{-1}	170	300	130	285	225	225	150	–
ω_{max}, $\chi = 0.3$, s^{-1}	250	470	155	375	270	275	195	–
γ_{max}, $\chi = 0$, cm^{-1}	16.5	26	11	30	29	18	8.5	–
γ_{max}, $\chi = 0.3$, cm^{-1}	24	38	19.5	46	47	33	18	–
ω_{pd}, s^{-1}	397	715	219	578	470	564	431	±30%

Note: The following notation is used in the table: p is the pressure of plasma-forming neon; η is the viscosity of dust particles in neon; n_e and T_e are the electron concentration and temperature, respectively, in the region of dust waves; n_i and T_i are the ion concentration and effective temperature; r_D is the Debye radius; n_d and Z_d are the concentration and charge of dust particles; v_{DAI}, λ_{DAI}, ω_{DAI}, and k_{DAI} are the measured phase velocity, wavelength, cyclic frequency, and wave number of the DAI wave; ω_{max} is the calculated DAI frequency for which $\gamma_{DAI}(\omega) = \max$ for $\chi = 0$ and 0.33 (Eq. (19)); γ_{DAI} is the measured value of the increment of the increase in the DAI wave amplitude; γ_{max} is the calculated value of the increment of the increase in the DAI wave amplitude; and (*) indicates low concentration of dust particles.

its radius R. The range of displacement of the probe was 5 cm along the tube axis and from the center of the tube to its wall along the radius. Since an inductive discharge does not require the presence of electrodes in the plasma, we specially introduced a large plane counterelectrode to be able to use the electron branch of the current–voltage characteristic of the probe. The employment of the electron branch of the *IV* characteristic is dictated by the necessity to determine the potential of the space and the complexity of determining the concentration of charge particles from the ion current. In order to reduce the perturbing action on the plasma, the counterelectrode was mounted on the other side of the inductor relative to the region of measurements. The counterelectrode cross-sectional area of 24 cm^2 was sufficient for transmitting the electron current of the movable probe, but still insufficient for disregarding the plasma–counterelectrode resistance. For this reason, the potential of the movable probe was measured relative to another fixed (reference) electrode located near the inductor under the action of a floating potential. In order to reduce the effect of stray rf currents between the movable probe and the measuring system as well as between the reference probe and the measuring system, we connected *LC* filter–plugs tuned to the frequency of the generator feeding the inductor. The scanning of the voltage across the Langmuir probe (in the voltage range from –40 to + 20 V) and the recording of the probe current (the minimum detectable current was 2 nA) were controlled by a PC. Twenty *IV* characteristics were averaged during processing.

The electron concentration n_e was determined from the magnitude of the probe current at the inflection point of the electron branch $I = I(U)$ of the current–voltage characteristic, and the potential U_s of the space was determined relative to the reference electrode from the voltage across the probe at the point of inflection of the *IV* characteristic. The electron temperature T_e was determined from the relation

$$U_s(x) = T_e \ln n_e(x) + \text{const}, \quad (1)$$

where $U_s(x)$ is the potential of space at point x. The results of probe measurements of the quantities n_e and T_e in the region of suspension of dust particles under pressures of 10, 15, 20, 30, and 50 Pa of plasma-forming neon are given in the table. By way of an example, the results of measurements of the profiles of $n_e(x)$ and

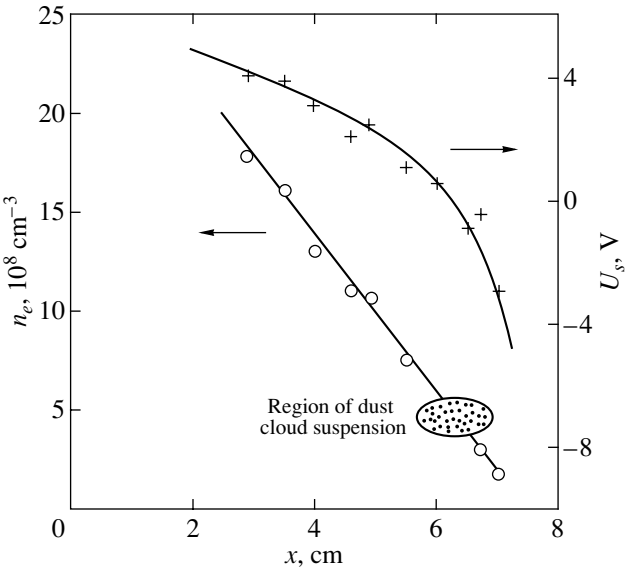


Fig. 4. Dependence of the electron concentration n_e (circles) and the potential U_s of space (crosses) at the tube axis on the distance x to the inductor for the pressure p = 50 Pa of the plasma-forming neon.

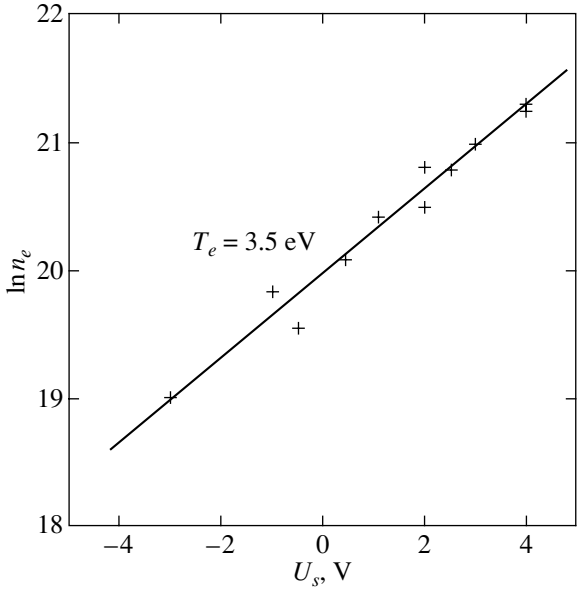


Fig. 5. Dependence of the logarithm of electron concentration on the space potential in different regions of the rf discharge under the plasma-forming neon pressure p = 50 Pa. The straight line approximates the experimental data obtained by the least squares method and corresponds to T_e = 3.5 eV.

$U_s(x)$ are presented in Fig. 4 for a pressure p = 50 Pa of plasma-forming neon. Figure 5 shows the measured dependence of the logarithm of electron concentration on the potential of space. The fact that all experimental points lie on the same straight line indicates the existence of a universal electron temperature in the entire range of probe measurements. The electric field strength was calculated as $E = -\text{grad}\,U_s(x)$. On the other hand, the measurements of $n_e(x)$ revealed (Fig. 4) that dust structures hover on the region of almost linear decrease in the electron concentration. The linear dependence of the electron concentration on the coordinate, $n_e(x) = \alpha x$, determines the relation between the electric field of ambipolar diffusion and the electron concentration:

$$E = \alpha\frac{T_e}{en_e}. \qquad (2)$$

The values of the electric field strength calculated as $E = -\text{grad}\,U_s(x)$ and by formula (2) correlate well. In the suspension region of dust structures, $E \sim 4$ V/cm for all values of neon pressure. Thus, the probe measurements of the profiles of $T_e(x)$, $n_e(x)$, and $n_e(r)$ proved high spatial homogeneity of the main plasma parameters in the rf discharge region, where the dust cloud hovers and DAI develops. In this respect, the given discharge is advantageous as compared, for example, with the dc glow discharge [16, 17] for which considerable gradients of the $T_e(x)$ and $n_e(x)$ profiles observed in the suspension region of the dust cloud complicate an analytic description of the phenomenon.

The neon ion concentration n_i was calculated from the quasi-neutrality relation for a dust plasma,

$$n_i = n_e + Z_d n_d. \qquad (3)$$

The ion temperature is virtually equal to the temperature of neutral neon atoms (300 K), but in the case of relatively low pressures (p < 25 Pa), it increases due to heating in the ambipolar diffusion field and amounts to 2/3 of the mean energy of ions drifting in an electric field:

$$T_i = \frac{2}{9}Ee\lambda_i. \qquad (4)$$

The charge Z_d of dust particles was calculated numerically taking into account the charge exchange of neon ions in the Debye sphere around a dust particle [23, 24]. The results of calculation of Z_d are given in the table.

3. ANALYTIC MODEL

Since the observed phenomenon, viz., dust–acoustic instability, had the form of traveling waves of dust particle concentration density, its theoretical interpretation was carried out in the framework of the theory of a dust–acoustic wave. The possibility of the existence of dust–acoustic waves with extremely low phase velocity and frequency in a nonmagnetized dust plasma was predicted for the first time in 1990 [26]. Since then, many publications have been devoted to this problem, taking into account some aspects of the physics of this phenomenon: the extent of imperfection of the dust

plasma [28], concentration density of dust particles [29], frictional force [30], varying charge of dust particles [16], etc. The number of publications devoted to an analysis of possible reason for DAI buildup is much smaller. At the present time, the following three reasons leading to the excitation of DAI are mainly considered: the drag force by an ion flow [22], charge-dependent variable electric forces exciting the vibrational motion of dust particles [16], and ionization processes [30]. Possible reasons for buildup of vibrations are investigated through an analysis of the imaginary component of the corresponding dispersion equation. The effect of ionization processes was disregarded by us since the dust cloud was outside the region of energy evolution from the inductive rf discharge.

It was noted above that the DAI parameters ν_d, λ_d, ω_d, and k_d were measured in the upper part of the dust cloud, where the amplitude of vibrations is minimal, so that a comparison of experimental data with the results of the theory of linear dust–acoustic waves, which is required for interpreting experimental data, was possible. In order to find the dispersion relation for a dust–acoustic wave,

$$k(\omega) = k_{\mathrm{Re}}(\omega) + ik_{\mathrm{Im}}(\omega) = k_{\mathrm{Re}}(\omega) - i\gamma(\omega), \quad (5)$$

we solved the Poisson equation linearized for small harmonic perturbations,

$$\delta\varphi \propto \exp(ikx - i\omega t),$$

of the electrostatic potential of the dust–acoustic wave [16, 26, 30],

$$k^2\delta\varphi = 4\pi e(-\delta n_e + \delta n_i - Z_d\delta n_d - n_d\delta Z_d), \quad (6)$$

where δn_e, δn_i, and δn_d are the perturbations of number densities of electrons, ions, and dust particles, respectively, and δZ_d is the perturbation of the dust particle charge. Since the geometry of the observed phenomenon is one-dimensional (the coordinate axis is directed along the x axis of the tube), the theoretical analysis of the problem will also be one-dimensional. The choice of the space–time dependence

$$\delta\varphi \sim \exp(ikx - i\omega t)$$

is dictated by the fact that the observed density perturbation δn_d of dust particles was successfully approximated by this dependence.

Let us determine the dependences of δn_e, δn_i, δZ_d, and δn_d on $\delta\varphi$, assuming that the wave amplitude is small. In accordance with probe measurements, the energy distribution of electrons is successfully described by the Boltzmann equation

$$n_e = n_{e0}\exp\left(\frac{e\varphi}{T_e}\right). \quad (7)$$

Since $T_e \gg e\delta\varphi$, for a small harmonic component $\delta\varphi$, Eq. (7) gives

$$\delta n_e = \left(\frac{n_{e0}e}{T_e}\right)\delta\varphi. \quad (8)$$

Since $T_i \ll T_e$, the wave potential $\delta\varphi$ can be comparable with T_i, and approximation (8) for ions will be incorrect. In this case, the ion density perturbation in the field of a harmonic electrostatic wave is determined in the flux approximation from the continuity equation for particle and momentum fluxes:

$$\frac{\partial n_i}{\partial t} + \frac{\partial(n_i u_i)}{\partial x} = 0, \quad (9)$$

$$n_i m_i\left(\frac{\partial u_i}{\partial t} + u_i\frac{\partial u_i}{\partial x} + u_i\nu_{in}\right) = n_i eE - T_i\frac{\partial n_i}{\partial x}, \quad (10)$$

where n_i is the concentration of ions, u_i is the velocity of directional motion (drift) of ions along the x axis, m_i is the ion mass, ν_{in} is the frequency of ion collisions with neutrals, and $E = E_0 - \partial\delta\varphi/dx$, E_0 is the constant electric field strength. It should be noted that the frequencies of self-excited vibrations amount to tens of hertz, which is much lower than the electron as well as the ion plasma frequency and also the rate of stabilization of the dust particle charge. For this reason, the time derivatives $\partial n_i/\partial t$ and $\partial u/\partial t$ are approximately equal to zero. Equations (9) and (10) in the linear approximation for harmonic corrections lead to

$$ik(u\delta n_i + n_{i0}\delta u) = 0, \quad (11)$$

$$E_0 e\delta n_i - ikeen_{i0}\delta\varphi - ikT_i\delta n_i = 0, \quad (12)$$

which gives

$$\delta n_i = i\frac{ken_{i0}}{eE_0 - ikT_i}\delta\varphi. \quad (13)$$

The continuity equation for the dust particle flux and Newton's second law have the form

$$\frac{\partial n_d}{\partial t} + \frac{\partial(n_d u_d)}{\partial x} = 0, \quad (14)$$

$$\frac{\partial u_d}{\partial t} + u_d\frac{\partial u_d}{\partial x} = -\frac{Z_d eE_0}{m_d} + \frac{F_\Sigma}{m_d} - \eta u_d, \quad (15)$$

where u_d is the velocity of a dust particle along the x axis; F_Σ is the sum of forces (force of gravity, thermophoretic force, and ion drag force) balancing the electric force on the average during the period of vibrations,

$$\bar{F}_\Sigma = Z_{d0}eE_0;$$

and η is the coefficient of viscous friction of a particle against the gas. In the linear approximation, we obtain

from Eqs. (14) and (15) the following expressions for harmonic corrections:

$$ikn_{d0}\delta v_d = i\omega\delta n_d, \tag{16}$$

$$-i\omega\delta n_d = -\frac{eE_0}{M_d}\delta Z_d + i\frac{keZ_0}{M_p}\delta\varphi - \eta\delta v_d, \tag{17}$$

which gives

$$\delta n_d = -n_{d0}\frac{k^2 Z_d e\delta\varphi + ikeE_0\delta Z_d}{M_d\omega(\omega + i\eta)}. \tag{18}$$

It remains for us to find δZ_d. The charge Z_{d0} of a dust particle depends on the ratio of the electron and ion concentrations. Accordingly, we can write

$$\begin{aligned}\delta Z_d &= Z_{d0}\chi\left(\frac{\delta n_e}{n_{e0}} + \frac{\delta n_i}{n_{i0}}\right)\\ &= Z_{d0}\chi\left(\frac{e\delta\varphi}{T_e} + \frac{ike\delta\varphi}{ikT_i - eE_0}\right),\end{aligned} \tag{19}$$

where $\chi(n_{i0}/n_{e0})$ is the logarithmic derivative of the dust particle charge with respect to the ratio n_{e0}/n_{i0}. Using the expression for the dust particle charge Z_d in the framework of the orbital approximation [31], we can derive the following formula:

$$\frac{\delta Z_d}{Z_{d0}} = \frac{1 + \dfrac{T_i}{e\varphi_d}}{1 + \dfrac{T_i}{T_e} + \dfrac{e\varphi_d}{T_e}}\frac{\delta(n_e/n_i)}{(n_e/n_i)}, \tag{20}$$

where $\varphi_d \approx eZ_{d0}/R_d$ is the potential on the surface of a dust particle. For $T_e/T_i \approx 100$ of the neon plasma, we have $\varphi_d e/T_e \approx 2$ and $\chi \approx 0.33$. It is important to note that unperturbed concentrations of electrons and ions appear in formula (20), while Eq. (6) contains volume-averaged quantities. If we assume that the charge of each dust particle is completely screened by the ion cloud, and the distance between particles is much larger than the screening radius, a change in the number density of dust particles will lead to a change in the mean concentration of ions, but the unperturbed concentration remains unchanged. For this reason, the value of χ may in fact be much smaller than 0.33; i.e., $0 < \chi \ll 0.33$. Physically, the case $\chi = 0$ corresponds to invariability of the charge of particle in the DAI wave.

Substituting expressions (8), (11), (14), and (15) into Eq. (6) and disregarding small terms, we obtain the generalized dispersion equation

$$1 + \frac{\chi P}{1+P} + \tilde{k}^2 + i\tilde{E}\tilde{k} = \frac{\tilde{k}^2 + i\tilde{E}\tilde{k}(1+\chi)}{\tilde{\omega}(\tilde{\omega} + i\tilde{\eta})} \tag{21}$$

with the dimensionless parameters

$$\tilde{k} = kr_{Di}, \quad \tilde{E} = eE_0 r_{Di}/T_i, \quad \tilde{\omega} = \omega/\omega_{pd},$$

$$\tilde{\eta} = \eta/\omega_{pd}, \quad P = Z_{d0}n_{d0}/n_{e0},$$

where r_{Di} and r_{De} are the ion and electron Debye radii and

$$\omega_{pd} = 2eZ_{d0}(\pi n_d/m_d)^{-1/2}$$

is the dust plasma frequency. This equation was solved numerically with the help of the Mathcad-2000 utility relative to

$$\tilde{k}_{\mathrm{Re}} = \tilde{k}_{\mathrm{Re}}(\tilde{\omega}) = k_{\mathrm{Re}}(\omega/\omega_{pd})r_{Di}$$

and

$$k_{\mathrm{Im}} = k_{\mathrm{Im}}(\omega) = -\gamma(\omega)$$

for the quantities r_{Di}, E_0, T_i, and ω_{pd} corresponding to the experimental conditions from the table. The results of these calculations, their comparison with the experimental data, and an analysis of the results are presented in the next section.

4. COMPARISON OF THE ANALYTIC MODEL WITH EXPERIMENTAL DATA

In our experiments, it is quite difficult to determine the dispersion relation $k_{\mathrm{Re}} = k_{\mathrm{Re}}(\omega)$ in a wide frequency range, since the experimenter cannot deliberately change the value of ω as, for example, in the case of excitation of vibrations by an electric potential [32] or by the force of pressure exerted by a laser beam on particles [33]. In our case, there exist two main ways of affecting the DAI frequency: by changing the value of n_d and by varying the pressure of the plasma-forming gas. In this case, however, the plasma–dust frequency ω_{pd} also changes. For this reason, the experimental dispersion relation $k_{\mathrm{Re}} = k_{\mathrm{Re}}(\omega)$ can be plotted only in the reduced coordinates

$$\tilde{\omega} = \omega/\omega_{pd}$$

and

$$\tilde{k}_r = k_r r_{Di}.$$

A change in the power of the inductive rf discharge (other parameters remaining the same) does not lead to a noticeable change in the parameters of observed DAI since the dust cloud for any discharge power was always at the edge of the discharge (see Fig. 1), where the plasma parameters were always approximately the same under a given pressure of the plasma-forming gas. Since the value of the logarithmic derivative χ is not determined exactly and may vary in the range from 0 to 0.33, we calculated $\tilde{k}_{\mathrm{Re}} = \tilde{k}_{\mathrm{Re}}(\omega)$ and $k_{\mathrm{Im}} = k_{\mathrm{Im}}(\omega)$ for the limiting cases $\chi = 0$ and 0.33. The results of calculations of the dispersion curves $\tilde{k}_{\mathrm{Re}} = \tilde{k}_{\mathrm{Re}}(\tilde{\omega})$ under the conditions of seven experiments (see table) together with seven experimental points are presented in Fig. 6. An analysis of the distribution of experimental points

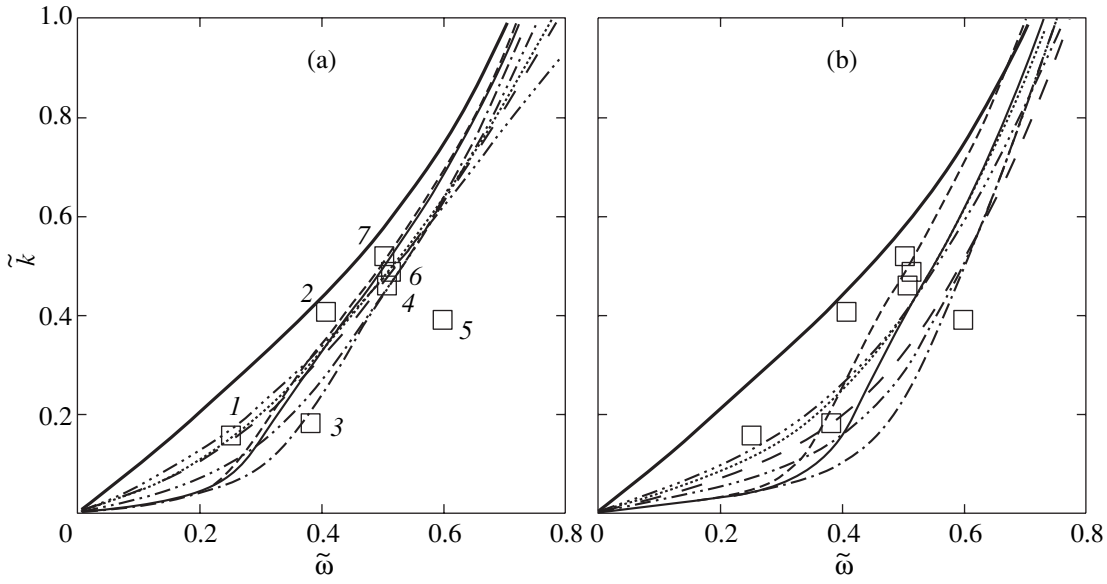


Fig. 6. Measured (squares) and calculated (curves) dispersion dependences of the reduced wave number $\tilde{k} = kr_D$ on the reduced frequency $\tilde{\omega} = \omega/\omega_{pd}$ of the dust–acoustic instability for $\chi = 0$ (a) and 0.33 (b) from Eq. (19). The numbers on the squares correspond to the number of the experiment from the table. The bold curve describes the dispersion dependence for collisionless dust sound, calculated by formula (22). Thin curves correspond to numerical calculations based on formula (21) for the conditions of experiments no. 1 (solid curve), no. 2 (short-dashed curve), no. 3 (long-dashed curve), no. 4 (dot-and-dashed curve), no. 5 (double-dot-and-dashed curve), no. 6 (dashed curve), and no. 7 (triple-dot-and-dashed curve).

on the $(\tilde{\omega}, \tilde{k}_{\mathrm{Re}})$ plane shows that the most effective method of changing the DAI reduced frequency $\tilde{\omega}$ is to change the concentration n_d of dust particles in the cloud. It can be seen that the results of measurements $(\tilde{\omega}, \tilde{k}_{\mathrm{Re}})$ for a high concentration of particles (experiments 2, 4–7) are grouped separately from the results of measurements of $(\tilde{\omega}, \tilde{k}_{\mathrm{Re}})$ for a low concentration of particles (experiments 1 and 3), while the change in the pressure of the plasma-forming gas from 10 to 50 Pa did change the value of $\tilde{\omega}$ significantly. A certain spread in the experimental points can be attributed to the error in determining the values of ω_{pd} and r_{Di}. Figure 6 also shows the bold dispersion curve corresponding to the "classical" collisionless dust sound [26]

$$\tilde{k} = \tilde{\omega}/\sqrt{1-\tilde{\omega}^2}. \qquad (22)$$

It can be seen from Fig. 6 that all the seven calculated dependences $\tilde{k}_{\mathrm{Re}} = \tilde{k}_{\mathrm{Re}}(\tilde{\omega})$ are grouped compactly on the plane $(\tilde{\omega}, \tilde{k}_{\mathrm{Re}})$ and correlate with the experimental data better than the dispersion curve corresponding to collisionless dust sound (22). The difference in calculations of $\tilde{k}_{\mathrm{Re}} = \tilde{k}_{\mathrm{Re}}(\tilde{\omega})$ for $\chi = 0$ and 0.33 is insignificant.

Let us now analyze the results of calculations of the growth increment $\gamma = \gamma(\omega)$ presented in Fig. 7 also for two limiting values of $\chi = 0$ and 0.33. A comparison of these theoretical curves with the DAI experimental parameters from the table leads to the following conclusions.

1. The frequencies $\omega_{\max}$ corresponding to the peaks $\gamma_{\max} = \gamma(\omega_{\max})$ of the theoretical curves correlate with the experimentally measured DAI frequencies.

2. The calculated values of the growth increment $\gamma_{\max} = \gamma(\omega_{\max})$ correlate with the values of the growth increment γ estimated in experiments.

3. The width of the "amplification band" of the calculated dependence $\gamma = \gamma(\omega)$ increases upon a decrease in the density of neutrals (cf. curve *2* and curves *5*, *6*, and *7*), which is, in fact, observed in experiments: dust waves become less regular upon a decrease in the pressure of the plasma-forming gas.

4. As the density of neutrals increases, the maximum calculated value $\gamma_{\max} = \gamma(\omega_{\max})$ becomes smaller and smaller and assumes a negative value at pressures above 80 Pa, which is actually observed in experiments: no DAI is observed for $p > 60$ Pa.

5. Other conditions being equal, the calculated value of $\gamma_{\max} = \gamma(\omega_{\max})$ is proportional to the concentration of dust particles (cf. curves *3* and *4*), which is indeed observed in experiments: self-excitation of DAI occurs only when n_d exceeds a certain threshold value, which is determined by the strength of the background constant electric field E and viscosity η of the medium.

6. The calculated value of $\omega_{\max}$ increases with the number density n_d of dust particles, which is observed in experiments (cf. curves *3* and *4*).

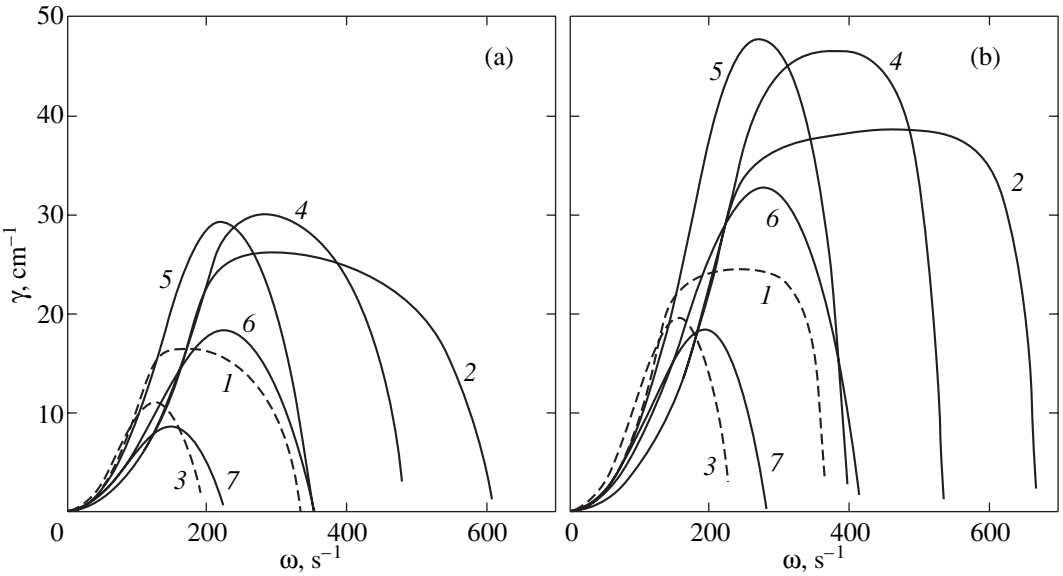


Fig. 7. Calculated dependences of the increment of increase $\gamma(\omega)$ in the amplitude of a dust–acoustic instability wave for $\chi = 0$ (a) and 0.33 (b) obtained from Eq. (19) for seven experimental conditions from the table. The figures on the curves correspond to the number of the experiments from the table, whose condition were used in Eq. (21) for calculating $\gamma = \gamma(\omega)$.

7. The presence of a constant electric field is a necessary condition for the self-excitation of dust waves: forced displacement of the dust cloud by 5 mm in the upward direction by the gas flow towards the region with lower values of the electric field strength led to wave damping.

8. An analysis of Eq. (2) shows that its solutions exist only for $E_0 k > 0$, which is indeed observed in experiments: DAI waves propagate only in the direction of the electric field.

Thus, we can conclude that the experimental results and the predictions of the proposed analytical model are in good qualitative (and in many cases quantitative) agreement in a wide range of experimental conditions. It follows hence that DAI is a dust–acoustic wave excited in a constant electric field of ambipolar diffusion. As regards the presence or absence of electric charge fluctuations δZ_d, it is difficult to draw any definite conclusion in view of a considerable error in the measurement of the background plasma parameters. Theoretically, the existence of the maximum possible fluctuation δZ_d for $\chi = 0.33$ increases the value of $\gamma_{max} = \gamma(\omega_{max})$ approximately by a factor of 1.5, but does not change anything qualitatively (see Figs. 6 and 7).

5. CONCLUSIONS

In this work, we have carried out a complex experimental and theoretical analysis of dust–acoustic instability developing spontaneously in the plasma of an rf inductive low-pressure gas discharge. We studied the DAI experimentally in the region of the diffusive edge of the discharge, i.e., the region of transition from the discharge region to the neutral gas. Under certain experimental conditions, a dust cloud formed by micrometer-size particles hovered in the region of this edge, and dust particle density waves were excited. The parameters of these waves were measured in a wide range of experimental conditions. The background plasma parameters (electron concentration and temperature, electric field strength) were determined from probe measurements. These measurements proved that the background plasma parameters within the dust cloud are homogeneous enough for carrying out analytical investigations of DAI.

The analytic description of the observed waves of dust particle density was obtained on the basis of the theory of dust sound in a collisional dust plasma with particles having a variable electric charge. We derived a generalized dispersion equation including explicitly the external electric field strength. Numerical solutions were obtained for the real and imaginary components of this equation as applied to the given experimental conditions. The charge of dust particles was calculated through determining the flux of electrons and ions at the surface of a dust particle in the approximation of confined orbits as well as by using numerical methods for a real collisional plasma on the basis of the measured parameters of the background plasma. We analyzed the effect of the variable charge of dust particles on the magnitude of amplitude buildup increment in the dust waves. It is shown that the maximum possible charge variation leads to an increase in the increment by a factor of 1.5. Thus, the existence of the variable charge on dust particles facilitates the development of instability, but is not, however, a necessary condition for the development of the given type of instability. On the whole,

good correlation is observed between the experimental data and the main conclusions of the model used in a wide range of experimental conditions. It is shown that the necessary condition leading to the development of DAI is the presence of a constant electric field. It can be concluded that the physical mechanism of this type of dust–acoustic instability is as follows. DAI is generated in the upper part of the cloud from random fluctuations of the dust particle number density. As the wave propagates over the dust cloud, these fluctuations are enhanced in the acoustic mode in the ambipolar diffusion field, and the selection of their mode composition takes place in view of nonuniform amplification of the wave at different frequencies. As the amplitude increases, the mode of the formed wave changes from the acoustic to the nonlinear regime.

ACKNOWLEDGMENTS

We thank the scientists from the Institute of Structural Macrokinetics, Russian Academy of Sciences (Chernogolovka), W.I. Goldshleger and S.D. Amosov, who kindly permitted us to use the Redlake 500 high-speed video camera.

This study was carried out in the framework of the complex research program “Physics and Chemistry of Extreme States of Matter” of the Russian Academy of Sciences.

REFERENCES

1. V. V. Zhakhovskiĭ, V. I. Molotkov, A. P. Nefedov, *et al.*, Pis’ma Zh. Éksp. Teor. Fiz. **66**, 392 (1997) [JETP Lett. **66**, 419 (1997)].
2. O. S. Vaulina, S. A. Khrapak, A. P. Nefedov, *et al.*, Phys. Rev. E **60**, 5959 (1999).
3. I. V. Schweigert, V. A. Schweigert, A. Melzer, and A. Piel, J. Phys. IV **10**, Pr5-417 (2000).
4. R. A. Quinn and J. Goree, Phys. Rev. E **61**, 3033 (2000).
5. R. A. Quinn and J. Goree, Phys. Plasmas **7**, 3904 (2000).
6. S. Nunomura, T. Misawa, N. Ohno, *et al.*, Phys. Rev. Lett. **83**, 1970 (1999).
7. D. Samsonov and J. Goree, Phys. Rev. E **59**, 1047 (1999).
8. G. E. Morfill, H. M. Thomas, U. Konopka, *et al.*, Phys. Rev. Lett. **83**, 1598 (1999).
9. V. E. Fortov, V. I. Molotkov, and V. M. Torchinsky, in *Fronties in Dusty Plasmas*, Ed. by Y. Nakamura, T. Yokota, and P. K. Shukla (Elsevier, Amsterdam, 2000), p. 445.
10. V. I. Vladimirov, L. V. Deputatova, A. P. Nefedov, *et al.*, Zh. Éksp. Teor. Fiz. **120**, 353 (2001) [JETP **93**, 313 (2001)].
11. G. E. Morfill, H. M. Thomas, U. Konopka, *et al.*, in *Proceedings of the PKE-Nefedov Symposium, Munich, 2001*, Oral Report.
12. A. Barkan, R. L. Merlino, and N. D’Angelo, Phys. Plasmas **2**, 3563 (1995).
13. R. L. Merlino, A. Barkan, C. Thompson, and N. D’Angelo, Phys. Plasmas **5**, 1607 (1998).
14. V. I. Molotkov, A. P. Nefedov, V. M. Torchinskiĭ, *et al.*, Zh. Éksp. Teor. Fiz. **116**, 902 (1999) [JETP **89**, 477 (1999)].
15. S. Iizuka, R. Ozaki, G. Uchida, and N. Sato, in *Fronties in Dusty Plasmas*, Ed. by Y. Nakamura, T. Yokota, and P. K. Shukla (Elsevier, Amsterdam, 2000), p. 453.
16. V. E. Fortov, A. G. Khrapak, S. A. Khrapak, *et al.*, Phys. Plasmas **7**, 1374 (2000).
17. A. A. Samaryan, A. V. Chernyshev, O. F. Petrov, *et al.*, Zh. Éksp. Teor. Fiz. **119**, 524 (2001) [JETP **92**, 454 (2001)].
18. O. S. Vaulina, A. P. Nefedov, O. F. Petrov, and V. E. Fortov, Zh. Éksp. Teor. Fiz. **118**, 351 (2000) [JETP **91**, 307 (2000)].
19. R. K. Varma, Phys. Plasmas **8**, 3154 (2001).
20. J. H. Chu, Ji-Bin Du, and Lin I, J. Phys. D **27**, 296 (1994).
21. N. D’Angelo, J. Phys. D **28**, 1009 (1995).
22. M. Rosenberg, Planet. Space Sci. **41**, 229 (1993).
23. A. V. Zobnin, A. P. Nefedov, V. A. Sinel’shchikov, *et al.*, Zh. Éksp. Teor. Fiz. **118**, 554 (2000) [JETP **91**, 483 (2000)].
24. M. Lampe, V. Gavrishchaka, G. Ganguli, and G. Joyce, Phys. Rev. Lett. **86**, 5278 (2001).
25. A. V. Zobnin, A. P. Nefedov, V. A. Sinel’shchikov, *et al.*, Fiz. Plazmy **26**, 445 (2000) [Plasma Phys. Rep. **26**, 415 (2000)].
26. N. N. Rao, P. K. Shukla, and M. Y. Yu, Planet. Space Sci. **38**, 543 (1990).
27. P. K. Shukla, Phys. Plasmas **8**, 1791 (2001).
28. D. Winske, M. S. Murillo, and M. Rosenberg, Phys. Rev. E **59**, 2263 (1999).
29. N. N. Rao, Phys. Plasmas **7**, 795 (2000).
30. A. V. Ivlev, D. Samsonov, J. Goree, and G. Morfill, Phys. Plasmas **6**, 741 (1999).
31. J. E. Allen, Phys. Scr. **45**, 497 (1992).
32. J. B. Pieper and J. Goree, Phys. Rev. Lett. **77**, 3137 (1996).
33. A. Homman, A. Melzer, S. Peters, and A. Piel, Phys. Rev. E **56**, 7138 (1997).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 440–446.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 513–520.

FLUIDS

The Distribution Function for a Subsystem Experiencing Temperature Fluctuations

A. G. Bashkirov[a, *] and A. D. Sukhanov[b]

[a]*Institute of Dynamics of Geospheres, Russian Academy of Sciences, Moscow, 117334 Russia*
[b]*Russian University of Peoples' Friendship, Moscow, 117198 Russia*
**e-mail: abas@idg.chph.ras.ru*

Received February 15, 2002

Abstract—A nonlinear generalization of the Landau–Lifshitz theory of hydrodynamic fluctuations for the simplest case in which only energy flux and temperature fluctuations are observed is used to derive the distribution function for a subsystem with a fluctuating temperature, which coincides with the Levy distribution taken to be one of the main results of the so-called Tsallis's nonextensive statistics. It is demonstrated that the same distribution function is obtained from the principle of maximum of information entropy if the latter is provided by Renyi's entropy, which is an extensive quantity. The obtained distribution function is to be used instead of the Gibbs distribution in constructing the thermodynamics of systems with significant temperature fluctuations. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

When standard methods of statistical mechanics are used, the smallness of the mean square temperature fluctuation of the system is assumed, as a rule, which is estimated as [1]

$$\delta T/T_0 = \left(\frac{k_B}{C_V}\right)^{1/2}.$$

This assumption is justified if the heat capacity C_V of the system is high enough and the temperature T_0 is not too low. Examples in which this condition may be invalid include the atomic nucleus, for which the notion of temperature was successfully introduced by Landau, Frenkel, and Weisskopf [2], and low-temperature systems in which $T_0 \longrightarrow 0$ and, at the same time, $C_V \longrightarrow 0$ [3]. In this paper, we treat the temperature fluctuations of a small subsystem placed in a thermostat and use the Landau–Lifshitz theory of hydrodynamic fluctuations [4] to derive the gamma distribution for these fluctuations. The dispersion of the thus obtained gamma distribution depends on C_V and produces the foregoing estimate for δT. Then, this distribution function is used for averaging the Gibbs canonical distribution over temperatures, which brings about the Levy distribution or q distribution. In application to interpreting nuclear collisions, such an approach was treated by Wilk and Wlodarczyk [5], who explained the resultant Levy distribution in the light of so-called nonextensive statistics based on the variational principle of extremality of Tsallis's information entropy [6]. In view of the rapid development of nonextensive statistics covering the widest scope of problems (from cosmology to intranuclear processes), we found it necessary to dwell in brief on the problem of choosing the form of information entropy. Best justified is the Renyi one-parameter family of entropies (or simply Renyi's entropy). When the Renyi entropy parameter q is unity, the entropy transforms to the well-known Boltzmann–Shannon entropy. Renyi's entropy is additive; however, in the case of linearization in the neighborhood of $q \approx 1$, it loses additivity and changes to Tsallis' entropy. The application of the principle of extremality of information entropy to Renyi's entropy leads to precisely the same Levy distribution which is obtained during averaging of the Gibbs distribution over the temperature. This fact enables one to have a fresh view of the physical meaning of the Renyi parameter q and Renyi's entropy proper.

2. LANGEVIN EQUATION FOR TEMPERATURE

We will treat a subsystem which is a minor part of a large equilibrium system and experiences fluctuations of both energy and temperature. This is the radical difference of the suggested approach from the Gibbs approach traditionally employed in statistical physics, in which the temperature is preassigned by the constant characterizing the thermostat.

In order to analyze the temperature fluctuations, we will invoke the Landau–Lifshitz theory of hydrodynamic fluctuations, in which respective fluctuation analogs are added to regular flows of mass, momentum, and energy entering the set of hydrodynamic equations.

No flows of mass and momentum are present in the case treated by us; however, an energy flux must be observed because of the temperature fluctuations.

Then, the equation of conservation of energy density of the system $E(\mathbf{r}, t)$ takes the form

$$\frac{\partial E(\mathbf{r}, t)}{\partial t} = -\mathrm{div}(\mathbf{q}^R(\mathbf{r}, t) + \mathbf{q}^F(\mathbf{r}, t)), \quad (1)$$

where $\mathbf{q}^R(\mathbf{r}, t)$ describes a regular flow of energy density, and $\mathbf{q}^F(\mathbf{r}, t)$ represents the flow fluctuations.

We will single out in the system a subsystem of preassigned volume V. We integrate Eq. (1) with respect to the volume V and use the Gauss–Ostrogradsky formula to derive the equation of conservation for the energy of this subsystem,

$$\frac{d\bar{E}(t)}{dt} = -Q^R(t) - Q^F(t),$$
$$Q^{R,F} = \int_A dA\mathbf{q}^{R,F}, \quad (2)$$

where the surface area A of the singled-out subsystem is introduced.

We will further restrict ourselves to taking into account the fluctuations of only one parameter, namely, temperature, and represent the energy $\bar{E}(t)$ in the form of $\bar{E}(t) = C_V T(t)$, where C_V is the heat capacity of the subsystem. In addition, the flux Q^R may be conveniently written in the form of the heat-transfer equation

$$Q^R(t) = A\kappa(T(t) - T_0), \quad (3)$$

where κ is the heat-transfer coefficient, and T_0 is the average temperature of the system. Then, Eq. (2) takes the form

$$C_V\frac{dT(t)}{dt} = -A\kappa(T(t) - T_0) - Q^F(t). \quad (4)$$

This equation is the Langevin equation for the temperature which characterizes the singled-out part of the system and fluctuates under the effect of a random energy flux $Q^F(t)$ through the boundary of the discontinuous system being treated.

In the nonequilibrium linear thermodynamics [7], the thermodynamic force conjugate to the flux Q^R is provided by the quantity

$$\left(\frac{1}{k_B T_0} - \frac{1}{K_B T(t)}\right) \approx \frac{1}{k_B T_0^2}(T(t) - T_0) \quad (5)$$

rather than by the temperature difference.

Accordingly, the kinetic coefficient of the heat-transfer equation must have the form $k_B T_0^2 A\kappa$. Then, according to the Landau–Lifshitz theory of hydrodynamic fluctuations, in a linear approximation with respect to deviation from equilibrium, the stochastic properties of random flux have the form

$$\langle Q_l^F(t)\rangle = 0,$$
$$\langle Q_l^F(t)Q_l^F(t')\rangle = 2k_B T_0^2 A\kappa\delta(t - t'). \quad (6)$$

The second one of these expressions indicates that, within the linear theory, $Q_l^F \propto T_0$. This fact suggests a simple way of including the nonlinearity by replacing $Q_l^F(t)$ by $Q^F(t) = T(t)\xi(t)$, where $\xi(t)$ is a random function of time satisfying the relations

$$\langle\xi(t)\rangle = 0, \quad \langle\xi(t)\xi(t')\rangle = 2k_B A\kappa\delta(t - t'). \quad (7)$$

As a result, Eq. (4) takes the form of nonlinear stochastic Langevin equation,

$$\frac{dT(t)}{dt} = -\frac{1}{\tau}(T(t) - T_0) - \frac{1}{C_V}T(t)\xi(t), \quad (8)$$

where $\tau = C_V/A\kappa$.

3. FLUCTUATIONS OF THE SUBSYSTEM TEMPERATURE

Corresponding to the derived stochastic Langevin equation with δ-correlated noise is the Fokker–Planck kinetic equation for the temperature distribution function $f(T, t)$,

$$\frac{\partial f(T, t)}{\partial t} = -\frac{\partial}{\partial T}W_1(T)f(T, t) + \frac{1}{2}\frac{\partial^2}{\partial T^2}W_2(T)f(T, t). \quad (9)$$

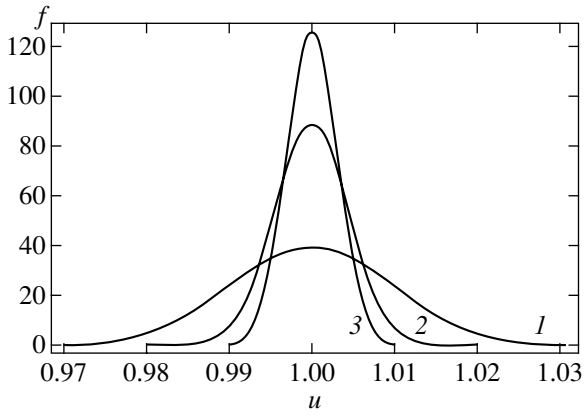
The coefficients $W_1(T)$ and $W_2(T)$ of this equation are expressed in terms of the first $\langle T(t) - T(t + \tau)\rangle$ and second $\langle(T(t) - T(t + \tau))^2\rangle$ conditional moments of stochastic equation (8), which correspond to some preassigned value of $T(t)$. For a linear stochastic Langevin equation, these moments are determined quite simply (see, for example, [8]). For nonlinear equations of the type of Eq. (8), the solution to this problem is also known and used in various applications of the theory of random processes [9]. For the case treated by us, the coefficients of the Fokker–Planck equation take the form

$$W_1(T) = -\frac{1}{\tau}(T - T_0) + k_B T\frac{1}{\tau^2 A\kappa}, \quad (10)$$

$$W_2(T) = 2k_B T^2\frac{1}{\tau^2 A\kappa}. \quad (11)$$

Because these coefficients are not explicitly dependent on time, a steady-state solution to Eq. (9) exists,

$$f(T) = \frac{K}{W_2(T)}\exp\left(2\int^T\frac{W_1}{W_2}dT\right). \quad (12)$$



Gamma distribution $f(u)$ of the temperature ratio $u = T_0/T$ for different values of the parameter $\gamma = C_V/k_B = 10^4$ (*1*), 0.5×10^5 (*2*), and 10^5 (*3*).

The constant K will be determined from the normalization condition; as a result, the choice of the lower limit of integration in the exponent is arbitrary and may be omitted. We substitute expressions (10) and (11) into (12) to derive

$$f(T) = \frac{K}{T}\exp\left(\frac{\tau A\kappa}{k_B}\int^{T}\frac{-T + T_0}{T^2}dT\right), \qquad (13)$$

whence follows

$$f(T) = KT^{-1-\gamma}\exp\left(-\frac{\gamma T_0}{T}\right), \qquad (14)$$

where the dimensionless constant $\gamma = C_V/k_B$ is introduced. Note that the resultant steady-state solution does not depend on either the heat-transfer coefficient κ or the surface area A. This fact suggests that the obtained distribution may be more general than the treated model of heat transfer.

In what follows, we will be interested in the distribution function with respect to the quantity $\beta = 1/k_BT$, rather than in the distribution function with respect to the temperature. In view of the relation

$$d\beta = -\frac{dT}{k_BT^2},$$

we derive

$$f(\beta) = K\beta^{\gamma-1}e^{-\gamma k_BT_0\beta}. \qquad (15)$$

The constant K is determined from the normalization condition reduced to

$$K^{-1} = \int_0^\infty \beta^{\gamma-1}e^{-\gamma k_BT_0\beta}d\beta = (\gamma k_BT_0)^{-\gamma}\Gamma(\gamma), \qquad (16)$$

whence we finally derive

$$f(\beta) = \frac{(\gamma k_BT_0)^\gamma}{\Gamma(\gamma)}\beta^{\gamma-1}e^{-\gamma k_BT_0\beta}. \qquad (17)$$

This function may also be represented in the form of the distribution of the temperature ratio $u = k_BT_0\beta = T_0/T$,

$$f(u) = \frac{\gamma^\gamma}{\Gamma(\gamma)}u^{\gamma-1}e^{-\gamma u}, \qquad (18)$$

or of the distribution of the dimensionless quantity $z = \gamma k_BT_0\beta = \gamma T_0/T$,

$$f(z) = \frac{1}{\Gamma(\gamma)}z^{\gamma-1}e^{-z}. \qquad (19)$$

Therefore, the thus derived distribution function of the inverse temperature of the subsystem in the dimensionless form of (19) is a gamma distribution. In concrete calculations below, we will largely use $f(\beta)$ or $f(u)$; for brevity, we will refer to them as gamma distributions as well.

Note that, if the mean energy of the singled-out volume $\bar{E}_0 = C_VT_0$ is introduced, the expression for $f(\beta)$ takes the form

$$f(\beta) = \frac{(\gamma\beta/\beta_0)^\gamma}{\beta\Gamma(\gamma)}e^{-\beta\bar{E}_0}. \qquad (20)$$

By its form, this expression is close to the Gibbs distribution; however, unlike the latter, it reflects the inclusion of the temperature fluctuation of the subsystem with the preassigned energy $\bar{E}_0$. As was to be expected, the temperature fluctuation described by this distribution is of the order of $\delta T/T_0 \approx (C_V/k_B)^{-1/2}$, which coincides with the estimate of the temperature fluctuation in equilibrium statistical physics [1]. The figure shows the form of gamma distribution $f(u)$ at $\gamma = 10^4$, 0.5×10^5, 10^5. One can see that the dispersion of the inverse temperature of the subsystem decreases abruptly with increasing γ. However, small values of γ correspond to very small subsystems. Indeed, for an ideal monatomic gas at normal conditions, $\gamma = 3N/2$, where N is the number of particles in the volume; according to the Avogadro law, $N \approx 0.3 \times 10^{20}V$. Then, for the singled-out volume with the characteristic size of the order of the free path length (about 10^{-5} cm), we have the value of $\gamma = C_V/k_B = 0.5 \times 10^5$. One can see in the figure that the temperature dispersion in this case is of the order of 0.005, which coincides with the result of the thermodynamic theory of fluctuations [1] $\delta T/T_0 \approx (k_B/C_V)^{1/2}$.

It appears to be more promising to apply the thus derived relations to heterogeneous systems in which the size of small particles (for example, atomic nucleus [2]) may not exceed several angstroms, as well as to low-temperature systems with $C_V \longrightarrow 0$ at $T_0 \longrightarrow 0$.

4. DISTRIBUTION FUNCTION FOR A SUBSYSTEM

In order to describe a subsystem in contact with a large thermally equilibrium system (thermostat), the Gibbs canonical distribution is used in statistical physics (here and below, the factor G_i allowing for the number of states of energy H_i is omitted for brevity),

$$\rho_i = Q_0^{-1} e^{-\beta H_i}, \tag{21}$$

where H_i is the energy of the subsystem (the subscript i may indicate the number of the discrete energy level or the totality of the values of coordinates and momenta of molecules of the subsystem), and Q is the partition function or statistical integral. In so doing, the inverse temperature $\beta = 1/k_B T$ is taken to be a known preassigned quantity.

As was demonstrated above, the temperature may fluctuate. In view of this, the question arises as to how the Gibbs distribution is modified under the effect of temperature fluctuations. The answer to this question may be obtained by way of averaging the distribution given by Eq. (21) with the gamma distribution for temperature T (or β) obtained in the previous section.

For further treatment, ρ_i may be conveniently represented in an equivalent form,

$$\rho_i = Q^{-1} e^{-\beta \Delta H_i}, \quad Q = \sum_i e^{-\beta \Delta H_i}, \qquad \Delta H_i = H_i - \langle H \rangle, \tag{22}$$

where the symbol $\sum_i$ may indicate both the summation and integration over a continuous totality of the values of coordinates and momenta.

The temperature dependence of ρ_i is defined both by the factor β in the exponent and, in the general case, by the unknown temperature dependence of partition function $Q(\beta)$ (in the simplest case of classical ideal gas, $Q \propto \beta^{-3N/2}$, where N is the number of molecules). Therefore, we will use the mean value theorem and represent the Gibbs distribution averaged over β in the form

$$\bar{\rho}_i = \int_0^\infty d\beta\, f(\beta) \rho_i = \frac{1}{Q^*} \int_0^\infty d\beta\, f(\beta) e^{-\beta \Delta H_i}, \tag{23}$$

where Q^* lies in the range of possible variation of $Q(\beta)$ from $Q(0)$ to $Q(\infty)$. From the conditions of normalization to unity of the distributions of $f(\beta)$ and ρ, we have

$$\frac{1}{Q^*} \sum_i \int_0^\infty d\beta\, f(\beta) e^{-\beta \Delta H_i} = 1, \tag{24}$$

whence we find

$$Q^* = \sum_i \int_0^\infty d\beta\, f(\beta) e^{-\beta \Delta H_i}. \tag{25}$$

Therefore, it is sufficient to calculate only the average value of the exponent,

$$\int_0^\infty d\beta\, f(\beta) e^{-\beta \Delta H_i} = \frac{(\gamma k_B T_0)^\gamma}{\Gamma(\gamma)} \beta^{\gamma - 1} \times \int_0^\infty d\beta \beta^{\gamma - 1} e^{-\beta(\gamma k_B T_0 + \Delta H_i)} = \left(1 + \frac{\beta_0}{\gamma} \Delta H_i\right)^{-\gamma}. \tag{26}$$

Finally, the averaged Gibbs distribution takes the form

$$\bar{\rho}_i = \frac{\left(1 + \frac{\beta_0}{\gamma}(H_i - \langle H \rangle)\right)^{-\gamma}}{\sum_i \left(1 + \frac{\beta_0}{\gamma}(H_i - \langle H \rangle)\right)^{-\gamma}}. \tag{27}$$

In the $\gamma \longrightarrow \infty$ limit corresponding to a high heat capacity of the singled-out subsystem, $\bar{\rho}_i$ goes to ρ_i.

5. INFORMATION ENTROPY

The information entropy, or simply entropy, is the measure of uncertainty in information in the case of statistical (incomplete) description of a system using the distribution of probabilities $p = \{p_i\}$, $0 \le p_i \le 1$, $i = 1, \ldots, N$. The best known is the representation of entropy in the Boltzmann–Shannon form,

$$S_B = -\sum_i^N p_i \ln p_i. \tag{28}$$

In the case when the subscripts i indicate dynamic microstates in the Gibbs phase space and the distribution p_i corresponds to the macroscopic equilibrium state of the system, the entropy S_B coincides with the thermodynamic entropy.

On the other hand, the Gibbs distribution provides for the extremality of the Boltzmann–Gibbs (or Boltzmann–Shannon) entropy, which makes up the content of the variation principle of maximum of information entropy (see, for example, [10]) providing for the uniqueness of the Gibbs distribution under appropriate additional conditions. In view of this, the need arises for critical analysis of the basic provisions of this principle.

Consider how justified the choice of this particular form of entropy is. Formally, the information entropy in

the Boltzmann–Shannon form is uniquely defined by four axioms of Khinchin [11]:

(i) $S(p)$ is defined by the probabilities p_i alone and does not depend on any other properties of the ith states.

(ii) $S(p)$ is maximal in the case of uniform distribution of probabilities, $p_i = 1/N$.

(iii) $S(p)$ does not vary when the number of states N is increased by one or more states with zero probability, i.e., $S(p_1, \ldots, p_N) = S(p_1, \ldots, p_N, 0)$.

(iv) The fourth axiom treats a system consisting of two subsystems A and B, so that p_{ij} of the composite system is represented in the form $p_{ij} = Q(j|i)\,p_i$, where the subscript i relates to A and the subscript j to B, and $Q(j|i)$ is the conditional probability of finding B in the jth state if A is in the ith state. Then, the quantity $S(p^{A+B})$ must satisfy the relation

$$S(p^{A+B}) = S(p^A) + \sum_i p_i^A S(Q|i),$$

where the conditional entropy $S(Q|i)$ has the form

$$S(Q|i) = -\sum_j Q(j|i)\ln Q(j|i).$$

It is this latter axiom that provides for a unique determination of information entropy in the Boltzmann–Shannon form. Note that, in order to validate such a formulation of this axiom (number 2^o in [11]), Khinchin [11] used the expression for conditional entropy $S(Q|i)$ derived by substituting $S(p^{A+B})$ into the *a priori* known formula (28) for the Boltzmann–Shannon entropy. Therefore, axiom (iv) turns out to be an artificial restriction for preferring the information entropy in the Boltzmann–Shannon form. Therefore, the theorem of existence and uniqueness of the Boltzmann–Shannon entropy, proved by Khinchin [11] using axioms (i)–(iv), appears quite natural. We have not come across any other validation in the literature of such a formulation of axiom (iv).

A critical analysis of the similar Shannon theorem, which validates the unique derivation of entropy in the Boltzmann–Shannon form, was made by Uffink [12]. In this case as well, the axiom associated with conditional entropy turns out to be an obstacle.

Most convincing appears to be the system of axioms of Shore and Johnson [13] leading to Renyi's one-parameter family of entropies [14]; for the distribution $\{p_i\}$ normalized to unity, this family is written in the form

$$S_R^{(q)}(p) = \frac{1}{1-q}\ln\sum_i^N p_i^q, \quad \sum_i^N p_i = 1, \qquad (29)$$

where q is an arbitrary positive number (it cannot be less than zero, because $\{p_i\}$ may include zero values).

If Khinchin's fourth axiom is moderated, and only the independent subsystems A and B are treated, for which $p_{ij}^{A+B} = p_i^A p_j^B$, the fourth axiom takes the form (iv')

$$S(p) = S(p^A) + S(p^B).$$

The combinations of axioms (i)–(iii) and (iv') are satisfied both by the Boltzmann–Shannon entropy and by a more general form of Renyi's information entropy. One can readily see that, at $q = 1$, Renyi's entropy goes to the Boltzmann–Shannon entropy,

$$S_R^{(q=1)}(p) = S_B(p).$$

Various properties of Renyi's entropy are discussed, in particular, in the monographs [14–16].

In the case of $\left|1 - \sum_i p_i^q\right| \ll 1$ (which, in view of normalization of the distribution $\{p_i\}$, corresponds to the condition $|1-q| \ll 1$), one can restrict oneself to the linear term of log expansion in the expression for $S_R^{(q)}(p)$ over this difference, and $S_R^{(q)}(p)$ changes to

$$S_T^{(q)}(p) = -\frac{1}{1-q}\left(1 - \sum_i^N p_i^q\right). \qquad (30)$$

Such a linearization of Renyi's entropy was first suggested by Daroczy [17]; at present, this expression for entropy came to be known as Tsallis' entropy [6].

The log linearization results in the entropy becoming nonextensive. This property is widely used by Tsallis and by the international scientific school that has developed around him for the investigation of diverse nonextensive systems [5, 6, 18–20]. In so doing, the above-identified restriction $|1-q| \ll 1$ is disregarded, as a rule. In our opinion, the attempt by Abe [20] at independent validation of this form of entropy appears unconvincing, because it is based on the axiomatic introduction of such a form for conditional entropy which uniquely provides for obtaining Tsallis' entropy.

6. EXTREMALITY OF ENTROPY

According to the variation principle of extremality of information entropy, in the case of statistical (incomplete, from the dynamic standpoint) description of the system, its distribution function must provide for correct values of those few average quantities which appear in the statistical description; otherwise, it must be as indeterminate as possible. Such an approach in application to equilibrium thermodynamic systems (isolated or weakly interacting with the thermostat) has long been used to construct equilibrium statistical thermodynamics (see, for example, [21] or [22]); however, it was only after studies by Jaynes [23] that it came to be firmly established as a principle validating (at least, on the physical level of rigor) the use of Gibbs ensembles in statistical description of thermodynamic sys-

tems (see, for example, [10]). The information entropy is traditionally taken to mean the Boltzmann–Shannon entropy.

Here, the principle of extremality of information entropy will be applied to Renyi's entropy.

We are interested in the distribution of probabilities $\{p_i\}$ providing for the extremality of information entropy with an additional condition which consists in preassigning the average value $\langle H\rangle$ of the random quantity H_i. This requirement must be taken into account in searching for the conditional extremum $S_R^{(q)}(p)$ along with the requirement of normalization of p_i.

Then, the distribution of probabilities $\{p_i\}$, which provides for the extremality of $S_R^{(q)}(p)$ given two conditions specified above, must be determined from the extremum of the functional

$$L(p) = \frac{1}{1-q}\ln\sum_i^N p_i^q - \alpha\sum_i^N H_i p_i - \Phi\sum_i p_i, \quad (31)$$

where α and Φ are Lagrange multipliers. Note that, in the $q \longrightarrow 1$ limit, it changes to the well-known [10] functional; its extremum is ensured by the Gibbs canonical distribution, in which $\alpha = \beta_0 = 1/k_B T_0$ is the inverse temperature and Φ is the free energy.

Prior to beginning to solve the variation problem in the conditional extremum of Renyi's entropy, we will make a remark.

The presence of the Renyi parameter q brings about a variation of the relative contributions by various probabilities p_i to Renyi's entropy. Indeed, at $q > 1$, the importance of p_i with the maximal values increases, and at $q < 1$, of p_i with the minimal values. In view of this, it proved most fruitful to introduce the so-called escort distribution [15, 16, 18]

$$P_i = \frac{p_i^q}{\sum_i p_i^q}.$$

In what follows, we will refer to the p distribution as the starting distribution. The meaning of the escort distribution may be explained as follows. If the starting distribution is represented in the form $p_i = \exp(-b_i)$, then

$$P_i = \exp[q(\Psi - b_i)], \quad \Psi = -\frac{1}{q}\ln\sum_i p_i^q.$$

In view of this meaning, it is safe to say that the transition to the escort distribution is similar to the transition to the canonical distribution: $1/q$ serves as the temperature of sorts, and Ψ as the free energy. The consistency of this transition is partly secured by the condition of conservation of the preassigned average value of energy $\langle H\rangle = \sum_i H_i P_i$. A detailed analysis of transition to this form of averaging in application to Tsallis' entropy is made in [18]. Functional (31) is represented as

$$L(p) = \frac{1}{1-q}\ln\sum_i^N p_i^q - \alpha\frac{\sum_i H_i p_i^q}{\sum_i p_i^q} - \Phi\sum_i p_i. \quad (32)$$

We equate its functional derivative to zero to derive

$$\frac{\delta L(p)}{\delta p_i} = \frac{q}{1-q}\frac{p_i^{q-1}}{\sum_j p_j^q} - \alpha q(H_i - \langle H\rangle)\frac{p_i^{q-1}}{\sum_j p_j^q} - \Phi = 0. \quad (33)$$

Then it follows from Eq. (33) that

$$p_i = (1 - \alpha(1-q)(H_i - \langle H\rangle))^{1/(1-q)} \times\left(\frac{1-q}{q}\Phi\sum_j^N p_j^q\right)^{-1/(1-q)}.$$

The condition of normalization of $\sum_i p_i = 1$ gives

$$\left(\frac{1-q}{q}\Phi\sum_j^N p_j^q\right)^{1/(1-q)} = \sum_i^N (1 - \alpha(1-q)(H_i - \langle H\rangle))^{1/(1-q)}$$

and, finally,

$$p_i = \frac{(1 - \beta_0(1-q)(H_i - \langle H\rangle))^{1/(1-q)}}{\sum_i (1 - \beta_0(1-q)(H_i - \langle H\rangle))^{1/(1-q)}}. \quad (34)$$

Here, we utilized the fact that, at $q \longrightarrow 1$, the distribution $\{p_i\}$ becomes the Gibbs canonical distribution in which the constant $\alpha = \beta_0$ is the reciprocal of the temperature.

Now, if we represent γ as $\gamma = -(1-q)^{-1}$, expression (27) will take the form

$$\bar{\rho}_i = \frac{(1 - \beta_0(1-q)(H_i - \langle H\rangle))^{1/(1-q)}}{\sum_i (1 - \beta_0(1-q)(H_i - \langle H\rangle))^{1/(1-q)}}. \quad (35)$$

The full identity of this expression (obtained by averaging the Gibbs distribution over temperature fluctuations) with the probability density p_i given by Eq. (34) and ensuring the extremality of Renyi's entropy enables

one to take a new view of the physical meaning of the Renyi entropy and parameter,

$$q = 1 + \frac{1}{\gamma} = \frac{C_V + k_B}{C_V}. \quad (36)$$

So, the Renyi parameter differs significantly from unity only in the case where the heat capacity of the singled-out system is of the same order of magnitude as the Boltzmann constant k_B. The thermodynamics of such systems must be constructed on the basis of Renyi's entropy and of the distribution function $\{\bar{\rho}_i\}$ or $\{p_i\}$ which is referred to in the literature on nonextensive thermodynamics as the Levy distribution function or q distribution. We will emphasize once again that, in this case, this function was obtained without invoking any additional considerations as to the nonextensiveness of the systems being treated.

Note in conclusion that, if we assume a Hamiltonian in the form of the power function $H(x) = hx^r$, where h is a constant, then, for $q \neq 1$ and sufficiently significant fluctuations of x exceeding the minimal value

$$x_{\min} \gg \left(\langle x^r \rangle + \frac{1}{\beta_0 h(q-1)}\right)^{1/r}, \quad (37)$$

expression (34) transforms to the power law distribution

$$p(x) \sim x^{-s}, \quad s = \frac{r}{q-1}. \quad (38)$$

So, we derived the Zipf–Pareto power distribution as a particular form of distribution (34) ensuring the extremality of Renyi's entropy at $q \neq 1$. Note that, according to expression (37), $x_{\min}$ increases abruptly at $q \longrightarrow 1$, so that the range of values of x in which the power distribution is valid goes to zero.

When applied to general stochastic systems, including biological, economic, linguistic, and others, in which the Zipf–Pareto distribution is observed, the Renyi parameter may be taken to be arbitrary, because relation (36) does not extend to these situations. In [24], the variation principle for q was formulated. According to this principle, the preferred values of q depend on $x_{\min}$ and r and lie in the range from 1.5 to 3.

ACKNOWLEDGMENTS

We are grateful to A.V. Vityazev for useful discussions of the problems treated in this paper.

REFERENCES

1. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 5: *Statistical Physics* (Nauka, Moscow, 1976; Pergamon, Oxford, 1980), Part 1.
2. L. D. Landau, Zh. Éksp. Teor. Fiz. **7**, 819 (1937); Ya. I. Frenkel', *Principles of the Nuclear Theory* (Akad. Nauk SSSR, Moscow, 1950); V. Weisskopf, in *Lecture Series in Nuclear Physics* (US Govt. Print. Off., Washington, 1947; Inostrannaya Literatura, Moscow, 1952).
3. J. Wu and A. Widom, Phys. Rev. E **57**, 5178 (1998).
4. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 5: *Statistical Physics* (Nauka, Moscow, 1978; Pergamon, Oxford, 1980), Part 2.
5. G. Wilk and Z. Wlodarczyk, Phys. Rev. Lett. **84**, 2770 (2000).
6. C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
7. S. R. de Groot and P. Mazur, *Nonequilibium Thermodynamics* (North-Holland, Amsterdam, 1962; Mir, Moscow, 1964).
8. S. Chandrasekhar, *Stochastic Problems in Physics and Astronomy* (American Inst. of Physics, New York, 1943; Inostrannaya Literatura, Moscow, 1947).
9. S. A. Akhmanov, Yu. E. D'yakov, and A. S. Chirkin, *Introduction to Statistical Radio Physics and Optics* (Nauka, Moscow, 1981).
10. D. N. Zubarev, *Noneguilibrium Statistical Thermodynamics* (Nauka, Moscow, 1971; Consultants Bureau, New York, 1974).
11. A. Ya. Khinchin, Usp. Mat. Nauk **8** (3), 3 (1953).
12. J. Uffink, Stud. Hist. Philos. Mod. Phys. B **26**, 223 (1995).
13. J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **IT-26**, 26 (1980).
14. A. Renyi, *Probability Theory* (North-Holland, Amsterdam, 1970).
15. C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems* (Cambridge Univ. Press, Cambridge, 1993).
16. Yu. L. Klimontovich, *Statistical Theory of Open Systems* (Yanus, Moscow, 1995; Kluwer, Dordrecht, 1995).
17. Z. Daroczy, Inf. Control **16**, 36 (1970).
18. C. Tsallis, R. S. Mendes, and A. R. Plastino, Physica A (Amsterdam) **261**, 534 (1998).
19. A. R. Plastino and A. Plastino, Physica A (Amsterdam) **222**, 347 (1995).
20. S. Abe, Phys. Lett. A **271**, 74 (2000).
21. E. Schrödinger, *Statistical Thermodynamics* (Cambridge Univ. Press, Cambridge, 1946; Inostrannaya Literatura, Moscow, 1948).
22. A. Katz, *Principles of Statistical Mechanics* (Freeman, San Francisco, 1967).
23. E. T. Jaynes, Phys. Rev. **106**, 620 (1957); Phys. Rev. **108**, 171 (1957).
24. A. G. Bashkirov and A. V. Vityazev, Physica A (Amsterdam) **177**, 136 (2000).

Translated by H. Bronstein

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 447–454.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 521–529.

FLUIDS

The Turbulence of Capillary Waves on the Surface of Liquid Hydrogen

M. Yu. Brazhnikov, G. V. Kolmakov, and A. A. Levchenko*

Institute of Solid-State Physics, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia

**e-mail: levch@issp.ac.ru*

Received April 6, 2002

Abstract—It is experimentally demonstrated that the surface excitation of liquid hydrogen at a low frequency results in the turbulent mode in a system of capillary waves. The experimental results are in good agreement with the theory of weak wave turbulence. The pair correlation function of the surface deviations is described by the exponential function ω^m. The exponent m decreases in magnitude from $m = -3.7 \pm 0.3$ to -2.8 ± 0.2 when the pumping at a single resonant frequency changes to broadband noise excitation. Measurements are made of the dependence of the boundary frequency ω_b of the upper edge of the inertial range in which the Kolmogorov spectrum is formed on the wave amplitude η_p at the pumping frequency. It is demonstrated that the obtained data are well described by a function of the form $\omega_b \propto \eta_p^{4/3}\omega_p^{23/9}$. © *2002 MAIK "Nauka/Interperiodica".*

1. MOTIVATION

A highly excited state of a system with numerous degrees of freedom, which is characterized by the presence of a directional (in the k space) energy flux, is referred to as turbulent. In the turbulent mode, a system finds itself away from its thermodynamic equilibrium and is characterized by a significant nonlinear interaction of the degrees of freedom, as well as by the dissipation of energy. The nonlinear interaction brings about an effective redistribution of energy between the degrees of freedom (modes).

The turbulence may be observed in systems in which the frequencies of excitation (energy pumping) and dissipation of energy are widely spaced apart on the frequency scale. Such systems include wind waves on the ocean surface [1] and large-scale flows in the Earth's atmosphere [2]. It is the interaction of these two powerful nonlinear systems that largely defines the weather. Such systems further include spin waves in solids [3] and waves in plasma [4]. Studies into the propagation of energy in such systems are of great interest from the standpoint of both fundamental nonlinear physics and practical applications.

Capillary waves on the surface of a liquid represent yet another object for studies into turbulence. The theory of weak turbulence was developed in the late 1960s [5]. However, in spite of the large number of experimental investigations of the nonlinear dynamics of surface waves, just a few reports have been published recently of the experimental observations of isotropic spectra of capillary waves on the surface of water, the results of which may be compared with the theoretical predictions.

This paper gives the results of investigation of nonlinear capillary waves on the surface of liquid hydrogen. Liquid hydrogen is a suitable object for experiments in turbulence, because it is characterized by a relatively low value of the kinematic viscosity coefficient ν and a high value of the coefficient $V \sim (\alpha/\rho^3)^{1/4}$ characterizing nonlinearity of capillary waves (α is the surface tension coefficient, and ρ is the density of liquid hydrogen). For hydrogen at a temperature $T = 15$ K, we have $\nu = 2.6 \times 10^{-3}$ cm^2/s and $V = 9$ cm$^{3/4}$/s g, and for water, $\nu = 10^{-2}$ cm^2/s and $V = 3$ cm$^{3/4}$/s g at $T = 20$°C. This enables one to examine the turbulent mode in a wide frequency range. In addition, owing to low density, an external force required to excite oscillations on the surface of liquid hydrogen is several times less than that in the case of water. This fact proved to be decisive in using the procedure in which the waves on the surface are excited by electric forces. The previous experiments have revealed [6] that one can charge the surface of liquid hydrogen with charges injected into the bulk of the liquid, hold the charges in the vicinity of the surface for a long period of time, and excite surface waves using a variable electric field. An important advantage of this procedure for the observation of capillary turbulence is the possibility of directly affecting the surface of a liquid by an external force, virtually without acting on the bulk of the liquid, as well as the high degree of isotropism of the exciting force, which enabled one to study the turbulence under well-controlled experimental conditions.

1063-7761/02/9503-0447$22.00

2. INTRODUCTION

It is known that capillary waves on the surface of a liquid represent an example of nonlinearly interacting waves and are characterized by the dissipation of energy mainly at high frequencies because of the loss due to viscosity. The theory of homogeneous capillary turbulence was described by Zakharov and Filonenko [7]. They have demonstrated that an ensemble of weakly interacting capillary waves may be described within a kinetic equation similar to the Boltzmann equation of gas dynamics [8, 9],

$$\frac{dn_k}{dt} + 2\gamma_k n_k = \mathrm{St}(n_k) + F(t),$$

where n_k is the wave distribution function in the wave vectors k, $\mathrm{St}(n_k)$ is the collision integral, γ_k is the damping coefficient, and $F(t)$ is the pumping term.

The main problem involved in the investigation of wave turbulence is that of finding the law of distribution of the energy of a system of waves with respect to frequency, i.e., the stationary spectrum of the turbulent energy E_ω. The energy E per unit surface of liquid may be written in the form

$$E = \int \omega_k n_k dk = \int \omega n(\omega) d\omega = \int E_\omega d\omega, \qquad (1)$$

where ω_k is the frequency of a wave with the vector $\mathbf{k}$.

The capillary wave dispersion law

$$\omega = (\sigma/\rho)^{1/2} k^{3/2}$$

is of the decay type ($\omega'' > 0$) and, therefore, the main contribution to the wave interaction is made by three-wave processes such as the decay of a wave into two with the conservation of the overall wave vector and overall frequency, as well as the reverse process of coalescence of two waves into one. A frequency range (inertial range) exists in a system of capillary waves on the surface of a liquid, which is limited from below by the pumping frequency ω_p and at high frequencies by viscous damping, in which the energy distribution E_ω has the exponential form

$$E_\omega \propto \omega^m.$$

According to the present-day theory [4], when the surface of a liquid is excited at low frequencies belonging to a fairly wide band $\omega_p \pm \Delta\omega$ ("wideband pumping," $\Delta\omega \approx \omega_p$), a constant energy flux Q towards high frequencies, i.e., direct cascade, sets in the k space. The theory of homogeneous capillary turbulence predicts the power law dependence on frequency for the wave distribution function $n(\omega)$ and the energy distribution E_ω (Kolmogorov spectrum) within the inertial range,

$$n(\omega) \propto Q^{1/2} \rho^{3/2} \sigma^{-1/4} \omega^{-15/6}, \qquad (2)$$

which corresponds to

$$n_k \propto Q^{1/2} \rho^{3/4} \sigma^{-1/4} k^{-17/4} \qquad (2a)$$

in the k representation.

The steady-state distribution of the surface wave energy in the inertial range may also be equivalently described by the pair correlation function in the Fourier representation

$$I_\omega = \langle |\eta_\omega|^2 \rangle$$

for a departure of the surface from the planar state $\eta(r, t)$,

$$I_\omega \propto \rho^{-17/6} \sigma^{-7/12} n(\omega) \omega^{-1/3}. \qquad (3)$$

From the experimental standpoint, it is most convenient to investigate the correlation function I_ω rather than the energy distribution E_ω, because the deviations of the surface from the planar state $\eta(r, t)$ may be measured directly in the experiment.

When the surface oscillations are excited in a wide frequency range, the correlation function is predicted by the theory in the form [3]

$$I_\omega = \mathrm{const}\omega^{-17/6}. \qquad (4)$$

The theoretical prediction of relation (4) is supported by the results of numerical calculations of the evolution of nonlinear capillary waves, performed directly from the first principles using the hydrodynamic equations [8, 9].

In the case of "narrowband pumping" ($\Delta\omega < \omega_p$), it was demonstrated by the results of calculations by Fal'kovich and Shafarenko [10] that a system of equidistant peaks at frequencies which were multiples of the pumping frequency was formed on the I_ω curve. The frequency dependence of the peak height is described by an exponential function with an exponent of (–21/6),

$$I_\omega = \mathrm{const}\omega^{-21/6}. \qquad (5)$$

Note that relations (4) and (5) were derived for systems of capillary waves with a continuous spectrum of wave vectors, i.e., for an idealized infinite surface of liquid. However, under experimental conditions, with a limited size of the surface, the $\omega(k)$ spectrum is discrete. This fact must be taken into account in comparing the real correlation function with theoretical prediction. The effect of discreteness decreases with increasing frequency ω, because the resonance width defined by the quality factor increases faster than the distance between the resonances: the spectrum becomes quasi-continuous. Pushkarev and Zakharov [9] used numerical methods to demonstrate that, for discrete systems at a fairly high level of excitation, relation (4) is also valid. As the pumping amplitude decreases, a threshold level is attained below which the system is in the state of frozen turbulence when the oscillation energy is concentrated in a finite frequency range, and the energy flux in the region of high values of k is zero.

We will define the high-frequency edge of the inertial range (boundary frequency) as the frequency ω_b at which the time τ_ν of viscous damping is comparable by the order of magnitude with the characteristic time τ_n of nonlinear interaction, τ_n, $\tau_\nu \sim \text{const}\,\tau_n$ (the kinetic time of relaxation in a turbulent wave system), where const is some dimensionless constant.

The characteristic time τ_n of nonlinear interaction in a turbulent cascade wave system is defined by the parameters of the liquid, as well as by the capillary wave distribution function $n(\omega)$, and may be written as

$$1/\tau_n \propto |V_\omega|^2 n(\omega), \tag{6}$$

where $V_\omega \approx (\sigma/\rho^{3/2})\omega^{3/2}$ is the coefficient of nonlinearity of capillary waves. The value of τ_n defines the characteristic scale of times of relaxation of perturbation over the cascade.

It is known [11] that the time of viscous damping of capillary waves decreases with increasing frequency as

$$1/\tau_\nu = 2\nu\omega^{4/3}(\sigma/\rho)^{2/3}. \tag{7}$$

Relations (6) and (7) enable one to derive the dependence of the wave frequency ω_b on the wave amplitude η_p at the pumping frequency ω_p at which the times of viscous damping and nonlinear interaction become equal in order of magnitude (the boundary frequency of the upper edge of the inertial range),

$$\omega_b \sim \eta_p^\beta \omega_p^\gamma. \tag{8}$$

The values of the exponents β and γ are defined by the frequency dependence of the correlation function

$$I_\omega \propto \eta_p^2(\omega/\omega_p)^\alpha.$$

In the case of excitation of surface oscillations in a wide frequency band, the exponent of the distribution function $\alpha = -17/6$, with $\beta = 2.4$ and $\gamma = 19/5$. When the surface oscillation is excited by spectrally narrow pumping with $\alpha = -21/6$, the exponent β decreases to 4/3, and $\gamma = 23/9$.

During the last decade, several reliable experiments with water were performed to study the capillary turbulence. The power law dependence on frequency for the correlation function at frequencies up to 1 kHz was observed in experiments [12] involving measurements of the power spectrum of radiation transmitted through a layer of water whose surface was excited at a low frequency. The exponent of the power function turned out to be close to the predicted value of –17/6. In the experiments of Wright *et al.* [13], who also investigated waves on the surface of water, the exponent in the correlation function was close to –3/2. Quite recently, experiments were performed [14] in which it was possible to observe the correlation function with the power dependence given by Eq. (4) in the frequency range of approximately 100 to 8000 Hz with resonant pumping at low frequencies. In this study, a new procedure was used: the oscillation amplitude of a spot produced by a laser on the surface of water with a fluorescent impurity was measured as a function of time.

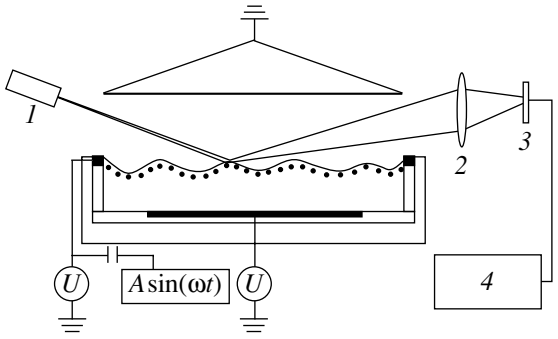


Fig. 1. Schematic of the experimental cell: (*1*) laser, (*2*) lens, (*3*) photodetector, (*4*) ADC.

Our investigations have shown [15] that, when the charged surface of liquid hydrogen is excited by an external periodic electric force at the resonance frequency of the cell, a power law dependence on frequency is observed for the correlation function in the frequency range from 100 Hz to 10 kHz. In this case, the exponent in the correlation function was close to -3.7 ± 0.3. When the surface was excited simultaneously at two resonant frequencies, the exponent decreased in magnitude and amounted to -3.0 ± 0.2 [16].

The boundary frequency of the upper edge of the inertial range could be experimentally determined for the first time in [17]. As the wave amplitude η_p at the pumping frequency ω_p increases, the boundary frequency shifts, by the power law given by Eq. (8), towards high frequencies with the exponent $\beta = 4/3$, as it must for the case of pumping in a "narrow band."

This paper deals with a more complete investigation of the dependence of the boundary frequency of the inertial range on the wave amplitude for three pumping frequencies. Experimental proof is given of the effect of the pattern of low-frequency pumping on the exponent in the correlation function.

3. EXPERIMENTAL PROCEDURE

The experiments were performed in an optical cell located in a helium cryostat. The experimental scheme is given in Fig. 1. A plane horizontal capacitor was placed inside the cell. A radioactive plate was located on the bottom capacitor plate. Hydrogen was condensed into a sleeve formed by the bottom capacitor plate and a guard ring 25 mm in diameter and 3 mm high. The layer of liquid was 3 mm thick. The top capacitor plate (a collector 25 mm in diameter) was located at a distance of 4 mm above the surface of the liquid. The temperature of the liquid in the experiments was 15–16 K.

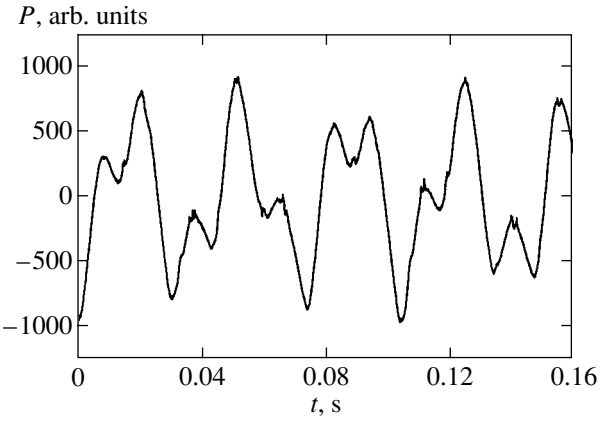


Fig. 2. A fragment of the time dependence $P(t)$ of the voltage across the photodetector in the case of excitation of the surface of liquid hydrogen at two frequencies of 28 and 67 Hz.

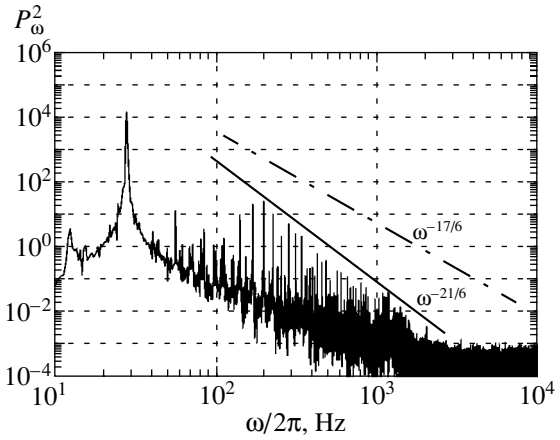


Fig. 3. The P_ω^2 distribution in the case of pumping at the frequency of 28 Hz.

The free surface of the liquid was charged with the aid of the radioactive plate emitting β electrons into the bulk of the liquid. The electrons emitted by the radioactive plate ionized the thin layer of liquid in the vicinity of this plate. The dc voltage U was applied between the capacitor plates. The sign of the charges forming a quasi-two-dimensional layer below the surface of the liquid was defined by the voltage polarity. In these experiments, the oscillation of a positively charged surface was studied. The metal guard ring installed around the radioactive plate prevented the charges from escaping from under the surface to the container walls.

The oscillations of the surface of liquid hydrogen (standing waves) were excited with the aid of the ac voltage applied to the guard ring in addition to the dc voltage at one of the resonant frequencies.

The oscillations of the surface of liquid hydrogen were recorded by the variation of the power of a laser beam reflected from the surface. The beam reflected from the oscillating surface was focused by a lens onto a photodetector. The voltage across the photodetector, which was directly proportional to the beam power $P(t)$, was recorded within several seconds by a computer with the aid of a high-speed 12- or 16-bit analog-to-digital converter. We analyzed the frequency spectrum P_ω of the total power of reflected laser beam, which was obtained by Fourier transformation in time of the $P(t)$ dependence being recorded. Figure 2 gives an example of recording a $P(t)$ signal when pumping the surface at two resonant frequencies ω_p equal to 28 and 67 Hz.

A laser beam 0.5 mm in diameter incident on the surface of the liquid at a graying angle of about 0.2 rad was used in the experiments. The axes of the ellipse of the light spot on the surface of the liquid were 2.5 and 0.5 mm. As was observed by Henry *et al.* [12], given this size of the light spot, the square of the Fourier amplitude P_ω^2 of the measured signal is directly proportional to the correlation function in the frequency representation, $I_\omega \propto P_\omega^2$, at frequencies above 50 Hz.

The procedures for excitation of surface oscillation and its recording, as well as the procedure of processing the experimental data, are described in [18].

After the cell was completely filled with liquid hydrogen, the dc voltage was switched on between the capacitor plates. Then, the maximal ac voltage of frequency ω was determined, at which the laser beam angle of deflection had a value known from geometric considerations (when an oscillating laser beam came in contact with the guard ring), and the maximal wave amplitude was calculated. It was found in preliminary experiments that the wave amplitude at the pumping frequency depended linearly on the amplitude A of applied dc voltage. Therefore, the wave amplitudes η smaller than the maximal value were calculated by the known experimentally obtained values of dc voltage.

4. EXPERIMENTAL RESULTS

4.1. The Effect of the Type of Pumping on the Frequency Dependence of the Correlation Function

As follows from relations (4) and (5), the exponent m in the correlation function $I_\omega \propto \omega^m$ must vary from $m = -21/6$ to $-17/6$ during transition from the narrow-band to wideband pumping. The measurement accuracy proved to be sufficient to form a reliable opinion of the variation of the exponent m. The experimental capabilities of the procedure made it possible to obtain and compare the frequency dependences of correlations functions for three types of excitation of charged surface, namely, at a single resonant frequency, at two resonant frequencies, and by noise in a band covering several resonances.

Figure 3 gives the dependence P_ω^2 in the case of excitation of a surface at the resonant frequency of 28 Hz. One can see that the amplitude of the principal

peak exceeds those of peaks of harmonics by three orders of magnitude. In the frequency range from 28 to 200 Hz, a dip in the P_ω^2 curve is observed. The nonmonotonicity of the P_ω^2 dependence may be attributed to special features of the employed method of optical detection of surface oscillations [18]. As was revealed by our observations, this may be due, among other things, to the fact that, in the spectrum P_ω of the signal being recorded, the amplitudes of Fourier harmonics at low frequencies depend on the position of the laser spot on the surface of liquid. At the same time, the shape of the high-frequency part of the spectrum does not depend on the position of the laser beam, because the spot size exceeds the wavelengths significantly.

At frequencies above 200 Hz, the height of peaks decreases monotonically on the average. In the frequency range of 0.2–2.0 kHz, the P_ω^2 dependence may be well described by a power function. The exponent obtained by averaging over ten measurements is $m = -3.7 \pm 0.3$. For comparison, the solid line in the figure indicates the function $\omega^{-21/6}$, and the dot-and-dash line indicates the function $\omega^{-17/6}$.

It turned out that the pattern of excitation of surface wave oscillation largely defines the frequency dependence of P_ω^2. When the surface is excited at two resonant frequencies, the experimentally obtained P_ω^2 dependences are well described by a power function with the exponent $m = -2.8 \pm 0.2$, which is close to the predicted value of $m = -17/6$. Figure 4 gives the dependence measured in the case of excitation of the surface at two resonant frequencies $\omega_1 = 28$ Hz and $\omega_2 = 67$ Hz (the recording of the $P(t)$ signal is given in Fig. 2). In Fig. 4, the dot-and-dash line corresponds to the $\omega^{-17/6}$ dependence, and the solid line, to the $\omega^{-21/6}$ dependence. In addition to the principal peaks, the P_ω^2 curve exhibits peaks corresponding to the combination frequencies $p\omega_2 \pm q\omega_1$, where p and q are integers.

When low-frequency noise is used to excite surface oscillation, the P_ω^2 distribution turns out to be close to the predicted dependence given by Eq. (4), as in the case of excitation at two frequencies. Figure 5 gives the P_ω^2 distribution in the case of surface excitation by noise in the frequency band of approximately 1 to 30 Hz. The solid curve indicates the distribution of the square of Fourier harmonics of the ac voltage applied to the guard ring, expressed in arbitrary units. The dot-and-dash line corresponds to the function $\omega^{-17/6}$. Figure 5 gives the result obtained by averaging over three files of the P_ω^2 distribution. The time of recording the $P(t)$ signal in these experiments was two seconds. The P_ω^2 distribution may be well described by a power function of frequency with the exponent $m = -2.8 \pm 0.2$. One can see that the experimentally obtained dependences turn out to be close to $\omega^{-17/6}$. This agrees well with the results of numerical calculation by Fal'kovich and Shafarenko [10] of the dependence of the exponent m on the pattern of pumping at low frequencies.

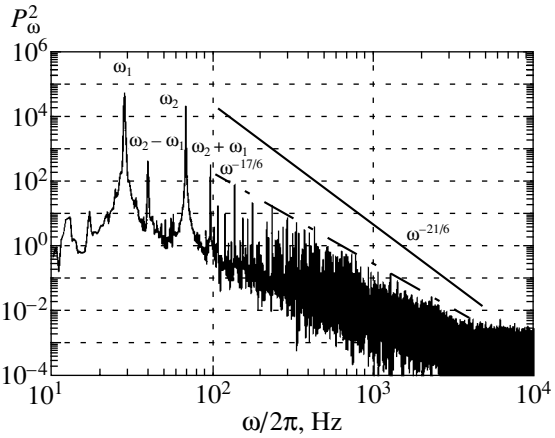


Fig. 4. The P_ω^2 distribution in the case of pumping at two frequencies of 28 and 67 Hz.

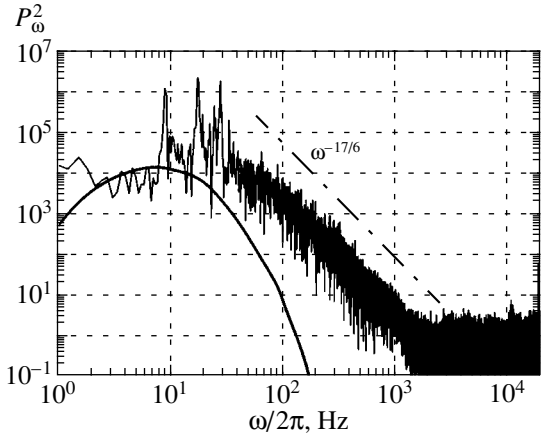


Fig. 5. The P_ω^2 distribution in the case of pumping by noise at low frequencies. The solid line describes the distribution of the square of Fourier harmonics of the ac voltage applied to the guard ring (in arbitrary units).

4.2. Dependence of the Boundary Frequency on the Wave Amplitude at the Pumping Frequency

As will be shown below, the P_ω^2 distribution depends both on the type of pumping and on its amplitude. Figures 6 and 7 gives two frequency dependences of the square of the Fourier amplitude P_ω^2 of the $P(t)$ signal, measured during surface excitation at the frequency $\omega_p = 290$ Hz. In Fig. 6, the wave amplitude η_p

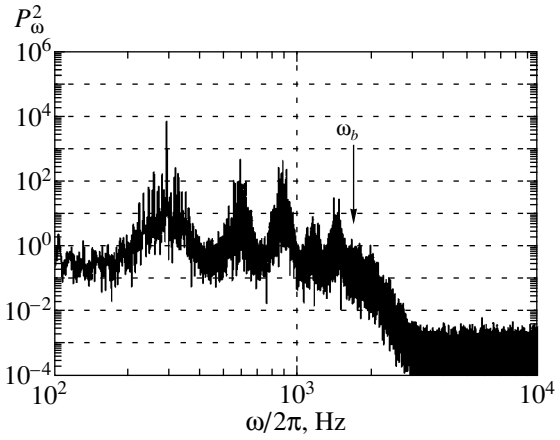


Fig. 6. The P_ω^2 distribution with the wave amplitude of 0.0015 mm at the pumping frequency of 290 Hz.

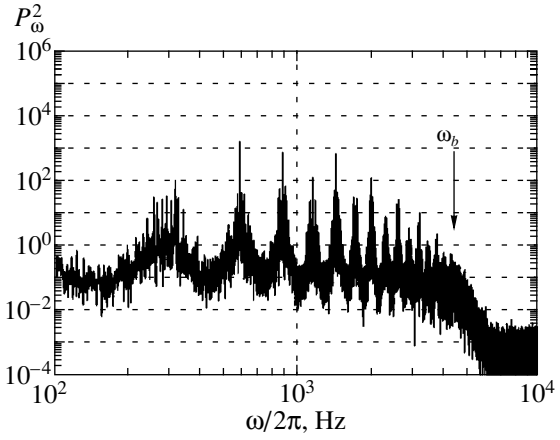


Fig. 7. The P_ω^2 distribution with the wave amplitude of 0.0079 mm at the pumping frequency of 290 Hz.

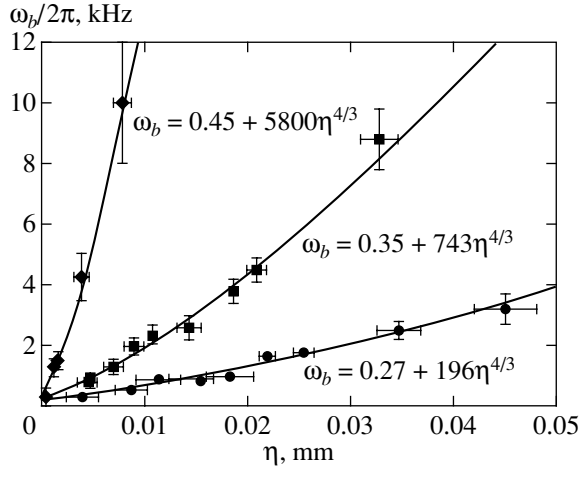


Fig. 8. The boundary frequency ω_b as a function of the wave amplitude at the pumping frequencies of 83 (●), 135 (■), and 290 (◆) Hz (in linear coordinates).

at the pumping frequency was 0.0015 ± 0.0002 mm; in Fig. 7, $\eta_p = 0.0079 \pm 0.0008$ mm with the wavelength $\lambda = 1.39$ mm. The arrows indicate the frequencies at which an abrupt variation of the P_ω^2 dependence occurs—the edge of the inertial range. For the pumping frequency $\omega_p = 135$ Hz, similar results are given in [17]. In Fig. 6, the boundary frequency of the edge of the inertial range is $\omega_p = 1700 \pm 200$ Hz, and, in Fig. 7, $\omega_b = 4200 \pm 1000$ Hz. One can clearly see that, as the wave amplitude increases, the boundary frequency of the inertial range shifts towards higher frequencies.

When the pumping wave amplitude is not high, a cascade consisting of only several harmonics of the pumping frequency ω_p is realized in the P_ω^2 spectrum. When the pumping wave amplitude increases, the inertial range is expanded, and the P_ω^2 spectrum comes to be made up of tens and even hundreds of harmonics.

Figure 8 gives three dependences of the boundary frequency of the edge of the inertial range ω_b on the wave amplitude η_p at the pumping frequencies of 83, 135, and 290 Hz. The ordinates of the points (frequencies) shown in the figure were estimated from the experimentally obtained curves similar to the curves given in Figs. 6 and 7. The pumping wave amplitudes were calculated by the known values of ac voltage applied to the guard ring. One can see that the experimentally obtained dependences $\omega_b(\eta_p)$ may be described by power functions with the exponent close to unity.

The solid curves in the figure correspond to the power law dependences of the boundary frequency of the inertial range ω_b on the amplitude η_p, with the exponent of 4/3. For better agreement between the predicted curve and experimental data, relation (8) includes the constant term independent of the wave amplitude at the pumping frequency. From simple physical considerations, it is clear that the boundary frequency ω_b cannot be less than the pumping frequency ω_p. The results of fitting are given in Fig. 8. One can see that adequate agreement is reached between the experimental points and the predicted dependence. The constant term turned out to exceed the pumping frequency ω_p by a factor of 2–3.

The amplitude dependence of the boundary frequency ω_b (as given by Eq. (8)) implies the existence of scaling with respect to the pumping frequency ω_p: the experimental points for ω_b, irrespective of the pumping frequency ω_p, must fit a single straight line in the coordinates $\omega_b/\omega_p^{23/9}$ and $\eta^{4/3}$ with $m = -21/6$. Figure 9 gives the result of plotting of experimental data in such reduced coordinates. One can state that the experimental points for three pumping frequencies fit the straight line well. This supports the validity of our assumption of the determining effect of viscosity when estimating

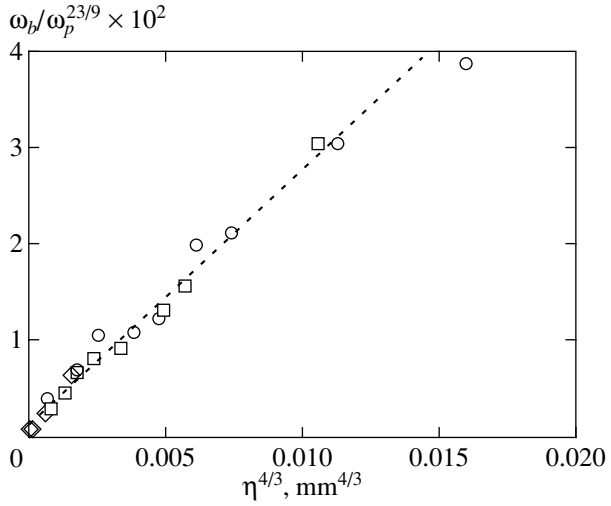


Fig. 9. The same as in Fig. 8, with the pumping frequencies of 83 (○), 135 (□), and 290 Hz (◇) in reduced coordinates.

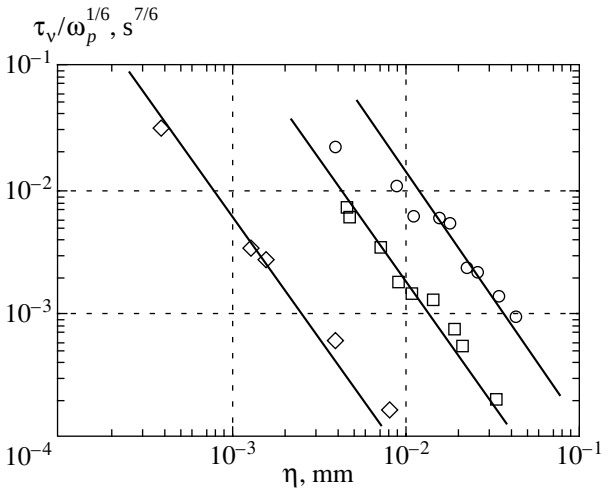


Fig. 10. The dependence of the function $\tau_\nu/\omega_b^{1/6}$ on the wave amplitude at the pumping frequencies of 83 (○), 135 (□), and 290 Hz (◇) in logarithmic coordinates.

the value of the frequency of the high-frequency edge of the inertial range.

Note that, in the case of low pumping frequencies $\omega_p < 60$ Hz, we failed to derive reliable P_ω^2 dependences with a clearly visible termination of the inertial range. This may be associated with the insufficiently wide experimental dynamic range of measurement of the $P(t)$ signal. We failed to fully utilize the amplitude range of a 16-bit ADC because of instrument noise.

5. DISCUSSION

We recall that, under conditions of our investigations, the correlation function is directly proportional to the square of Fourier harmonic of the signal being measured, $I_\omega \sim P_\omega^2$. Therefore, the results given in Figs. 3–5 demonstrate that the frequency dependence of the correlation function is defined by the pattern of surface excitation at low frequencies. In the case of pumping at a single fixed frequency, the correlation function is described by the power law dependence with the exponent $m = -3.7 \pm 0.3$. The exponent decreases in magnitude to $m = -2.8 \pm 0.2$ in the cases of pumping at two resonant frequencies and by noise in a band of frequencies. This variation of the frequency dependence of the correlation function agrees both qualitatively and quantitatively with the predictions of theory (formulas (4) and (5)).

The dependence of the boundary frequency on the wave amplitude at the pumping frequency is well described by relation (8) given below with the exponent $\beta = 4/3$, derived for the case of narrowband pumping. It is noteworthy that, in reduced coordinates, all experimental points adequately fit a single curve. This agrees with the predictions of the model of [10]. Note that, in the case of a low pumping frequency, $\omega_p = 28$ and 67 Hz, one cannot reliably record the high-frequency edge of the inertial range.

One can use relations (3) and (5), write the correlation function in the case of narrowband pumping as

$$I_\omega \propto \eta_p^2(\omega/\omega_p)^{-21/6},$$

and derive the relation

$$\tau_n/\omega^{1/6} \propto \eta_p^{-2}. \quad (9)$$

In the experiment, only the wave amplitude η_p and the boundary frequency ω_b are measured, with the time of nonlinear interaction τ_n estimated using the assumption made above (according to this assumption, at the boundary frequency ω_b, the time of nonlinear interaction is comparable in order of magnitude with the time of viscous damping, $\tau_n \sim \tau_\nu$). Therefore, in relation (9) at a frequency $\omega = \omega_b$, the time τ_n may be replaced by τ_ν, which gives

$$\tau_\nu/\omega_b^{1/6} \propto \eta_p^{-2}. \quad (9a)$$

The time of viscous damping τ_ν at the frequency ω_b is calculated by the known values of the parameters of the liquid by formula (7). The calculated values of $\tau_\nu/\omega_b^{1/6}$ as a function of the wave amplitude η_p are given in Fig. 10. One can see that the experimental points are adequately described by relation (9a). This agreement is better at the pumping frequencies of 290 and 135 Hz than at the pumping frequency of 83 Hz.

When a surface is excited in a wide frequency band, the exponent is $m = -17/6$, and relation (9a) transforms to

$$\tau_\nu\omega_b^{1/2} \propto \eta_p^{-2}.$$

The latter relation differs significantly from the experimentally observed dependence given in Fig. 10. Therefore, one can conclude that the experimental results obtained in the case of surface pumping at a single resonant frequency are well described within the model of [10].

6. CONCLUSION

It has been experimentally demonstrated for the first time that the pattern of excitation of surface oscillations at a low frequency affects the frequency dependence of the correlation function of the surface deviation from equilibrium and, consequently, the energy distribution over the surface oscillation frequencies. In the case of pumping at a single resonant frequency, the correlation function is described by the power function of frequency with the exponent $m = -3.7 \pm 0.2$, which is close to the predicted value of $m = -21/6$. This corresponds to the stationary spectrum of turbulence $E_\omega \propto \omega^{-13/6}$.

In the case of wideband pumping or excitation of a surface at two resonant frequencies, the observed exponent is $m = -2.8 \pm 0.2$, and theory gives $m = -17/6$. With this value of the exponent, the known law of energy distribution is derived, $E_\omega \propto \omega^{-3/2}$.

We have experimentally observed the boundary frequency of the inertial range for developed capillary turbulence on the surface of liquid hydrogen. It has been found that the inertial range expands towards high frequencies with increasing wave amplitude at the pumping frequency. The wave amplitude dependence of the boundary frequency may be well described by a power function with the exponent of 4/3. It has been demonstrated that the experimental data agree well with the existing theory of weak wave turbulence.

ACKNOWLEDGMENTS

We are grateful to L.P. Mezhov-Deglin for his attention and interest in our work, to V.N. Khlopinskiĭ for assistance in preparing the experiments, and to M.T. Levinsen for valuable discussions.

The investigations received partial support from the Russian Foundation for Basic Research (project no. 00-15-96703) and INTAS (grant no. 2001-0618).

REFERENCES

1. Proc. R. Soc. London, Ser. A **299** (1456) (1967).
2. S. D. Danilov and D. Gurariĭ, Usp. Fiz. Nauk **170**, 921 (2000) [Phys. Usp. **43**, 863 (2000)].
3. V. S. L'vov, *Nonlinear Spin Waves* (Nauka, Moscow, 1987).
4. V. Zakharov, V. L'vov, and G. Fal'kovich, *Kolmogorov Spectra of Turbulence*, Vol. 1: *Wave Turbulence* (Springer-Verlag, Berlin, 1992).
5. V. E. Zakharov, Zh. Éksp. Teor. Fiz. **51**, 688 (1966) [Sov. Phys. JETP **24**, 455 (1966)].
6. A. A. Levchenko and L. P. Mezhov-Deglin, Fiz. Nizk. Temp. **22**, 210 (1996) [Low Temp. Phys. **22**, 162 (1996)].
7. V. E. Zakharov and N. N. Filonenko, Zh. Prikl. Mekh. Tekh. Fiz. **5**, 62 (1967).
8. A. N. Pushkarev and V. E. Zakharov, Phys. Rev. Lett. **76**, 3320 (1996).
9. A. N. Pushkarev and V. E. Zakharov, Physica D (Amsterdam) **135**, 98 (2000).
10. G. E. Fal'kovich and A. B. Shafarenko, Zh. Éksp. Teor. Fiz. **94**, 172 (1988) [Sov. Phys. JETP **67**, 1393 (1988)].
11. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 6: *Fluid Mechanics* (Pergamon, New York, 1987; Nauka, Moscow, 1988).
12. E. Henry, P. Alstrom, and M. T. Levinsen, Europhys. Lett. **52**, 27 (2000).
13. W. Wright, R. Hiller, and S. Putterman, J. Acoust. Soc. Am. **92**, 2360 (1992).
14. M. Lommer and M. T. Levinsen, J. Fluoresc. **12**, 45 (2002).
15. M. Yu. Brazhnikov, A. A. Levchenko, G. V. Kolmakov, and L. P. Mezhov-Deglin, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 439 (2001) [JETP Lett. **73**, 398 (2001)].
16. M. Yu. Brazhnikov, A. A. Levchenko, G. V. Kolmakov, and L. P. Mezhov-Deglin, Fiz. Nizk. Temp. **27**, 1183 (2001) [Low Temp. Phys. **27**, 876 (2001)].
17. M. Yu. Brazhnikov, A. A. Levchenko, G. V. Kolmakov, and L. P. Mezhov-Deglin, Pis'ma Zh. Éksp. Teor. Fiz. **74**, 660 (2001) [JETP Lett. **74**, 583 (2001)].
18. M. Yu. Brazhnikov, A. A. Levchenko, and L. P. Mezhov-Deglin, submitted to Prib. Tekh. Éksp. (2002) (in press).

Translated by H. Bronstein

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 455–461.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 530–537.

FLUIDS

The Instability of the Surface of Helium Crystal in a Superfluid Flow

L. A. Maksimov and V. L. Tsymbalenko*
Russian Research Centre Kurchatov Institute, pl. Kurchatova 1, Moscow, 123182 Russia
**e-mail: vlt@isssph.kiae.ru*
Received April 24, 2002

Abstract—The problem on crystal growth under conditions of normal incidence of fluid on the boundary is investigated for stability. The threshold velocity of the emergence of instability is found; at low temperatures, this velocity proves to be much lower than the sound velocity. The stability is examined of the shape of cylindrical crystal in a fluid flow parallel to the crystal axis. The behavior of the atomically rough surface of crystal helium is experimentally investigated in a jet of fluid in the temperature range from 1 to 1.4 K, where the emergence is observed of an instability of the type previously predicted by Kagan, as well as by Nozieres and Uwaha. Observations reveal that, below the roughening transition, the (0001) basal face is stable in a jet of fluid. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The equilibrium form of a crystal in contact with a stationary fluid in the absence of the gravitational field is defined by the minimum of the surface energy α [1]. At a temperature above that of the roughening transition, the surface is in the atomically rough state, the angular dependence of the surface energy exhibits no singularities, and the crystal has a smooth rounded surface. The effect of additional factors (hydrostatic pressure gradient, internal stresses, pressure of electrons localized above the interface on the crystal surface, flows of heat or fluid) brings about a variation of the steady-state shape of crystal. In this case, two conditions must be valid at each point of the surface, namely, the equality of pressures (with due regard for the curvature) and of chemical potentials. It follows from the latter requirement that, during the times available in the experiment, a steady-state shape may only be attained in the case of a sufficiently high rate of surface growth. Of known substances, only ^{4}He crystals fit this criterion, because, as was theoretically predicted [2] and experimentally demonstrated [3], the rate of growth of the atomically rough surface of helium is high.

The first type of interfacial instability was discovered by Bodensohn *et al.* [4] on the surface of a helium crystal, above which electrons were localized. As the pressing field increased, an electrocapillary instability developed on the charged crystal surface [5], which was similar to the instability of the charged surface of superfluid liquid helium (see review [6]). This type of instability is not specific to the phase boundary. The second type of instability was also described in [4]. This instability characteristic just for the phase boundary was predicted by Grinfel'd [7]. It is caused by mechanical stresses arising in a crystal during its cooling as a result of thermal compression. The instability leads to the emergence of a periodic structure with a step of the order of the capillary constant on the surface. In all of these cases, the crystal surface is in contact with a stationary fluid.

The surface instability arising during the flow of heat through a surface was theoretically studied by Bowley and Nozieres [8]. They have shown that a surface becomes unstable when the heat flux from the fluid to crystal exceeds the critical value. It follows from the estimates of Bowley and Nozieres [8] that the instability of this type, owing to the high mobility of the surface of helium crystal, develops at such a high temperature gradient that is hardly attainable in real experiments, as was pointed out by the authors themselves.

A qualitatively different situation is observed in the case when a superfluid flows parallel to the crystal surface. The case of $T = 0$ was treated by Nozieres and Uwaha [9] and Kagan [10], who demonstrated that a tangential flow of fluid with a finite kinetic coefficient of growth brought about an absolute instability of the surface. In the gravitational field, an addition associated with the hydrostatic pressure gradient stabilized the surface. The minimal critical velocity of fluid flow was approximately 4 cm/s. It was assumed in the studies cited above that the parameters of the surface in contact with the moving fluid did not change significantly. Note that the instability of tangential discontinuities in conventional hydrodynamics exhibits similar features [11].

It was demonstrated by Andreev [12] that the assumption of the invariability of the parameters of the interface in contact with the flow is not valid for surfaces which differ from the atomically smooth basal face by the presence of a low concentration of steps (vicinal faces). The variation of the thermodynamic

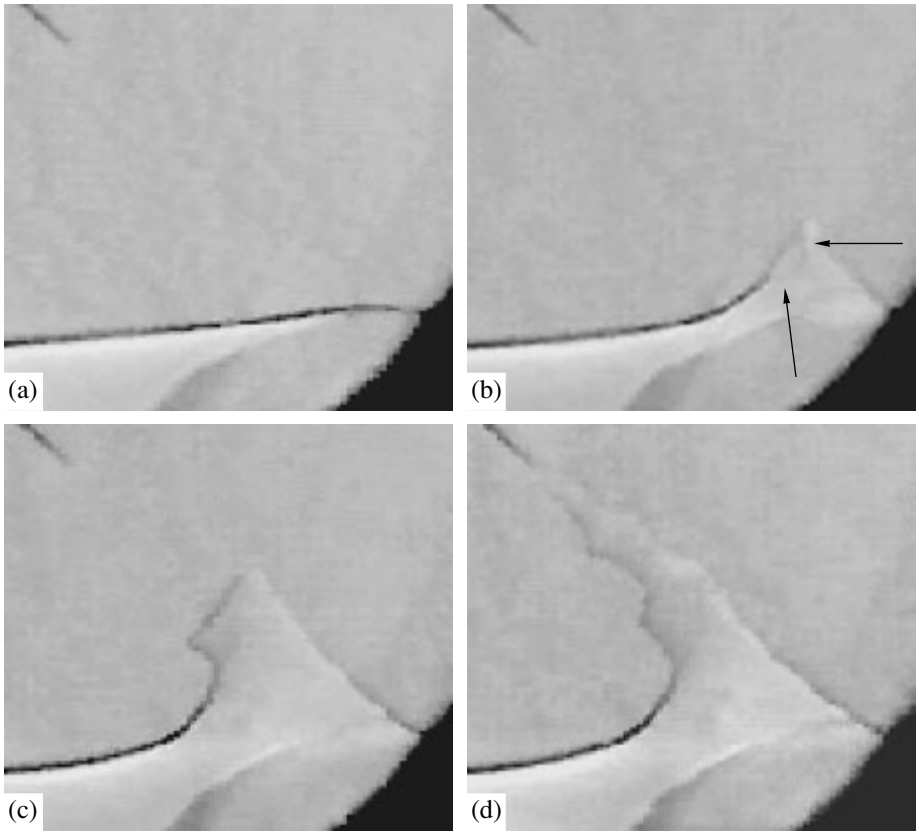


Fig. 1. The development of instability at $T = 1.412$ K. At the moment of time $t = 0$ (a), the emission from the needle point begins. The field of vision is 5.5×4.9 mm^2; $t = 1.25$ (b), 2 (c), and 2.3 s (d). The arrows in frame (b) indicate the recess formed on the growing crystal "stalagmite." On the frame (d) one can see the development of cylindrical instability. Waists are formed on the surface, which break the crystal "stalagmite."

characteristics of the surface turned out to be so strong as to cause a variation of the surface profile. With the velocity of fluid motion being as low as is desired, these faces are unstable relative to spontaneous formation of pairs of steps oriented in the direction normal to the velocity of liquid. Such an instability results in a peculiar cylindrical faceting of crystal. This result was obtained assuming that the general variation of the crystal shape was moderate.

The experimental verification of predictions was made difficult by the fact that the development of a controlled flow of fluid in a closed container is a complex procedural problem. It is known that heat fluxes in superfluid helium bring about the motion of the normal component and the emergence of a counter superfluid flow. However, it is shown by estimates that, in order to attain the critical velocity of tangential instability in acceptable geometry (for example, a flow in a slot of $1 \times 10 \times 10$ mm^3), a power of the order of 10 mW is required, which is hard to combine with the low helium temperature. Note that distortions of the surface by flows of heat and helium were observed in some studies (for example, [13]), where they presented a secondary interfering factor. We develop a flow of helium by entraining a fluid by charges emitted by a sharp point. In the case of incidence of a jet of fluid onto a surface in the atomically rough state, an instability of the crystal–liquid helium interface was observed. In Section 2, the experimental procedure is described and the results of observations are given. In Section 3, possible reasons for such a phenomenon are examined. Section 4 describes the observations of the interaction between a jet and the (0001) crystal face below the roughening transition.

2. DESCRIPTION OF THE EXPERIMENT AND THE BASIC RESULTS

The experiments were performed in an optical container [14] placed in a ^{3}He optical cryostat [15]. The crystal was grown from thermomechanically purified helium. A jet of helium was developed using the motion of electrons emitted by a tungsten needle located within the container (see Fig. 1). The sufficiently high voltage resulted in a cold emission of electrons entraining liquid helium. The observation was performed using a CCD module generating a black-and-white TV signal which was then recorded on a videotape recorder and

simultaneously digitized and stored in a computer [16]. The crystal was illuminated by a pulse infrared light-emitting diode. The light flash duration was 15 μs.

In order to study the instability, a crystal was grown filling the container such that the distance from the tip to the surface was 1–2 mm (see Fig. 1a). The crystal orientation was selected such that the surface was in the atomically rough state at given temperatures. Then, a voltage in the range from 1 to 2.5 kV was applied to the needle. The surface profile did not vary before the emission began. The increase in the pressure associated with the electrostatic field at a distance of 1 mm is too small for appreciably shifting the surface. Then, when the emission begins, but the flow is still low, a protuberance is formed opposite the needle point. For constant voltage, the protuberance remains on the surface without increasing in height. Its minor shifts and surface vibrations are observed; this is apparently associated with the unsteady-state pattern of current and, as a consequence, of the fluid flow. Given some value of the voltage, the protuberance starts growing in the direction towards the needle (see Fig. 1b). Note that, at the beginning of the development of instability, the end of the cylindrical "snout" sometimes has a hemispherical recess in the middle; in Fig. 1, this recess is indicated by arrows. Then, the protuberance narrows in its top portion and continues to move towards the needle point until contact. In the process of crystal growth towards the flow of fluid and after coming in contact with the needle, waists are observed on the cone-shaped surface of the crystal (Fig. 1d), which decrease with time, sometimes up to complete separation of the crystal from the base.

Therefore, in the case of low emission currents and fluid flows, one observes (with reservations specified above) a stable variation of the surface profile; after some critical value of the velocity is exceeded, the surface becomes unstable. One must emphasize that the emergence of protuberance is not associated with the electron pressure on the surface under the effect of the electrostatic field of the needle. As was experimentally observed by Leiderer [4], the electron pressure brings about a surface deflection; i.e., this effect has the opposite sign. Note that a surface deflection under the effect of the electron pressure was observed in a stationary fluid as well [17]. In [17], as well as in [4], this effect was utilized for measuring the kinetic coefficient of growth of the liquid–solid interface. Hence, it follows that this phenomenon is kinetic rather than static. The crystal growth towards the needle is not associated with the heat flux to the surface either. At the investigated temperatures, the crystallization heat is other than zero and positive; therefore, the heat flux must have melted the surface, i.e., must have formed a recess on the phase boundary.

This pattern of development of instability on an atomically rough surface is typical of all investigated temperatures from 1.01 to 1.41 K. With cooling, because of the increase in the kinetic coefficient of growth, the time of instability development decreases. Note that in these experiments the crystal shape after coming in contact with the needle does not become stationary. One can assume that the reason for this has to do both with the nonstationarity of the cold emission current and with the nature of instability. As was revealed by direct measurements, the emission current from the needle in these experiments varies strongly with time: by one (sometimes two) orders of magnitude from 1 to 100 nA within several seconds. Evidently, the additional electrostatic pressure in the vicinity of the needle point leads to the solidification of helium on the needle and blocking of the electron motion, because the mobility of negative charges in the crystal at this temperature is much less than the charge mobility in the fluid. For these reasons, we failed to determine with acceptable accuracy the threshold value of current causing the instability. The range of critical velocity is estimated in Subsection 3.1.

3. MECHANISMS OF INSTABILITY

3.1. General Remarks

We will assume for simplicity that liquid helium is an ideal incompressible liquid whose flow is potential, and the surface energy α and the coefficient G of kinetic growth of crystal are isotropic and independent of the fluid velocity. This restricts the region to be analyzed to that of flow velocities which are much lower than the sound velocity.

The emergence and growth of a protuberance at the site of jet incidence is attributed to the high pressure at the jet center. At the critical point on the surface, where the velocity is zero, the pressure exceeds the ambient pressure by $\rho v^2/2$ ([11], p. 38). The emergence of a protuberance leads to an increase in the surface curvature at the critical point, to a rise of pressure in the crystal, to a decrease in the chemical potential difference, and, as a result, to a deceleration of growth. If the "stalagmite" height and the surface curvature compensate for the chemical potential difference, the crystal growth will cease and the surface shape will become stationary. If the equality is not attained, the crystal growth continues until coming in contact with the needle point. As is revealed by the experiment, the surface comes to be quasi-stationary with the protuberance size $R \le 1$ mm, which gives the following estimate for the jet velocity:

$$v \le \sqrt{\frac{\alpha}{R\Delta\rho}} \sim 10 \text{ cm/s}. \quad (1)$$

Here, ρ and ρ' denote the helium density, $\Delta\rho = \rho' - \rho$; here and below, the prime indicates the solid phase. In the experiment, the surface curvature in the vicinity of the top of the crystal "stalagmite" reaches 0.03 mm, which gives the estimate of ~50 cm/s for the velocity. The surface symmetry reflects the jet symmetry, which

is close to axial. As the jet spreads along the crystal surface, the fluid velocity decreases.

3.2. The Stability of the Surface in Contact with Moving Fluid

3.2.1. The instability of a flat surface under conditions of crystal growth. We will treat the incidence of a fluid flow at a constant velocity v_0 with an arbitrary angle θ onto the surface of a crystal. We will direct the z axis normally to the surface. The normal component of flow v_{0z} brings about the crystal growth at a constant rate V. The perturbed surface has the z coordinate $Vt + \zeta(x, t)$. The normal **n** to the surface and the unit tangent vector **m** are given by the expressions

$$\mathbf{n} = \left(-\frac{\partial\zeta}{\partial x}, 1\right), \quad \mathbf{m} = \left(1, \frac{\partial\zeta}{\partial x}\right). \tag{2}$$

It is demonstrated in [9, 10] that the boundary conditions are derived from the laws of conservation of mass j, momentum $\Pi_{\alpha\beta}$, and energy Q,

$$n_\alpha j'_\alpha = n_\alpha j_\alpha, \tag{3}$$

$$n_\alpha Q'_\alpha = n_\alpha Q_\alpha, \tag{4}$$

$$\begin{aligned} n_\alpha \Pi'_{\alpha\beta} n_\beta &= n_\alpha \Pi_{\alpha\beta} n_\beta, \\ n_\alpha \Pi'_{\alpha\beta} m_\beta &= n_\alpha \Pi_{\alpha\beta} m_\beta. \end{aligned} \tag{5}$$

The conservation of mass flow given by Eq. (3) in a linear approximation with respect to $\partial\zeta/\partial x$ gives the expressions which relate the rate of crystal growth to the velocity of fluid flow under conditions of steady growth and the relation for minor surface deviations from the plane,

$$V = -\frac{\rho}{\Delta\rho} v_{0z}, \quad \Delta\rho\frac{\partial\zeta}{\partial t} = -\rho\left(v_z - v_x\frac{\partial\zeta}{\partial x}\right). \tag{6}$$

The conservation of the flow of momentum (5) and energy (4) (for more detail, see [9]) leads to the following expression for the chemical potential difference:

$$\delta\mu = \frac{\Delta\rho}{\rho\rho'}\delta p - \frac{\alpha}{\rho'}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) + \frac{1}{2}v^2. \tag{7}$$

It is assumed that the crystal is stationary and the surface energy is isotropic, $R_{1,2}$ denotes the principal surface curvatures, and δp is the deviation of the fluid pressure from the equilibrium pressure in the case of a plane interface. The rate of surface growth is defined by the expression

$$\frac{\partial\zeta}{\partial t} = K\delta\mu, \tag{8}$$

where K is the kinetic coefficient of growth. The motion of ideal fluid is described by the Laplace equation,

$$\Delta\varphi = 0, \quad \mathbf{v} = \nabla\varphi, \quad \delta p = -\rho\left(\frac{\partial\varphi}{\partial t} + \frac{1}{2}v^2 + g\zeta\right). \tag{9}$$

We will introduce minor oscillations of the surface and fluid,

$$\begin{aligned} \zeta(x, t) &= \zeta_0 \exp(-i\omega t + ikx), \\ \varphi(x, z, t) &= \varphi_0 \exp(-i\omega t + ikx - kz), \end{aligned} \tag{10}$$

satisfying the Laplace equation. In view of the boundary conditions given by Eqs. (6) and (7) and of relations (8) and (9), we derive the dispersion equation for crystallization waves,

$$\begin{aligned} &\omega^2 + \omega k\left[2\frac{\rho}{\Delta\rho}v_{0x} + i\left(\frac{\rho}{\Delta\rho}v_{0z} + \frac{1}{K}\frac{\rho\rho'}{\Delta\rho^2}\right)\right] \\ &+ k^2\left(\frac{\rho}{\Delta\rho}\right)^2 v_{0x}(v_{0x} + iv_{0z}) - \frac{\alpha\rho}{\Delta\rho^2}k^3 - \frac{g}{\Delta\rho}k = 0. \end{aligned} \tag{11}$$

For a stationary fluid, Eq. (11) gives the known law of dispersion of crystallization waves [2]. For a slip flow of fluid ($v_{0z} = 0$), the equation derived in [9, 10] is obtained, which gives the instability at velocities higher than 4 cm/s. For a purely normal incidence of fluid ($v_0 \equiv v_{0z}$), when a monotonic growth (or melting) of crystal occurs, Eq. (11) takes the form

$$\omega^2 + i\omega k\left(\frac{\rho}{\Delta\rho}v_0 + \frac{1}{K}\frac{\rho\rho'}{\Delta\rho^2}\right) - \frac{\alpha\rho}{\Delta\rho^2}k^3 - \frac{g}{\Delta\rho}k = 0. \tag{12}$$

It was observed by Nozieres and Uwaha [9] that a normal flow does not change the law of dispersion of crystallization waves, as is confirmed by the form of the equation. However, one can see in Eq. (12) that, under certain conditions, an instability arises, which was not mentioned by Nozieres and Uwaha [9]. While the expression in brackets is positive, the imaginary parts of the equation roots are negative, and minor perturbations decay with time. When the condition

$$v_0 < v_c = -\frac{1}{K}\frac{\rho'}{\Delta\rho} \tag{13}$$

is valid, the imaginary parts of both roots become positive, and the perturbations increase exponentially with time. It follows from condition (13) that the flow velocity is negative, i.e., the given instability is asymmetric about the processes of melting and solidification and develops just under conditions of crystal growth. The temperature dependence of the magnitude of boundary velocity v_c, calculated by the values of the kinetic coefficient of growth, is given in Fig. 2. The supersaturation values are calculated by the same formula. Note that the conservation of the linear dependence between the growth rate and supersaturation may be disturbed in the case of low temperatures and significant deviations from equilibrium. For this reason, the supersaturation

values given in Fig. 2 must be treated as rather rough estimates. One can see in the plot that, in the range from 1 to 1.4 K, the flow velocity necessary for the given instability to develop is in the range from 10 to 100 m/s, which corresponds to the crystal growth rate of 100–1000 m/s. No such rates were attained in our experiment.

3.2.2. Cylindrical crystal in a slip flow of fluid. The emergence of waists on the crystal "stalagmite" is outwardly similar to the process of capillary instability of a fluid jet [18]. However, the processes responsible for the collapse of the cylindrical shape differ considerably; therefore, the results of [18] cannot be applied to our case. Suffice it to say that the displacement of the crystal surface is caused by the kinetics of the phase boundary; this is fundamentally different from the case of a fluid jet.

We will treat the stability of a cylindrical crystal of radius R along whose surface and in parallel with the axis a superfluid flows at a velocity v. In the axisymmetric case, all quantities depend on the z coordinate along the crystal axis and on the distance r to the axis. We will introduce minor deviations ζ of the radius with the wave vector k along the cylinder axis,

$$\begin{aligned}\zeta(z, t) &= \zeta_0 \exp(-i\omega t + ikz),\\ \varphi(z, r, t) &= \varphi_0 K_0(kr)\exp(-i\omega t + ikz),\end{aligned} \tag{14}$$

where K_n is the Bessel function of the nth order of an imaginary argument (see Eqs. (9)). In view of conditions (6)–(9) and (14), we derive the dispersion equation for minor oscillations,

$$\begin{aligned}&\omega^2 + \omega k\left[2\frac{\rho}{\Delta\rho}v + i\frac{1}{K}\frac{\rho\rho'}{\Delta\rho^2}\frac{K_1(kR)}{K_0(kR)}\right]\\ &+ k^2\left(\frac{\rho}{\Delta\rho}\right)^2 v^2 - \frac{\alpha\rho}{\Delta\rho^2}\left(k^3 - \frac{k}{R^2}\right)\frac{K_1(kR)}{K_0(kR)} = 0,\end{aligned} \tag{15}$$

which is a generalization of Eq. (11) for a cylindrical surface disregarding the gravitational force. The cylindrical geometry leads to a variation of the law of dispersion of crystallization waves in the absence of fluid flow. For perturbations with the wave vectors $k < 1/R$, roots with a negative imaginary part, i.e., exponentially growing with time, are observed. Therefore, the cylindrical shape of crystal is unstable, similarly to the fluid jet instability. At $T = 0$ ($1/K \longrightarrow 0$), the fluid flow brings about a wave frequency shift (Doppler effect), as in the case of a flat surface. The instability in this case, as in the previous one, develops at low values of wave vectors. As the flow velocity increases, the instability boundary shifts towards the region of high values of wave vectors; i.e., the instability occurs on smaller scales. Therefore, a flow of fluid along the crystal "stalagmite" surface promotes the development of instability associated with shape. As was experimentally observed, an increase in the amplitude of cylindrical oscillation in some cases leads to the breaking of the "stalagmite."

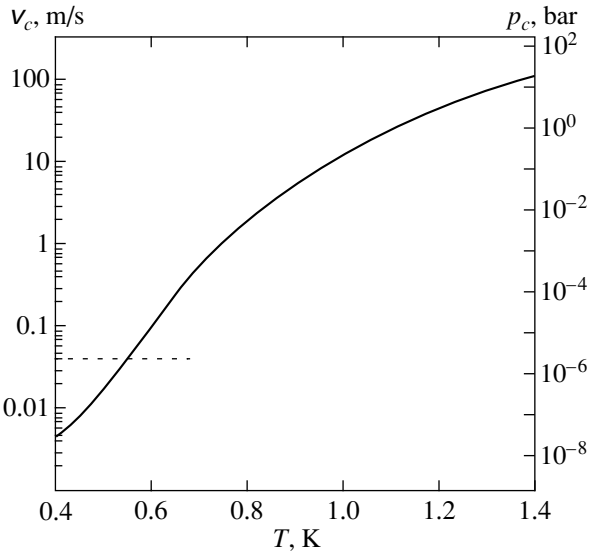


Fig. 2. The critical velocity as a function of temperature under conditions of normal incidence of flow onto the interface. The dashed curve indicates the value of critical velocity for tangential flow of fluid.

4. INTERACTION BETWEEN A JET AND CRYSTAL FACE

At a temperature below approximately 1.3 K, (0001) equilibrium faces of basal planes arise and, below 0.9 K, $(10\bar{1}0)$ side faces. The behavior of a face in a jet differs qualitatively from that of an atomically rough surface. Shown in Fig. 3 are stages in the development of instability of the crystal shape at 0.781 K, when a flow of fluid is directed to the basal plane oriented almost in parallel with the container bottom, so that the facet plane is parallel to the line of sight. The effect of the jet, as in the previous case, brings about the crystal "attraction" to the needle. However, at all of these stages of the process, one can see the basal face plane. Even in the strongest flow in the vicinity of the needle point, this plane is observed (Fig. 3, the last four frames). Unlike the atomically rough surface, no contact between the needle and crystal is observed. The flow of fluid directed at an angle to the facet entrains the latter, so that a flat crystal bridge is formed which accompanies the jet to the container wall. One can see in Fig. 3 that this bridge preserves the crystal orientation, i.e., the jet does not destroy the faceting of the basal plane even in the strongest possible flow. The observed pattern is not stationary; this is apparently due to the reasons specified above. After the crystal bridge comes in contact with the wall, it melts down. A face remains on the crystal, which serves a nucleus for generation of a new bridge. Therefore, the effect produced by the jet at velocities attained in the experiment is insufficient to destroy the basal faceting. Note that we failed to observe the cylindrical faceting predicted by Andreev [12] for tangential flow of fluid. This fact does

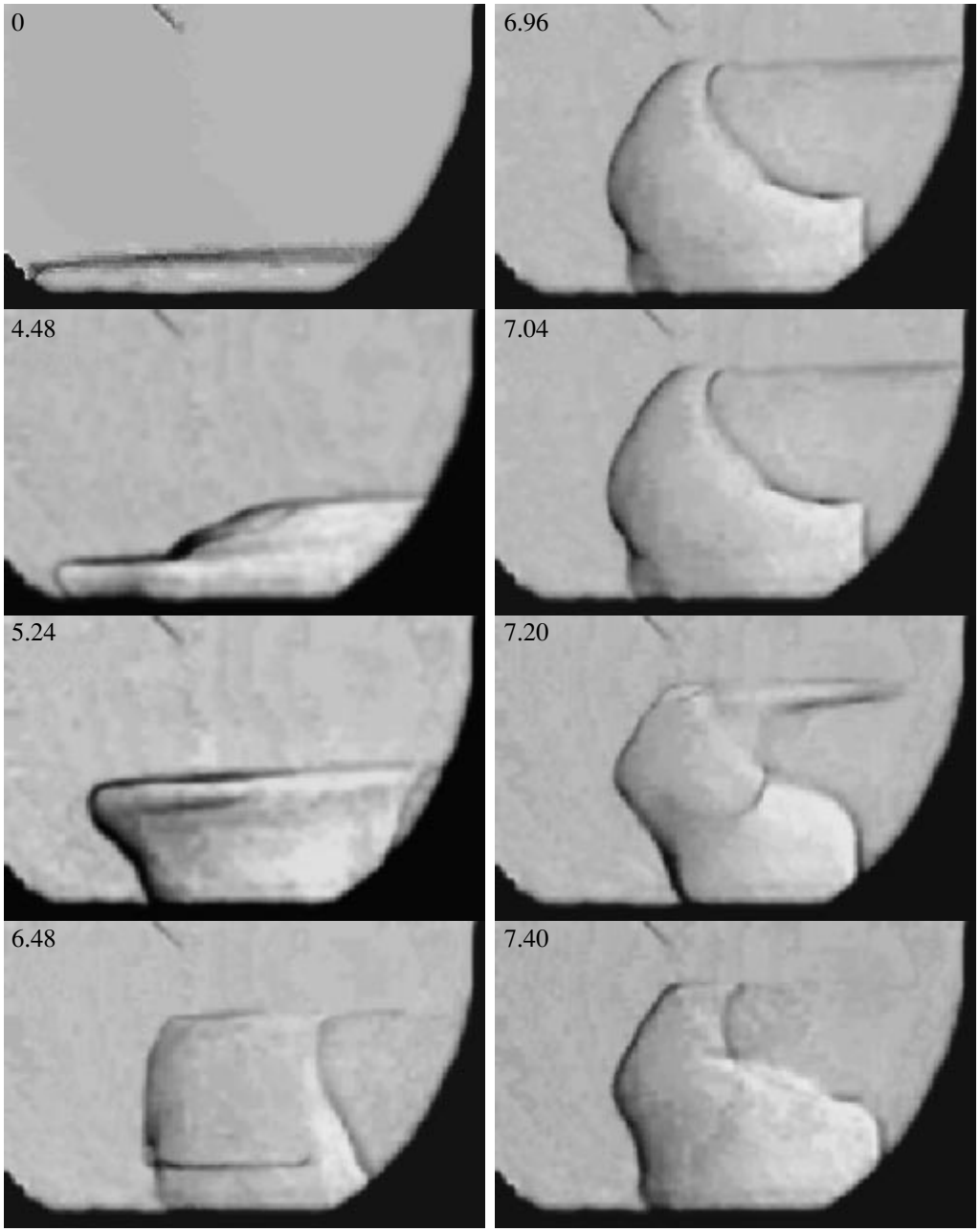


Fig. 3. The interaction between a jet of fluid and crystal below both roughening transitions, $T = 0.781$ K. The crystal lies on the basal plane. The field of vision is 8.9×5.6 mm^2. The numerals on the frames give the time in seconds from the moment the flow-inducing voltage is applied. The basal face is retained in the case of maximal flows. One can see the formation of a bridge from the crystal to the side internal wall of the container under the effect of the jet.

not contradict the results of [12], because this theoretical inference was made by Andreev assuming a minor variation of the entire crystal shape, and this assumption was clearly not valid in the experiment.

5. CONCLUSIONS

Thus, we managed to observe in the experiment a hydrodynamic instability of the phase boundary, which was apparently close to that predicted by Kagan [10] and Nozieres and Uwaha [9]. Theoretical analysis revealed a new type of instability showing up during the crystal growth under conditions of normal incidence of flow onto the boundary. It has been further demonstrated that the high rate of surface growth brings about an instability of the cylindrical shape of crystal, with the flow of fluid shifting the instability scale into the region of smaller lengths.

We have not yet found a theoretical explanation for the experimentally observed behavior of the basal facet in the flow of fluid below the roughening transition.

ACKNOWLEDGMENTS

We are grateful to A.Ya. Parshin for valuable remarks. This study was supported by the Russian Foundation for Basic Research (project no. 02-02-16772).

REFERENCES

1. A. A. Chernov, in *Modern Crystallography*, Vol. 3: *Crystal Growth*, Ed. by B. K. Vainshtein, A. A. Chernov, and L. A. Shuvalov (Nauka, Moscow, 1980; Springer-Verlag, Berlin, 1984).
2. A. F. Andreev and A. Ya. Parshin, Zh. Éksp. Teor. Fiz. **75**, 1511 (1978) [Sov. Phys. JETP **48**, 763 (1978)].
3. K. O. Keshishev, A. Ya. Parshin, and A. V. Babkin, Pis'ma Zh. Éksp. Teor. Fiz. **30**, 63 (1979) [JETP Lett. **30**, 56 (1979)].
4. J. Bodensohn, P. Leiderer, and K. Nicolai, Z. Phys. B **64**, 55 (1986).
5. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8: *Electrodynamics of Continuous Media* (Nauka, Moscow, 1982; Pergamon, New York, 1984).
6. V. S. Édel'man, Usp. Fiz. Nauk **130**, 675 (1980) [Sov. Phys. Usp. **23**, 227 (1980)].
7. M. A. Grinfel'd, Dokl. Akad. Nauk SSSR **290**, 1358 (1986) [Sov. Phys. Dokl. **31**, 831 (1986)].
8. R. M. Bowley and P. Noziéres, J. Phys. I **2**, 433 (1992).
9. P. Noziéres and M. Uwaha, J. Phys. (Paris) **47**, 263 (1986).
10. M. Yu. Kagan, Zh. Éksp. Teor. Fiz. **90**, 498 (1986) [Sov. Phys. JETP **63**, 288 (1986)].
11. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 6: *Hydrodinamics* (Pergamon, New York, 1987; Nauka, Moscow, 1988).
12. A. F. Andreev, Zh. Éksp. Teor. Fiz. **106**, 1219 (1994) [JETP **79**, 660 (1994)].
13. A. V. Babkin, D. B. Kopeliovich, and A. Ya. Parshin, Zh. Éksp. Teor. Fiz. **89**, 2288 (1985) [Sov. Phys. JETP **62**, 1322 (1985)].
14. V. L. Tsymbalenko, Cryogenics **36**, 65 (1996).
15. V. L. Tsymbalenko, Prib. Tekh. Éksp., No. 4, 161 (1997).
16. V. L. Tsymbalenko, Prib. Tekh. Éksp., No. 3, 77 (1997); No. 2, 78 (1999).
17. V. L. Tsymbalenko, Fiz. Nizk. Temp. **23**, 619 (1997) [Low Temp. Phys. **23**, 464 (1997)].
18. V. G. Levich, *Physicochemical Hydrodynamics* (Akad. Nauk SSSR, Moscow, 1959).

Translated by H. Bronstein

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 462–471.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 538–548.

SOLIDS
Structure

Viscoelastic Properties of Inhomogeneous Media with a Fractal Structure

V. V. Novikov[a, *] and K. W. Wojciechowski[b]

[a]*Odessa National Polytechnic University, Odessa, 65044 Ukraine*
**e-mail: novikov@Te.Net.Ua*
[b]*Institute of Molecular Physics, Polish Academy of Sciences, 60-179, Poznan Poland*

Received December 28, 2001

Abstract—The viscoelastic properties of a two-phase medium with a chaotic structure are calculated over the entire concentration range. The conditions for a monotonic and singular behavior of effective viscoelastic properties are established. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

If a body is subjected to an external force, strains appear in it, and the body itself becomes stressed. If this stress always exists during the action of the force and instantly disappears when this action ceases, then the body is ideally elastic. In this case, the relation between the stress tensor σ and the strain tensor ε is described by Hook's law [1]

$$\sigma = C\varepsilon, \quad \varepsilon = S\sigma, \tag{1.1}$$

where C is the elastic modulus tensor and S is the compliance modulus tensor.

When the strain is irreversible, i.e., when a body exhibits percolation, the stress decreases rapidly and recovers again because of the displacement of structural elements. If the shape and state of the structural elements do not undergo any variations in this case, then the body is ideally viscous, and its behavior is described by the Newton equation [2, 3]

$$\sigma = \eta\frac{d\varepsilon}{dt}, \tag{1.2}$$

where η is the liquid viscosity.

Most real bodies are viscoelastic and obey laws (1.1) and (1.2) only under certain conditions. Because of this, the concept of the stress decay time or the relaxation time τ is introduced to characterize a strain-stress state of real bodies. For absolutely elastic bodies, $\tau \longrightarrow 0$, whereas, for ideally viscous bodies, $\tau \longrightarrow \infty$. Real viscous, anomalous viscous, and viscoelastic media are located in the interval $0 < \tau < \infty$.

If an external force is periodic, the stress and strain change as [2, 3]

$$\sigma_{ij}(t) = \sigma_{ij}^0 e^{i\omega t}, \quad \varepsilon_{kl}(t) = \varepsilon_{kl}^0 e^{i(\omega t - \varphi)}, \tag{1.3}$$

where σ_{ij}^0 and ε_{ij}^0 are the stress and strain amplitudes, ω is the circular frequency, and φ is the loss angle.

If only one relaxation process proceeds in the body, then the amplitudes σ_{ij}^0 and ε_{ij}^0 satisfy Hook's law, i.e.,

$$\sigma_{ij}^0 = c_{ijkl}(\omega)\varepsilon_{kl}^0, \quad \varepsilon_{ij}^0 = s_{ijkl}(\omega)\sigma_{kl}^0, \tag{1.4}$$

where c_{ijkl} are components of the elastic modulus tensor and s_{ijkl} are components of the compliance tensor.

In an ideal elastic medium, $\varphi = 0$, and the relation between the stress tensor σ and the strain tensor ε is described by Hook's law (1.1) for an elastic medium.

If $\varphi \neq 0$, then, taking (1.3) into account, we obtain

$$\varepsilon_{ij} = S_{ijkl}(\omega)e^{-i\varphi}\sigma_{kl}. \tag{1.5}$$

In this case, it is convenient to introduce the concept of the complex compliance $S^*(\omega)$,

$$S^*(\omega) = S(\omega)e^{-i\varphi} = S(\omega)(\cos\varphi - i\sin\varphi), \tag{1.6}$$

and represent Hook's law in the form

$$\varepsilon = S^*(\omega)\sigma, \tag{1.7}$$

where

$$S^*(\omega) = S'(\omega) - iS''(\omega), \tag{1.8}$$

$S'(\omega)$ is the accumulation compliance, and $S''(\omega)$ is the loss compliance.

We can show that the relative scattering of the elastic energy is related only to the imaginary component $S''(\omega)$ of elastic moduli [2, 3]. For this reason, φ is usually called the internal friction of a material or the loss angle.

Below, we will consider isotropic media, for which, similarly to (1.8), the concept of a complex bulk elastic modulus $K^*(\omega)$ can be introduced [3]:

$$K^*(\omega) = K'(\omega) + iK''(\omega) = K'(\omega)(1 + i\tan\varphi), \tag{1.9}$$

where

$$\tan\varphi = K''(\omega)/K'(\omega). \tag{1.10}$$

1063-7761/02/9503-0462$22.00

The complex shear modulus μ^* and the complex viscosity η^* can be written in the form [2, 3]

$$\mu^*(\omega) = \mu'(\omega) + i\mu''(\omega), \quad (1.11)$$

$$\eta^* = \mu^*/i\omega, \quad (1.12)$$

$$\eta^*(\omega) = \eta'(\omega) - i\eta''(\omega). \quad (1.13)$$

The relation between $\mu'(\omega)$, $\mu''(\omega)$, and $\eta'(\omega)$, $\eta''(\omega)$ has the form

$$\mu'(\omega) = \omega\eta''(\omega), \quad \mu''(\omega) = \omega\eta'(\omega). \quad (1.14)$$

For a medium representing a Newton liquid, we have

$$\mu^*(\omega) = i\omega\eta'(\omega). \quad (1.15)$$

Viscoelastic media have been described by a variety of models involving a combination of a spring and a piston in a viscous liquid. In this (one-dimensional) case, Hook's and Newton's laws have the form [2]

$$F_H = kx, \quad (1.16)$$

$$F_N = \eta\frac{dx}{dt}. \quad (1.17)$$

A sequential combination of these elements corresponds to the Maxwell model, while their parallel combination corresponds to the Kelvin–Voigt model (Fig. 1).

The inverse transition from the models to a continuous medium is performed by replacing the force F and displacements x by stresses σ and strains ε.

1.1. The Maxwell Model

The Maxwell model is shown in Fig. 1a. Here, the total strain ε consists of the elastic (ε_0) and viscous (ε_m) components:

$$\varepsilon = \varepsilon_0 + \varepsilon_m. \quad (1.18)$$

The rate of variation of the elastic strain is

$$\frac{d\varepsilon_0}{dt} = \frac{1}{\mu_\infty}\frac{d\sigma}{dt}, \quad (1.19)$$

because $\varepsilon_0 = \sigma/\mu_\infty$, where $\mu_\infty = \mu(\omega)|_{\omega\to\infty}$ is the nonrelaxed value of the shear modulus ($\omega \longrightarrow \infty$). A variation of the viscous strain with time is related to the stress by the expression

$$\frac{d\varepsilon_m}{dt} = \frac{\sigma}{\eta}, \quad (1.20)$$

where η is the viscosity of the medium.

Thus, we obtain

$$\frac{d\varepsilon}{dt} = \frac{1}{\mu_\infty}\frac{d\sigma}{dt} + \frac{\sigma}{\eta}. \quad (1.21)$$

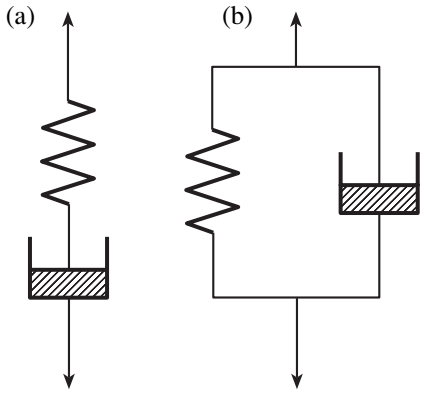


Fig. 1. Models of viscoelastic properties: (a) Maxwell model; (b) Kelvin–Voigt model.

For the time-independent strain

$$\frac{d\varepsilon}{dt} = 0, \quad (1.22)$$

we obtain the solution of Eq. (1.21)

$$\sigma(t) = \sigma_0\exp[-(t/\tau_\varepsilon)], \quad (1.23)$$

which characterizes the dependence of the stress on time t, i.e., the relaxation of the body. The ratio $\tau_\varepsilon = \eta/\mu_\infty$ is the Maxwell stress relaxation time.

In the case of the periodic action $\varepsilon = \varepsilon_0 e^{i\omega t}$, taking (1.21), we obtain

$$\sigma_0\mu_\infty e^{i\omega t} + i\omega\eta\sigma_0 e^{i\omega t} = \eta\mu_\infty i\omega\varepsilon e^{i\omega t},$$

which yields

$$\sigma = \frac{\mu_\infty i\omega\tau_\varepsilon\varepsilon}{1 + i\omega\tau_\varepsilon}; \quad (1.24)$$

i.e., the complex shear modulus for the Maxwell medium is

$$\mu^*(\omega) = \frac{\mu_\infty i\omega\tau_\varepsilon}{1 + i\omega\tau_\varepsilon}, \quad \tau_\varepsilon = \frac{\eta}{\mu_\infty}. \quad (1.25)$$

1.2. The Kelvin–Voigt Model

In the Kelvin–Voigt model (Fig. 1b), the elastic (σ_1) and viscous (σ_2) elements are connected in parallel. In this case,

$$\sigma = \sigma_1 + \sigma_2, \quad (1.26)$$

or

$$\sigma = 2\mu_0\left(\varepsilon + \tau_\sigma\frac{d\varepsilon}{dt}\right), \quad \tau_\sigma = \frac{\eta}{\mu_0}, \quad (1.27)$$

which yields

$$\varepsilon = \frac{\sigma}{\mu_0[1 - \exp(-t/\tau_\sigma)]}. \quad (1.28)$$

Here, $\mu_0 = \mu(\omega)|_{\omega=0}$ is the relaxed value of the shear modulus ($\omega \longrightarrow 0$).

If the action is periodic, then, taking (1.27) into account, we obtain for the Kelvin–Voigt medium

$$\sigma_0 e^{i\omega t} = \mu_0(\varepsilon_0 e^{i\omega t} + \tau_\sigma i\omega\varepsilon_0 e^{i\omega t}).$$

Therefore,

$$\sigma = \mu_0(1 + i\omega\tau_\sigma), \quad \tau_\sigma = \eta/\mu_0; \tag{1.29}$$

i.e., the complex shear modulus for the Kelvin–Voigt medium is

$$\mu^* = \mu_0(1 + i\omega\tau_\sigma). \tag{1.30}$$

1.3. The Zener Model

The main disadvantage of the Maxwell model is that the static shear modulus μ_0 vanishes in this model, while the drawback of the Kelvin–Voigt model is that it cannot describe the stress relaxation.

The Zener model (the model of a standard linear body) [2] is devoid of these disadvantages. This model combines the Maxwell and Kelvin–Voigt models and describes strains closely to the real process.

The elasticity equation for a standard linear body can be written in the form [2]

$$\sigma + \tau_\varepsilon \frac{d^\alpha \sigma}{dt^\alpha} = \mu\left(\varepsilon + \tau_\sigma \frac{d^\alpha \varepsilon}{dt^\alpha}\right), \tag{1.31}$$

where

$$\mu_0 = \mu(\omega)|_{\omega=0}, \quad \mu_\infty = \lim_{\omega\to\infty}\mu(\omega), \quad \tau_\varepsilon/\tau_\sigma = \mu_0/\mu_\infty,$$

and ω is the frequency of the action on a sample. By applying the Fourier transform to (1.31), we obtain [2]

$$\bar\sigma + (i\omega\tau)^\alpha\bar\sigma = 2\mu_0(\bar\sigma + (i\omega\tau)^\alpha\bar\varepsilon),$$

where $\bar\sigma$ and $\bar\varepsilon$ are the Fourier transforms of σ and ε. It follows from this that the complex shear modulus for a standard linear body is

$$\mu^*(\omega) = \mu_\infty - \frac{\mu_\infty - \mu_0}{1 + (i\omega\tau_\varepsilon)^\alpha}. \tag{1.32}$$

Taking (1.11) into account, we obtain

$$\frac{\mu_\infty - \mu'(\omega)}{\mu_\infty - \mu_0} = \frac{1 + (\omega\tau)^\alpha\cos(\pi\alpha/2)}{1 + (\omega\tau)^\alpha[2\cos(\pi\alpha/2) + (\omega\tau)^\alpha]}, \tag{1.33}$$

$$\frac{\mu_\infty - \mu''(\omega)}{\mu_\infty - \mu_0} = \frac{(\omega\tau)^\alpha\sin(\pi\alpha/2)}{1 + (\omega\tau)^\alpha[2\cos(\pi\alpha/2) + (\omega\tau)^\alpha]}. \tag{1.34}$$

If the Fourier transform of the function $\mu(t)$ is known, then the Fourier transform of the distribution $f(\tau)$ of relaxation times has the form [2]

$$\bar f\left(\frac{1}{\omega}\right) = \pm\frac{1}{\pi}\mathrm{Im}\mu[\omega\exp(\pm i\pi)]. \tag{1.35}$$

By using (1.34) and (1.35), we can determine the normalized density of the distribution $f_0(\tau)$ of relaxation times (Fig. 2c):

$$f_0(\tau) = \frac{\sin(\alpha\pi)}{2\pi\{\cosh[\alpha\ln(\tau/\tau_\varepsilon)] + \cos(\alpha\pi)\}}, \tag{1.36}$$

where

$$f_0(\tau) = \frac{f(\tau)}{\mu_\infty - \mu_0}.$$

The dependence of the dispersion γ^2 of the relaxation time of the chaotic dynamics on the parameter α has the form

$$\gamma^2 = \int_{-\infty}^{\infty}\ln^2\left(\frac{\tau}{\tau_\varepsilon}\right)f_0(\tau)d\ln\tau = \frac{\pi^2}{3}\frac{1-\alpha^2}{\alpha^2}. \tag{1.37}$$

It has been shown in paper [4] that Eq. (1.31) with fractional derivatives can be obtained by assuming that a set of relaxation times has a fractal nature. The parameter α in Eqs. (1.31), (1.33), and (1.34) (Figs. 2a, 2b) is equal [4] to the fractional dimensionality of a fractal set of relaxation times and characterizes the localization (spread) of the relaxation spectrum.

2. THE STRUCTURAL MODEL

Below, we consider the viscoelastic properties of a model inhomogeneous medium with a chaotic fractal structure.

Consider a hierarchy model of the two-phase medium structure, whose variation with increasing volume concentration p of the first phase can be qualitatively described in the following way. Initially, isolated clusters from the first phase are formed in a continuous medium of the second phase. Then, as the volume concentration p of the isolated clusters increases, they aggregate to form the so-called infinite cluster consisting of the first phase.

The chaotic structure of the inhomogeneous medium can be simulated on the basis of grids with randomly distributed parameters [5–7]. The nodes of a grid simulate the spatial distribution of phases, while the connections between the nodes simulate their contacts with neighbors. Below, each connection was represented by a spring and a piston connected in parallel (Fig. 3).

The basic set $\Omega_n(l_0, p_0)$ of connections was obtained with the help of iterations. In the first stage of calculations, we studied a grid of a finite size with the edge of length l_0 and the probability p_0 of the connection

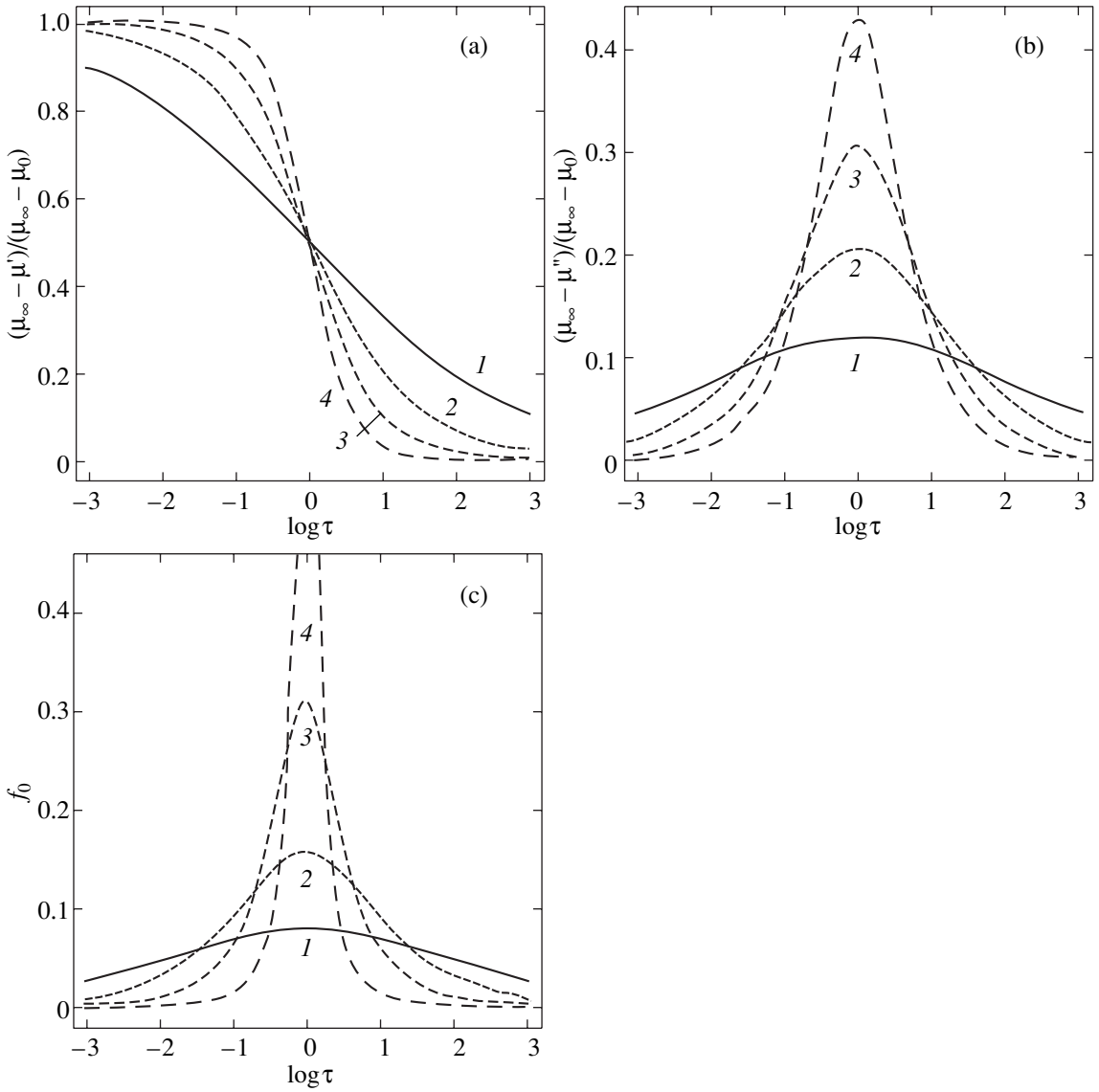


Fig. 2. Dependences of viscoelastic properties of a standard linear body on $\log\tau$ for $\alpha = 0.2$ (curves *1*); 0.4 (*2*); 0.7 (*3*); 0.9 (*4*). (a) The real part of the relative shear modulus; (b) the imaginary part of the relative shear modulus; (c) normalized distribution function of the relaxation time.

belonging to the first phase. In the next stage (k = 1, 2, …, n), each connection in the grid was replaced by the grid obtained in the previous step (Figs. 3, 4). The iteration process was terminated when the properties of the grid became independent of the iteration number k.

Therefore, the set $\Omega_n(l_0, p_0)$ of connections found by iterations depend on the size l_0 of the initial grid and the probability p_0, and is a self-similar, i.e., fractal, set [5–7].

3. VISCOELASTIC PROPERTIES

Consider a two-phase system with the distribution function

$$P_0(C) = p_0\delta(C - C_1^{(0)}) + (1 - p_0)\delta(C - C_2^{(0)}), \quad (3.1)$$

where $\delta(x)$ is the Dirac delta; the given local region has the property $C_1^{(0)}$ with the probability p_0 and the property $C_2^{(0)}$ with the probability $1 - p_0$.

After k iteration steps, the density function takes the form

$$P_k(C) = p_k\delta(C - C_c^{(k)}) + (1 - p_k)\delta(C - C_n^{(k)}). \quad (3.2)$$

Here, $C_c^{(k)}$ and $C_n^{(k)}$ are the properties of the connected and disconnected sets, respectively, at the kth iteration step; $p_k = R(l_{k-1}, p_{k-1})$ is the density of the set of connections; the function $R(l_{k-1}, p_{k-1})$ is equal to the ratio

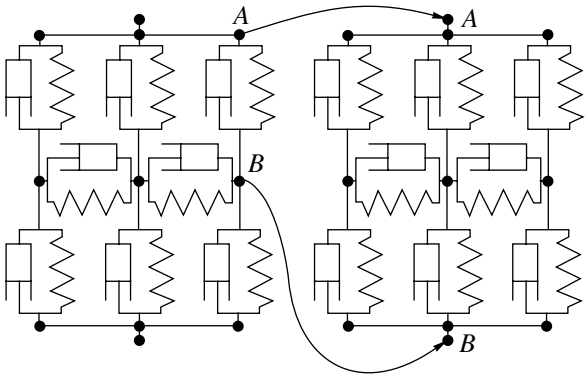


Fig. 3. Construction of the basic fractal set $\Omega_n(l_0, p_0 = 1)$ of viscoelastic elements for $l_0 = 2$.

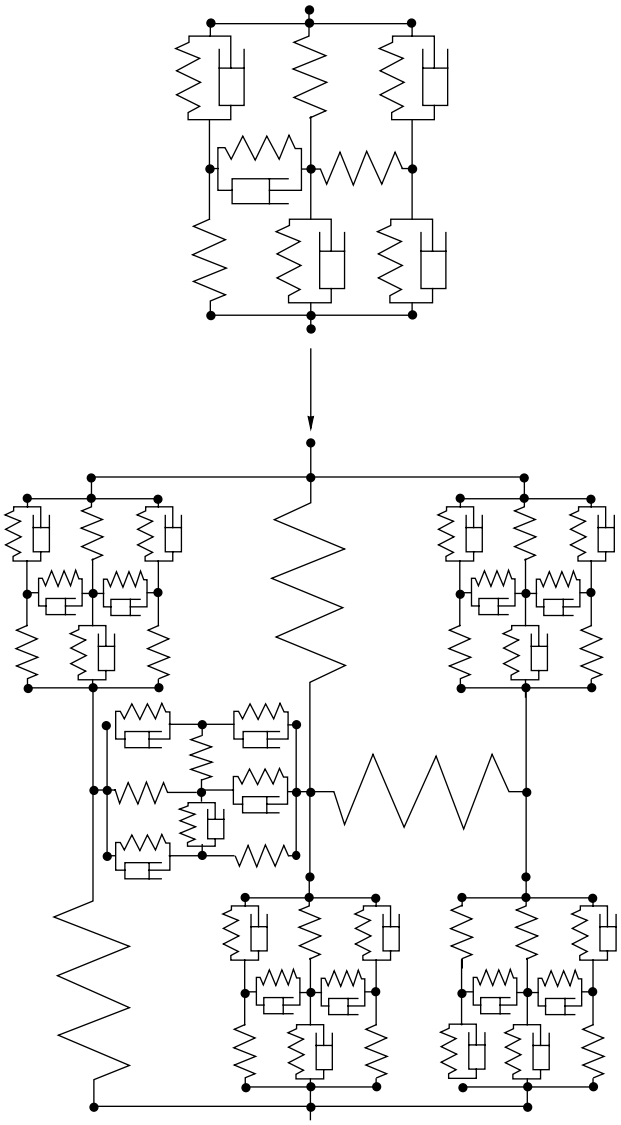


Fig. 4. Construction of the set $\Omega_n(l_0, p_0)$ for $l_0 = 2$, $p_0 = 3/8$ at the second iteration step ($k = 2$).

of the number of connected sets to the number of all spreads ("colorings") on the grid.

For $k \longrightarrow \infty$,

$$\lim_{k\to\infty} P_k(C) = \begin{cases} \delta(C - C_c^{(\infty)}), & p_0 > p^* \\ \delta(C - C_n^{(\infty)}), & p_0 < p^*. \end{cases} \quad (3.3)$$

In this case,

$$C_c^{(k)} \geq C \geq C_n^{(k)},$$
$$\lim_{k\to\infty} C_c^{(k)} = \lim_{k\to\infty} C_n^{(k)} = C. \quad (3.4)$$

In each iteration step, the structures of the connected and disconnected sets were simulated by a composite "drop" [5, 6]. The former represents a continuous mass of the first phase containing a sphere (drop) from the second phase (Fig. 5a); the latter represents a continuous mass of the second phase containing a sphere (drop) from the first phase (Fig. 5b). We assumed in this case that the volume elastic modulus K_1 of the first phase and the shear modulus μ_1 were greater than the corresponding quantities K_2 and μ_2 for the second phase. Thus, the effective viscoelastic properties of fractal structures are determined by iterations using several analytic dependences, namely, the probability function $R(l, p)$ and dependences of the viscoelastic properties of connected and disconnected sets on the properties and concentration of phases of the inhomogeneous medium.

The probability function $R(l, p)$ determines the probability that the set of connections will be connected for specified l and p. Analysis of numerical calculations of the function $R(p)$ [7] for the $2 \times 2 \times 2$ grid showed that the function [8]

$$R(p) = p^2(4 + 8p - 14p^2 - 40p^3 + 16p^4 + 288p^5 - 655p^6 + 672p^7 - 376p^8 + 112p^9 - 14p^{10}) \quad (3.5)$$

agrees well with the numerical calculations.

According to (3.5), the percolation threshold p_c is 0.2084626828; i.e., a disconnected set transforms to a connected set for $p_c \approx 0.208462$.

We calculated the elastic properties of connected and disconnected sets using the Hashin–Shtrikman formulas, which are based on a structural "sphere" model in a homogeneous medium [9, 10] (Fig. 5).

The Hashin–Shtrikman formulas were obtained based on the principle of minimal additional energy using the variational method for calculating the effective elastic moduli in an inhomogeneous medium (p. 120 in [2]). They determine the upper (K_c, μ_c) and lower (K_n, μ_n) boundaries of the effective elastic moduli:

$$K_c = K_1 + \frac{(1-p)(K_2 - K_1)}{1 + pa_1(K_2 - K_1)}, \quad (3.6)$$

$$\mu_c = \mu_1 + \frac{(1-p)(\mu_2-\mu_1)}{1+pb_1(\mu_2-\mu_1)}, \quad (3.7)$$

where

$$a_1 = \frac{3}{3K_1+4\mu_1}, \quad b_1 = \frac{6(K_1+2\mu_1)}{5\mu_1(3K_1+4\mu_1)}. \quad (3.8)$$

Expressions for K_n and μ_n can be obtained from (3.6)–(3.8) by replacing indices $c \longrightarrow n$, $1 \rightleftharpoons 2$, and $p \rightleftharpoons 1-p$.

As shown above, elastic static solutions can be transformed to viscoelastic solutions for steady-state harmonic oscillations by replacing the elastic moduli K and μ by corresponding elastic complex moduli K^* and μ^*.

By using this correspondence principle for a connected set, the complex volume elastic modulus K_c^* and the complex shear modulus μ_c^* at the $(j+1)$th step can be written in the form [5, 6]

$$K_c^{*(i+1)} = K_c^{*(i)} + \frac{(1-p_i)(K_n^{*(i)}-K_c^{*(i)})}{1+p_i a_c^{(i)}(K_n^{*(i)}-K_c^{*(i)})}, \quad (3.9)$$

$$\mu_c^{*(i+1)} = \mu_c^{*(i)} + \frac{(1-p_i)(\mu_n^{*(i)}-\mu_c^{*(i)})}{1+p_i b_c^{(i)}(\mu_n^{*(i)}-\mu_c^{*(i)})}, \quad (3.10)$$

where

$$a_c^{(i)} = \frac{3}{3K_c^{*(i)}+4\mu_c^{*(i)}}, \qquad b_c^{(i)} = \frac{6(K_c^{*(i)}+2\mu_c^{*(i)})}{5\mu_c^{*(i)}(3K_c^{*(i)}+4\mu_c^{*(i)})}, \quad (3.11)$$

where $K_c^{*(0)} = K_1^*$ and $\mu_c^{*(0)} = \mu_1^*$ are the complex volume elastic modulus and the complex shear modulus for the first phase of the inhomogeneous medium, respectively; $K_n^{*(0)} = K_2^*$ and $\mu_n^{*(0)} = \mu_2^*$ are the complex volume elastic modulus and the complex shear modulus for the second phase, respectively; and $p_{i+1} = R(p_i)$ according to (3.5).

The elastic properties $K_n^{*(i+1)}$ and $\mu_n^{*(i+1)}$ for a disconnected set are determined from expressions that can be obtained from (3.9)–(3.11) after replacements $c \rightleftharpoons n$ and $p_i \rightleftharpoons 1-p_i$.

4. RESULTS OF CALCULATIONS

Viscoelastic media. The calculations were performed for a two-phase (two-component) inhomogeneous medium assuming that volume strains are elastic, while shear strains are viscoelastic. The ratio K_1'/K_2' of local volume moduli was set equal to 10^4.

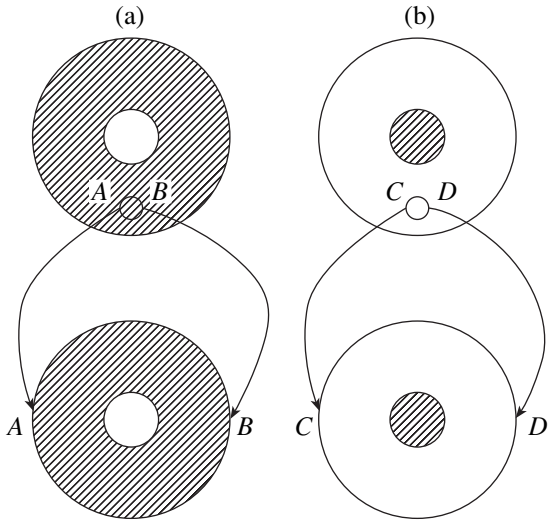


Fig. 5. Simulation of (a) a connected and (b) a disconnected set.

For convenience of calculations, the local shear moduli (phase shear moduli) were written in the form

$$\mu_1^* = \mu_2' x(1+iay), \quad (4.1)$$

$$\mu_2^* = \mu_2'(1+iy), \quad y = \tan\varphi_2 = \mu_2''/\mu_2', \quad (4.2)$$

$$x = \mu_1'/\mu_2' = \eta_1''/\eta_2'', \quad a = \tan\varphi_1/\tan\varphi_2. \quad (4.3)$$

The complex viscosity is

$$\eta_j^*(\omega) = \eta_j'(\omega) - i\eta_j''(\omega),$$

where

$$\mu_j' = \omega\eta_j''(\omega), \quad \mu_j''(\omega) = \omega\eta_j'(\omega), \quad j = 1, 2.$$

Figure 6 shows the dependences of the logarithm of the relative effective viscosity of an inhomogeneous fractal medium $\eta'/\eta_2' = \mathrm{Im}[\mu^*(\omega)]/\mathrm{Im}[\mu_2^*(\omega)]$ (η_2' is the viscosity of the second phase) on the concentration p of the first phase calculated for different values of a.

The calculations were performed for the ratio of the real parts of the shear modulus $x = 10^4$ and $y = 10^{-2}$, 10, 10^2, and 10^3.

It follows from Fig. 6a ($a = 0.1$, $\mu_1''/\mu_1' \ll 1$, $\mu_2''/\mu_2' \ll 1$) that the concentration dependence of the relaxed viscosity ($\omega \longrightarrow 0$) is described by a monotonic curve and is independent of the ratio μ_2''/μ_2'. For $a = 0.01$ and $\mu_2''/\mu_2' = 0.01$ (Fig. 6b), a local maximum and a local minimum appear near the percolation threshold, which strongly depend on the ratio μ_2''/μ_2'

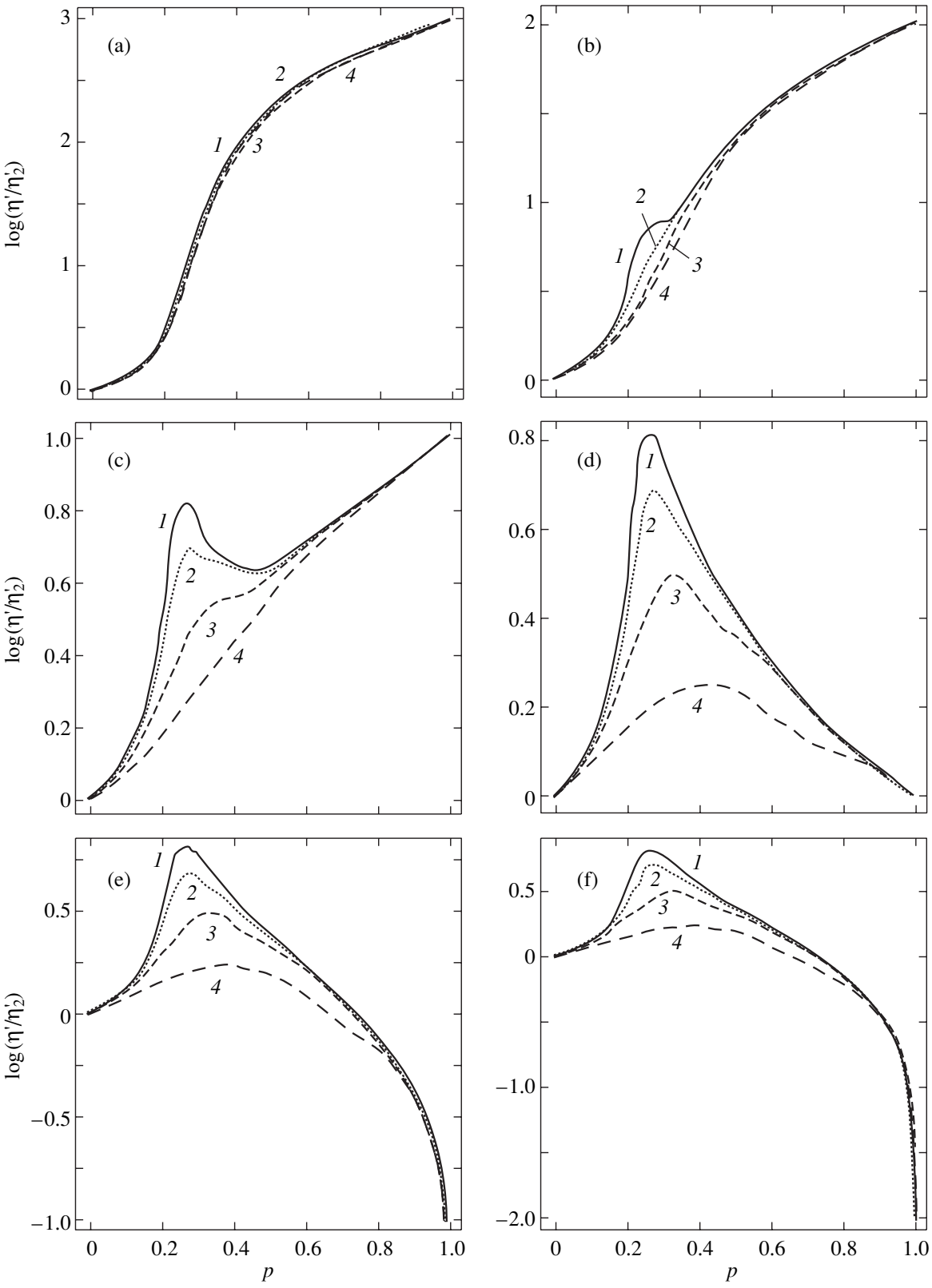


Fig. 6. Dependence of the logarithm of the effective relative viscosity $\eta'/\eta_2' = \mathrm{Im}[\mu^*(\omega)]/\mathrm{Im}[\mu_2^*(\omega)]$ on the concentration p of the first phase for $x = 10^4$ and $a = 10^{-1}$ (a), 10^{-2} (b), 10^{-3} (c), 10^{-4} (d), 10^{-5} (e), and 10^{-6} (f). Calculations were performed for $y = 10^{-2}$ (*1*), 10 (*2*), 10^2 (*3*), and 10^4 (*4*).

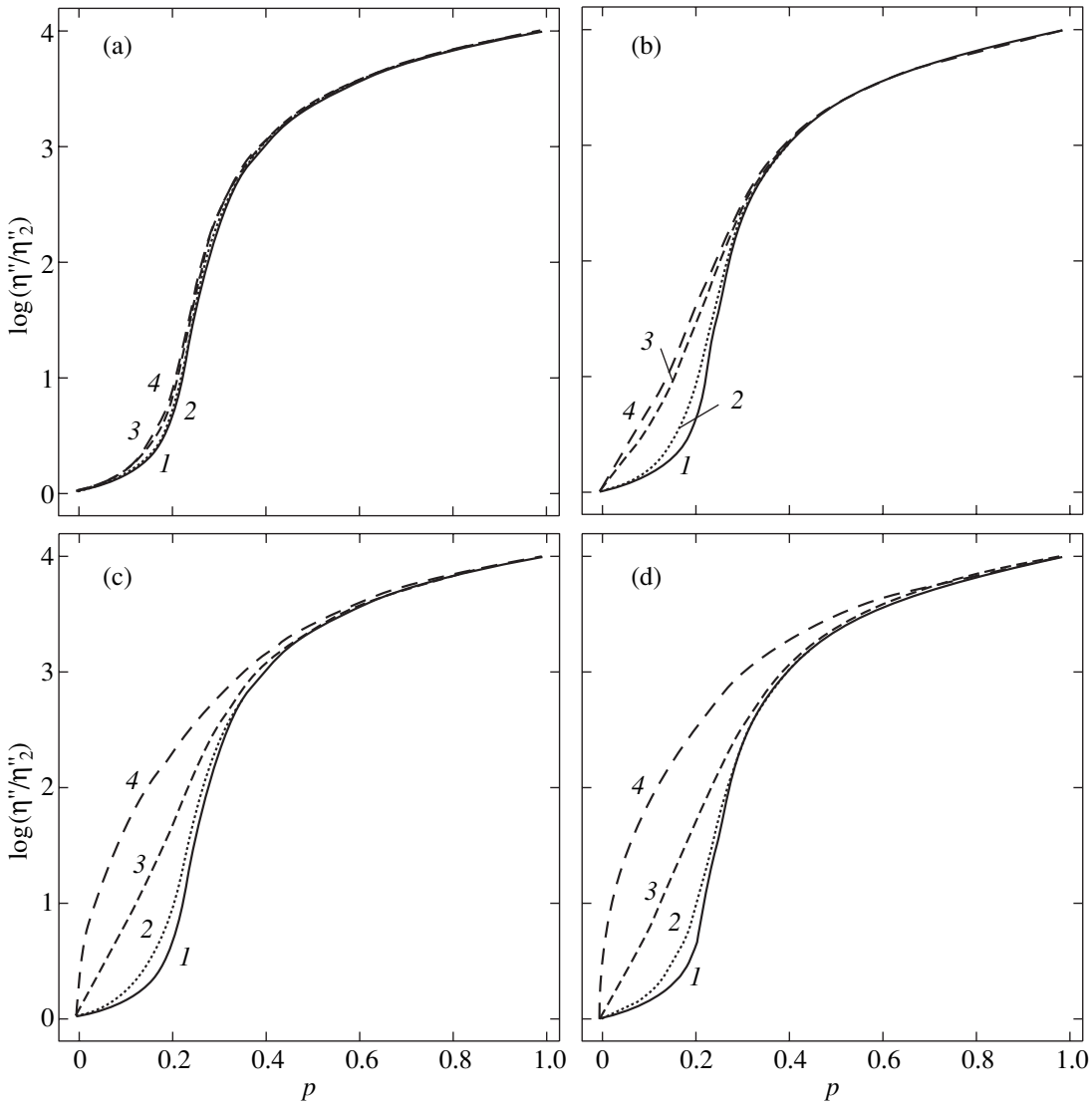


Fig. 7. Dependences of the logarithm of the effective relative viscosity η''/η_2'' = Re[μ*(ω)]/Re[$\mu_2^*(\omega)$] on the concentration p of the first phase for the same values of x, a, and y as in Fig. 6.

for $a < 0.01$ (Figs. 6c–6f). The type of the dependence hardly changes before the percolation threshold ($p < p_c$) (Figs. 6c–6f), whereas, after the percolation threshold ($p > p_c$), the concave curve (Figs. 5c–5e) becomes convex (Fig. 5f, $\mu_1''/\mu_1' \ll 1$, $\mu_2''/\mu_2' \gg 1$). For $a \le 10^{-4}$, the minimum disappears and only the maximum remains in the vicinity of the percolation threshold, which also disappears for $\mu_2''/\mu_2' \longrightarrow \infty$ (Figs. 6e, 6f). These results show that the dependence $\log|\eta'/\eta_2'|$ on the concentration p of phases of the fractal structure becomes convex with a single maximum when $ax \sim 1$, i.e., when $\mu_1'' \sim \mu_2''$ for $x \gg 1$ ($\mu_1' \gg \mu_2'$).

Figure 7 shows the calculated dependences of the logarithm of the relative effective viscosity η''/η_2'' = Re[μ*(ω)]/Re[$\mu_2^*(\omega)$] on the concentration p of the first phase. They show that, for $a \le 10^{-3}$ (Figs. 7c, 7d), the relative effective viscosity η''/η_2'' is virtually independent of a. For $p < p_c$, the type of the dependence changes when $a \longrightarrow 0$, while, for $p > p_c$, it does not change.

Viscoelastic media with a negative shear modulus. In [11–13], an inhomogeneous material was studied with one of its components having a negative shear modulus (negative rigidity). The authors found that composites containing inclusions with negative shear moduli in viscoelastic media had a higher rigidity and a higher mechanical damping compared to the components comprising the composite.

Figure 8 shows the calculations of the shear modulus of a viscoelastic inhomogeneous medium with a

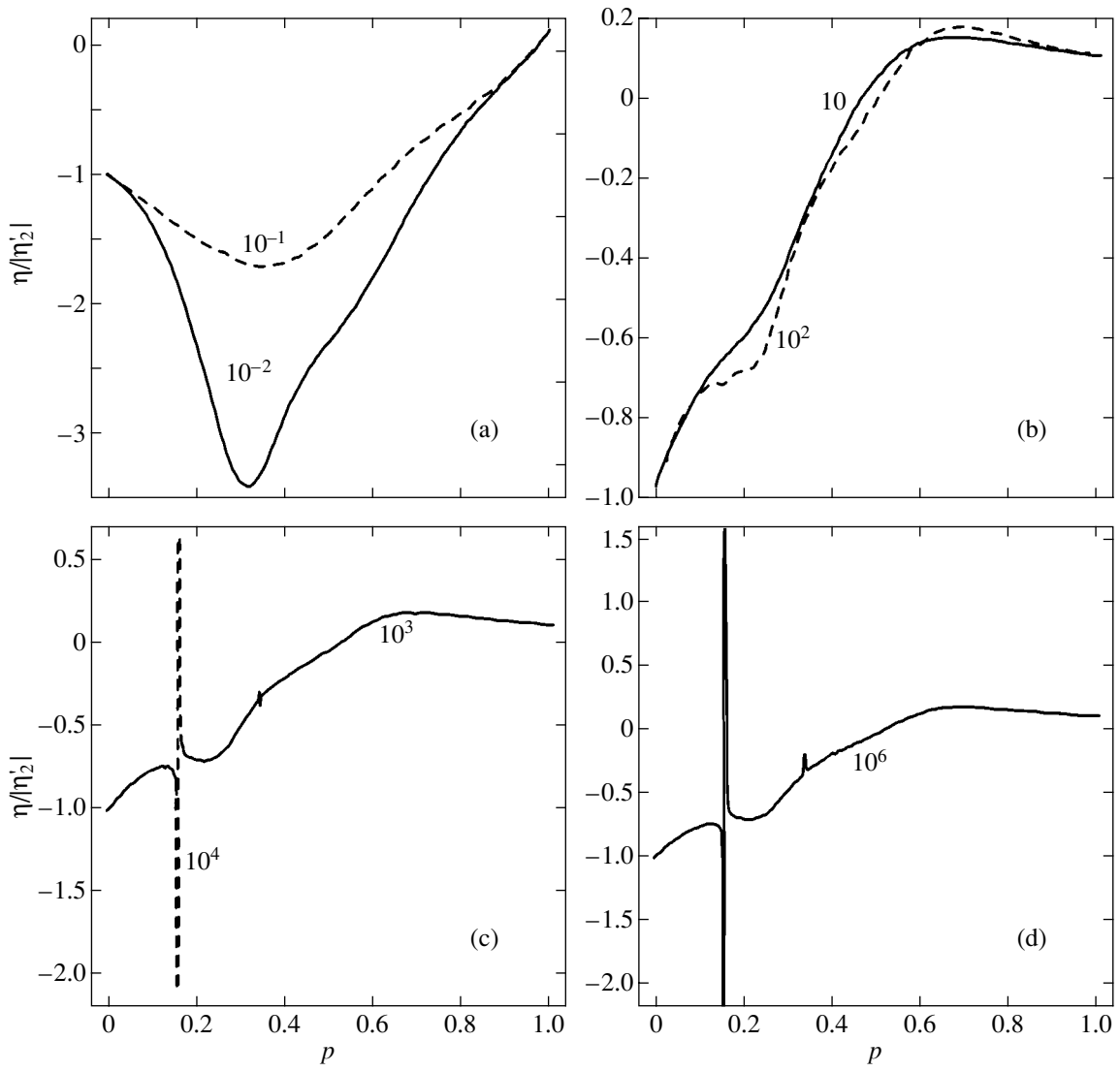


Fig. 8. Dependences of the effective relative viscosity on the concentration p of the first phase when the second phase has a negative viscosity. The values of y are shown at the curves.

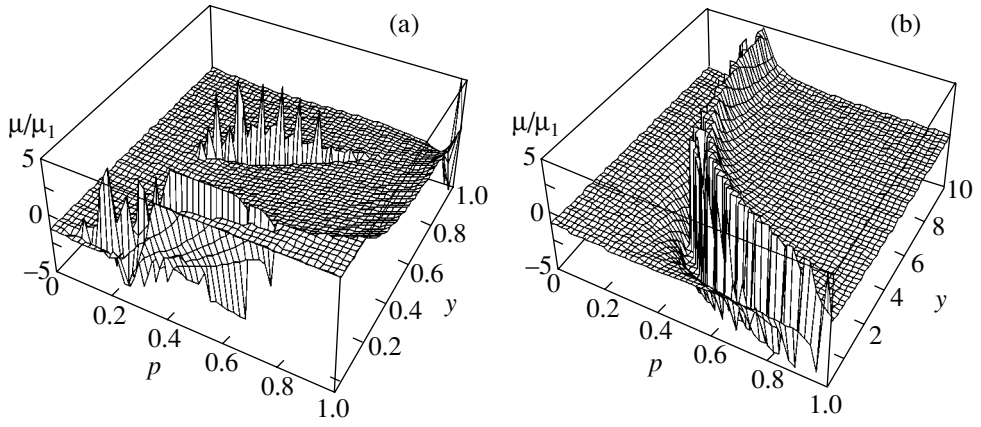


Fig. 9. Dependences of the relative effective shear modulus on the concentration p of the first phase when the phase shear moduli are $\mu_1^* = 1$ and $\mu_2^* = -y$. The values of y were varied from 0.01 to 1 (a) and from 1 to 10 (b).

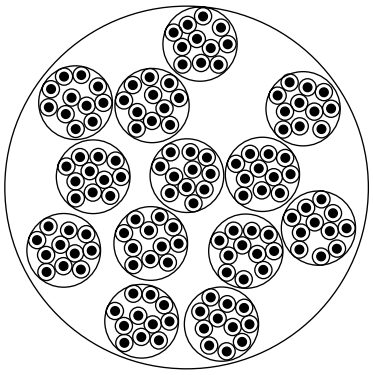


Fig. 10. An example of fabricating a material with a fractal structure.

chaotic fractal structure in which the first phase has the complex shear modulus

$$\mu_1^* = 1 + 0.1iy, \tag{4.4}$$

and the second phase has the negative shear modulus

$$\mu_2^* = -iy. \tag{4.5}$$

It follows from the calculations (Fig. 8) that the dependence of the relative effective viscosity on the concentration of phases in an inhomogeneous medium exhibits singularities in the vicinity of the percolation threshold at $y \geq 10$ ($\mu_1'' \geq \mu_1'$), which disappear for $y \ll 1$ ($|\mu_2''| \ll \mu_1'$). When $y \ll 1$ ($|\mu_2''| \ll \mu_1'$), the dependence of the effective relative viscosity $\eta'/|\eta_2'|$ on p has a minimum in the vicinity of the threshold p_c, which disappears for $y \longrightarrow 10$ ($\mu_1'' \longrightarrow \mu_1'$).

Figure 9 shows the effective shear modulus calculated for phase shear moduli $\mu_1^* = 1$ and $\mu_2^* = -y$. One can see that, for $y \gg 1$, singularities are localized and shift to the region of low concentrations p. For $y \ll 1$, singularities appear over the entire concentration range p.

The distribution function $f(t)$ of the relaxation time of a viscoelastic medium with a fractal structure can be analyzed similarly to the distribution function of the relaxation time of dielectric properties [14].

In conclusion, let us note that materials with a fractal structure, which will have viscoelastic properties that are appropriate to the above model calculations, can be fabricated in the following way. At the first stage, "pellets" are fabricated, for example, for the production of a material with good damping properties. The pellet consists of a polymer coating with inclusions of single domains of a ferromagnetic material. At the second stage, a pellet is fabricated with inclusions representing pellets fabricated at the first stage, etc. (Fig. 10).

5. CONCLUSION

The viscoelastic properties of inhomogeneous media with a chaotic fractal structure have been studied. The conditions are revealed for a nonmonotonic behavior of the effective viscosity in viscoelastic media and a singular behavior of the effective viscosity and the effective shear modulus when the shear (viscosity) modulus of one of the phases of the inhomogeneous medium is negative.

REFERENCES

1. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 7: *Theory of Elasticity* (Pergamon, New York, 1986; Nauka, Moscow, 1987).
2. T. D. Shermergor, *Theory of Elasticity of Micro-inhomogeneous Media* (Nauka, Moscow, 1977).
3. R. M. Christensen, *Mechanics of Composite Materials* (Wiley, New York, 1979; Mir, Moscow, 1982).
4. V. V. Novikov and K. W. Wojciechowski, Prikl. Mekh. Tekh. Fiz. **41**, 162 (2000).
5. V. V. Novikov, K. W. Wojciechowski, D. V. Belov, and V. P. Privalko, Phys. Rev. E **63**, 036120 (2001).
6. V. V. Novikov and K. W. Wojciechowski, Fiz. Tverd. Tela (St. Petersburg) **41**, 2147 (1999) [Phys. Solid State **41**, 1970 (1999)].
7. V. V. Novikov and V. P. Belov, Zh. Éksp. Teor. Fiz. **106**, 780 (1994) [JETP **79**, 428 (1994)].
8. J. Bernasconi, Phys. Rev. B **18**, 2185 (1978).
9. Z. Hashin and S. Shtrikman, J. Mech. Phys. Solids **10**, 335 (1962).
10. Z. Hashin, J. Appl. Mech. **50**, 481 (1983).
11. R. S. Lakes, Phys. Rev. Lett. **86**, 2897 (2001).
12. R. S. Lakes, T. Lee, A. Bersie, and Y. C. Wang, Nature **410**, 565 (2001).
13. R. S. Lakes, Philos. Mag. Lett. **81**, 95 (2001).
14. V. V. Novikov and V. P. Privalko, Phys. Rev. E **64**, 031504 (2001).

Translated by M. Sapozhnikov

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 472–479.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 549–557.

SOLIDS
Structure

Kossel Lines as a New Type of X-ray Source

A. M. Afanas'ev[a], M. V. Koval'chuk[b], M. A. Chuev[a, *], and P. G. Medvedev[b]

[a]*Institute of Physics and Technology, Russian Academy of Sciences, Moscow, 117218 Russia*
**e-mail: chuev@ftian.oivta.ru*
[b]*Institute of Crystallography, Russian Academy of Sciences, Moscow, 117333 Russia*
Received February 5, 2002

Abstract—The radiation intensity distribution within the Kossel line corresponding to the extremely asymmetric pattern of X-ray diffraction has an anomalous form of a clearly manifested peak exceeding the background intensity by more than two orders of magnitude. A detailed theoretical analysis of this effect is carried out and versions of experimental observation of anomalous Kossel lines are proposed. The possibility of the employment of the effect for obtaining a new source of X rays with a narrow angular collimation is discussed. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

Although the method of Kossel lines has been known since 1930s [1], the basic physical phenomena observed when this type of radiation emerges from a crystal have not been studied comprehensively. It was noted in our recent communication [2] that Kossel lines acquire an extremely anomalous shape differing significantly from the standard form in the case of diffraction of an X-ray beam in an extremely asymmetric scheme. In such a diffraction scheme, a strongly compressed X-ray beam propagating parallel to the crystal surface is formed at a certain depth of the crystal, and the compression ratio may be as high as several hundred [3]. This prediction was confirmed when the method of standing X-ray waves was used with the detection of photoelectrons [4, 5] using a specially developed proportional gas counter [6]. It is natural to expect that the strong compression of the X-ray beam will be manifested in the radiation intensity distribution in the Kossel line. It will be proved below that the intensity distribution of the emerging X-ray beam in the Kossel line corresponding to strongly asymmetric diffraction has the form of a clearly manifested peak whose intensity is more than two orders of magnitude higher than the background intensity [2].

The process of dynamic scattering in an extremely asymmetric diffraction scheme, when the specular reflection of the incident beam or the beam reflected from the crystal must be taken into account along with diffraction scattering, is analyzed in detail in Section 2. In Section 3, the properties of dynamic diffraction in the vicinity of the so-called degenerate point, when the roots of the dispersion equation turn out to be triply degenerate, are considered. It is at this point that the maximum degree of compression of the X-ray beam occurs in the crystal. In Section 4, the features of the angular distribution of X rays in the Kossel line are analyzed in the vicinity of the degeneracy point. Section 5 is devoted to the discussion of several versions of experimental observation of the effect and the possibility of using Kossel lines for creating the sources of X rays collimated over angles.

2. EQUATIONS OF DYNAMIC DIFFRACTION TAKING INTO ACCOUNT SPECULAR REFLECTION

Extremely asymmetric diffraction is realized in the so-called Bragg–Laue geometry [7], when the diffracted wave propagates at an angle φ_h almost parallel to the crystal surface; by varying slightly the angle of incidence φ_0 of the X-ray beam on the crystal, it is possible to change the beam diffraction from the Laue geometry to the Bragg geometry. Such schemes with angles φ_0 of the order of unity and with $\varphi_h \ll 1$ can easily be realized in experiments (see Fig. 1 and the table given below).

In order to find the distribution of wave fields in the crystal with such a diffraction scheme, we must take into account specular reflection in addition to the diffraction scattering. Since $\varphi_0 \approx 1$, we can disregard the specular reflection of the incident wave and take into account the specular reflection only for the diffracted wave. We will seek the wave fields outside the crystal, $\mathbf{E}(\mathbf{r})$, and within it, $\mathbf{D}(\mathbf{r})$, in the form

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0 \exp(i\boldsymbol{\kappa} \cdot \mathbf{r}) + \mathbf{E}_h \exp(i\boldsymbol{\kappa}_h \cdot \mathbf{r}), \quad (1)$$

$$\mathbf{D}(\mathbf{r}) = \mathbf{D}_0 \exp(i\mathbf{k}_0 \cdot \mathbf{r}) + \mathbf{D}_h \exp(i\mathbf{k}_h \cdot \mathbf{r}). \quad (2)$$

Here, $\boldsymbol{\kappa}$ and $\mathbf{k}_0$ are the wave vectors of the incident wave in vacuum and in the crystal, respectively, and $\mathbf{k}_h$ is the wave vector of the diffracted wave in the crystal:

$$\mathbf{k}_h = \mathbf{k}_0 + \mathbf{K}_h, \quad (3)$$

where $\mathbf{K}_h$ is the reciprocal lattice vector. As regards vector $\boldsymbol{\kappa}_h$, it can be determined from the conditions

$$\boldsymbol{\kappa}_{h\|} = \boldsymbol{\kappa}_{\|} + \mathbf{K}_{h\|}, \quad \kappa^2 = \kappa_h^2, \tag{4}$$

where $\boldsymbol{\kappa}_{h\|}$ and $\boldsymbol{\kappa}_{\|}$ are the projections of the corresponding vectors on the crystal surface.

The amplitudes D_0 and D_h of the wave fields in the crystal must satisfy the well-known system of dynamic equations [8]

$$\begin{aligned} \frac{k_0^2 - \kappa^2}{\kappa^2} D_0 &= \chi_0 D_0 + C\chi_{\bar{h}} D_h, \\ \frac{k_h^2 - \kappa^2}{\kappa^2} D_h &= \chi_0 D_h + C\chi_h D_0, \end{aligned} \tag{5}$$

where $\kappa = 2\pi/\lambda$, λ being the radiation wavelength; χ_0, χ_h, and $\chi_{\bar{h}}$ are the Fourier components of the crystal polarizability corresponding to the preset reflection; and C is the factor determined by the polarization vectors of incident and diffracted waves [7]. The wave vector $\mathbf{k}_0$ of the incident wave in the crystal differs from the corresponding vector $\boldsymbol{\kappa}$ in vacuum; it is convenient to seek this vector in the form

$$\mathbf{k}_0 = \boldsymbol{\kappa} + \frac{\kappa\chi_0}{2\gamma_0}\mathbf{n} + \kappa y\mathbf{n}, \tag{6}$$

where $\mathbf{n}$ is the normal to the crystal surface. The last two terms in this expression describe conventional refraction and the correction due to diffraction, while

$$\gamma_0 = \sin\varphi_0.$$

Taking into account Eq. (6), we can write the system of dynamic equations (5) in the form

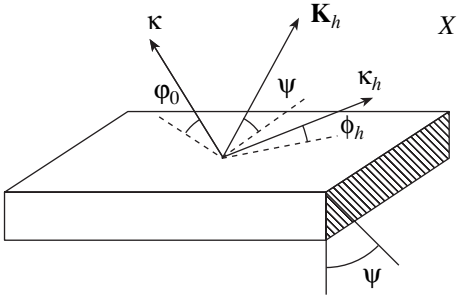


Fig. 1. Diffraction scheme in extremely asymmetric Bragg–Laue geometry.

$$\begin{aligned} [(y + \chi_0/2\gamma_0)^2 + 2\gamma_0 y]D_0 + C\chi_{\bar{h}} D_h &= 0, \\ C\chi_h D_0 + [(y + \chi_0/2\gamma_0)^2 & \\ + 2(\gamma_0 - \tilde{\psi})(y + \chi_0/2\gamma_0) + \alpha - \chi_0]D_h &= 0, \end{aligned} \tag{7}$$

where

$$\alpha = -2\sin(2\theta_B)\Delta\theta, \quad \tilde{\psi} = 2\sin\psi\sin\theta_B,$$

θ_B is the Bragg angle, $\Delta\theta$ is the deviation from the Bragg angle, and ψ is the angle between the crystal surface and vector $\mathbf{K}_h$.

For standard (symmetric or weakly asymmetric) diffraction schemes, when the angles of incident and reflected waves are quite large, we can disregard the terms quadratic in parameter y in Eqs. (7), and the system of equations (7) itself can be reduced to a form defining two sets of wave fields, (1) and (2), in the crystal (see, for example, [7]). In the case of an extremely asymmetric scheme, angle $\varphi \approx 1$, and we can disregard the terms quadratic in y in the first equation in system (7). At the same time, conditions (4) reflecting the conservation of the tangential component of the wave vector

Intensity peak Y_m of the anomalous Kossel line and its angular divergence in an extremely asymmetric diffraction scheme for different reflections of the GeK_α radiation from a germanium crystal with various cut planes

Cut plane	Reflection	θ_B	ψ	φ^*_{B-L}	FWHM(φ)	FWHM(θ)	Y_m
(111)	220	18.2°	54.7°	30.8°	17′	1″	200
	113	21.5°	60.5°	39.8°	12′	0.3″	200
	$\bar{1}13$	21.6°	46.9°	32.5°	14′	0.5″	170
(100)	220	18.2°	45.0°	26.3°	14′	1″	175
	113	21.6°	17.5°	12.8°	12′	0.5″	85
	111	11.1°	35.3°	12.8°	12′	1.5″	90
(110)	133	28.9°	40.4°	38.8°	14′	0.2″	160
	113	21.5°	25.2°	18.3°	15′	0.5′	120
	111	11.1°	54.7°	18.3°	12′	1.2″	120

Note: $\varphi^*_{B-L} = \arcsin\tilde{\psi}$ corresponds to a transition from the Bragg geometry to the Laue geometry. The sixth and seventh columns contain values of half-height widths of the line for angles φ and θ.

during the passage through the crystal surface and the elastic nature of diffraction scattering imply that

$$\varphi_h^2 = (\gamma_0 - \tilde{\psi})^2 + 2\sin(2\theta_B)\Delta\theta, \tag{8}$$

where $\varphi_h \ll 1$; i.e., in the second equation of system (7), the terms quadratic in y cannot be disregarded any longer in view of the smallness of the term containing $\gamma_0 - \tilde{\psi} \approx 0$. As a result, system (7) can be reduced to a simpler form:

$$\begin{aligned} 2\gamma_0 y D_0 + C\chi_{\bar{h}} D_h &= 0, \\ C\chi_h D_0 + [(y + \eta_h)^2 - \tilde{\varphi}_h^2] D_h &= 0, \end{aligned} \tag{9}$$

where

$$\tilde{\varphi}_h = \sqrt{\varphi_h^2 + \chi_0}, \quad \eta_h = \gamma_0 - \tilde{\psi} + \frac{\chi_0}{2\gamma_0}.$$

The condition for the existence of solutions of the homogeneous system (9) leads to the following dispersion equation:

$$y[(y + \eta_h)^2 - \tilde{\varphi}_h^2] - \delta_0^3 = 0, \tag{10}$$

where

$$\delta_0^3 = \frac{C^2 \chi_h \chi_{\bar{h}}}{2\gamma_0}.$$

The dispersion equation (10) has three roots determining the extinction modes of the wave field that are excited in the crystal. Thus, the number of modes excited in the crystal in the case of the extremely asymmetric diffraction scheme increases as compared to the standard schemes from two to three.

In a semi-infinite crystal, only modes satisfying the condition $\mathrm{Im}(y_{1,2} + \chi_0/2\gamma_0) > 0$ are excited, and the resultant wave field has the form

$$\begin{aligned} &D(\mathbf{r}) \\ &= \sum_{j=1,2} [D_0^{(j)} \exp(i\mathbf{k}_{0j} \cdot \mathbf{r}) + D_h^{(j)} \exp(i\mathbf{k}_{hj} \cdot \mathbf{r})]. \end{aligned} \tag{11}$$

The amplitudes $D_0^{(j)}$ and $D_h^{(j)}$ of the modes excited in the crystal can be determined from the boundary conditions of continuity of the incident wave at the crystal surface and from the conditions of continuity of the field and its derivative for the diffracted wave:

$$\begin{aligned} D_0^{(1)} + D_0^{(2)} &= E_0, \\ D_h^{(1)} + D_h^{(2)} &= E_h, \\ y_1 D_h^{(1)} + y_2 D_h^{(2)} &= -(\eta_h + \varphi_h) E_h. \end{aligned} \tag{12}$$

Using these boundary conditions and Eqs. (9) for the wave field amplitudes $D_0^{(j)}$ and $D_h^{(j)}$, we can easily find that

$$\begin{aligned} D_h^{(1)} &= \frac{\varphi_h + \eta_h + y_2}{y_2 - y_1} E_h, \\ D_h^{(2)} &= -\frac{\varphi_h + \eta_h + y_1}{y_2 - y_1} E_h, \\ D_0^{(j)} &= \frac{C\chi_{\bar{h}}}{2\gamma_0 y_j} D_h^{(j)}, \end{aligned} \tag{13}$$

where

$$E_h = \frac{C\chi_h}{y_3(\varphi_h - \eta_h - y_3)} E_0 \tag{14}$$

is the amplitude of the reflected wave. Here, y_3 is the root of the dispersion equation (10), having a negative imaginary component. Formulas (13) and (14) make it possible to calculate the wave fields in the crystal and in vacuum.

3. DYNAMIC DIFFRACTION IN THE VICINITY OF A DEGENERATE POINT

It can be seen from formulas (13) that the wave field amplitudes $D_h^{(j)}$ in the crystal considerably exceed the amplitude E_h of the reflected wave if the roots of the dispersion equation (10) are close. It is interesting to note that the amplitude of the reflected wave emerging from the crystal has no such singularities in accordance with formula (14). Obviously, the case when the roots of Eq. (10) coincide, (i.e., are degenerate) is of special interest.

We can readily see that if we disregard absorption, there exist directions of incidence of the X-ray beam for

$$\varphi_0 = \arcsin\left(\tilde{\psi} - \frac{3}{2}\delta_0\right) \text{ and } \Delta\theta = \frac{\varphi_c^2 - 3\delta_0^2}{2\sin(2\theta_B)} \tag{15}$$

(here, $\varphi_c = \sqrt{-\mathrm{Re}\chi_0}$ is the specular reflection angle) for which Eq. (10) has a single triply degenerate root:

$$y_1 = y_2 = y_3 = \delta_0. \tag{16}$$

In this case, we have

$$\eta_h = -\frac{3}{2}\delta_0, \quad \varphi_h^2 = \varphi_c^2 - \frac{3}{4}\delta_0^2. \tag{17}$$

The amplitudes $D_h^{(1)}$ and $D_h^{(2)}$ of the wave fields diverge in this case; however, the total amplitude of the diffracted wave in the crystal,

$$D_h(z) = D_h^{(1)} \exp(i\kappa y_1 z) + D_h^{(2)} \exp(i\kappa y_2 z), \tag{18}$$

remains finite everywhere on account of formula (13), although it increases indefinitely during the propagation of the wave to the bulk of the crystal:

$$D_h(z) = E_h + i\frac{C\chi_h}{\delta_0}\kappa z E_0. \tag{19}$$

This result shows that the amplitude of the diffracted wave and, hence, the flux density of the diffracted wave increase indefinitely with z. The increase in the flux density indicates that the X-ray beam is compressed unlimitedly in the bulk of the crystal.

The compression of the beam as a result of diffraction with the asymmetric scheme with the asymmetry parameter

$$\beta = \frac{\gamma_0}{\gamma_h} = \frac{\sin\varphi_0}{\sin\varphi_h} > 1$$

can easily be explained on the basis of simple geometrical considerations. It can be seen from Fig. 2 that the cross sectional area of the reflected radiation beam decreases by a factor of β relative to the cross-sectional area of the incident beam for moderate values of the asymmetry parameter ($\varphi_h \gg \varphi_c$). In this case, in view of the equality of the fluxes of reflected and incident beams, the radiation flux density in the reflected beam increases by a factor of β as compared to the flux density in the incident beam. This phenomenon is effectively used for obtaining X-ray beams with a small cross section. As the angle φ_h decreases, specular reflection starts playing a noticeable role, and a larger and larger part of the reflected beam passes to the bulk of the crystal. In this case, the intensity of the reflected beam decreases, and the effect of diffraction-induced reflection becomes of no physical significance at first glance. It can be seen from formula (19), however, that a decrease in angle φ_h leads to a further beam compression with conservation of the total radiation flux, but in the bulk of the crystal rather than at its surface. It is this phenomenon that underlies the anomalous Kossel effect.

Let us take into account the actual absorption in the crystal. In most cases, it turns out to be small so that

$$\mathrm{Im}\chi_0 < \delta_0^2. \tag{20}$$

In this case, the roots of Eq. (10) have the form

$$y_1 = \delta_0\left(1 - i\eta a - \frac{\eta^2 a^2}{3}\right),$$

$$y_2 = \delta_0\left(1 - i\eta^2 a - \frac{\eta a^2}{3}\right), \tag{21}$$

$$y_3 = \delta_0\left(1 - ia - \frac{a^2}{3}\right),$$

where

$$\eta = e^{2\pi i/3}, \quad a = (\mathrm{Im}\chi_0/\delta_0^2)^{1/3}. \tag{22}$$

In this case, the amplitude of the diffracted wave has the form

$$D_h(z) = \left[-\frac{2i}{\sqrt{3}}\frac{\varphi_c}{\delta_0 a}\sin\frac{\sqrt{3}z}{L_0} + \cos\frac{\sqrt{3}z}{L_0}\right]\exp\left(-\frac{z}{L_0}\right)E_h, \tag{23}$$

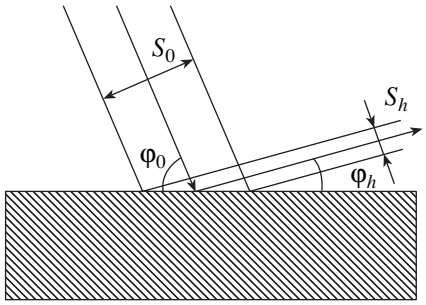


Fig. 2. X-ray beam compression in the asymmetric diffraction scheme ($S_h = S_0/\beta$).

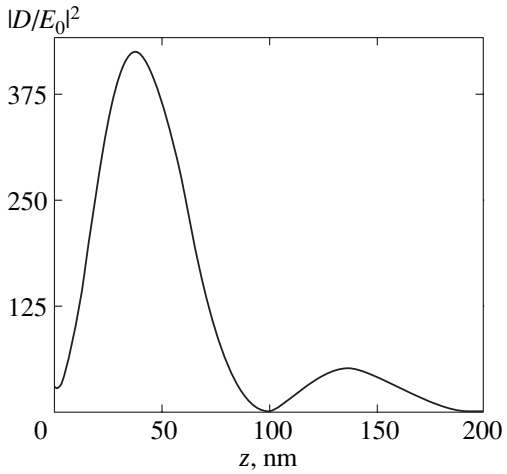


Fig. 3. Intensity distribution for the diffracted wave over the depth of the germanium crystal cut along the (111) plane for the (220) reflection of the GeK_α radiation ($\psi = 54.74°$, $\varphi_0 = 30.8°$).

where

$$L_0 = \lambda/\pi a\delta_0. \tag{24}$$

Thus, the inclusion of the absorption effect removes the degeneracy of the roots of Eq. (10) and limits the indefinite growth of the amplitude $D_h(z)$. It follows from Eq. (23) that the peak of compression of the diffracted wave lies at the depth L_0.

Figure 3 shows the field distribution of the diffracted wave in the bulk of the crystal in the case of (220) reflection of the characteristic GeK_α radiation from the germanium crystal and the direction of the incident wave corresponding to the degenerate point defined by expressions (15) and (17), taking into account all real factors (including absorption). It can be seen that even when all real factors are taken into account, the radiation flux density in the case under investigation increases almost by a factor of 500. It should be noted that the compression ratio actually attained for external reflected beams in asymmetric diffraction schemes amounts to only 15–20.

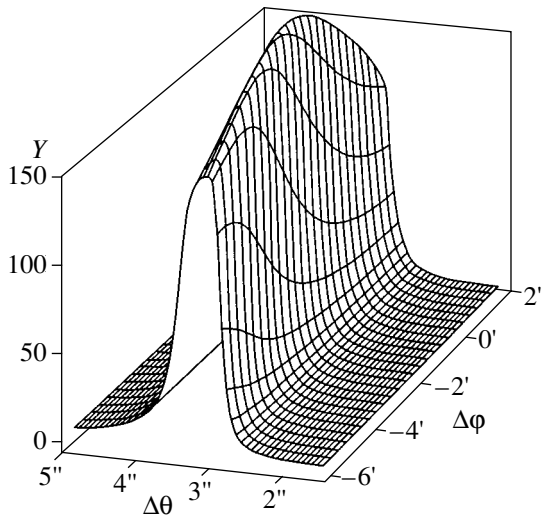


Fig. 4. Anomalous Kossel effect for the (220) reflection of the GeK_α radiation from a germanium crystal cut along the (111) plane for an exponential distribution of excited atoms over the depth for L = 60 nm, $\Delta\varphi = \varphi - \varphi_{B-L}$ (see table).

4. ANOMALOUS KOSSEL LINES

It can be seen from the results of the previous section and Fig. 3 that the field intensity at the atoms lying at a certain depth of the crystal may considerably exceed the field intensity for incidence angles differing considerably from the Bragg angle for certain directions of the incident beam corresponding to extremely asymmetric Bragg diffraction. By virtue of the reciprocity principle, we can expect that an spike of radiation intensity must emerge as a result of excitation of atoms at a small depth in a crystal in directions corresponding to extremely asymmetric diffraction. This phenomenon will be referred to as the anomalous Kossel effect. In order to determine the radiation intensity distribution in a Kossel line, we must first calculate the radiation field intensity at crystal atoms depending on the direction of the incident radiation. This problem can be solved on the basis of the results of previous sections. For the radiation field intensity at the sites of location of atoms in the crystal, we have

$$I_a(z, \Delta\theta, \varphi) = \left| \sum_{j=1,2} [D_0^{(j)} + D_h^{(j)} \exp(i\mathbf{K}_h \cdot \boldsymbol{\rho}_a)] \exp\left[i\left(y_j + \frac{\chi_0}{2\gamma_0}\right)\kappa z \right] \right|^2, \quad (25)$$

where $\boldsymbol{\rho}_a$ determines the position of atoms of group a in the unit cell. Using formulas (10), (13), and (14), we can easily carry out specific calculations of this intensity.

In accordance with the reciprocity principle [9], if X rays incident on the crystal at angles φ and θ create a field of intensity (25) at crystal atoms at depth z with coordinates $\boldsymbol{\rho}_a$ in a unit cell, the excited atom located in the bulk of the crystal at the same depth z with the same coordinates $\boldsymbol{\rho}_a$ emits in the direction (φ, θ) with intensity I_a^{out} which is connected with the quantity (25) through the relation

$$I_a^{\text{out}}(z, \Delta\theta, \varphi) = \text{const} I_a(z, \Delta\theta, \varphi). \quad (26)$$

If atoms in the crystal are excited with a distribution defined by the function $P(z)$, the shape of the radiation intensity distribution is given by

$$Y(\Delta\theta, \varphi) = \frac{1}{A}\int_0^\infty dz P(z) I_a^{\text{out}}(z, \Delta\theta, \varphi), \quad (27)$$

where

$$A = \int_0^\infty dz P(z) I_a^{\text{out}}(z, \Delta\theta \longrightarrow \infty, \varphi) = E_0^2 \int_0^\infty dz P(z) \exp\left(-\frac{\mu z}{\gamma_0}\right) \quad (28)$$

is the coefficient ensuring the renormalization of the background intensity, $Y(\Delta\theta \longrightarrow \infty, \varphi) = 1$, for large deviations from the Bragg angle and $\mu = \kappa \text{Im}\chi_0$ is the linear coefficient of radiation absorption in the crystal.

Figure 4 shows the intensity distribution for GeK_α radiation in the Kossel line for the (220) reflection from the germanium crystal cut along the (111) plane, which is calculated using formulas (25), (27), and (28) with an exponential distribution of excited atoms over the depth,

$$P(z) = e^{-z/L}, \quad (29)$$

for L = 60 nm. Such a distribution of excited atoms over the depth can be ensured by using the characteristic MoK_α radiation incident on the crystal at a small angle close to the specular reflection angle. The anomalous radiation in the Kossel line emerges from the crystal at an angle $\varphi = 30.8°$ to its surface (see table), and, hence, such radiation can easily be detected. It can be seen from Fig. 4 that the radiation intensity distribution in the Kossel line has a clearly pronounced peak exceeding the background intensity by more than two orders of magnitude. The depth L = 60 nm of exponential decrease in the distribution of excited atoms ensures the maximum value of radiation intensity in the Kossel line, corresponding to the field distribution presented in Fig. 3. Considerable excess of the radiation intensity over the background embraces in this case a small range of angles in the vicinity of the degenerate point (only a few angular seconds) in angle θ, while in angle φ this region is much wider (of the order of ten angular minutes). It should be noted that the intensity distribution in the standard Kossel lines exhibits only insignificant deviations relative to the background, but embraces the entire range of variation of φ.

The table contains the results of corresponding calculations of the intensity peak Y_m for the anomalous Kossel line and its angular divergences in θ and φ for some reflections of the GeK_α radiation from a germanium crystal cut along the principal crystallographic planes. It can be seen from these data that, for any crystal, there exist a large number of crystallographic directions along which anomalous Kossel lines differing in the maximum values and shape of the radiation intensity distribution, but preserving the main qualitative feature (these lines have the form of a narrow collimated radiation beam with a divergence in angle θ of the order of angular seconds and in angle φ of the order of ten angular minutes).

The simplest way of detecting anomalous Kossel lines is to use a crystal analyzer (Fig. 5) which must be made of the same material as the sample under investigation. In this case, in order to be able to disregard dispersion, use should be made of the same reflection for which the anomalous Kossel effect must be observed. In addition, the crystal analyzer should be oriented so that the reciprocal lattice vectors $\mathbf{K}_h$ for the sample under investigation and the crystal analyzer lie in the plane of reflection. Rotating the crystal analyzer about a vertical axis, we can measure in this scheme the dependence of the reflected wave intensity on the angle of rotation $\Delta\theta_A$ of the crystal analyzer. The calculation of the diffraction-induced reflection (rocking) curve obtained in this experiment is reduced to the calculation of the convolution of the radiation intensity distribution function within the Kossel line (27) with the reflection curve $P_R^A(\theta, \varphi)$ of the crystal analyzer in the angular interval determined by the system of gaps:

$$P_R(\Delta\theta_A) = \frac{1}{B}\iint Y(\Delta\theta, \varphi)P_R^A(\Delta\theta_A - \Delta\theta, \varphi)d\Delta\theta d\varphi, \quad (30)$$

where

$$B = \iint P_R^A(\Delta\theta, \varphi)d\Delta\theta d\varphi \quad (31)$$

is the coefficient ensuring the normalization to the background intensity. In order to avoid the suppression of the observed effect due to the background enhancement, we must cut a certain angular region of the order of ten angular minutes in angle φ in the vicinity of the radiation intensity peak in the Kossel line with the help of a slit in front of the crystal analyzer and, in addition, choose the extremely asymmetric reflection in the analyzer to make the asymmetry coefficient β large enough to ensure the formation of a narrow detection window in angle θ in the crystal analyzer. In addition, the angle at which radiation emerges from the analyzer must be much larger than the specular reflection angle to ensure relatively simple detection of this radiation, preserving the maximum value of reflectivity.

Figure 6a shows the curve of reflection from the germanium crystal analyzer for asymmetric (220) reflection with the asymmetry coefficient β = 14. Such a geometry can be ensured by choosing a germanium single crystal cut so that the crystal surface forms an angle of 16° with the (220) planes. In this case, the angle φ_h^A at which the reflected wave emerges amounts approximately to 2°. Figure 6b shows the diffraction-induced reflection curve corresponding to the experimental setup depicted in Fig. 5 and calculated by formulas (30) and (31). The corresponding curve has a clearly manifested peak whose intensity at the maximum exceeds the background intensity by a factor greater than 30. Since the reflection curve for the analyzer is much broader than the distribution in the initial Kossel line (see Fig. 4), the diffraction-induced reflection curve in Fig. 6b is more blurred as compared to the initial anomalous Kossel line; however, the effect remains strong enough for its experimental detection.

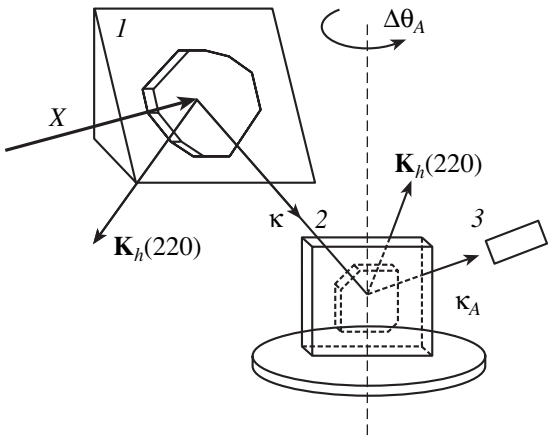


Fig. 5. Schematic diagram of experiment for observing the anomalous Kossel effect: *X* stands for X-ray beam; *1*, sample; *2*, crystal analyzer; *3*, detector.

For comparison, the dashed curve in Fig. 6b presents the profile of the diffraction-induced reflection curve for the standard Kossel line corresponding to a symmetric diffraction scheme. In this case, relatively small changes in the radiation intensity are observed, when the peak of the distribution is higher than the background intensity only by a factor of several units.

5. NEW TYPE OF X-RAY SOURCE

It can be seen from Fig. 4 that the radiation intensity in the Kossel line considerably exceeds the background intensity only in a small angular region θ of a few angular seconds. However, this is precisely the angular range which is used in diffraction studies of crystals with a high degree of perfection. In a conventional X-ray tube, radiation is distributed quite uniformly over a wide angular interval of the order of several degrees. However, in order to analyze a crystal, just an angular interval of the order of a few seconds must be cut with the help of a monochromator crystal, while the remain-

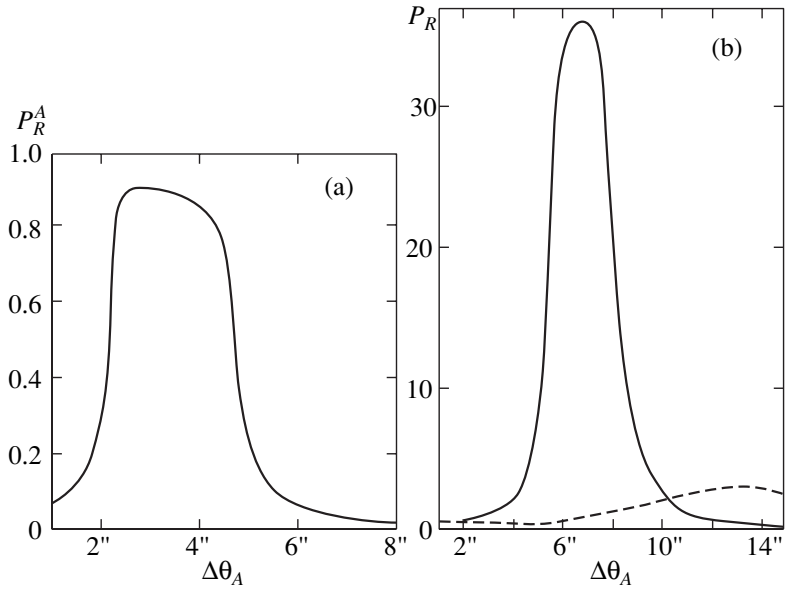


Fig. 6. (a) Curve describing the reflection of the GeK_α radiation from the (220) planes of the germanium crystal analyzer with the asymmetry coefficient $\beta = 14$ ($\psi_A = 74°$ is the angle between the normal to the surface and reflecting planes). (b) Curves describing diffraction-induced reflection for the anomalous Kossel line, presented in Fig. 4, in the experiment shown in Fig. 5 (solid curve) and for the Kossel line corresponding to the symmetric diffraction scheme (dashed curve).

ing large part of radiation is not used. Moreover, the crystal is usually collimated in the vertical angle φ within a degree with the help of a system of slits. As a result, a negligibly small (of the order of 10^{-7}–10^{-6}) fraction of X-ray radiation of the tube is used in each specific experiment. In view of this circumstance, anomalous Kossel lines may serve as a new type of X-ray source since the radiation emerging from the crystal has an angular distribution exactly meeting the requirements of diffraction experiments, but the background intensity in this case is lower than the peak intensity by two or three orders of magnitude. In other words, such a source intended for ensuring the X-ray radiation intensity required for diffraction intensity studies would have a considerably lower power as compared to a standard X-ray tube.

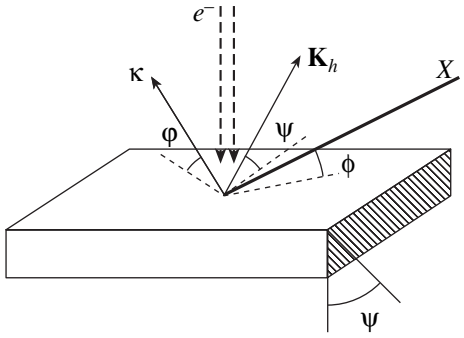


Fig. 7. Two ways of exciting atoms in a thin surface layer by an X-ray beam (X) and by an electron beam (e^-).

A Kossel line can be excited in a crystal in two ways illustrated in Fig. 7. The first method involves the employment of X-ray radiation with energy exceeding the excitation energy for the corresponding K_α line, incident at a small angle ϕ to the crystal surface. This method is most convenient for detecting the effect of anomalous Kossel lines proper. The second method is the excitation by an electron beam that is normally used in experiments on observation of Kossel lines. In the case of the anomalous Kossel effect under investigation, we must ensure the excitation of atoms in a thin layer near the crystal surface. This can easily be achieved by selecting the energy E_e of the electron beam. The energy E_e of electrons should not be much higher than the excitation energy for atoms in the crystal. For example, in the case of GeK_α radiation considered by us here, the values of E_e, according to estimates, lie in the interval 13–15 keV. In standard X-ray tubes, a much higher voltage of the order of 40–60 kV is required for obtaining energies in this range. This gives grounds to expect that the anomalous Kossel effect can be used for creating a compact and low-power source of X-ray radiation, ensuring the required intensity for carrying out diffraction experiments.

ACKNOWLEDGMENTS

This study was supported financially by the Russian Foundation for Basic Research (project no. 99-02-16665) and Charity Foundation for Supporting Science in Russia.

REFERENCES

1. W. Kossel, V. Loeck, and H. Voges, Z. Phys. **94**, 139 (1935).
2. A. M. Afanas'ev, M. V. Koval'chuk, and M. A. Chuev, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 309 (2001) [JETP Lett. **73**, 271 (2001)].
3. A. M. Afanas'ev and A. V. Esayan, Phys. Status Solidi A **126**, 303 (1991).
4. A. M. Afanas'ev, R. M. Imamov, É. Kh. Mukhamedzhanov, *et al.*, Fiz. Tverd. Tela (Leningrad) **32**, 650 (1990) [Sov. Phys. Solid State **32**, 383 (1990)].
5. A. M. Afanas'ev, R. M. Imamov, and É. Kh. Mukhamedzhanov, Kristallografiya **40**, 567 (1995) [Crystallogr. Rep. **40**, 521 (1995)].
6. N. Hertel, M. V. Kovalchuk, A. M. Afanas'ev, and R. M. Imamov, Phys. Lett. A **75**, 501 (1980).
7. A. M. Afanas'ev, P. A. Aleksandrov, and R. M. Imamov, *X-ray Diffraction Diagnostics of Submicron Layers* (Nauka, Moscow, 1989).
8. Z. G. Pinsker, *Dynamical Scattering of X-rays in Crystals* (Nauka, Moscow, 1974; Springer-Verlag, Berlin, 1978).
9. M. von Laue, *Roentgenstrahlen-interferentzen* (Akademische Verlagsgesellschaft, Frankfurt am Main, 1960).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 480–490.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 558–569.

SOLIDS
Structure

The Effect of Normal Phonon–Phonon Scattering Processes on the Maximum Thermal Conductivity of Isotopically Pure Silicon Crystals

I. G. Kuleev* and I. I. Kuleev

Institute of Metal Physics, Ural Division, Russian Academy of Sciences, ul. S. Kovalevskoĭ 18, Yekaterinburg, 620219 Russia

**e-mail: kuleev@imp.uran.ru*

Received February 18, 2002

Abstract—The effect of normal phonon–phonon scattering processes on the thermal conductivity of silicon crystals with various degrees of isotope disorder is considered. The redistribution of phonon momentum in normal scattering processes is taken into account within each oscillation branch (the Callaway generalized model), as well as between different oscillation branches of the phonon spectrum (the Herring mechanism). The values of the parameters are obtained that determine the phonon momentum relaxation in anharmonic scattering processes. The contributions of the drift motion of longitudinal and transverse phonons to the thermal conductivity are analyzed. It is shown that the momentum redistribution between longitudinal and transverse phonons in the Herring relaxation model represents an efficient mechanism that limits the maximum thermal conductivity in isotopically pure silicon crystals. The dependence of the maximum thermal conductivity on the degree of isotope disorder is calculated. The maximum thermal conductivity of isotopically pure silicon crystals is estimated for two variants of phonon momentum relaxation in normal phonon–phonon scattering processes. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In view of the intense development of modern semiconductor technologies, the studies of the physical properties of isotopically enriched crystals of germanium, silicon, and diamond, whose thermal conductivity considerably increases as the isotope disorder decreases, have become especially important [1–11]. The use of isotopically enriched crystals as substrates for integrated circuits can substantially increase the operational stability of microprocessors and considerably increase the density of elements in integrated circuits due to the faster heat removal. The experimental investigations of the thermal conductivity and the thermoelectric power [2–5] carried out on germanium crystals with various degrees of isotope disorder have shown that the maximum values of thermal conductivity in isotopically pure samples with 99.99% ^{70}Ge isotope concentration are an order of magnitude greater than those in crystals with the natural isotope composition and the absolute values of thermoelectric power in the former crystals are greater than those in the latter by a factor of greater than two. As compared with natural crystals, the maximum thermal conductivity in silicon crystals with 99.8588% ^{28}Si isotope concentration is greater by a factor of six [6]. At room temperatures, this factor was 60% [6]. Since germanium and silicon are widely used in semiconductor microelectronics, the investigation of the physical properties of isotopically pure crystals and the understanding of microscopic relaxation processes of quasiparticles in these materials is very important from the viewpoint of applications. Therefore, one of the aims of this study is the investigation of the effect of normal phonon–phonon scattering processes on the maximum values of the thermal conductivity that can be attained in isotopically enriched silicon crystals. Proper consideration of normal phonon scattering processes is especially important for isotopically enriched crystals of germanium, silicon, and synthetic diamond at sufficiently low temperatures when the umklapp processes of phonon–phonon scattering are largely frozen and the normal processes of scattering play a critical role in the phonon momentum relaxation [1–12]. However, as shown in [13], this case requires that the role of normal scattering processes for the phonons belonging to different oscillation branches should be analyzed more carefully than was done in [1–12].

The role of normal phonon–phonon scattering processes in the theory of lattice thermal conductivity has been sufficiently well studied [14–19]. These scattering processes must be taken into account under the conditions when the phonon relaxation rate $\nu_{phN}(q)$ in normal processes is either greater than or comparable with the relaxation rate $\nu_{phR}(q)$ in resistive scattering processes, which are associated with the phonon relaxation in umklapp processes, on the boundaries, impurities, and electrons. The normal scattering processes do not directly contribute to the phonon momentum relaxation and, hence, to the thermal resistivity. However, these

processes redistribute energy and momentum between different phonon modes and thereby form a nonequilibrium phonon distribution and bring a phonon system to a drift locally equilibrium distribution [14–19]. When the normal scattering processes were taken into account without separating the contributions of longitudinal (L) and transverse (T) phonons in the Callaway model [14] with the use of the isotropic Debye model, one obtained overestimated values of thermal conductivity near the maximum for germanium and silicon crystals with the natural isotope composition. Later [20], it was shown that one should separate the contributions of L and T phonons when calculating the thermal conductivity of germanium and silicon crystals, because T phonons have strong dispersion and, therefore, the Debye temperatures for the two oscillation branches are strongly different. A further development of the theory of lattice thermal conductivity was restrained by the lack of the correct analysis of the role of normal scattering processes for the phonons belonging to different oscillation branches. In the generalized Callaway model, which was widely used for calculating the thermal conductivity of isotopically enriched crystals of germanium, silicon, and diamond [1–12], it was assumed that the momentum relaxation of phonons only occurs within each branch of the phonon spectrum and that phonons of different polarizations contribute additively to the thermal conductivity. This model gave considerably overestimated values of thermal conductivity for isotopically pure crystals of ^{70}Ge (99.99%) near the maximum [2]. The introduction of an additional, dislocation, mechanism of phonon scattering could not remedy the situation because, according to [21], the concentration of dislocations was four orders of magnitude lower than what was necessary in [2] for fitting calculated data to measured values of the thermal conductivity of ^{70}Ge (99.99%) near the maximum.

In [13, 22], the authors considered the effect of the phonon momentum redistribution in normal phonon–phonon scattering processes both within each oscillation branch (the Simons mechanism) and between different oscillation branches (the Herring mechanism) on the effective relaxation rate of electrons, the phonon-drag thermoelectric power, and the lattice thermal conductivity of conductors. It was demonstrated that the above parameters essentially depend on the character of momentum relaxation of phonons in normal scattering processes and have different forms for the Herring [23] and Simons [24] relaxation mechanisms. The analysis of the thermal conductivity of germanium crystals with different degrees of isotope disorder [13] for two variants of phonon momentum relaxation in normal processes has shown that the generalized Callaway model corresponds to the Simons relaxation mechanism [24] and is not correct when the dominant mechanism of phonon momentum relaxation in normal processes is the Herring mechanism [23], as is the case in germanium and silicon. In this case, the momentum redistribution between L and T phonons in normal Herring processes leads to a substantial suppression of the drift motion of longitudinal phonons in isotopically pure samples of Ge (99.99%) primarily due to the relaxation of T phonons. Therefore, this mechanism of phonon momentum redistribution in normal processes leads to a substantial decrease in the maximal value of the total thermal conductivity κ_{max} of isotopically enriched germanium crystals. As a result, the maximum value of the thermal conductivity of germanium crystals $\kappa_{max}(g)$ versus the isotope disorder parameter g reaches a saturation level as g decreases below 10^{-6} (which corresponds to 99.9% concentration of ^{70}Ge isotope). However, these maximum values essentially depend on the magnitude and character of boundary scattering of phonons. Thus, a further increase in the isotopic purity of germanium crystals may lead to an increase in κ_{max} by less than 1% as compared with the values attained for ^{70}Ge (99.99%) [2]. This result of [13] is of great practical importance. The model that we proposed for the phonon momentum redistribution in normal Herring scattering processes [23] provided a more adequate interpretation of the experimental data on the thermal conductivity of germanium crystals with various degrees of isotope disorder as compared with the generalized Callaway model used earlier [1, 2, 6–12]. In addition, our model does not require the introduction of an additional scattering mechanism of phonons by dislocations for ^{70}Ge (99.99%) [2] or an additional fitting parameter of the theory.

In this paper, we apply the method described in [13, 22] to the calculation of the thermal conductivity of silicon crystals with various isotope compositions for two variants of phonon momentum relaxation in normal scattering processes. Based on the results of [1, 2, 6, 25], we find the parameters that determine the phonon relaxation rates in resistive and normal scattering processes. We analyze the contributions of the drift motion of L and T phonons to the thermal conductivity of silicon with various degrees of isotope disorder. Let us estimate what maximum values of the thermal conductivity can be obtained in perfect, chemically pure and isotopically homogeneous, silicon crystals.

2. NORMAL PHONON–PHONON SCATTERING PROCESSES AND THE LATTICE THERMAL CONDUCTIVITY OF SILICON CRYSTALS WITH VARIOUS ISOTOPE COMPOSITIONS

As shown in [13], the lattice thermal conductivity with separate contributions from different branches of the phonon spectrum is expressed as

$$\kappa(T) = \sum_{\lambda} \frac{k_B s_\lambda^2 q_{T\lambda}^3}{6\pi^2} \times \int_0^{z_{d\lambda}} dz_q^\lambda \frac{(z_q^\lambda)^4}{\tilde{\nu}_{\lambda^{ph}}^{(S,H)}(q)} N_{q\lambda}^0 (N_{q\lambda}^0 + 1), \quad (1)$$

where

$$z_q^\lambda = \frac{\hbar\omega_{q\lambda}}{k_BT} = \frac{q}{q_{T\lambda}}, \quad q_{T\lambda} = \frac{k_BT}{\hbar s_\lambda},$$

$N^0_{q\lambda}$ is the Planck function, s_λ is the speed of acoustic phonons with polarization λ, $z_{d\lambda} = \hbar\omega_{d\lambda}/k_BT$ ($\omega_{d\lambda}$ is the Debye frequency of phonons with polarization λ), and $\tilde{\nu}^{(S,H)}_{\lambda^{ph}}(q)$ is the effective phonon momentum relaxation rate. An allowance for normal phonon–phonon scattering processes reduces to the renormalization of the phonon momentum relaxation rate entering the lattice thermal conductivity. The character of this renormalization depends on whether the normal processes redistribute the phonon momentum within each oscillatory branch (the Simons mechanism [24]) or the momentum is predominantly redistributed between different oscillation branches (the Herring mechanism [23]):

$$\tilde{\nu}^{(S,H)}_{\lambda^{ph}}(q) = \nu^\lambda_{ph}(q)\left(1 + \nu^\lambda_{phN}(q)\beta_{(S,H)}\right)^{-1}, \qquad (2)$$

$$\beta_S = \frac{\Psi^\lambda_N}{\Psi^\lambda_{NR}}, \quad \beta_H = \left(\frac{s_L}{s_\lambda}\right)^2 \frac{\Psi^L_N + 2S^3_*\Psi^T_N}{\Psi^L_{NR} + 2S^5_*\Psi^T_{NR}}. \qquad (3)$$

Here,

$$\Psi^\lambda_N = \left\langle \frac{\nu^\lambda_{phN}(q)}{\nu^\lambda_{ph}(q)} \right\rangle_{z_{d\lambda}}$$

$$\equiv \int_0^{z_{d\lambda}} dz^\lambda_q (z^\lambda_q)^4 \frac{\nu^\lambda_{phN}(q)}{\nu^\lambda_{ph}(q)} N^0_{q\lambda}(N^0_{q\lambda}+1), \qquad (4)$$

$$\Psi^\lambda_{NR} = \left\langle \frac{\nu^\lambda_{phR}(q)\nu^\lambda_{phN}(q)}{\nu^\lambda_{ph}(q)} \right\rangle_{z_{d\lambda}},$$

$S_* = s_L/s_T$, $\nu^\lambda_{ph}(q) = \nu^\lambda_{phN}(q) + \nu^\lambda_{phR}(q)$ is the total relaxation rate of phonons with wave vector q and polarization λ, $\nu^\lambda_{phN}(q)$ is the relaxation rate of phonons in normal scattering processes, and $\nu^\lambda_{phR}(q) = \nu^\lambda_{phU}(q) + \nu^\lambda_{phi}(q) + \nu^\lambda_{phB}(q)$ is the relaxation rate of phonons in resistive scattering processes associated with umklapp phonon scattering processes ($\nu^\lambda_{phU}(q)$), scattering by impurities ($\nu^\lambda_{phi}(q)$), and scattering by the boundaries of a sample ($\nu^\lambda_{phB}(q)$). According to [26], for a silicon sample with orientation [111], we have $s_L = 9.35\times10^5$ cm/s, $s_T = 5.09\times10^5$ cm/s, and $S_* = 1.84$; for orientation [100], we have $s_L = 8.43\times10^5$ cm/s, $s_T = 5.85\times10^5$ cm/s, and $S_* = 1.44$.

The Simons relaxation mechanism [24] involves phonons of the same polarization. In this scattering mechanism ($\nu^\lambda_{phN} \approx B_\lambda T^4\omega_\lambda$), the momentum conservation law in normal processes holds for each branch of the phonon spectrum, and the drift velocities of L and T phonons are different [13]. In this case, the expression for the lattice thermal conductivity represents an additive sum of contributions from phonons with different polarizations:

$$\kappa(T) = \sum_\lambda \frac{k_B}{6\pi^2 s_\lambda}\left(\frac{k_BT}{\hbar}\right)^3 \int_0^{z_{d\lambda}} dz\, f(z)\left(1 + \nu^\lambda_{phN}(q)\frac{\Psi^\lambda_N}{\Psi^\lambda_{NR}}\right), \qquad (5)$$

$$f(z) = \frac{z^4 e^z}{(e^z-1)^2\nu^\lambda_{ph}(q)}.$$

Expression (5) corresponds to the generalized Callaway model, which was widely used for calculating the thermal conductivity of isotopically enriched crystals of germanium, silicon, and diamond [1, 2, 6–12]. It is valid when normal processes redistribute the phonon momentum only within each separate oscillation branch and corresponds to the Simons mechanism [24].

However, the dominant mechanism of normal three-phonon scattering processes in germanium and silicon crystals is the Herring mechanism [23], which redistributes the phonon momentum among different oscillation branches. In this mechanism, the relaxation rate of transverse phonons ($\nu^T_{phN} \approx B_T T^4\omega_T$) is determined by three-phonon scattering processes ($T + L \rightleftharpoons L$) in which one T and two L phonons take part [16]. The relaxation rate of L phonons in the anisotropic continuum model ($\nu^L_{phN} \approx B_L T^3\omega^2_L$) is determined either by three-phonon decay processes, in which an L phonon is decomposed into two T phonons belonging to different branches, or by a fusion of two T phonons that gives rise to an L phonon ($L \rightleftharpoons T_1 + T_2$) [15–17]. Thus, phonons of different polarizations take part in the normal Herring processes, and this relaxation mechanism guarantees the redistribution of the drift momentum between longitudinal and transverse phonons. Thus, the Herring three-phonon processes in a nonequilibrium phonon system tend to establish a locally equilibrium distribution with a unique mean drift velocity of phonons for both polarizations: $u_L = u_T = u_H$. In this case ($u_T = u_L = u_{ph}$), the solution of the kinetic equation for the phonon distribution function yields the following expression for the contributions of longitudinal and transverse phonons to the lattice thermal conductivity [13]:

Table 1

	B_N^L, 1/K^4	B_N^T, 1/K^3	B_U^L, s	B_U^T, s	C_U^L, K	C_U^T, K	γ [16] (method I)	γ [16] (method II)	γ [28]
Ge	2×10^{-21}	2×10^{-13}	5×10^{-19}	1×10^{-19}	180	55	0.67	0.76	0.50
Si	2.39×10^{-22}	1.99×10^{-14}	2.2×10^{-19}	4.1×10^{-20}	308	98	0.51	0.47	0.42
ξ	0.12	0.1	0.44	0.41					

$$\kappa^L(T) = \frac{k_B}{6\pi^2 s_L}\left(\frac{k_B T}{\hbar}\right)^3 \times \int_0^{z_{dL}} dz\, f(z)\left(1 + \nu_{phN}^L \frac{\Psi_N^L + 2S_*^3\Psi_N^T}{\Psi_{NR}^L + 2S_*^5\Psi_{NR}^T}\right),$$
$$\kappa^T(T) = \frac{k_B}{3\pi^2 s_T}\left(\frac{k_B T}{\hbar}\right)^3 \times \int_0^{z_{dT}} dz\, f(z)\left(1 + S_*^2\nu_{phN}^T \frac{\Psi_N^L + 2S_*^3\Psi_N^T}{\Psi_{NR}^L + 2S_*^5\Psi_{NR}^T}\right). \quad (6)$$

Therefore, we will apply formulas (6) to calculating the lattice thermal conductivity of silicon crystals with various degrees of isotope disorder and improve the results of the analysis carried out in [6] within the framework of the generalized Callaway model. Formulas (6) show that the contribution of T phonons to the drift terms increases due to the factor S_*, whose value for a silicon sample with orientation [111] amounts to 1.84.

3. PHONON RELAXATION RATES IN SILICON CRYSTALS

Based on the results obtained for germanium in [1, 2, 13], let us find parameters that determine the relaxation rates of phonons in anharmonic scattering processes in silicon. According to the experiments on neutron scattering, the phonon dispersion curves in the units of ion plasma frequency for germanium and silicon are close over the whole Brillouin zone [27]. The parameters of the effective force interaction also have close values. The main difference between germanium and silicon as regards the oscillation spectra and the parameters of anharmonic relaxation processes is attributed to the difference between the masses of vibrating atoms. It is this factor that is primarily responsible for the difference in the relevant effective Debye temperatures: Θ(Si)/Θ(Ge) = 1.75. Therefore, one can determine the relaxation parameters for silicon based on the results for germanium. According to [1], parameter A_N, which determines the phonon relaxation rates in normal scattering processes, is given by

$$A_N \propto \frac{\gamma^2}{Ma^2\Theta^5}, \quad (7)$$

where M is the atomic mass, a is the lattice constant, and γ is the Grunisen parameter. Then, we obtain the following expression for the coefficient ξ_N^λ, which determines the phonon relaxation rate in normal scattering processes in silicon in terms of the corresponding parameter for germanium:

$$A_N^\lambda(\mathrm{Si}) = \xi_N^\lambda A_N^\lambda(\mathrm{Ge}),$$
$$\xi_N^\lambda = \left(\frac{\gamma_{\mathrm{Si}}}{\gamma_{\mathrm{Ge}}}\right)^2\left(\frac{M_{\mathrm{Ge}}}{M_{\mathrm{Si}}}\right)\left(\frac{a_{\mathrm{Ge}}}{a_{\mathrm{Si}}}\right)^2\left(\frac{\Theta_{\mathrm{Ge}}^\lambda}{\Theta_{\mathrm{Si}}^\lambda}\right)^5. \quad (8)$$

If we neglect the variation in the Grunisen parameter while passing from germanium to silicon, as was assumed in [1], and make use of the effective Debye temperatures Θ(Si) = 650 K and Θ(Ge) = 376 K [20], then we obtain the following transition parameter: $\xi_N = 0.17$. If we take into account that the Debye temperatures for L and T phonons are different (Θ^L(Si) = 570 K, Θ^T(Si) = 210 K, Θ^L(Ge) = 333 K, and Θ^T(Ge) = 118 K), we obtain the following estimates for the parameter ξ_N^λ:

$$\xi_N^L \approx 0.19, \quad \xi_N^T \approx 0.16. \quad (9)$$

Let us estimate the variation in ξ_N^λ when the difference between the Grunisen parameters of germanium (γ_{Ge}) and silicon (γ_{Si}) is taken into consideration. Note that the values of the Grunisen parameters γ_{Ge} and γ_{Si} strongly depend on the method of their determination [16, 28]. A direct measurement of the third-order elasticity constants yields $\gamma_{\mathrm{Ge}} = 0.67$ and $\gamma_{\mathrm{Si}} = 0.51$ (see Table 1) [16] (method I); then, from (8) we obtain

$$\xi_N^L \approx 0.11, \quad \xi_N^T \approx 0.091. \quad (9a)$$

It follows from thermodynamic relations [16] that $\gamma_{\mathrm{Ge}} = 0.76$ and $\gamma_{\mathrm{Si}} = 0.47$ (see Table 1) [16] (method II); then, taking into account (8), we obtain

$$\xi_N^L \approx 0.073, \quad \xi_N^T \approx 0.06. \quad (9b)$$

Table 2

	A_i^L, s^{-1}	A_i^T, s^{-1}	A_N^L, s^{-1}	A_N^T, s^{-1}	A_U^L, s^{-1}	A_U^T, s^{-1}	C_U^L, K	C_U^T, K
Ge	37.1×10^6	167×10^6	3.4×10^6	2.6×10^3	8.6×10^6	1.72×10^6	180	55
Si	5.8×10^6	35.7×10^6	0.41×10^6	0.26×10^3	3.77×10^6	0.70×10^6	308	98

Note: There is a misprint in [13] in the values of A_U^λ: the parameters A_U^L and A_U^T are interchanged. The correct values are given in Table 2.

From the formulas for the phonon relaxation rates in normal scattering processes [16] (formulas (1.3.6) and (1.3.13)) for various values of the Grunisen parameters (see Table 1), we obtain

$$0.05 \le \xi_N^L \le 0.13, \quad 0.05 \le \xi_N^T \le 0.14. \tag{9c}$$

Table 1 shows that the values of the parameter ξ_N^λ obtained by fitting the temperature dependences of the thermal conductivity of silicon (see below) lie in the interval given by estimates (9)–(9c).

The phonon relaxation rate in umklapp processes is expressed as

$$\nu_{phU}^\lambda = B_U^\lambda \omega^2 T \exp\left(-\frac{C_\lambda}{T}\right), \tag{10}$$

where $C_\lambda = \Theta^\lambda/\alpha^\lambda$. Taking into account that the parameter α^λ is almost identical for germanium and silicon crystals [6, 25] ($\alpha^T = 2.15$ and $\alpha^L = 1.85$) and the Debye temperatures are $\Theta^L = 570$ K and $\Theta^T = 210$ K [6], we obtain $C^L = 308$ K and $C^T = 98$ K for silicon. The variation in the parameter B_U can roughly be estimated as follows: two of the three phonons taking part in umklapp processes have energies on the order of the Debye temperature $\Theta^{(i)}$. Therefore, $A_u^{(i)} \propto (\Theta^{(i)})^{-2}$ [1]. This relation yields the following estimates:

$$A_U^\lambda(\mathrm{Si}) = \xi_U^\lambda A_U^\lambda(\mathrm{Ge}), \quad \xi_U^\lambda \approx (\Theta_{\mathrm{Ge}}^\lambda/\Theta_{\mathrm{Si}}^\lambda)^2, \qquad \xi_U^L \approx 0.34, \quad \xi_U^T \approx 0.32. \tag{11}$$

Parameter A_u can also be estimated in a different way. At sufficiently high temperatures $T \sim$ (300–400) K, the dominant contribution to the phonon momentum relaxation is made by umklapp phonon–phonon processes. In this case, from the experimental data on the thermal conductivity of germanium and silicon [6, 25], one obtains $\xi_U \approx 0.42$. The analytic expression $\kappa \propto M\Theta_D^3 V_0^{1/3}/\gamma^2 T$ [16] for the thermal conductivity at high temperatures implies that

$$\xi_U \approx \frac{\kappa(\mathrm{Ge})}{\kappa(\mathrm{Si})} \approx \frac{M_{\mathrm{Ge}}}{M_{\mathrm{Si}}}\left(\frac{\Theta_D(\mathrm{Ge})}{\Theta_D(\mathrm{Si})}\right)^3 \frac{a_0^{\mathrm{Ge}}}{a_0^{\mathrm{Si}}} \approx 0.5. \tag{12}$$

One can see from Table 1 that the values of ξ_U^λ obtained from fitting the temperature dependence of the thermal conductivity of silicon [6, 25] lie within the interval 0.3–0.5 and virtually coincide with the estimate obtained from the high-temperature data on the thermal conductivity of germanium and silicon with the natural isotope compositions [6, 25].

As shown in [13], it is more convenient to use the parameters determining the relaxation rates in s^{-1}. Table 2 presents the calculated constants for germanium and silicon.

For the values of the fitting parameters presented in Table 1, the relaxation rate of transverse phonons in normal process is three orders of magnitude lower than that for longitudinal phonons in silicon crystals:

$$\nu_{phN}^T[\mathrm{s}^{-1}] \approx 2.6\times10^2\left(\frac{T}{10}\right)^5 z_T, \qquad \nu_{phN}^L[\mathrm{s}^{-1}] \approx 4.1\times10^5\left(\frac{T}{10}\right)^5 z_L^2. \tag{13}$$

In the case of scattering by isotope disorder, the relaxation rate of transverse phonons is nearly six times higher than that for longitudinal phonons ($S_*^3 \approx 6.2$):

$$\nu_{phi}^T[\mathrm{s}^{-1}] \approx 35.7\times10^6 g\left(\frac{T}{10}\right)^4 z_T^4, \qquad \nu_{phi}^L[\mathrm{s}^{-1}] \approx 5.8\times10^6 g\left(\frac{T}{10}\right)^4 z_L^4; \tag{14}$$

here, $g = 2.3\times10^{-6}$ for ^{28}Si (99.86%) and $g = 2.01\times10^{-4}$ for silicon of natural composition. In the case of scattering by the boundaries of a sample, we have

$$\nu_{phB}^\lambda[\mathrm{s}^{-1}] = \frac{s_\lambda}{L_C}\left\{\frac{1-P}{1+P} + \frac{L_C}{l}\right\} = C_{B\lambda}\times10^6, \qquad C_{BL} = C_{BT}S_*, \tag{15}$$

where L_C is the Casimir length, l is the sample length, and P is the probability of mirror reflection of phonons. The value of ν_{phB}^L in the silicon crystals under investigation was $(1–2)\times10^6$ s^{-1} [6]. Using the parameters of

Table 1, we obtain the following expression for the phonon relaxation rate in umklapp processes:

$$\nu_{phU}^{\lambda}[\mathrm{s}^{-1}] = A_{\lambda}\left(\frac{T}{10}\right)^3 \exp\left(-\frac{C_{\lambda}}{T}\right) z_{\lambda}^2. \qquad (16)$$

One can easily verify that the inequality $\nu_{phN}^{T}(q) \lll \nu_{phR}^{T}(q)$ holds for T phonons in the entire temperature range and the contribution of these phonons to the thermal conductivity is primarily determined by the diffusion motion. However, for L phonons, the ratio

$$\frac{\nu_{phN}^{L}}{\nu_{phR}^{L}} = \frac{0.41\left(\frac{T}{10}\right)^5 z_L^2}{C_{BL} + \frac{35.7}{S_*^3} g\left(\frac{T}{10}\right)^4 z_L^4 + 3.77\left(\frac{T}{10}\right)^3 \exp\left(-\frac{308}{T}\right) z_L^2} \qquad (17)$$

is greater than unity at $T > 13$ K and even much greater than unity in the temperature interval $20 < T < 100$ K. Therefore, the drift motion of phonons in isotopically enriched crystals of silicon largely determines the contribution of L phonons to the thermal conductivity, which, with regard to the above inequalities, can be represented as

$$\kappa^{L}(T) = \frac{1}{3} C_{VL} s_L^2 / \langle \nu_{phR} \rangle_L,$$
$$\langle \nu_{phR} \rangle_L = \{ \langle \nu_{phR}^{L} \rangle + 2S_*^5 \langle \nu_{phN}^{T} \rangle \} / J_L^{(4)}. \qquad (18)$$

Formulas (18) show that not only resistive scattering processes of L phonons but also normal scattering processes of T phonons, whose role is considerably increased due to the factor S_*^5, make a contribution to the effective relaxation rate of L phonons for the Herring relaxation mechanism (in contrast to the generalized Callaway model [1–12]). This fact results in a considerable suppression of the drift motion of L phonons and a decrease in the maximum values of thermal conductivity for the Herring relaxation mechanism in isotopically pure silicon crystals.

Next, we demonstrate that the procedure described for determining anharmonic parameters of silicon provides a satisfactory description of the experimental data for natural silicon and enriched silicon with 99.8588% ^{28}Si isotope concentration (SI284 [6]).

4. CALCULATION OF THE THERMAL CONDUCTIVITY OF SILICON CRYSTALS WITH VARIOUS ISOTOPE COMPOSITIONS

Below, we present the results of calculating the thermal conductivity $\kappa(T)$ of silicon samples with various isotope compositions for two variants of phonon momentum relaxation in normal processes (formulas (14) and (15)). The results were fitted by varying the parameters of anharmonic relaxation processes to obtain the best agreement between the calculated values of $\kappa(T)$ and experimental data both near the maximum of $\kappa(T)$ and at lower temperatures. The calculated data obtained with the use of the results of [1, 2, 6] are presented in Tables 1 and 2.

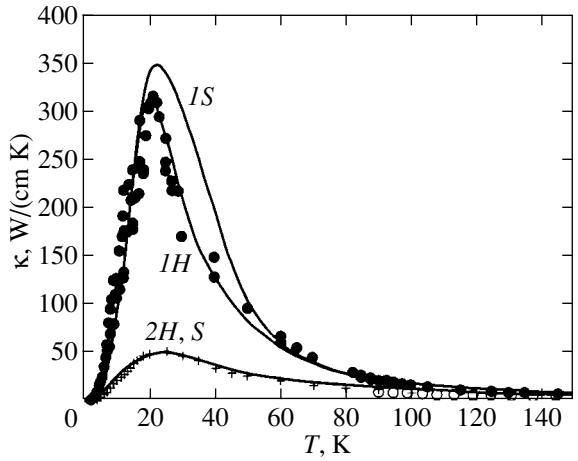


Fig. 1. Total thermal conductivity of silicon crystals with various isotope compositions versus temperature: (*1*) with 99.86% ^{28}Si isotope concentration ($C_{BL} = 1.2$), (*2*) with the natural isotope composition ($C_{BL} = 1.8$) for two mechanisms of momentum relaxation in normal phonon scattering processes (*H* corresponds to the Herring mechanism and *S* corresponds to the generalized Callaway model). Symbols represent the experimental data from [2, 6].

Figure 1 displays the thermal conductivity versus temperature calculated for silicon crystals with 99.8588% ^{28}Si isotope concentration (SI284 [6]) (curves *1H* and *1S*) and for silicon with the natural isotope composition [25] for two variants of phonon momentum relaxation in normal processes (curves *2H*, *S*): the Herring mechanism (*H*) and the generalized Callaway model (*S*). One can see that the difference between the calculated results for the two variants of phonon relaxation in natural silicon crystals is negligible. For ^{28}Si (99.86%), there is a small difference (within experimental error) between the positions of the maxima of thermal conductivity for these variants [6]. The maximum values $\kappa_{max}(T_{max})$ of thermal conductivity for the generalized Callaway model are appreciably greater (by 11%) than those for the Herring mechanism. However, by varying the parameters related to the anharmonicity of lattice oscillations (A_u^{λ} and A_N^{λ}), one can fit the calculated values of thermal conductivity to the measured values. Thus, the experimental results of [6] do not allow one to make a definite conclusion in favor of one or another model, although it has been generally recognized that the dominant relaxation mechanism of

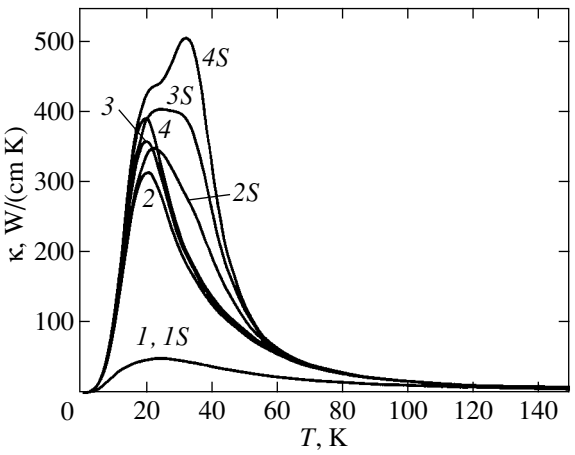


Fig. 2. Total thermal conductivity of silicon crystals versus temperature (curves *1–4*) for the Herring mechanism and (curves *1S*, *2S*, *3S*, and *4S*) for the generalized Callaway model. Curves *1* and *1S* correspond to silicon with the natural isotope composition and $g = 201 \times 10^{-6}$, curves *2* and *2S* correspond to SI284 [6] and $g = 2.33 \times 10^{-6}$, curves *3* and *3S* correspond to $g = 7 \times 10^{-7}$, and curves *4* and *4S* correspond to monoisotopic ^{28}Si and $g = 0$.

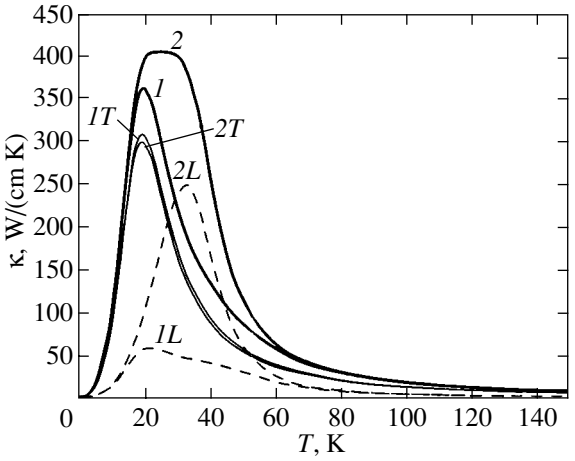


Fig. 3. Total thermal conductivity (curves *1* and *2*) and contributions of *L* and *T* phonons versus temperature for a silicon crystal ($C_{BL} = 1.2$) for $g = 7 \times 10^{-7}$ for two variants of phonon momentum relaxation in normal processes: (curve *1*) the Herring mechanism and (curve *2*) the generalized Callaway model.

thermal phonons in both germanium and silicon crystals is the Herring mechanism [1, 15–18]. It is the analysis of experimental [2] and calculated [13] temperature dependences of thermal conductivity for two variants of phonon relaxation in highly enriched ^{70}Ge (99.99%) that has allowed us to make a definite conclusion that the Herring mechanism in germanium is not only the dominant one but also an efficient factor that limits the maximal values of thermal conductivity in isotopically pure germanium crystals.

The situation is qualitatively changed as the degree of isotope disorder decreases further. Figure 2 shows the calculated values of κ for silicon crystals with different concentrations of the ^{28}Si isotope. One can see that the two variants of phonon momentum relaxation in normal processes give qualitatively different results for isotopically enriched silicon crystals. When $g < 10^{-6}$, the contribution of *L* phonons to the maximal value of the total thermal conductivity κ_{max} in the generalized Callaway model sharply increases as the isotope disorder degree *g* decreases, and, for $g < 7 \times 10^{-7}$, κ_{max} is predominantly determined by *L* phonons. For the Herring mechanism, the maximum thermal conductivity is predominantly determined by *T* phonons in the entire range of variation of *g* ($g < 10^{-6}$). The momentum redistribution between *L* and *T* phonons in normal Herring processes leads to a considerable suppression of the drift motion of *L* phonons and, accordingly, their contribution to $\kappa(T)$. Thus, the normal Herring phonon–phonon scattering processes represent an efficient mechanism that limits the maximum thermal conductivity in isotopically enriched crystals of silicon. The two variants of phonon relaxation in normal scattering processes for $g < 10^{-6}$ exhibit qualitatively different behaviors of $\kappa(T)$ in the neighborhood of the maximum: an increase in the contribution of *L* phonons near the maximum of $\kappa(T)$ changes the shape of the function $\kappa(T)$. For $g \approx 7 \times 10^{-7}$, the contributions of *L* and *T* phonons become approximately equal, and the function $\kappa(T)$ has a broad flat top for the generalized Callaway model (curve *3S*). As $g \longrightarrow 0$, κ_{max} in the generalized Callaway model is mainly attributed to *L* phonons (curve *4S*), while the temperature T_{max} is shifted by more than 10 K as compared with the Herring variant (curve *4*). This difference between the behaviors of $\kappa(T)$ in the neighborhood of maxima in the two variants of phonon relaxation in normal scattering processes can be verified experimentally.

Figure 3 represents the calculated thermal conductivity and the contributions of *L* and *T* phonons to $\kappa(T)$ for silicon crystals with $g = 7 \times 10^{-7}$ for the generalized Callaway model (curves *2*, *2L*, and *2T*) and for the Herring mechanism (curves *1*, *1L*, and *1T*). One can see that the contributions of transverse phonons to κ for two variants of phonon relaxation in normal processes differ insignificantly (curves *1T* and *2T*) because these contributions are mainly determined by the diffusion motion of phonons. However, the contribution of *L* phonons is determined by the drift motion, and, for the generalized Callaway model (curve *2L*), this contribution is about 4.5 times greater than that for the Herring mechanism (curve *1L*).

The positions of the maxima of the thermal conductivity, $T_{max}(g)$, also differ significantly for the two variants of phonon relaxation in normal scattering processes in isotopically enriched crystals (see Fig. 4). For the Herring mechanism, T_{max} decreases from about

22–23 K for silicon with the natural composition to 19 K for isotopically pure silicon. For the generalized Callaway model, a decrease in the isotope disorder g first leads to a decrease in T_{max} (for silicon with the natural composition, $T_{max} \approx 22$ K for $g \approx 10^{-5}$) and then sharply increases as g decreases to $g \approx 7 \times 10^{-7}$. Finally, for isotopically pure silicon ($g = 0$), we have $T_{max} \approx 32$ K. Such a behavior of $T_{max}(g)$ for the generalized Callaway model is associated with the fact that, instead of T phonons ($T^T_{max}(g)$), L phonons ($T^L_{max}(g)$) start to play the dominant role in thermal conductivity; the thermal conductivity associated with the latter phonons attains its maximum at higher temperatures (see curves *2L* and *1T*). This difference in the behavior of $T_{max}(g)$ for two variants of phonon relaxation in normal scattering processes admits experimental verification.

The analysis of the contributions of T phonons to the total thermal conductivity of silicon crystals with different degrees of isotope disorder has shown that (see Fig. 5) L phonons make the dominant contribution to the thermal conductivity of silicon crystals with the natural isotope composition (curves *1H* and *1S*). In the range of temperatures from 15 to 60 K, this contributions is greater than 80% of the total thermal conductivity. A decrease in the degree of isotope disorder leads to an increase in the drift velocity of longitudinal phonons and, consequently, to an increase in their contribution to the thermal conductivity. For isotopically enriched crystals of ^{28}Si (99.86%) (curves *2H* and *2S*), the relative contribution of T phonons to $\kappa(T)$ decreases for both models: up to 67% at 40 K for the Herring mechanism and up to 44% at 38 K for the generalized Callaway model. For isotopically pure ^{28}Si ($g = 0$) (curves *3H* and *3S*), the contribution of T phonons to $\kappa(T)$ for the Herring mechanism decreases insignificantly (by 2%) as compared with ^{28}Si (99.86%) and amounts to 65% of the total thermal conductivity. For the generalized Callaway model, the contribution of T phonons decreases in this case to 25% of the total thermal conductivity.

Now, let us consider the ratio $\kappa_2^\lambda/\kappa_1^\lambda$ of the drift and diffusion contributions to the thermal conductivity of silicon with various degrees of isotope disorder for both branches of the phonon spectrum. Figure 6 shows that, for the Herring mechanism, the contribution of the drift motion of L phonons to the thermal conductivity of ^{28}Si (99.86%) near the maximum decreases by about 2.5 times as compared with the result obtained for the Callaway model. However, it is two orders of magnitude greater than the contribution of the diffusion motion of L phonons in this case too. Note that the contribution of the drift motion of L phonons to the thermal conductivity of silicon with 99.86% ^{28}Si isotope concentration for the Herring mechanism attains its maximum at a temperature of 52 K; in this case, it is two orders of magnitude greater than the diffusion contribution. On the other hand, the contribution of the drift motion of T phonons near the maximum thermal conductivity amounts to about 1% of the total thermal conductivity κ^T for the generalized Callaway model and about 4% for the Herring mechanism. For isotopically pure ^{28}Si (curves *3S* and *3H*), the drift motion of T phonons makes a small contribution to the thermal conductivity for both variants of phonon relaxation in normal processes. However, for the Herring mechanism

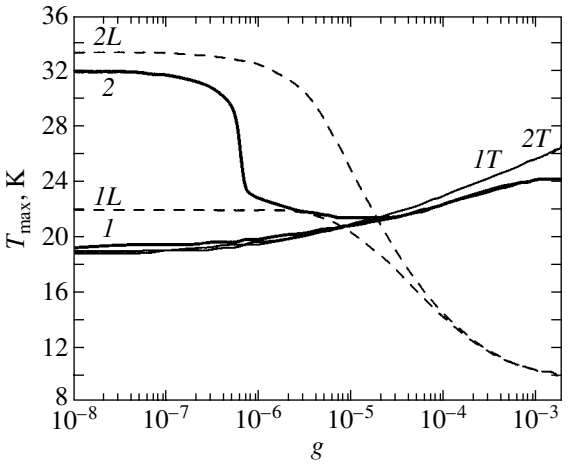


Fig. 4. Temperature T_{max} of the maximum total thermal conductivity versus the degree of isotope disorder g for two variants of phonon momentum relaxation in normal scattering processes (curve *1* corresponds to the Herring mechanism and curve *2* to the generalized Callaway model) and the contributions of L and T phonons.

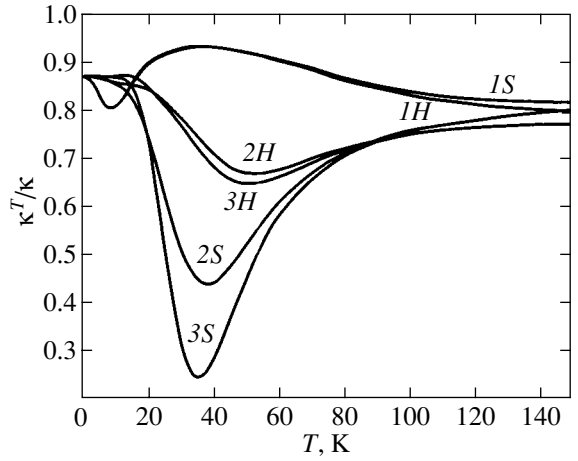


Fig. 5. The ratio of the contributions of T phonons to the total thermal conductivity for silicon samples with various degrees of isotope disorder (curve *1* corresponds to silicon with the natural isotope composition, curve *2* to ^{28}Si(99.86%), and curve *3* to monoisotopic ^{28}Si) calculated for (*H*) the Herring model and (*S*) the generalized Callaway model.

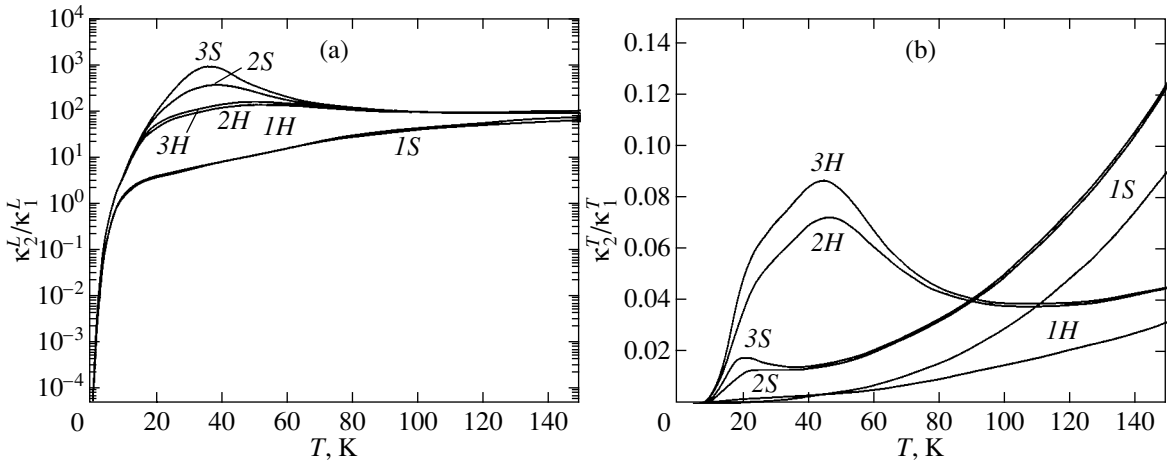


Fig. 6. The ratio $\kappa_2^\lambda/\kappa_1^\lambda$ of the drift and diffusion contributions to the thermal conductivity versus temperature for (a) *L* and (b) *T* phonons in silicon crystals with various degrees of isotope disorder: curve *1* corresponds to silicon with the natural isotope composition, curve *2* corresponds to ^{28}Si (99.86%), and curve *3* corresponds to monoisotopic silicon ^{28}Si; *S* and *H* indicate the generalized Callaway and Herring models, respectively.

near the maximum, this contribution is three times higher, which is attributed to the transfer of the drift momentum from longitudinal to transverse phonons in the Herring normal processes. On the other hand, the contribution of the drift motion of *L* phonons to the thermal conductivity of ^{28}Si crystals is three orders of magnitude higher than that of the diffusion motion for the generalized Callaway model. The redistribution of the drift momentum from *L* to *T* phonons in normal scattering processes reduces this contribution by nearly an order of magnitude, although it is still greater than the contribution of the diffusion motion by two orders of magnitude. As the isotope disorder increases, the ratio $\kappa_2^\lambda/\kappa_1^\lambda$ decreases for both branches of the phonon spectrum, and the difference between the generalized Callaway model and the Herring mechanism for Si with the natural isotope composition becomes negligible.

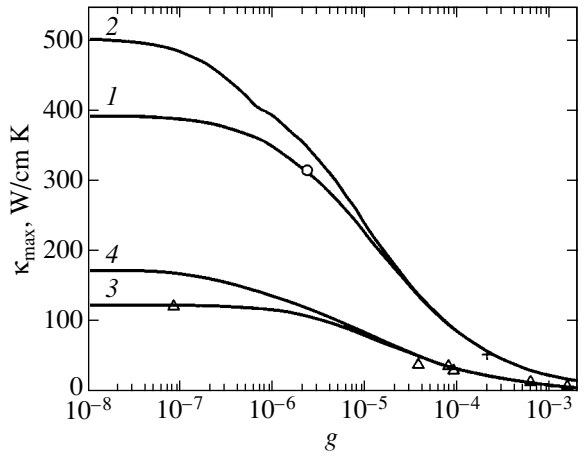


Fig. 7. Maximum values of thermal conductivity κ_{max} versus the degree of isotope disorder *g* for (curves *1* and *2*) silicon and (curves *3* and *4*) germanium crystals for the two variants of phonon momentum relaxation in normal scattering processes; curves *1* and *3* correspond to the Herring model, and curves *2* and *4*, to the generalized Callaway model. Symbols represent experimental results: ○ corresponds to ^{28}Si (99.86%) [6], + corresponds to silicon with the natural composition [25], and △ correspond to germanium crystals with various isotope compositions [2, 5].

Figure 7 illustrates the temperature dependence of the maximum thermal conductivity κ_{max} on the degree of isotope disorder *g* for the two variants of phonon momentum relaxation in normal scattering processes for the same values of boundary scattering as in [6] (C_{BL} = 1.2). One can see that these graphs are essentially different for small values of *g*. For the Herring mechanism, $\kappa_{max}(g)$ virtually attains saturation (curve *1*) at the values of *g* less than 7×10^{-7} (which corresponds to 99.96% ^{28}Si isotope concentration). For the generalized Callaway model (curve *2*), κ_{max} continues to increase for values of *g* less than 7×10^{-7}. According to the experimental data of [6] for ^{28}Si (99.86%) and the estimates given in the present paper, the maximum thermal conductivity in monoisotopic ^{28}Si for the Herring mechanism is $\kappa_{max}(g = 0) \approx 393$ W/cm K; i.e., one can increase the thermal conductivity by 25% as compared with ^{28}Si (99.86%). For $g = 10^{-7}$ (which corresponds to 99.99% ^{28}Si isotope concentration), the maximum thermal conductivity is only 1.4% less than $\kappa_{max}(g = 0)$, whereas, for $g = 7 \times 10^{-7}$, κ_{max} is less than $\kappa_{max}(g = 0)$ by 8.6%. Therefore, the limit of isotopic enrichment of ^{28}Si (99.99%) can be regarded as the optimal value for obtaining the maximum thermal conductivity. It is obvious that the normal Herring phonon–phonon scat-

tering processes represent an efficient mechanism that limits the maximum thermal conductivity of isotopically enriched silicon crystals.

5. CONCLUSION

We have carried out a detailed analysis of the lattice thermal conductivity of silicon crystals with various isotope compositions. Based on the experimental data for silicon with the natural isotope composition and ^{28}Si (99.86%), we have determined the parameters responsible for the phonon momentum relaxation in anharmonic phonon–phonon scattering processes. We have considered two variants of phonon momentum relaxation in normal phonon–phonon scattering processes. The calculated values of the thermal conductivity for silicon with the natural isotope composition and for ^{28}Si (99.86%) for the Herring relaxation mechanism are in a good quantitative agreement with the experimental data. We have demonstrated that, for a sample of ^{28}Si (99.86%) ($g = 2.33 \times 10^{-6}$), the degree of isotope disorder ($g = 7 \times 10^{-7}$) is yet higher than that necessary for the two variants of phonon relaxation in normal processes to give qualitatively different results.

The analysis has shown that the thermal conductivities for the generalized Callaway model and our model, based on the specific features of phonon relaxation in the normal Herring processes, are qualitatively different for isotopically enriched crystals of ^{28}Si (99.96%) for $g = 7 \times 10^{-7}$. In this case, the maximum thermal conductivity κ_{max} is mainly determined by T phonons for the Herring mechanism, whereas, for the generalized Callaway model, it is predominantly determined by L-phonons. In this case, the functions $\kappa(T)$ in the neighborhood of the maximum are qualitatively different for the two variants of phonon relaxation in normal processes for $g < 10^{-6}$, while the positions of the maxima of the thermal conductivity for the generalized Callaway model are shifted by about 10 K to higher temperatures with respect to the temperatures obtained by the Herring mechanism. Such differences in the behavior of $T_{max}(g)$ and $\kappa(T)$ near the maximum for the two variants of phonon relaxation in normal scattering processes admit experimental verification.

We have also calculated the maximum thermal conductivity $\kappa_{max}(g)$ as a function of the isotope disorder parameter g for the two variants of phonon momentum relaxation in normal processes. For monoisotopic ^{28}Si with the same boundary scattering parameters as ^{28}Si (99.86%) [6], we obtained $\kappa_{max}(g = 0) \approx 393$ W/cm K and $T_{max} \approx 19.2$ K for the Herring model and $\kappa_{max}(g = 0) \approx 505$ W/cm K and $T_{max} \approx 32$ K for the generalized Callaway model. Since the dominant mechanism of normal scattering processes in silicon crystals is the Herring mechanism, based on the results of the present analysis, we can assume that the maximum thermal conductivity of silicon can be increased by 25% as compared with that achieved for ^{28}Si (99.86%) [6]. It should be noted that the function $\kappa_{max}(g)$ attains saturation for $g \approx (1–2) \times 10^{-7}$, and the values of $\kappa_{max}(g)$ are only (1–2)% less than $\kappa_{max}(g = 0)$. Therefore, by enriching silicon from 99.86% to 99.99%, one can increase the maximum thermal conductivity of silicon by 24% for samples of the same sizes and the same surface treatment. Note that our estimates apply to chemically pure perfect silicon crystals. The presence of impurities, especially electrically charged ones, may substantially reduce κ_{max}.

ACKNOWLEDGMENTS

We are grateful to T. Ruf for kindly presenting experimental data, A.V. Inyushkin for discussing the problems considered in this paper, and A.P. Tankeev for discussing the results of this study.

This work was supported by the Russian Foundation for Basic Research, project no. 00-02-16299.

REFERENCES

1. A. P. Zhernov and A. V. Inyushkin, *Isotope Effects in Solids* (Ross. Nauchn. Tsentr "Kurchatovskiĭ Institut," Moscow, 2001).
2. M. Asen-Palmer, K. Bartkowski, E. Gmelin, *et al.*, Phys. Rev. B **56**, 9431 (1997).
3. V. I. Ozhogin, A. V. Inyushkin, A. N. Taldenkov, *et al.*, Pis'ma Zh. Éksp. Teor. Fiz. **63**, 463 (1996) [JETP Lett. **63**, 490 (1969)].
4. A. N. Taldenkov, A. V. Inyushkin, V. I. Ozhogin, *et al.*, in *Proceedings of the IV Conference "Physicochemical Processes under Atomic and Molecular Selection,"* Zvenigorod 1999, Nauka, Moscow (1999).
5. T. H. Geballe and G. W. Hull, Phys. Rev. **110**, 1773 (1958).
6. T. Ruf, R. W. Henn, M. Asen-Palmer, *et al.*, Solid State Commun. **115**, 243 (2000).
7. R. Berman, Phys. Rev. B **45**, 5726 (1992).
8. W. S. Capinski, H. J. Maris, E. Bauser, *et al.*, Appl. Phys. Lett. **71**, 2109 (1997).
9. Lanhua Wei, P. K. Kuo, R. L. Thomas, *et al.*, Phys. Rev. Lett. **70**, 3764 (1993).
10. J. E. Graebner, M. E. Reiss, L. Seibles, *et al.*, Phys. Rev. B **50**, 3702 (1994).
11. J. R. Olson, R. O. Pohl, J. W. Vandersande, *et al.*, Phys. Rev. B **47**, 14 850 (1993).
12. A. P. Zhernov and D. A. Zhernov, Zh. Éksp. Teor. Fiz. **114**, 1757 (1998) [JETP **87**, 952 (1998)]; A. P. Zhernov, Fiz. Tverd. Tela (St. Petersburg) **41**, 1185 (1999) [Phys. Solid State **41**, 1079 (1999)].
13. I. G. Kuleev and I. I. Kuleev, Zh. Éksp. Teor. Fiz. **120**, 649 (2001) [JETP **93**, 568 (2001)].
14. J. Callaway, Phys. Rev. **113**, 1046 (1959).
15. R. Berman, *Thermal Conduction in Solids* (Clarendon, Oxford, 1976; Mir, Moscow, 1979).

16. B. M. Mogilevskiĭ and A. F. Chudnovskiĭ, *Thermal Conductivity of Semiconductors* (Nauka, Moscow, 1972),
17. V. S. Oskotskiĭ and I. A. Smirnov, *Defects in Crystals and Thermal Conductivity* (Nauka, Leningrad, 1972), p. 205.
18. B. H. Armstrong, Phys. Rev. B **32**, 3381 (1985).
19. J. A. Krumhansl, Proc. Phys. Soc. London **85**, 921 (1965).
20. M. G. Holland, Phys. Rev. **132**, 2461 (1963).
21. K. Itoh, *Low Temperature Carrier Transport Properties in Isotopically Controlled Germanium*, PhD Thesis (University of California, Berkeley, 1994).
22. I. G. Kuleev, Fiz. Tverd. Tela (St. Petersburg) **44**, 215 (2002) [Phys. Solid State **44**, 223 (2002)].
23. C. Herring, Phys. Rev. **95**, 954 (1954).
24. S. Simons, Proc. Phys. Soc. London **82**, 401 (1963); **83**, 749 (1964).
25. G. A. Slack and C. J. Glassbrenner, Phys. Rev. **120**, 782 (1960).
26. B. Truel, C. Elbaum, and B. B. Chick, *Ultrasonic Methods in Solid State Physics* (Academic, New York, 1969; Mir, Moscow, 1972).
27. G. Nilsson and G. Nelin, Phys. Rev. B **6**, 3777 (1972).
28. J. P. Srivastava, J. Phys. Chem. Solids **41**, 357 (1980).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 491–501.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 570–581.

SOLIDS
Structure

Rashba Splitting in MIS Structures HgCdTe

V. F. Radantsev[a, *] and A. M. Yafyasov[b]

[a]*Ural State University, pr. Lenina 51, Yekaterinburg, 620083 Russia*
**e-mail: victor.radantsev@usu.ru*
[b]*St. Petersburg State University, St. Petersburg, 198504 Russia*

Received February 22, 2002

Abstract—The measured parameters of spin–orbit spectral splitting in HgCdTe-based MIS structures with positive and negative Kane gap E_g are compared with the parameters calculated using the three- and four-band Kane model. The disregard of the finite spin–orbit splitting Δ of the valence band in calculations leads to exaggerated values of Rashba splitting (especially for $E_g < 0$) even for small ratios $|E_g|/\Delta$, although the subband parameters averaged over two spin branches of the spectrum in the two-, three-, and four-band Kane approximations for the same concentrations are practically identical. In the zero-gap HgCdTe, the measured as well as calculated values are noticeably higher, but the four-band approximation leads to values of splitting for both materials which are 20–40% lower than the experimental value. The inclusion of the interband interaction reduces these discrepancies, but does not eliminate them completely. It is shown that the approximations of the 2D spectrum with spin–orbit splitting linear in quasimomentum, which are conventionally used in the analysis, may lower the effective Rashba parameter by a factor of 2–4. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

A strong increase in the interest in theoretical [1] and experimental [2] investigations of the spin–orbit splitting of the 2D spectrum in asymmetric quantum wells (Rashba effect) during the last 2–3 years is due to widespread discussions of the ideas of creating a quantum computer and a spin analog of the field transistor proposed by Datta and Das [3] on the basis of the Rashba effect (it should be noted, however, that theoretical investigations of the problem have been carried out for more than forty years [4–6], while the first purposeful experiments were conducted in 1989–1990 [7–9]). In the simplest phenomenological Rashba model, the quadratic 2D spectrum of asymmetric quantum wells is supplemented with an additional term linear in the 2D wave vector k with the Rashba parameter α:

$$E^{\pm} = \frac{\hbar^2 k^2}{2m^*} \pm \alpha k. \qquad (1)$$

If the splitting $\Delta_R = E^+ - E^-$ corresponds to the precession frequency $\omega_R = \Delta_R/\hbar$, the polarization vector rotates through the angle $\theta = \omega_R L/v = \Delta_R L/\nabla_k E$ during the motion of an electron with velocity $v = \nabla_k E/\hbar$ in a channel of length L. It should be noted that we are speaking of the circular ("chiral") polarization since the Rashba subbands, being mixtures of "spin-up" and "spin-down" states, are not polarized in the conventional sense even in the one-band analysis, and additional complications appear in the case of Kane semiconductors we are interested in. We shall not touch upon this aspect of the problem nor the problem of "spin-selective" sink and shall source (which were considered in [10–12]) and shall confine our analysis to the Rashba splitting proper. In accordance with Eq. (1), we have $\Delta E = 2\alpha k$ and $\theta \approx 2\alpha m^* L/\hbar^2$; i.e., these quantities are determined not by the energy splitting proper, but by the Rashba parameter α (this statement formulated in [3] is valid, however, only in the simplest case of a parabolic subband (1); see below), which is normally used as a measure of the effect.

Being a relativistic-type effect, Rashba splitting is ultimately due to band mixing by an electrostatic surface potential confining electrons at the boundary, which plays a dual role in the problem under investigation. On the one hand, it leads to size quantization and, as a result, to "two-dimensionalization" of the electron spectrum; on the other hand, being asymmetric and perpendicular to the 2D quasimomentum, it leads to spin–orbit splitting. In the framework of the two-band model taking into account only the interaction of bands Γ_6 and the light branch Γ_8 in the lowest (third) order of perturbation theory, $\alpha = \langle \alpha_m dV/dz \rangle$, where we have singled out the factor

$$\alpha_m = P^2/3E_g^2, \qquad (2)$$

determined by material parameters, viz., the Kane gap E_g and the Kane matrix element P which are virtually identical for all Kane semiconductors. The quantity α (which is a phenomenological parameter in the one-band model) is determined both by band parameters and by the form of potential $V(z)$ and the wave function (for a symmetric potential, $\langle \alpha_m dV/dz \rangle = 0$ and Rashba splitting is absent). The implementation of the idea of the spin transistor requires a high degree of splitting ensuring the precession angle $\theta \approx \pi$ over the spin coher-

ence length as well as the possibility of modulation $\delta\theta \approx \pi$ by an external electric field.

The results of different studies (asymmetric quantum wells in the InGaAs system were mainly investigated [13, 14]; in recent years, publications have appeared on MIS structure of the InAs type [15]) differ not only in the values of α, but also (which is most significant) in its dependence on the applied external field. In [13], a decrease in α upon an increase in the positive bias voltage at the field electrode (upon an increase in the subband concentrations n_i) was observed, while the authors of [14, 15] observed the opposite behavior. On the other hand, it was proved earlier in [7] both experimentally and in the framework of a semiphenomenological analysis that, although the splitting Δ_R in surface quantum wells based on narrow-gap materials increases with the electric field, parameter α does not change in a typical experimental situation and has the universal value $\alpha \approx e^2/\varepsilon$, which does not depend on the depth of the well or on the band parameters of the material. Independence of α of the external electric field was also observed for InAs/AlSb quantum wells [16]. As regards the theory, the estimates available for InGaAs quantum wells lead to values of α which are 2–4 times higher than the experimental values.

It will be proved below that, to a certain extent, the above discrepancy can be due to the fact that the experimentally measured quantity is not α, but the splitting $\Delta k_F = \sqrt{4\pi}(\sqrt{n_-} - \sqrt{n_+})$ of the Fermi "surfaces" ($n_\pm$ are 2D concentrations in two "spin" subsubbands of the size quantization subband, which are determined directly from magneto-oscillation effects) or the energy splitting Δ_R (measured by optical methods and using weak antilocalization effects). However, the value of the Rashba parameter depends considerably on the models used for describing the spectrum in 2D subbands. Almost in all publications, the analysis is carried out on the basis of either the parabolic approximation (1) or the subband dispersion relation of the Kane type, but with Rashba splitting linear in k as in relation (1):

$$E = \sqrt{(s_i\hbar k)^2 + (m_i s_i^2)^2} - m_i s_i^2 \pm \alpha_i k, \qquad (3)$$

where the parameters approximating the spectrum in the ith 2D subband are the rest mass m_i and the Kane velocity s_i (in surface quantum wells based on narrow-gap semiconductors, several size-quantization subbands are filled, as a rule, even for insignificant band bendings). It should be emphasized that the application of the parabolic dispersion relation is inadequate in the given experimental situation since clearly manifested Rashba splitting is observed just for materials with a narrow gap and a strong spin–orbit interaction (in complete agreement with relation (2) derived in the limit $\Delta \longrightarrow \infty$), in which nonparabolicity cannot be ignored.

However, it was shown even in [7] that the Kane Hamiltonian in fact corresponds to the dispersion relation

$$E = \sqrt{(s_i\hbar k)^2 + (m_i s_i^2)^2 \pm 2m_i s_i^2 \alpha_i k} - m_i s_i^2 \qquad (4)$$

with saturated splitting $\Delta_R^{\max} = 2\alpha_i m_i s_i/\hbar$, which was later confirmed by numerical calculations [17, 18]. It will be shown below that, when different approximations are used for the subband dispersion relations, the values of α extracted from the experimental values of splitting,

$$\alpha = \Delta_R/2k\sqrt{1 + [s_i^2\hbar^2k^2 - (\Delta_R/2)^2]/(m_i s_i^2)^2}$$

(for approximation (4)), or from the occupancies of subsubbands, $\alpha = \sqrt{\pi}\hbar^2(\sqrt{n_-} - \sqrt{n_+})/m_i$ (in approximation (3), an additional factor equal approximately to $1 - \pi\hbar^2(n_+ + n_-)/m_i^2 s_i^2$ appears), may differ by a factor of several units.

In order to eliminate the above contradictions and to find the materials, structures, and conditions for which the Rashba effect is manifested most clearly, further experimental and theoretical investigations are required. In this work, we report on the results of such an investigation in a 2D electron gas of MIS structures based on the narrow-gap semiconductor HgCdTe with direct and inverse band structure, in which the Rashba splitting must be manifested most strongly in view of the small width of the Kane gap, strong spin–orbit interaction, and extreme asymmetry of quantum wells in MIS structures. As regards the comparison with the theory, it is important that the potential distribution (and the spin–orbit splitting which is most sensitive to it) in the surface channels of MIS structures can be reliably calculated in the framework of a self-consistent procedure, while the confining potential in semiconducting heterostructures is not known exactly as a rule.

2. EXPERIMENT

We studied MIS structures with a typical area of 5×10^{-4} cm^2, prepared by anode oxidation on substrates made of the ternary compound $Hg_{1-x}Cd_xTe$ with two compositions (x = 0.135 and 0.195) corresponding to energy gaps E_g = –50 meV (semimetallic sample SM) and E_g = +50 meV (semiconducting sample, SC), which are close to the band-inversion point, but have opposite signs, and with concentrations $N_A - N_D = 7 \times 10^{16}$ cm^{-3} (SM) and $N_A - N_D = 5 \times 10^{16}$ cm^{-3} (SC) of uncompensated acceptors, leading to close values of charge of the depleted layer for the same n_i. We studied magneto-oscillations of differential capacity of the space-charge region in quantizing magnetic fields H up to 6 T for voltages across the field electrode ranging from –1 to 10 V, which corresponds to the concentration range $n_s \sim (0–8) \times 10^{12}$ cm^{-2} of electrons induced

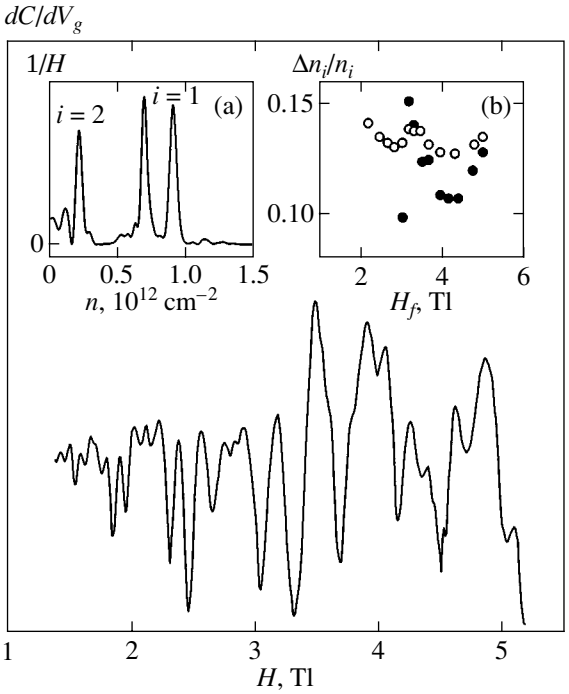


Fig. 1. (a) Magneto-oscillations of capacitance and their Fourier spectrum $1/H$ for the SC sample for $n_s = 7.9 \times 10^{12}$ cm^{-2}. Inset (b) shows the values $\Delta n_i/n_i$ of obtained from the Fourier analysis in the magnetic field interval H_l–H_h for $H_l = 1$ (light circles) and 2.2 T (dark circles). Fourier curves for $i = 0$ correspond to the region $n \approx 2.7 \times 10^{12}$ cm^{-2}.

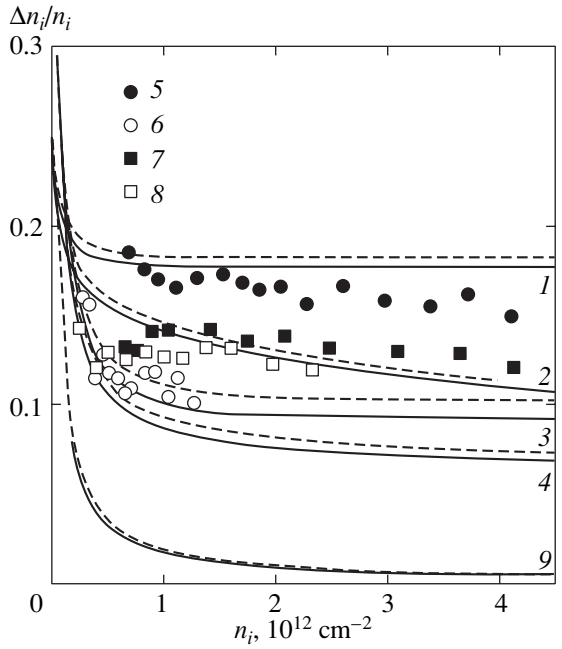


Fig. 2. Experimental (symbols) and calculated (curves) dependences of $\Delta n_i/n_i$ on the subband concentrations n_i in SM (*1, 2, 5, 6*) and SC samples (*3, 4, 7, 8*) calculated for the ground (*5, 7*, and solid curves) and the first excited subbands (*6, 8*, and dashes curves) in the three- (*1, 3*) and four-band (*2, 4*) approximations. Curves *9* present the ratio N_{dep}/n_i for the SC sample.

in the inversion layer. The application of the method of magnetocapacitive spectroscopy which is not critical to the gap width makes it possible to study the features of Rashba splitting in materials with different signs of E_g by using the same experimental approach.

Both types of structures in the subband concentration range $n_i = (0.5\text{–}4.0) \times 10^{12}$ cm^{-2} display clearly manifested oscillation beats for all the three observed size quantization subbands (Fig. 1; all the results in this work correspond to $T = 4.2$ K), indicating the formation of Rashba spin subsubbands and the clearly manifested splitting of the Fourier spectra $1/H$. The latter spectra were used to determine the populations $n_i^{\pm}$ of subsubbands and the "degree of polarization" $\Delta n_i/n_i = (n_i^- - n_i^+)/(n_i^- + n_i^+)$, which is plotted in Fig. 2 for the first two subbands as a function of n_i. This quantity appears as the most suitable characteristic of the magnitude of the effect not only because it can be measured directly in experiments, but also because it provides a description of "polarization" evolution in the spin transistor channel. Indeed, since the value of Δ_R is quite small (third-order effect in perturbation theory), we have $\Delta_R = \nabla_k E \Delta k_F$, and the precession angle $\theta = L\Delta k_F \approx (\Delta n_i/n_i)\sqrt{2\pi n_i}$ is determined only by the "degree of polarization" irrespective of the type of dispersion relation.

It should be noted that, for a small number of beat nodes, the experimental values of $\Delta n_i/n_i$ depend strongly and nonmonotonically on the range of magnetic fields under investigation: variations may reach 50% and higher (see Fig. 1b). This circumstance should be specially emphasized since only one beat was observed in a typical experimental situation in which the Rashba splitting modulated by the field electrode was investigated. Together with the weak splitting of Fourier lines in this publication, this may crucially affect the splitting parameters and (which is most significant) their dependence (and even its type) on the voltage across the field electrode.

In both types of structure, oscillations were observed even for low concentrations down to $n_i \approx 5 \times 10^{10}$ cm^{-2}; however, the small number of oscillations does not allow one to determine reliably the value of $\Delta n_i/n_i$ in this region for the above reason. For the fundamental subband in the sample with the positive gap, for $n_s < 1 \times 10^{12}$ cm^{-2}, oscillations display individual spin components, Fourier frequencies are doubled, and it becomes practically impossible to determine $\Delta n_i/n_i$. In both samples, it is difficult to determine $\Delta n_i/n_i$ for $n_s > (3\text{–}4) \times 10^{12}$ cm^{-2} also in view of strong quenching

of oscillation amplitudes for the low-energy branch of the spectrum (due to large values of cyclotron masses in this subsubband [19]).

The same figure shows the results of calculations made in the framework of the six-band ($\Delta \longrightarrow \infty$) approximation [17]. The theory is in satisfactory agreement both as regards the order of magnitude of $\Delta n_i/n_i$ (the discrepancy does not exceed 40%, while for InGaAs quantum wells the theoretical and experimental values may differ by a factor of several units) and as regards the form of the concentration dependence of this quantity. For the fundamental subband of the SM sample, the theory gives a slightly exaggerated value of the effect (although the discrepancy exceeds the experimental error only slightly), while the calculated values of $\Delta n_i/n_i$ for the SC sample are noticeably lower. The theory leads to slightly higher values of $\Delta n_i/n_i$ for excited subbands as compared to the fundamental subband, while the reverse situation is observed in experiments (especially for $E_g < 0$). It should be noted that, as regards usually measured parameters of 2D subbands not associated with the Rashba splitting (carrier distribution over subbands, values of n_s corresponding to the starting points of subbands, the values of cyclotron mass, etc.), the discrepancy between the theory and experiments lie within the errors. Since the finite value of the spin–orbit splitting of the valence band must primarily be manifested in spin effects, it cannot be ruled out that taking into account the band Γ_7 in the initial Hamiltonian exactly may change significantly the extent of matching to experimental results as regards the spin–orbit splitting.

3. THEORETICAL MODEL

The theoretical description of the spectrum and parameters of subbands is carried out in the framework of the scenario [17] based on the reduction of the initial matrix equation to an equation of the Schrödinger type. In contrast of the methods based on direct numerical integration of the initial matrix equations which do not permit a clear physical interpretation, we can easily single out in such an approach the terms responsible for the effects of nonparabolicity, spin–orbit splitting, and resonant interband mixing and can clearly see specific features of materials with direct and inverse band structure. An important advantage of this method is the possibility of comparison with the results of experimental and theoretical investigations based on an approximate description of the spin–orbit splitting by introducing a phenomenological term of the spin–orbit interaction with the Rashba parameter into the subband equations for effective mass.

An analysis [17] based on the three-band model does not provide the dependence of the Rashba splitting on parameter Δ, while the effect itself is directly caused by the spin–orbit interaction. The correctness of a description ignoring the band Γ_7 (especially spinor-type effects) is most questionable in the case of 2D systems based on relatively wideband semiconductors HgTe and especially InAs, in which the values of E_g and Δ are virtually identical. For large values of the depth of a surface quantum well (as in the case of inversion layers in a strongly doped semiconductor), the correctness of such an approach is also dubious in the case of narrow-gap materials (especially with a negative gap) and at least requires verification.

In zero magnetic field, the inclusion of the Γ_7 band can be easily incorporated in the algorithm of the approach developed in [17]. After solving the standard 8×8 matrix equation for the wave function component pertaining to the Γ_6 band (s electrons in the surface layers of materials with $E_g > 0$) or to the light branch of the Γ_8 band (p electrons in the electron channel of a zero-gap semiconductor) and using the transformation eliminating first-order derivatives, we can reduce the problem to a Schrödinger-type equation which we present in standard form by introducing the bulk effective mass m_b at the bottom of the conduction band (actually, the spectrum is independent of this parameter) and the Kane velocity $s_b = \sqrt{2/3}\,P/\hbar$:

$$\frac{\hbar^2}{2m_b}\frac{d^2\varphi_\pm^{s,p}}{dx^2} + [\mathscr{E} - (U_{KG} + U_{r1}^{s,p} + U_{r2}^{s,p} \pm U_{so}^{s,p})]\varphi_\pm^{s,p} = 0, \tag{5}$$

where the effective energy is given by

$$\mathscr{E} = \frac{E(E+E_g)}{2m_b s_b^2}.$$

In the effective potential, we have singled out the spin-independent term responsible for nonparabolicity effects (an analog of the potential in the Klein–Gordon–Fock equation),

$$U_{KG} = \left(\frac{(2E+E_g)V - V^2}{2m_b s_b^2} - \frac{E(E+E_g)}{2m_b s_b^2}\right) \times \left[1 - \frac{E_\Delta}{3E_\Delta + 2}\right] + \frac{E(E+E_g)}{2m_b s_b^2} + \frac{\hbar^2}{2m_b}k^2 \tag{6}$$

(the last term in the (1 + 2)-dimensional system under investigation can be regarded as an analog of the certrifugal potential in the 3D centrosymmetric problem), and the spinor terms describing the effects of interband mixing by the electrostatic surface potential. In view of degeneracy of the Γ_8 band, the interactions of the Γ_6 band and the light branch of the Γ_8 band with the heavy branch of the Γ_8 band differ significantly, and the spin–orbit term U_{so} and the terms U_{r1} and U_{r2} responsible for the mixing of states in the surface channel and the states in the bulk of the semiconductor by the electric field (in the case of a narrow gap, 2D states are often resonant

states), as well as the dimensionless parameter E_Δ, differ for channels with *s* and *p* electrons. For electron layers with $E_g > 0$, we have

$$U_{so}^s = k\frac{\hbar^2}{4m_b}\frac{V'}{E_+}\left[1 - \frac{E_\Delta(3E_\Delta + 1)}{(E_\Delta + 1)(3E_\Delta + 2)}\right],$$

$$U_{r1}^s = \frac{\hbar^2}{4m_b}\frac{V''}{E_+}\left[1 - \frac{E_\Delta}{(E_\Delta + 1)(3E_\Delta + 2)}\right],$$

$$U_{r2}^s = \frac{3\hbar^2}{8m_b}\left(\frac{V'}{E_+}\right)^2\left[1 - \frac{1}{3}\frac{E_\Delta(4 + 19E_\Delta + 18E_\Delta^2)}{(3E_\Delta + 2)^2(E_\Delta + 1)^2}\right],$$

$$E_\Delta = \frac{E_+}{\Delta},$$

while, for electron channels in materials with $E_g < 0$, we have

$$U_{so}^p = k\frac{\hbar^2}{4m_b}\frac{V'}{E_+}\frac{A_k}{B_k}$$

$$\times\left[1 - \frac{3E_\Delta}{(3E_\Delta + 2)}\left(1 - \frac{2}{3E_\Delta + 2}\frac{B_k}{A_k}\frac{E_+}{E_-}\right)\right],$$

$$U_{r1}^p = \frac{\hbar^2}{4m_b}\frac{V''}{E_+}\frac{A_k}{B_k},$$

$$U_{r2}^p = \frac{\hbar^2}{4m_b}\left(\frac{V'}{E_+}\right)^2\frac{A_k}{B_k}\left(1 + \frac{1}{2}\frac{A_k}{B_k} + \frac{B_{k1}}{B_k} - \frac{A_{k1}}{A_k}\right),$$

where

$$A_k = 1 + \frac{3(s_b\hbar k)^2}{4E_-^2}\left[1 + \left(\frac{3E_\Delta}{3E_\Delta + 2}\right)^2\right],$$

$$B_k = 1 - \frac{3(s_b\hbar k)^2}{4E_+E_-}\left[1 + \frac{3E_\Delta}{3E_\Delta + 2}\right],$$

$$A_{k1} = -\frac{3(s_b\hbar k)^2}{2E_-^2}\frac{E_+}{E_-}\left[1 + \left(\frac{3E_\Delta}{3E_\Delta + 2}\right)^3\right],$$

$$B_{k1} = \frac{3(s_b\hbar k)^2}{4E_-^2}$$

$$\times\left\{\left[1 + \left(\frac{3E_\Delta}{3E_\Delta + 2}\right)^2\right] + \frac{E_-}{E_+}\left[1 + \frac{3E_\Delta}{3E_\Delta + 2}\right]\right\},$$

$$E_\Delta = \frac{E_-}{\Delta}.$$

If the energies are measured from the bottom of the Γ_6 band for *s* electrons and of the Γ_8 band for *p* electrons, we have

$$E_+ = E - V(z) + |E_g|, \quad E_- = E - V(z).$$

For $\Delta \longrightarrow \infty$, the factors in the brackets describing the effects of interaction with the Γ_7 band are equal to unity except for notation, and we arrive at the equations derived in [17]. It can easily be seen, however, that, in the opposite limiting case ($\Delta \longrightarrow 0$), these factors (except the spin–orbit term) do not suffer radical changes. For the resonant terms U_{r1} and U_{r2}, in the case of *s* electrons, these factors are also equal to unity (for the most unfavorable values of $\Delta \approx E_g$, their value differs from unity by less than 15%), while for *p* electrons they increase not more than twofold (the effective potential contains the ratios of the terms containing such factors). For a Klein–Gordon potential, in the limit $\Delta \longrightarrow 0$, the factor depending on Δ is equal to 3/2, which corresponds to the renormalization of mass m_b for $\Delta \longrightarrow 0$ as compared to the case when $\Delta \longrightarrow \infty$. Thus, the inclusion of the contribution from the Γ_7 band should not lead to strong variations of 2D-subband parameters calculated without taking into account the spin–orbit interaction, which is confirmed by numerical calculations (see below).

Radical changes (for moderate values of Δ) as compared to the three-band model occur only in the spin–orbit terms $U_{so}^{s,p}$. For small ratios Δ/E_g, potentials $U_{so}^{s,p}$ are linear in Δ, and Rashba splitting vanishes in the limit $\Delta \longrightarrow 0$.

It can easily be seen that the structure of the spin–orbit term for *s* electrons is similar to (or coincides with in the limit of large E_g) the expression $U_{so}^s = \alpha_m k(dV/dz)$ obtained in the form of a correction to the one-band approximation in [18, 20] with the parameter

$$\alpha_m = \frac{\hbar^2}{2m_b}\frac{1}{E_g}\frac{\Delta(2E_g + \Delta)}{(3E_g + 2\Delta)(\Delta + E_g)} = \frac{P^2}{3E_g^2}\left[1 - \left(\frac{E_g}{E_g + \Delta}\right)^2\right], \quad (7)$$

where we have used (in the last equality) the expression for the effective mass $m_b = 3\hbar^2E_g(E_g + \Delta)/2P^2(3E_g + 2\Delta)$.

4. COMPARISON WITH EXPERIMENT AND DISCUSSION

Numerical self-consistent calculations were carried out using the following two approaches: (1) direct self-consistent integration of the Poisson equation and Eq. (5) in a block with a size considerably exceeding the Debye screening length with zero boundary conditions and (2) in the framework of a semiclassical approach both in the quantization of the spectrum and in the calculation of the surface potential. It was noted above that 2D states in inversion layers of narrow-gap semiconductors often overlap in the spectrum with the states of the valence band in the bulk; i.e. these states are formally resonance states. However, both experiments and numerical calculations [21] indicate their very weak blurring even in the limiting case $|E_g| = 0$.

The physical reason for such a behavior was found in [22]. For any attractive potential $V(r)$ in the Klein–Gordon potential U_{KG}, there exists a potential barrier separating 2D states from 3D states. For an overwhelming majority of electrons in the surface layer (except the states near the bottom for 2D subbands), such a barrier is completely impermeable, which is ultimately due to nonconservation of the transverse quasimomentum during tunneling to the bulk states [22].

When the spinor terms (U_{r1}, U_{r2}, and U_{so}) are taken into account, the effective potential acquires an additional contribution with an infinitely high potential wall on the side of the bulk states, which is associated with a singularity in $1/E_+(z)$ (for p electrons, the singularity which is closer to the surface in the function $1/B_k(z)$ dominates). This apparently corresponds to nonconservation of a certain "spinor" characteristic related to helicity upon a transition to bulk states in the case of Dirac electrons. As a result, the states at the bottom of 2D subbands are also found to be stationary. The contribution from spinor terms narrows the surface potential well in the effective potential and, as a result, leads to a resonant shift of energy levels towards higher energies (as compared to the Klein–Gordon approximation), which is most significant for states near the bottom of 2D subbands. For nonzero values of k, the positions of zeros of $E_+(z)$ and $B_k(z)$ and the width of the well in the effective potential are different for two spin branches due to the contribution of U_{so}, which causes the splitting of the spectrum. As the 2D quasimomentum increases, the Rashba potential U_{so} linear in k increases, but the poles in $1/E_+(z)$ and $1/B_k(z)$ in this case move away from the surface. As a result, the contribution to the effective potential in the surface region of the potential well associated with U_{r1}, U_{r2}, and U_{so} decreases together with the difference in the potentials for two spectral branches (and, hence, the Rashba splitting) starting from certain values of k and disappears in the limit of large k.

Since the effect of remote bands forming the dispersion of heavy holes was disregarded in the initial Hamiltonian of this study (we used the experimental values of masses for heavy holes in the Poisson equation), the resonance mixing of the electron branches of the channel with the states of heavy holes is negligibly small in the framework of the model under investigation. The latter, however, cannot make a noticeable contribution to the spectrum since it is significant only in an extremely small neighborhood of the bottom of 2D subbands [23]. Thus, 2D states are in fact stationary, and the tunnel exchange with the bulk states (the low-frequency nature of the volt-farad characteristics in both types of structures under investigation down to frequencies of ~10 MHz indicate a sufficiently high rate of this exchange) and the broadening of the levels are due to scattering processes rather than the resonance mechanisms.

A comparison of the results of numerical integration and semiclassical quantization is of dual interest. On the one hand, the adequacy of the semiclassical description of the size-quantized spectrum in the surface layers of narrow-gap semiconductors with $E_g > 0$ in the framework of the two-band model was emphasized in many publications (the corresponding arguments and references can be found in [17]). On the other hand, computer time expenditures for direct integration are several orders of magnitude higher, which is practically inadmissible for simulating the magneto-oscillatory effects. It should be noted from the very outset that the results of the two approaches for identical 2D concentrations are quite close, including those in the description of spin effects. The discrepancy between the results obtained using the three-band and four-band approximations does not exceed 1–2% even for the ground subband, which is much smaller than typical experimental errors.

Since the value of the charge of the depleted layer in the inversion layers of narrow- and zero-gap semiconductors is relatively small even in the case of strong doping, the contribution of electrons to the formation of the surface potential is significant even for small populations of 2D subbands. On the other hand, the thicknesses of the inversion and depleted layers in such structures are close in magnitude, and the conventionally used approach in which the surface potential is treated as the sum of the self-consistent potential formed by electrons from the inversion layer and the parabolic potential (often, its linear component) of the depleted layer is inapplicable (especially for calculating the spin splitting, which is most sensitive to the electric field). In our calculations, both contributions were taken into account exactly, including the possible charge exchange in the doping impurity in the surface layer.

If the term describing the spin–orbit interaction is disregarded or (which is the same in view of the smallness of the effect) if averaging is carried out over the two spin branches, the 2D spectra as well as all the experimentally measured parameters of 2D subbands calculated in the framework of the four- and three-band approximations for the same values of n_s are virtually identical (moreover, these parameters, in fact, do not differ from those calculated in the two-band approximation, i.e., taking into account U_{KG} only). In the four-band model, slightly larger (by 2–5%) band bends V_s generally correspond to the same values of n_s (a consequence of the lower density of states); however, since the value of V_s cannot be measured directly, a comparison makes sense just for the same values of n_s or n_i, which can be controlled experimentally. Thus, the adequacy of the three-band approximation for the structures under investigation with $E_g/\Delta \ll 1$ mentioned in Section 2 in connection with the description of experimental parameters of 2D subbands which are not associated with spin splitting is confirmed. However, in the

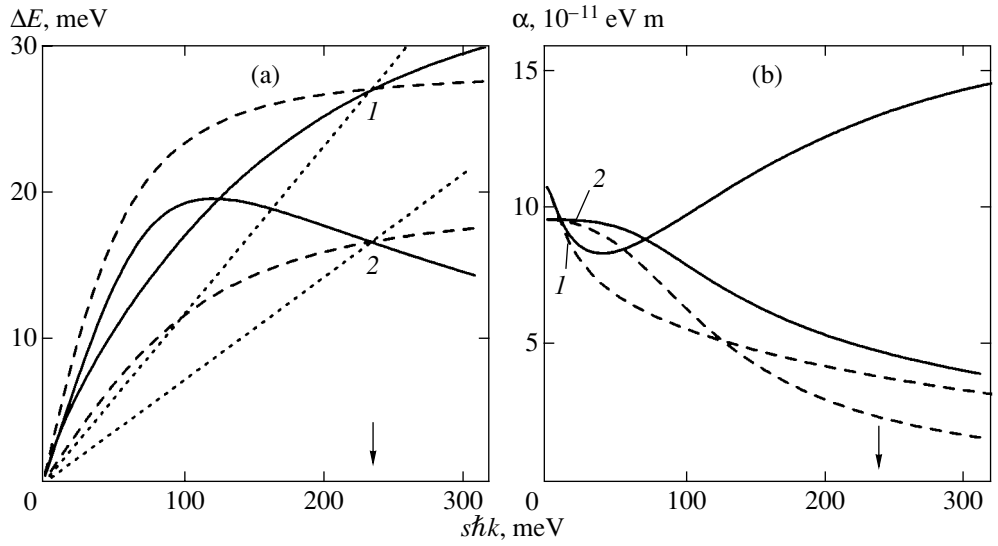


Fig. 3. (a) Energy splitting (solid curves correspond to exact calculations, dashed curves to approximation (4), and dotted curves to approximation (3)) and (b) the Rashba parameter (solid curves correspond to approximation (4) and dashed curves to approximation (3)) calculated in the four-band model as functions of the wave vector (in energy units) for the ground subband of the SM (*1*) and SC samples (*2*) for $n_0 = 2 \times 10^{12}$ cm^{-2}. Arrows correspond to $k = k_F$.

case of wideband materials (HgTe and especially InAs), the three-band approximation gives slightly lower values of the cyclotron masses of subbands without noticeably affecting the occupancy parameters.

For small values of $|E_g|$, the parabolic approximation (1) is found to be completely insufficient even for small band bends. At the same time, the dispersion relations for subbands, calculated without taking into account the spin splitting ($U_{so} = 0$), are correctly described by the Kane dispersion relation (Eqs. (3) and (4) with $\alpha = 0$) with the rest masses m_i and with the Kane velocities s_i of subbands (s_i differ from s_b only slightly) whose values can be determined unambiguously from the calculated subband values of the Fermi energies E_F, quasimomenta k_F, and cyclotron masses. The relative error introduced by such an approximation does not exceed 1–2% in the energy range of interest.

As regards spin splitting, it cannot be described even qualitatively using approximations (1) and (3) or when approximation (4) is employed. The splitting is not only nonlinear in k (for $E_g < 0$ even for very small values of k) and is not just saturated for large values of k as predicted by relation (4), but attains its maximum (for k considerably smaller than k_F for materials with $E_g > 0$ and for $k \sim 3k_F$ for materials with $E_g < 0$) and then decreases (Fig. 3a). In the limit of large k, the splitting virtually vanishes in accordance with the intuitive idea that the two spectral branches must coincide in the ultrarelativistic limit $k \longrightarrow \infty$. Figure 3a shows for comparison the dependences $\Delta_R(k)$ given by expressions (1), (3), and (4) with values of α obtained from joining the approximating and calculated spectra at the Fermi level. Obviously, for a correct description of the spectrum in the entire energy range, the phenomenological parameter α in expressions (1), (3), and (4) must be regarded as a function of the wave vector (different for different approximations) (Fig. 3b). Calculations show that this is also valid for InAs and HgTe, which are representatives of the class of narrow-gap materials with the widest bands. In semiconductors with a nearly parabolic spectrum, $E_g \gg \Delta$, and since $U_{so} \propto \Delta/E_g^3$ in this limit in accordance with relation (7), the expected splitting is at least 2 or 3 orders of magnitude smaller than in the materials under investigation and is hardly accessible for observations. Thus, in contrast to $\Delta n_i/n_i$, the Rashba parameter for the materials of interest is not a good characteristic of the effect in view of its strong dependence on energy in the models used. As regards its value at the Fermi level, which is used, as a rule, as a measure of the spin–orbit splitting, the above-mentioned ambiguity associated with inadequacy of the phenomenological approximations used for subband spectra should be borne in mind, especially while comparing the magnitudes of the effect in different materials and structures.

The parameters characterizing the spin–orbit splitting and calculated in the frameworks of three- and four-band models are compared in Figs. 2 and 4 for various populations of subbands. The inclusion of the Γ_7 band leads to a decrease in the energy splitting at the Fermi level and in the degree of polarization $\Delta n_i/n_i$, which is especially significant in the case of p electrons which interact with this band most strongly. As a result, the difference in the magnitudes of the effect between

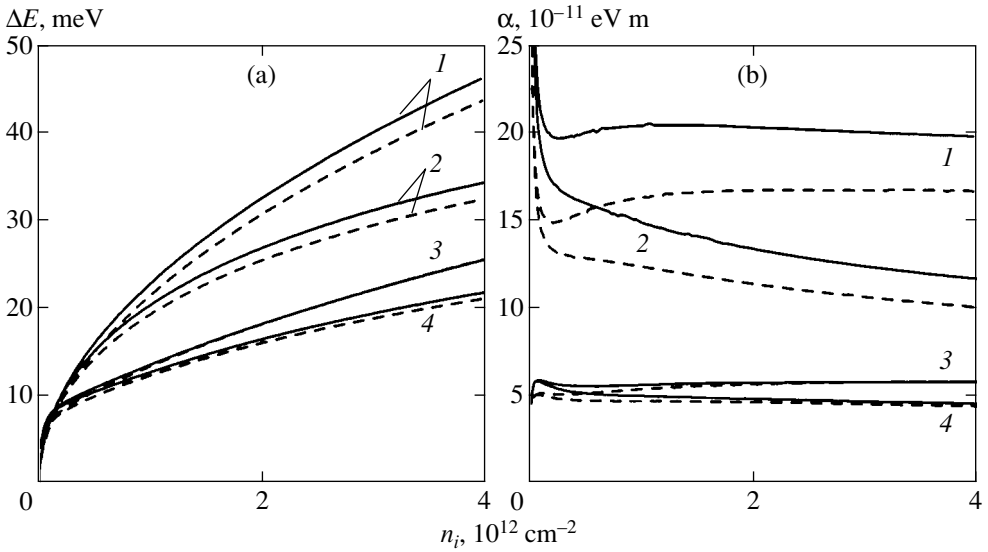


Fig. 4. Concentration dependences of (a) splitting and (b) Rashba parameter (in approximation (4)), calculated in the frameworks of the three-band (*1, 3*) and four-band (*2, 4*) models for the ground (solid curves) and excited (dashed curves) subbands for SM (*1, 2*) and SC (*3, 4*) samples.

materials with direct and inverse structure of energy bands decreases considerably as compared with the corresponding difference in the three-band model, while the discrepancy with the experiment becomes larger: the experimental values of $\Delta n_i/n_i$ turn out to be higher than the calculated values not only for $E_g > 0$, but also for $E_g < 0$.

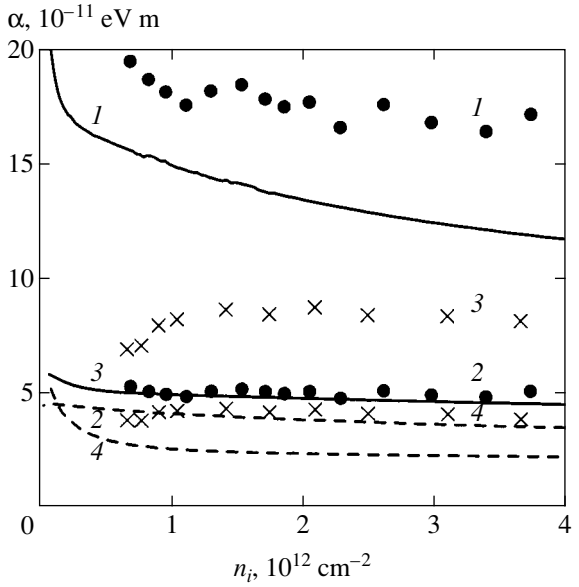


Fig. 5. Concentration dependences of the Rashba parameter for the ground subband calculated from the experimental (symbols) and theoretical (curves, four-band approximation) occupancies of the spin subbands using approximation (4) (*1, 3*) and (3) (*2, 4*) for the SM (*1, 2*) and SC (*3, 4*) samples.

The high values of $\Delta n_i/n_i$ and the Rashba parameter (here and below, we mean the value of α approximating the splitting at the Fermi level) in the region of not very high subband concentrations are due to the larger contribution of the electric field associated with the depletion layer in this concentration range; this is illustrated in Fig. 2 by a comparison with the ratio of the charge in the depletion layer N_{dep} (characterizing this contribution) to the subband concentrations. This is also confirmed by an increase in the calculated values of $\Delta n_i/n_i$ upon an increase in the concentration of acceptors. Outside this interval, $\Delta n_i/n_i$ and α attain the plateau in the framework of the three-band approximation, while, in the four-band case, these quantities slightly decrease upon an increase in n_i, especially for $E_g < 0$, which is in accord with the experimentally observed tendency (see Figs. 2, 5). At the same time, the energy splitting increases (although, sublinearly) upon an increase in n_i (however, the ratio Δ_R/E_F also decreases slightly). The inclusion of the Γ_7 band reduces the difference in the Rashba splitting for different subbands, without eliminating, however, the above discrepancy with the experimental values. It should be noted that, if the calculated values of $\Delta n_i/n_i$ in the excited subbands are slightly higher than in the ground subband for the same values of n_i, the values of Δ_R and α in these subbands are smaller due to slightly larger values of subband effective masses m_i.

Figure 5 shows the concentration dependences of the Rashba parameter at the Fermi level for the ground subband, calculated from the theoretical and experimental values of n_i^- and n_i^+ for different approximations for the subband dispersion relations. It can be seen

that the conventionally used approximation (3) which is linear in the wave vector (and the more so, the parabolic approximation (1)), leads to values of α smaller by a factor of 2–4 than the values given by a more adequate approximation (4). In the former case, the experimental values of parameter α for samples with direct and inverse structures are close, while, in the framework of approximation (4), the values of α for materials with $E_g < 0$ turn out to be twice as large. It should be noted that the values of the Rashba parameters in the system under investigation considerably exceed (for the same approximations) their values in the InGaAs quantum wells studied earlier.

Both experiment and theoretical analysis indicate a weak sensitivity of $\Delta n_i/n_i$ and the effective Rashba parameter to the magnitude of the electric field applied to the MIS structure, thus confirming the simple phenomenological considerations formulated in [7]. A noticeable dependence can be expected only in the range of moderate subband concentrations, where the contribution from the depletion layer is significant. A stronger doping of the semiconductor can increase the width of this interval. However, from the viewpoint of realization of a spin transistor in which the magnitude of the Rashba splitting should be varied without a significant change in the 2D concentration, it is preferable to control the magnitude of the depletion field by applying a bias voltage between the inversion layer and the bulk of the semiconductor. In the case of narrow-gap semiconductors, this can be realized in principle for low temperatures interesting for the problem under investigation also in view of a high rate of the tunnel exchange of carriers between the surface layer and the bulk of the semiconductor.

The reasons for the discrepancy between the calculated and measured values of the Rashba effect generally remain not quite clear, although there are several mechanisms which may lead to an increase in the spin–orbit splitting. It is well known that the magnitude of spin splitting (g factor) in 2D systems may increase significantly due to many-electron effects [24]. The Rashba splitting, which is similar in nature to the above splitting, must also be renormalized on account of correlation-exchange corrections. However, in view of the smallness of effective masses and high permittivities ε, the electron–electron interaction parameter $r_s = \sqrt{2}\, m_i e^2/\varepsilon\hbar^2 k_F$ for narrow-gap semiconductors is small (for the samples under investigation, $r_s \approx 0.2$ for a small Kane gap $m_i \propto k_F$ also; as a result, the value of r_s is virtually independent of the 2D concentration), and such a renormalization should not play a critical role. The calculations based on the results obtained in [25] lead to an increase in splitting by only 5% in both samples (this quantity, like r_s, is almost independent of n_i). It should be noted, however, that this estimate is valid for $\Delta_R/E_F \ll r_s$. The calculations corresponding to $\Delta_R/E_F \gg r_s$ ($r_s \ll 1$) give for both samples an increase in splitting by 20%, which is practically independent of n_i. The parameters of the samples under investigation correspond to the intermediate case $\Delta_R/E_F \approx r_s$, and a more stringent analysis is required for determining exact value of renormalization. It is important, however, that, in both limiting cases, additional "pushing apart" of the Rashba subsubbands due to the electron–electron interaction does not lead to an additional dependence of splitting on the electron density in contrast to predictions of the one-band model [26].

Another possible reason for the discrepancy with experimental results is the additional contribution to the splitting associated with the difference in the boundary conditions for different spin components of the wave function, which is assumed for heterostructures with narrow asymmetric quantum wells. However, in the case of MIS structures, the role of this mechanism cannot be significant in our opinion. In contrast to semiconducting heterostructures, the height of the barrier at the boundary with the insulator amounts to 2–3 eV, and the boundary conditions must be close to zero conditions, and the difference in the boundary conditions for different spin branches of the spectrum must the more so be small. Along with the large width of the surface quantum well in MIS structures based on narrow-gap materials and with a large magnitude of the effect due to the field in the well, this implies the smallness of the contribution associated with the surface.

This conclusion is in accord with the estimates obtained in [18] for wide quantum wells in heterostructures and is confirmed experimentally. We studied the structures prepared on the same substrates for different anodizing regimes, the structures with SiO_2, Al_2O_3, and with Blodgett–Lagnmuir films as a gate insulator with different thicknesses of dielectric layers and magnitudes of the charge implanted in the insulator. However, the degree of polarization $\Delta n/n$ for equal populations of the subbands was the same (the amplitudes of oscillations of the capacity of the space-charge region could differ by a factor of several units) and was determined only by the substrate material. It should be noted in this connection that, in view of amorphous or even organic nature of the dielectric layers in MIS structures, for which the description of the energy spectrum of the insulator on the basis of the symmetry classification of energy bands for the semiconducting substrate is essentially meaningless, the adequacy of application of new methods of calculations developed for semiconducting heterostructures and based on joining the Kane components of the wave functions appears dubious. In the framework of the approach used by us here (originating from [27, 28]) and based on the reduction of the initial matrix equation to an equation of the Schrödinger type, the properties of the spectrum of a Kane semiconductor (which are, in the long run, the properties of the lattice potential) are reflected in modification of the potential which "perceives" an electron during its motion and permits a clear physical interpretation (unless we con-

sider the formation and recombination of electron–hole pairs, this is a one-particle potential). In our opinion, the formulation of the boundary condition in this scheme must be based on the concepts which ultimately have one-particle origin and on the parameters which can be measured in principle (electron work function, barrier height for an electron, etc.).

In conclusion, we consider the advantages of the system studied by us here from the viewpoint of the creation of a spin quantum element (valve) for "spintronic" devices proposed for implementing the idea of creating a quantum computer, which is based on the Rashba effect. Narrow-gap compounds HgCdTe, which are characterized by the highest values of the spin–orbit interaction parameter among the known semiconductors and by extremely small effective masses, must exhibit (other conditions being the same) record-high values of the spin–orbit splitting of the 2D spectrum. In complete conformity with these intuitive considerations, the values of the splitting at the Fermi level calculated by us here, as well as the experimental values of $\Delta_R = \nabla_k E \Delta k_F$, amount to tens of millielectronvolts for typical concentrations, and the value of the effective Rashba parameter attains values of $(0.5–2.0) \times 10^{-10}$ eV m, which is almost an order of magnitude higher than the corresponding values for heterostructures studied earlier ($\Delta_R \sim (0.02–5)$ meV, $\alpha \sim (10^{-12}–10^{-11})$ eV m).The high values of splitting ensure not only large precession angles θ, but also the required modulation of $\theta \sim \pi$ for smaller relative variations $\Delta k_F(\alpha)$ due to a change in the voltage across the field electrode and/or the potential of the 2D channel relative to the quasi-neutral region. It is also clear that the materials investigated by us are most promising for the development of "spintronic" instruments operating at temperatures considerably higher than helium temperatures.

Another material parameter important from the viewpoint of realization of a spin transistor is the spin coherence length $l_s = \sqrt{l v_F \tau_s}$ (l is the mean free path and τ_s is the spin relaxation time). In this respect, narrow-gap semiconductors $A_{II}B_{VI}$ have advantages in view of high values of mobility and Fermi velocities close to their limiting value $s_b \approx 10^8$ cm/s for Kane semiconductors. Experimental estimates of τ for the 2D gas in HgCdTe have not been obtained to our knowledge; however, we have grounds to assume that their values must not differ strongly from those for InAs since large values of α in HgCdTe may be compensated by the smallness of effective masses. It should also be noted that the results presented above demonstrate not only larger values of parameters, but also their higher predictability in a HgCdTe-based spin transistor in the framework of the MIS architecture as compared to the architecture of devices on the basis of layered heterostructures.

ACKNOWLEDGMENTS

This study was supported financially by the Ministry of Education of the Russian Federation (project no. E00-3.4-278), US CRDF (grant no. REC-005), and the program "Universities of Russia."

REFERENCES

1. L. Wissinger, U. Rössler, R. Winkler, *et al.*, Phys. Rev. B **58**, 15 375 (1998); P. Pfeffer and W. Zawadzki, Phys. Rev. B **59**, R5312 (1999); P. Pfeffer, Phys. Rev. B **59**, 15902 (1999); R. Winkler, Phys. Rev. B **62**, 4245 (2000); D. Grundler, Phys. Rev. B **63**, 161 307 (2001); M. V. Entin and L. I. Magarill, Phys. Rev. B **64**, 085330 (2001).
2. X. C. Zhang, A. Pfeuffer-Jeschke, K. Ortner, *et al.*, Phys. Rev. B **63**, 245 305 (2001); C. H. Rowe, J. Nehls, R. A. Stradling, *et al.*, Phys. Rev. B **63**, 201 307 (2001); V. B. Bozhevol'nov, I. M. Ivankiv, V. F. Radan-tsev, and A. M. Yafyasov, Zh. Éksp. Teor. Fiz. **119**, 154 (2001) [JETP **92**, 135 (2001)]; C. M. Hu, J. Nitta, A. Jensen, *et al.*, Phys. Rev. B **63**, 125 333 (2001).
3. S. Datta and B. Das, Appl. Phys. Lett. **56**, 665 (1990).
4. É. I. Rashba, Fiz. Tverd. Tela (Leningrad) **2**, 1224 (1960) [Sov. Phys. Solid State **2**, 1109 (1960)].
5. F. J. Ohkawa and Y. Uemura, J. Phys. Soc. Jpn. **53**, 1325 (1974).
6. Yu. A. Bychkov and E. I. Rashba, J. Phys. C **17**, 6039 (1984).
7. V. F. Radantsev, Zh. Éksp. Teor. Fiz. **96**, 1793 (1989) [Sov. Phys. JETP **69**, 1012 (1989)].
8. J. Luo, H. Munekata, F. F. Fang, *et al.*, Phys. Rev. B **41**, 7685 (1990).
9. B. Das, S. Datta, and R. Reifenberger, Phys. Rev. B **41**, 8278 (1990).
10. M. A. Skvortsov, Pis'ma Zh. Éksp. Teor. Fiz. **67**, 118 (1998) [JETP Lett. **67**, 133 (1998)].
11. D. Grundler, Phys. Rev. Lett. **86**, 1058 (2001).
12. L. W. Molenkamp and G. Schmidt, Phys. Rev. B **64**, 121 202 (2001).
13. G. Engels, J. Lange, Th. Schäers, *et al.*, Phys. Rev. B **55**, 1958 (1997); J. Nitta, T. Akazaki, H. Takayanagi, *et al.*, Phys. Rev. Lett. **78**, 1335 (1997); C. M. Hu, J. Nitta, T. Akazaki, *et al.*, Phys. Rev. B **60**, 7736 (1999).
14. T. I. Deryabina, G. I. Kulaev, and V. F. Radantsev, Fiz. Tekh. Poluprovodn. (Leningrad) **24**, 1182 (1990) [Sov. Phys. Semicond. **24**, 746 (1990)]; M. Schultz, F. Heinrichs, U. Merkt, *et al.*, Semicond. Sci. Technol. **11**, 1168 (1996); D. Grundler, Phys. Rev. Lett. **84**, 6074 (2000).
15. T. Matsuyama, R. Kürsten, C. Meißner, *et al.*, Phys. Rev. B **61**, 15 588 (2000).
16. J. P. Heida, B. J. van Wees, J. J. Kuipers, *et al.*, Phys. Rev. B **57**, 11 911 (1998).
17. V. F. Radantsev, T. I. Deryabina, G. I. Kulaev, *et al.*, Phys. Rev. B **53**, 15 756 (1996).
18. E. A. De Andrada, E. Silva, G. C. La Rocca, *et al.*, Phys. Rev. B **55**, 16 293 (1997).
19. V. F. Radantsev, A. M. Yafyasov, and V. B. Bogevolnov, Semicond. Sci. Technol. **16**, 320 (2001).

20. L. G. Gerchikov and A. V. Subashiev, Fiz. Tekh. Poluprovodn. (St. Petersburg) **26**, 131 (1992) [Sov. Phys. Semicond. **26**, 73 (1992)].

21. A. Ziegler and U. Rossler, Europhys. Lett. **8**, 543 (1989).

22. V. F. Radantsev, Semicond. Sci. Technol. **8**, 394 (1993).

23. V. A. Larionova and A. V. Germanenko, Phys. Rev. B **55**, 13 062 (1997).

24. T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).

25. G. H. Chen and M. E. Raikh, Phys. Rev. B **60**, 4826 (1999).

26. W. Hausler, Phys. Rev. B **63**, 121 310 (2001).

27. Ya. B. Zel'dovich and V. S. Popov, Usp. Fiz. Nauk **105**, 403 (1971) [Sov. Phys. Usp. **14**, 673 (1971)].

28. A. B. Migdal, V. S. Popov, and D. N. Voskresenskiĭ, Zh. Éksp. Teor. Fiz. **72**, 834 (1977) [Sov. Phys. JETP **45**, 436 (1977)].

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 502–504.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 582–585.

SOLIDS
Structure

Porous Silicon-Based Ferroelectric Nanostructures

E. D. Mishina[a], K. A. Vorotilov[a], V. A. Vasil'ev[a], A. S. Sigov[a, *], N. Ohta[b], and S. Nakabayashi[b]

[a]*Moscow State Institute of Radio Engineering, Electronics, and Automatics, Moscow, 119454 Russia*
**e-mail: sigov@mirea.ru*
[b]*Department of Chemistry, Faculty of Science, Saitama University, Saitama, 338-8570 Japan*
Received April 24, 2002

Abstract—A procedure is suggested for the preparation of porous silicon-based ferroelectric nanostructures. It is demonstrated that the method of chemical deposition from solutions provides for the penetration of the initial components of the solution into the matrix pores, and subsequent annealing leads to the crystallization of the ferroelectric phase. The diagnostics of the ferroelectric properties is performed using the method of generation of second optical harmonic. The spectral characteristics of the prepared ferroelectric nanostructures are investigated. © *2002 MAIK "Nauka/Interperiodica".*

The size effects in thin ferroelectric films and ceramics have been investigated since the 1970s [1]. The Curie temperature, polarization, coercive field, and the switching rate of polarization, as well as the stability of these properties, depend on the film thickness and on the ceramic grain size [2]. It has been theoretically demonstrated that, at $T = 0$ K, the critical film thickness at which the ferroelectric polarization goes to zero (and grows with temperature) is 4 nm for $BaTiO_3$ and 8 nm for lead zirconate titanate $PbTi_{0.5}Zr_{0.5}O_3$ (LZT) [3]; the critical size of LZT nanocrystallite at room temperature is 6 nm [4]. Experimentally, the ferroelectric properties were revealed in perovskite films 6 nm thick [5], in polymer films two monolayers thick [6], and in LZT ceramic with the grain size of 7 nm [7]. The size effects in isolated nanocrystallites were not studied until presently for the lack of technology of their manufacture.

Porous membranes (porous silicon and alumina) are widely investigated at present as matrix materials for preparing nanostructures with various inclusions (magnetic, semiconductor, polymer, carbon—see review [8]). The methods of immersion and cathode deposition are employed for the introduction of a material into a matrix. Also used as matrixes are synthetic opals; however, their use is defined by the bulk properties, in particular, by the photon forbidden band arising in such photon crystals [9, 10]. Nevertheless, the procedures for the preparation and investigation of opals (three-dimensional structures) may be used in application to nanostructures on the basis of porous silicon matrixes (two-dimensional structures). One advanced application of porous silicon-based ferroelectric nanostructures may be their use as transformable photon crystals for optoelectronic devices. Another possible application of ferroelectric nanostructures may be their use in high-density memory devices.

We suggest a new method of preparing ferroelectric nanostructures with the transverse crystallite dimension of 10 nm. For this purpose, porous silicon is used as the matrix, with the precursor of ferroelectric material introduced into the matrix pores from a solution of organometallic compounds; the material acquires the ferroelectric properties in the process of annealing. The mechanism of filling nanopores that are so small is apparently based on the electrostatic interaction between the micelles of the initial sol and the substrate with a different charge [8].

The method of generation of second optical harmonic is used to investigate the ferroelectric properties. Because the second harmonic intensity in the centrally symmetric nonferroelectric phase is zero (in a dipole approximation), its increase during annealing is unique evidence of the transition of nanoparticles to the ferroelectric phase.

Porous silicon was prepared from platelets of *p*-Si with the resistivity of 0.01 Ω cm by anodic etching in a 15% solution of HF in ethanol. The anode current was 25 mA/cm^2 with the etching rate of 23 nm/s, which provides for the porosity of 66% (with respect to air). The average pore size in such structures was approximately 10 nm.

The starting solution for producing LZT with the composition of $PbZr_{0.53}Ti_{0.47}O_3$ was prepared by mixing a solution of titanium and zirconium isopropylates in methyl Cellosolve (for more detail, see [11]). Titanium and zirconium were introduced in stoichiometric amounts, and lead was taken with an excess of 10 wt. % to compensate for possible loss in the process of high-temperature annealing. Platelets of porous silicon were immersed in the thus prepared solution of organometallic compounds, after which they were dried at 200°C for 15 min and then annealed. The annealing temperature was varied from 300 to 700°C (20 min).

In order to investigate the second harmonic generation in the produced nanostructures, we used the radiation of an optical parametric amplifier pumped by a titanium–sapphire laser with amplifier (CPA-200, Klark Corporation). The radiation parameters were as follows: wavelength, 550–800 nm; pulse duration, 100 fs; pulse energy, 0.05 mJ; repetition rate, 1 kHz; size of focused spot on the sample, 100 μm. The radiation of the second harmonic was spectrally analyzed using color filters and a monochromator. The mirror-reflected and scattered radiation of the second harmonic (with the scattering angle of 30°) was recorded by a photomultiplier operating in the photon count mode. The second harmonic signal from the sample was normalized in a comparison channel identical with the measuring channel during the second harmonic generation from a platelet of crystalline quartz.

Figure 1 gives the dependence of the intensity of the second harmonic generated upon reflection from samples prepared at different annealing temperatures, compared with radiation in the case of an unfilled matrix of porous silicon (pumping radiation wavelength, 780 nm). At the annealing temperature $T \le 600$°C, the intensity of the second harmonic from a filled matrix increases significantly compared with an unfilled matrix (by a factor of ten) and almost does not vary with the annealing temperature. No scattered radiation of the second harmonic is observed in these samples. At $T = 700$°C, an abrupt increase (by an order of magnitude) in the radiation is observed in the direction of the mirror reflection of the second harmonic, and its scattered radiation emerges. The essentially different patterns of radiation of the second harmonic in samples obtained in the cases of low- and high-temperature annealing point to the different phase states of the ferroelectric material in these samples. The scattered radiation of the second harmonic arises in a medium in the presence of nonuniformities of the nonlinear optical properties which are less than the wavelength in size [12]. The absence of scattered radiation of the second harmonic in low-temperature samples points to the absence of such nonuniformities. At the same time, the radiation of the second harmonic increases compared with its radiation in an unfilled matrix. This means that such a medium behaves as quasi-homogeneous, and the radiation of the second harmonic is generated both in LZT-filled pores and in silicon proper. The increase in the second harmonic intensity compared with radiation in the case of unfilled porous silicon is associated with the variation of the local boundary conditions, local fields, and so on, and is quadrupole by its nature, because both materials are in the centrally symmetric phase. A dipole contribution by LZT nanocrystallite boundaries is also possible [13]. In high-temperature samples, a separate LZT nanocrystallite serves as the second harmonic emitter, compared with which the emission of porous silicon is negligibly low. It is the totality of noncentrally symmetric nanocrystallites that provides for the presence of high-intensity scattered radiation with a wide scattering diagram.

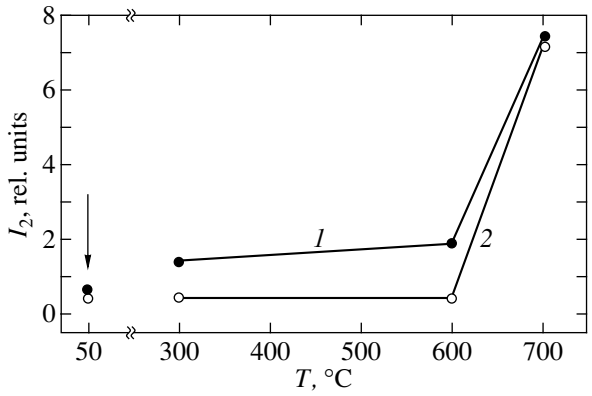


Fig. 1. The intensity of radiation of the second harmonic in ferroelectric nanostructures as a function of the annealing temperature: (*1*) mirror-reflected component, (*2*) diffuse component. The arrow indicates the unfilled matrix of porous silicon. The pumping radiation wavelength, 780 nm.

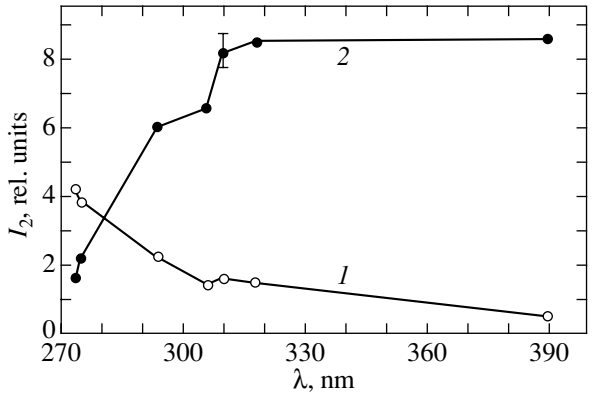


Fig. 2. Spectra of second harmonic radiation generated (*1*) by a porous silicon matrix and (*2*) by an LZT nanostructure annealed at $T = 700$°C.

The spectral singularities of the second harmonic radiation in a sample annealed at $T = 700$°C are given in Fig. 2 compared with the radiation of the second harmonic from the porous silicon matrix. The efficiency of the second harmonic generation on LZT nanocrystallites decreases with the pumping radiation wavelength; in porous silicon, on the contrary, it increases. The radiation spectrum of the second harmonic of porous silicon agrees with the absorption spectra of this material prepared under similar conditions [14]. The second harmonic wavelength (400 nm) lies on the edge of the absorption spectrum, with the resonant conditions for the second harmonic generation for this wavelength being valid only for the second harmonic radiation. As the wavelength decreases, both the pumping radiation and the second harmonic radiation fall in the region of resonant absorption; as a result, the resonant absorption intensity increases. The edge of the LZT absorption

band falls on a wavelength of the order of 300 nm [15]. At the same time, the maximum of the intensity of the second harmonic from the prepared nanostructures falls on the wavelength of 400 nm. This difference between the spectra of absorption and second harmonic points to the significant effect of the spectral dependences of the factors of local fields in rodlike nanocrystallites; it is this effect that must lead to a shift of the resonant frequency of radiation of the second harmonic [16].

So, we have demonstrated that the method of chemical deposition from solutions enables one to produce porous silicon-based ferroelectric nanostructures with ferroelectric nanocrystalites 10 to 20 nm in diameter. The second harmonic generation method was used to reveal the formation of the ferroelectric phase in LZT nanocrystallites.

ACKNOWLEDGMENTS

We are grateful to the Russian Foundation for Basic Research for financial support of this study (project no. 00-02-16557). The investigations were partly supported by CRDF and the Ministry of Education of the Russian Federation (grant no. VZ-010-0) and INTAS (grant no. 75-2002).

REFERENCES

1. W. R. Buessem, L. E. Cross, and A. K. Goswami, J. Am. Ceram. Soc. **49**, 33 (1966).
2. J. F. Scott and C. A. Paz de Araujo, Science **246**, 1400 (1989).
3. S. Li, J. A. Eastman, J. M. Vetrone, *et al.*, Jpn. J. Appl. Phys. **36**, 5169 (1997).
4. H. Huang, C. Q. Sun, Z. Tianshu, and P. Hing, Phys. Rev. B **63**, 184 112 (2001).
5. E. Mishina, N. Shersyuk, V. Mukhortov, *et al.*, in *Book of Abstracts, 1st International Meeting on Ferroelectric Access Memory FeRAM 2001, Gotemba, Japan, 2001*, p. 142.
6. S. Ducharme, S. P. Palto, L. M. Blinov, and V. M. Fridkin, AIP Conf. Proc. **535**, 354 (2000).
7. S. Chattopadhyay, P. Ayyub, V. R. Palkar, and M. Multani, Phys. Rev. B **52**, 13 177 (1995).
8. J. C. Hulteen and C. R. Martin, in *Nanoparticles and Nanostructured Films*, Ed. by J. H. Fendler (Wiley-VCH, Weinheim, 1998), p. 235.
9. V. G. Golubev, D. A. Kurdyukov, A. V. Medvedev, *et al.*, Fiz. Tekh. Poluprovodn. (St. Petersburg) **35**, 1376 (2001) [Semiconductors **35**, 1320 (2001)].
10. J. M. Dereppe, B. F. Borisov, E. V. Charnaya, *et al.*, Fiz. Tverd. Tela (St. Petersburg) **42**, 184 (2000) [Phys. Solid State **42**, 193 (2000)].
11. K. A. Vorotilov, M. I. Yanovskaya, E. P. Turevskaya, and A. S. Sigov, J. Sol-Gel Sci. Technol. **16**, 109 (1999).
12. O. A. Aktsipetrov, A. A. Fedyanin, D. A. Klimkin, *et al.*, Ferroelectrics **190**, 143 (1997).
13. A. Bürgel, W. Kleemann, and U. Bianchi, Phys. Rev. B **53**, 5222 (1996).
14. A. N. Obraztsov, V. Yu. Timoshenko, H. Okushi, and H. Watanabe, Fiz. Tekh. Poluprovodn. (St. Petersburg) **33**, 322 (1999) [Semiconductors **33**, 323 (1999)].
15. L. Pintilie and I. Pintilie, Mater. Sci. Eng. B **80**, 388 (2001).
16. C. K. Chen, T. F. Heinz, D. Ricard, and Y. R. Shen, Phys. Rev. B **27**, 1965 (1983).

Translated by H. Bronstein

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 505–510.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 586–592.

SOLIDS
Electronic Properties

The Plasma Oscillations Spectrum of a Periodically Inhomogeneous 2D Electron System near the Perforation Threshold

O. R. Matov, O. V. Polishchuk, and V. V. Popov*

Institute of Radio Engineering and Electronics, Russian Academy of Sciences (Saratov Branch), Saratov, 410019 Russia

**e-mail: popov@ire.san.ru*

Received October 16, 2001

Abstract—The transformation of the spectrum of plasma oscillations with the zero reduced wave vector in the spatially modulated two-dimensional electron system moving to the regime of isolated quasi-one-dimensional electron channels is theoretically investigated. The results provide an explanation of the well-known experimental observations of the plasma resonance transformation when a two-dimensional electron system crosses the continuity violation threshold. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In the papers [1, 2], the far-infrared Fourier spectroscopy technique was used to experimentally investigate the excitation of periodically inhomogeneous 2D electron plasma in heterostructures GaAs/AlGaAs near the threshold of crossover from the continuous 2D system with spatially modulated electron density to a system of isolated quasi-one-dimensional electron channels. The electron density in the 2D system was spatially modulated by applying a bias voltage $V_g < 0$ to a periodic gate electrode. A continuous semitransparent (for electromagnetic waves) conducting layer of NiCr with periodic corrugation was used as a gate electrode (see Fig. 1).

The far-infrared Fourier spectroscopy technique makes it possible to observe the plasma oscillations with the zero reduced wave vector $k = 0$ in the plane of a periodically inhomogeneous 2D system. In the papers [1, 2], the plasma resonance corresponding to the excitation of the principal (with the lowest frequency) plasma oscillation with $k = 0$ was investigated. An increase in the modulation depth of the electron density with increasing $|V_g|$ under the continuous operation of the 2D electron system results in decreasing the plasma resonance frequency. Moreover, the squared resonance frequency decreases more rapidly than the mean surface density of electrons in the 2D system. It was shown in [3] that this phenomenon is explained by the localization of the principal mode of plasma oscillations in the regions of the 2D plasma with lower concentration of electrons.

At $V_g < V_{tB}$, where V_{tB} is the threshold voltage corresponding to the complete depletion of the electron layer on the segment B (see Fig. 1), the continuity of the 2D electron system is violated (the perforation occurs) and a periodic system of isolated quasi-one-dimensional electron channels is formed. When $|V_g|$ increases further in the domain $V_g < V_{tB}$, an increase in the plasma oscillation frequency of electrons localized in the system of isolated quasi-one-dimensional channels is observed. This fact was given a theoretical explanation in [4].

The results of experiments [1, 2] testify that, when the continuity violation threshold of a 2D electron system is attained (at $V_g = V_{tB}$), the frequency of plasma resonance remains finite even in the absence of an external magnetic field. Moreover, the frequency of plasma resonance varies continuously when changing from a continuous spatially modulated 2D electron system to a system of isolated quasi-one-dimensional channels.

A theoretical treatment of the plasma oscillation spectrum transformation at a crossover from a homogeneous 2D electron system to a periodic system of isolated quasi-one-dimensional electron channels was

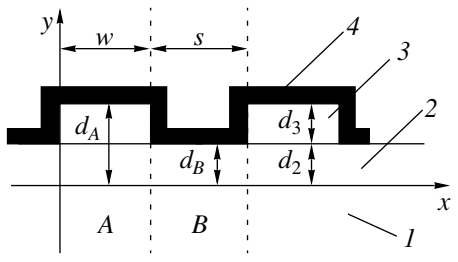


Fig. 1. A schematic outline of a structure with periodically inhomogeneous 2D electron plasma [1]: (*1*) GaAs; (*2*) AlGaAs; (*3*) photoresist; (*4*) gate electrode (NiCr). The spatially modulated electron 2D gas is placed at the interface between the media 1 and 2.

given in [5–7]. In [5, 6], the response of a periodically inhomogeneous 2D electron system was described using the local surface conductivity, while, in [7], a hydrodynamic approach in the Thomas–Fermi–Dirac–Weizsäcker model was used. The calculations in [5, 6] show that the frequencies of all (including principal ones) plasma modes with a zero reduced wave vector in a continuous periodically inhomogeneous 2D electron system decrease down to zero while approaching the continuity violation threshold. At the same time, the paper [7] suggests a conclusion that, as the depth of spatial modulation of the electron density in a 2D electron system with period L increases, the plasma mode of the homogeneous 2D electron system with the wave vector $k = 2\pi/L$ undergoes a continuous transformation to the principal dipole mode of plasma oscillations localized in isolated electron channels.

Thus, the physical mechanism underlying the transformation of plasma resonance when crossing the continuity violation threshold in a 2D electron system has remained unknown until the present time.

In the theoretical studies [5–7], plasma oscillations were described using the electrostatic approximation without taking into account their coupling to electromagnetic radiation fields. In [3], it was shown that it is important to take into account electrodynamic effects when considering plasma oscillations with a zero reduced wave vector in the plane of a periodically inhomogeneous 2D electron system. As has already been mentioned, just such oscillations occur in far-infrared Fourier spectroscopy experiments with low-dimensional electron systems.

In this paper, we use the rigorous electrodynamic theory developed in [3] to reveal the mechanism underlying the transformation of plasma oscillations with the zero reduced wave vector when changing from a continuous spatially modulated 2D system to a system of isolated quasi-one-dimensional electron channels. In Section 2, we discuss a model used to calculate the equilibrium periodic distribution of electron density in a 2D system depending on the voltage on the periodic gate electrode. In Section 3, the calculation results are presented and compared with the experimental data reported in [1, 2]. Conclusions are given in the last section.

2. EQUILIBRIUM ELECTRON DENSITY DISTRIBUTION IN A PERIODICALLY INHOMOGENEOUS 2D SYSTEM: THE PARALLEL PLANE CAPACITOR MODEL

In [1, 2], experimentally obtained values of the plasma resonance frequencies in the structure depicted in Fig. 1 are reported as a function of the gate voltage V_g. However, the theory developed in [3] requires as its input parameters the concentration of electrons $N_{A,B}$ on the segments A and B of the 2D system. Strictly speaking, the values $N_{A,B}$ can be found by solving the corresponding electrostatic problem for the structure with a periodically corrugated gate electrode. It is clear that the profile of the electron density distribution in the 2D system is generally different from the rectangular profile assumed in [3]. However, the shape of the gate electrode is not known precisely, since it cannot be effectively controlled when making the electrode. Moreover, no accurate value of the dielectric constant of AlGaAs is known for the range of frequencies studied in [1, 2] (it also depends on the content of aluminum). For these reasons, we use a simple approximate model for determining the equilibrium distribution of electrons in the 2D system. This model admits a direct application of the theory [3] for the description of plasma oscillations and makes it possible to "adjust" to the experimental conditions in [1, 2] using fitting parameters.

The surface concentration of electrons on the segments A and B (Fig. 1) is found by the plane capacitor formula

$$N_{A(B)} = \frac{V_g - V_{tA(B)}}{d_{A(B)}e}\varepsilon_{A(B)}\varepsilon_0 \quad (V_{tA(B)} \le V_g < 0), \quad (1)$$

where $d_{A(B)}$ is the distance from the 2D system to the gate electrode on segment $A(B)$, $V_{tA(B)}$ is the threshold gate voltage corresponding to the complete depletion of the 2D electron system on segment $A(B)$, e is the charge of electron ($e > 0$), and ε_0 is the electrical constant. The dielectric constant ε_B on segment B is assumed to be equal to the dielectric constant of AlGaAs. As in experimental samples in [1, 2], the capacitor on segment A is filled with the composite dielectric of thickness $d_A = d_2 + d_3$, where d_2 and d_3 are the thickness of the layers of AlGaAs and the photoresist, respectively. In this case, the effective dielectric constant ε_A involved in (1) has the following form for the plane capacitor model:

$$\varepsilon_A = \frac{\varepsilon_2\varepsilon_3(d_2 + d_3)}{d_2\varepsilon_3 + d_3\varepsilon_2},$$

where ε_2 and ε_3 are the dielectric constants of AlGaAs and the photoresist, respectively. The threshold voltage $V_{tB} = -0.5$ V corresponding to the formation of a system of isolated electron channels was found experimentally in [1] on the basis of CV and dc measurements. In our model, the voltage V_{tA} is determined from the condition $N_A = N_B$ at $V_g = 0$, which yields

$$V_{tA} = \frac{d_A}{d_B}\frac{\varepsilon_B}{\varepsilon_A}V_{tB}.$$

The plane capacitor model described above is often used for estimating the electron density in spatially modulated 2D systems [1, 2]. Obviously, this model yields correct results only for $d_A \ll w$ and $d_B \ll s$. For the structures investigated in [1, 2], these conditions hold only on segment B ($d_B \approx 50$ nm and $s \approx 250$ nm); however, they do not hold on segment A ($d_A \approx 170$ nm and $w \approx 250$ nm). Therefore, for segment A, formula (1)

can be used only as a fitting relation. The quantity d_3 can be used as a sole fitting parameter, which determines all other parameters (V_{tA} and ε_A) involved in formula (1) for N_A. Naturally, the fitted value of d_3 is generally different from the actual thickness of the photoresist *3*. The dielectric constants of the materials used in the electrostatic model were assumed to be $\varepsilon_1 = 12.8$, $\varepsilon_2 = 11.0$, and $\varepsilon_3 = 2.4$ [8].

3. CALCULATION RESULTS AND COMPARISON WITH THE EXPERIMENT

In this paper, the theoretical treatment is based on solving two separate problems. First, formula (1) is used to find the concentration of electrons on segments *A* and *B* as a function of the voltage applied to the gate. The choice of the fitting parameter d_3 is discussed below. Then, the values $N_{A,B}$ are used in the electrodynamic model [3].

The algorithm proposed in [3] makes it possible to calculate the frequency, radiative damping, and the distribution of amplitude of both the electric field and the charge density oscillations for any plasma mode in a periodically inhomogeneous 2D electron system with a rectangular profile of the electron density modulation:

$$N_s(x) = \begin{cases} N_A & \text{for } 0 < x < w \\ N_B & \text{for } w < x < L, \end{cases}$$

where $L = w + s$ is the structure period, for any modulation factor

$$\Delta n_s = \frac{N_A - N_B}{N_A + N_B} \le 1 \quad (N_A \ge N_B \ge 0).$$

The response of the 2D electron system to the action of a harmonic electric field $E\exp(i\tilde{\omega}t)$ is described (in the framework of the Drude model) by the local surface conductivity

$$\sigma_{A(B)} = \frac{e^2 N_{A(B)}}{m^*} \frac{\tau}{1 + i\tilde{\omega}\tau},$$

where m^* is the effective mass of an electron and τ is the phenomenological relaxation time of the electron momentum in the 2D system. The real part of the complex frequency $\tilde{\omega} = \omega + i\gamma$ corresponds to the frequency of the plasma oscillations, and the imaginary part γ is the damping of those oscillations due to both the dissipative processes in the system and the electromagnetic radiation from the structure. Obviously, in the case $1/\tau = 0$, we have $\gamma = \gamma_r$, where γ_r is the radiative damping.

It is assumed in the electrodynamic model that the 2D electron system is placed at the interface between two semi-infinite media with the dielectric constants ε_1 and ε_2. In this sense, the electrodynamic model differs from the actual structure depicted in Fig. 1, where a periodically corrugated semitransparent metal layer NiCr is placed on the top of the AlGaAs sample. However, using ε_2 as a sole fitting parameter in the electrodynamic problem, one can take into account the effect of the corrugated conducting gate on the screening of plasma oscillations in the 2D system.

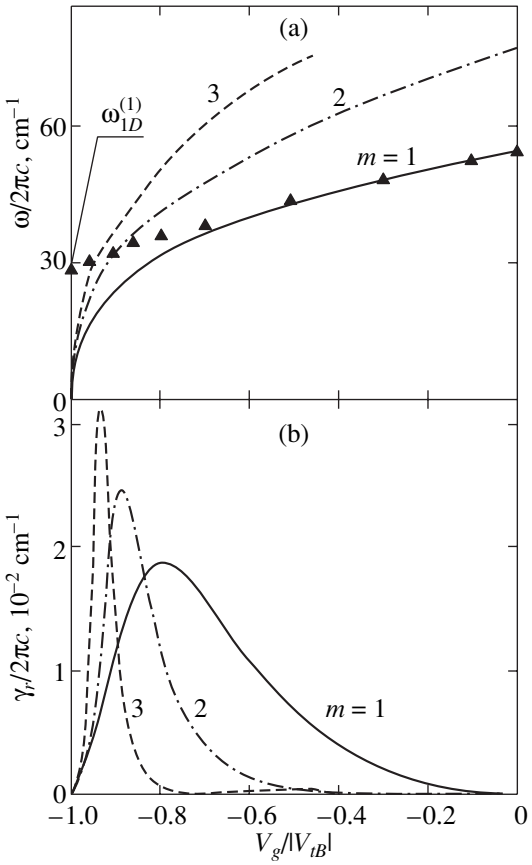


Fig. 2. The frequencies $\omega^{(m)}$ (a) and the radiative damping $\gamma_r^{(m)}$ (b) ($m = 1, 2, 3$) as functions of the gate voltage V_g. The triangles correspond to the experimental data reported in [1].

Figure 2 illustrates results of calculation of the frequencies $\omega^{(m)}$ and the radiative damping $\gamma_r^{(m)}$ for the three lower plasma modes with the zero reduced wave vector in the entire range of variation of the modulation factor $0 \le \Delta n_s \le 1$ for the characteristic parameters of the experiment reported in [1] for the case $1/\tau = 0$. Modes are denoted by the index *m* in ascending order of their frequency. From the physical point of view, the modes of plasma oscillations with different values of *m* differ in the number of variations of the electric field (and the charge density) oscillation amplitude over the period of the system. The calculation data are provided only for radiative modes, since only such modes can

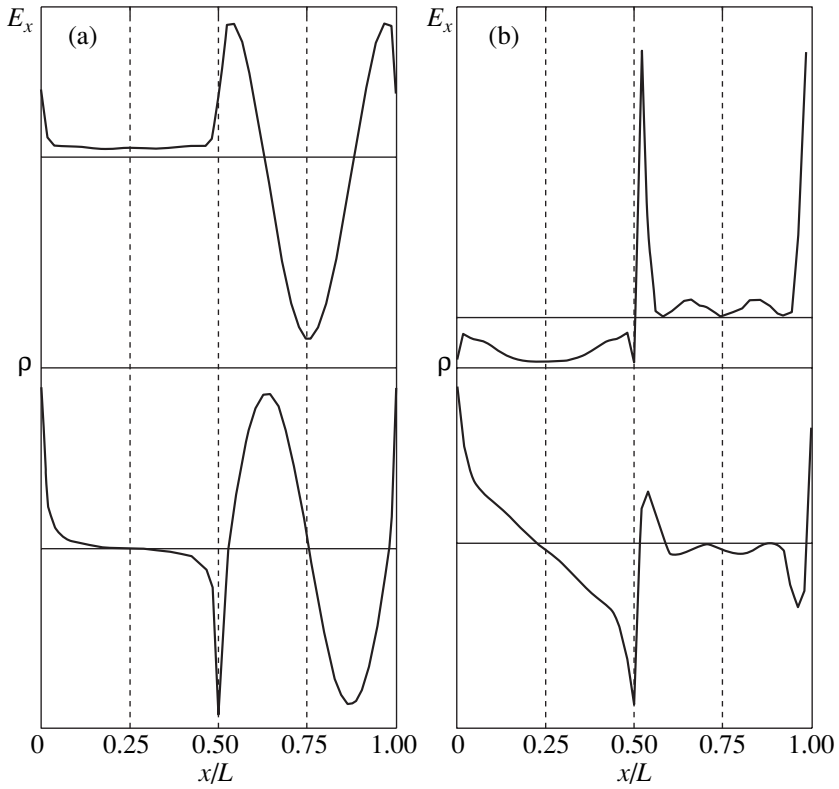


Fig. 3. The distribution of the longitudinal electric field E_x and the charge density ρ in the plane of the 2D system for two modes of plasma oscillations $m = 1$ (a) and $m = 3$ (b) at $\Delta n_s = 0.9$.

manifest themselves in the form of plasma resonances in experiments. It is seen in Fig. 2a that the frequencies of all modes decrease down to zero when the continuity violation threshold of the 2D system is reached at $V_g = V_{tB}$. As the mode index m increases, the decrease of frequency at $V_g \longrightarrow V_{tB}$ becomes steeper, so that the spectral branch of the mode with $m = 3$ approaches the line $V_g = V_{tB}$ on which the frequencies of dipole plasma oscillations $\omega_{1D}^{(m)}$ localized within electron strips of width w at $N_A\big|_{V_g = V_{tB}}$ are located. On the curve $\omega^{(3)}(V_g)$, a small kink at the frequency of the principal dipole mode ω_{1D} is clearly visible. This kink occurs due to the interaction of these modes.

In the electrodynamic problem, we used the fitted value $\varepsilon_2 = 16$, which was chosen to ensure the coincidence of the theoretical and experimental values of the principal mode frequency ($m = 1$) in the case of a homogeneous 2D system (at $V_g = 0$). The value of d_3 was chosen by fitting the theoretical value of the principal dipole mode frequency in isolated electron channels to the experimental data reported in [1] for the plasma resonance frequency at $V_g = V_{tB}$. Note that the corresponding frequency calculated by the evaluation formula (see [9, 10])

$$\omega_{1D}^{(1)} = \sqrt{\frac{2N_A\big|_{V_g = V_{tB}} e^2}{\varepsilon_0(\varepsilon_1 + \varepsilon_2)m^* w}}$$

for the classical isolated electron channel with a rectangular profile of electron density coincides with the exact theoretical value with an error less than 5%.

The dependence of the radiative damping of plasma modes on the gate voltage is nonmonotonic (see Fig. 2b). The maxima of the radiative damping for different modes occur at different values of V_g. It follows from the data presented in Fig. 2b that the mode with $m = 3$ has the maximum value of the radiative damping near the perforation threshold of the 2D system (as $V_g \longrightarrow V_{tB}$). Since the magnitude of the radiative damping controls the coupling of plasma oscillations to the external electromagnetic wave [11, 12], the excitation strength of one or another plasma mode will be different for different V_g.

Thus, the results presented in Fig. 2 show that the experimentally observed [1, 2] plasma resonance in a continuous 2D system with spatially modulated electron density at $V_{tB} < V_g \le 0$ (the corresponding experimental data are shown in Fig. 2a by triangles) is related

to the excitation of different plasma modes. Under weak modulation, the mode with $m = 1$ is excited. When approaching the continuity violation threshold of the 2D system, the mode with $m = 3$ is excited most effectively, producing resonance at the frequency $\omega^{(3)} \approx \omega_{1D}^{(1)}$ at $V_g \approx V_{tB}$ (see Fig. 2a). In the intermediate region, the series of experimental points can be explained by the excitation of the mode with $m = 2$.

The physical picture of the transformation of various plasma modes with changes in V_g is illustrated in Fig. 3. Here, the distribution of the longitudinal electric field E_x and the charge density in the plane of the 2D system for two modes of plasma oscillations ($m = 1, 3$) are shown for a deep modulation of the equilibrium electron density $\Delta n_s = 0.9$ ($V_g/|V_{tB}| = -0.97$). It is seen that, in this case, the plasma oscillations with $m = 1$ are localized in the region of the 2D system with smaller concentration of electrons, which results in decreasing the frequency and radiative damping of this mode. Calculations show that the distributions of amplitudes of both the electric field and the charge density oscillations in the plasma mode with $m = 2$ have similar behavior. At the same time, the mode with $m = 3$ demonstrates opposite behavior at the same modulation depth corresponding to the kink point of the spectral branch of this mode (see Fig. 2a). Charge oscillations are mainly localized in the region of the 2D system with large concentration of electrons. Moreover, the distributions of the field and charge density formed on segment A are similar to the corresponding distributions for the principal dipole oscillation in an isolated quasi-one-dimensional electron channel (cf. Figs. 3b and 4). This explains the gradual transformation of the plasma resonance when crossing the perforation threshold of the 2D system, which was observed in [1, 2]. It is quite natural that, as the gate voltage $|V_g|$ increases further, oscillations of the mode with $m = 3$ are also localized in regions of the 2D system with small concentration of electrons, which results in decreasing the frequency and radiative damping of this mode down to zero as $V_g \longrightarrow V_{tB}$.

4. CONCLUSIONS

In this paper, we use a rigorous electrodynamic approach to give an analysis of the plasma oscillation spectrum in a periodically inhomogeneous 2D electron system with a rectangular profile of the electron density distribution near the perforation threshold (the perforation leads to the occurrence of isolated quasi-one-dimensional electron channels). The frequencies of all plasma modes decrease down to zero when approaching the system perforation threshold, as is also the case in the electrostatic model [5, 6]. At the same time, the spectral branch of the high-frequency (third) plasma mode undergoes a kink near the perforation threshold. It is shown that this kink takes place at the frequency coinciding with that of the principal dipole plasma oscillation in isolated electron strips. Moreover, the distributions of amplitudes of both the electric field and the charge density oscillations for the third plasma mode in a strongly modulated 2D system are close to the corresponding distributions in isolated electron channels.

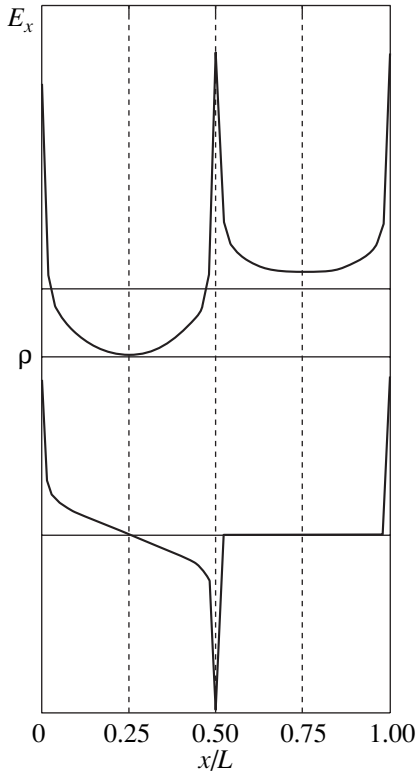


Fig. 4. The distribution of the longitudinal electric field and the charge density in a system of isolated electron channels.

The calculation results are fitted to the experimental data reported in [1, 2], which were obtained by observing plasma resonances in periodically inhomogeneous 2D electron systems. It is concluded that the dependence of the plasma resonance frequency on the modulation depth of the equilibrium electron density in a 2D system, which is observed experimentally, is explained by the excitation of different modes of plasma oscillations at different modulation depths. As a result, an explanation of the contradiction between the experimental and theoretical data concerning the behavior of the frequency of the plasma resonance when crossing the perforation threshold of the 2D system is proposed.

In conclusion, we note that a similar behavior of transformation of plasma oscillation spectrum has been recently discovered in paired electron wires with current coupling (see [13, 14]). Establishing a current coupling between two initially isolated electron wires leads to the formation of a broad electron channel with the equilibrium electron density that varies across the channel. In such an electron channel, the spectrum of plasma modes undergoes crowding when the current coupling between the electron wires goes down to zero. In the

absence of the current coupling, the spectrum of plasma oscillations is set in the system, which is characteristic of isolated electron wires. In contrast to [13, 14], in this paper the transition to the regime of isolated electron channels is accompanied by the spectrum crowding of plasma oscillations of the continuous 2D electron system with periodically modulated electron density.

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research, project no. 00-02-16440.

REFERENCES

1. J. P. Kotthaus, W. Hansen, H. Pohlmann, and M. Wassermeier, Surf. Sci. **196**, 600 (1988).
2. T. Demel, D. Heitmann, and P. Grambow, in *Proceedings of NATO ARW on Spectroscopy of Semiconductor Microstructures*, Ed. by G. Fasol, A. Fasolino, and P. Lugly (Plenum, New York, 1989), NATO ASI Series, Series B: Physics, Vol. 206, p. 75.
3. O. R. Matov, O. F. Meshkov, and V. V. Popov, Zh. Éksp. Teor. Fiz. **113**, 988 (1998) [JETP **86**, 538 (1998)].
4. V. B. Shikin, T. Demel', and D. Heitman, Zh. Éksp. Teor. Fiz. **96**, 1406 (1989) [Sov. Phys. JETP **69**, 797 (1989)].
5. V. Cataudella and V. M. Ramaglia, Phys. Rev. B **38**, 1828 (1988).
6. S. V. Meshkov, J. Phys.: Condens. Matter **3**, 1773 (1991).
7. B. P. van Zyl and E. Zaremba, Phys. Rev. B **59**, 2079 (1999).
8. R. W. Gruhlke, W. R. Holland, and D. S. Hall, Phys. Rev. Lett. **56**, 2838 (1986).
9. S. J. Allen, Jr., H. L. Störmer, and J. C. Hwang, Phys. Rev. B **28**, 4875 (1983).
10. J. Alsmeier, E. Batke, and J. P. Kotthaus, Phys. Rev. B **40**, 12 574 (1989).
11. M. V. Krasheninnikov and A. V. Chaplik, Zh. Éksp. Teor. Fiz. **88**, 129 (1985) [Sov. Phys. JETP **61**, 75 (1985)].
12. O. R. Matov, O. V. Polischuk, and V. V. Popov, Int. J. Infrared Millim. Waves **14**, 1455 (1993).
13. W. R. Frank, A. O. Govorov, J. P. Kotthaus, *et al.*, Phys. Rev. B **55**, 1950 (1997).
14. A. O. Govorov, W. R. Frank, and S. A. Studenikin, Fiz. Tverd. Tela (St. Petersburg) **40**, 542 (1998) [Phys. Solid State **40**, 499 (1998)].

Translated by A. Klimontovich

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 511–516.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 593–599.

SOLIDS
Electronic Properties

Magnetooptical Investigation of the Micromagnetic Structure and Magnetization Processes in $Co_{69}Fe_4Si_{12}B_{15}$ Amorphous Microwires

E. E. Shalygina[a, *], V. V. Molokanov[b], and M. A. Komarova[a]

[a] *Moscow State University, Vorob'evy gory, Moscow, 119899 Russia*

**e-mail: sahl@magn.phys.msu.su*

[b] *Baikov Institute of Metallurgy, Russian Academy of Sciences, Leninskiĭ pr. 49, Moscow, 117334 Russia*

Received November 12, 2001

Abstract—Magnetooptical investigation of the micromagnetic structure of $Co_{69}Fe_4Si_{12}B_{15}$ amorphous microwires 10–50 μm in diameter is carried out. The existence of domains with transverse circumferential magnetization is experimentally demonstrated in the near-surface region of microwires. The dependence of the width of circular domains on the length and diameter of wires is obtained. It is shown that the near-surface micromagnetic structure of amorphous wires is changed under a stretch-induced stress. It is proved that the magnetization reversal of microwires in a longitudinal magnetic field occurs due to the rotation of local magnetization vectors in circular domains. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

In spite of the fact that amorphous magnetic materials were discovered more than thirty years ago, the study of their structural, magnetic, and kinetic properties is still of interest. This is primarily associated with the fact that, quite recently, a giant variation in the microwave resistivity under a constant magnetic field (magnetic impedance) has been observed in soft amorphous magnetic materials fabricated in the form of ribbons and wires [1–5]. On the basis of this phenomenon, highly sensitive magnetic-field and voltage sensors and thin-film magnetoresistive probes have been developed. It is known [2, 3, 6] that the magnetic impedance is determined by the field-induced variation in the skin depth, which depends on the transverse (with respect to the applied magnetic field) magnetic permeability of a sample. Later on, it has been demonstrated [7] that the value of the magnetic impedance depends on the near-surface micromagnetic structure (equilibrium distribution of magnetization) of a ferromagnetic material. In view of this fact, the micromagnetic structure of amorphous ribbons and wires has become a subject of study by many researchers. Our attention has been focused on cobalt-doped amorphous microwires with negative magnetostriction λ_s, which are promising for applications. According to the existing views [2], there should exist domains with transverse circumferential orientation of the magnetization vector in the near-surface region of microwires with $\lambda_s < 0$. Certain experimental results corroborating this assumption were obtained in [8, 9]. The domain structure in cobalt-doped wires was observed by magnetooptic contrast in samples that were made semicylindrical by polishing; i.e., the original state of the material was certainly destroyed. In general, due to the small size of microwires, the study of their magnetic properties and the micromagnetic structure is associated with great difficulties. All the models proposed up to now for the distribution of magnetization in the bulk and the near-surface layer of microwires doped with either iron or cobalt are based on certain indirect data. In particular, they take into account internal stresses (longitudinal, radial, and circumferential) induced by the fabrication of amorphous wires and the magnetostriction constant and include the measurement of volume hysteresis loops and magnetization curves. Scanning Kerr microscopy, which has recently become especially popular, represents a direct method for investigating the near-surface micromagnetic structure of ferromagnetic materials. This method enables one to measure local magnetic properties and the magnetization components in the surface regions with areas of 1 μm^2 with a linear resolution of up to 0.2 μm.

The aim of the present paper is to investigate the near-surface micromagnetic structure and the magnetization reversal of amorphous $Co_{69}Fe_4Si_{12}B_{15}$ microwires by the method of scanning Kerr microscopy, as well as to study the effect of stretch-induced stresses on the local magnetic properties of microwires.

2. SAMPLES AND EXPERIMENTAL METHODS

Amorphous $Co_{69}Fe_4Si_{12}B_{15}$ microwires with a diameter of the metal part of 10–50 μm and a glass cladding of 5–10 μm were obtained by the modernized Taylor method [10]. The magnetostriction constant λ_s in these samples was on the order of -2×10^{-7}. The

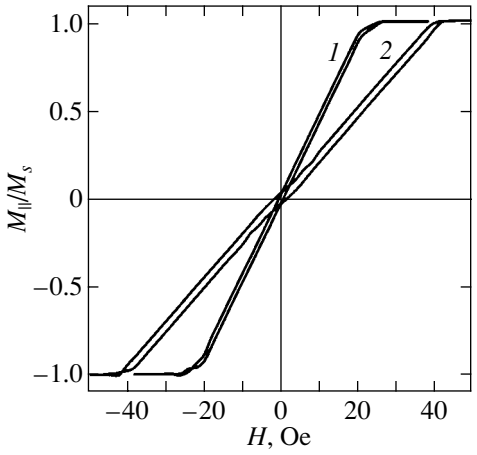


Fig. 1. Local hysteresis loops obtained in an axial magnetic field for the central part of a microwire 10 μm in diameter. Loops *1* and *2* were measured for 15- and 10-mm-long samples, respectively. Here, $M_{\|}(H)/M_s \propto \delta(H)/\delta_s$, where δ_s is the value of the transverse Kerr effect at $M = M_s$.

amorphous state of the samples was confirmed by the X-ray diffraction method. After removing the glass cladding, the microwires exhibited a cylindrical shape. The investigations were carried out on the samples with diameters D ranging within 5%. The microwires were cut into 6-, 10-, and 15-mm-long pieces.

The study of the micromagnetic structure of microwires was carried out on a magnetooptic micromagnetometer designed on the basis of a high-resolution MIM-8 microscope. The microscope provides a magnification of 1200. A light detector (multiplier phototube) was situated in the image plane of the microscope. The size of the microregion under examination was determined by the size of the slot placed before the light detector. The magnetization distribution in the sample and the local magnetic characteristics were measured while scanning the slot over the image of the object. In the case under consideration, the distribution of the tangential components of magnetization (parallel and perpendicular to the applied field **H**) and the local hysteresis loops were measured with the use of the transverse Kerr effect (TKE) while scanning a 0.5×2 μm^2 slot over the microwire image along its length L. Since we applied a modulation technique to increase the sensitivity when detecting magnetooptic signals, a sample was remagnetized by an ac magnetic field H with frequency of $f = 80$ Hz. The field **H** was applied along the centerline of the microwires. In this case, two signals were detected. The first signal U_- was proportional to the light intensity reflected from a nonmagnetized sample (I_0). The second signal was $U_\sim \propto \Delta = (I - I_0)$, where I is the intensity of light reflected from the magnetized sample. The difference Δ is attributed to the magnetooptic phenomenon associated with the magnetization reversal of the sample. The signals U_- and $U_\sim$ were measured by a dc microvoltmeter and a selective amplifier, respectively. The magnitude of the magnetooptic signal was determined by the relation

$$\delta = U_\sim/U_- = (I - I_0)/I_0.$$

An error in the detected values of δ was no greater than 5%. We measured the dependences $\delta(H, L)/\delta_s \propto M(H, L)/M_s$ (where δ_s is the value of the transverse Kerr effect for $M = M_s$, and M_s is the saturation magnetization). This allowed us to obtain information on the local magnetic properties and the micromagnetic structure of the samples under investigation. All the measurements were carried out on the central part of the samples in order to reduce end effects, in particular, to reduce the variation in the local demagnetizing factors. A tensile stress was applied along the centerline of the microwires.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Figure 1 shows typical local hysteresis loops observed in microwires of the same diameter but different lengths in an axial magnetic field. The parameters of local hysteresis loops (the initial permeability and the saturation field) measured in the central part of a microwire in different microregions differed by at most 10%. Figure 1 shows that the samples under study have hysteresis-free loops with a characteristic linear growth of magnetization with increasing magnetic field **H**. In a field perpendicular to the sample centerline, the samples exhibited virtually rectangular hysteresis loops. It should be pointed out that, in all the measurements, the external magnetic field was applied along the centerline of a wire parallel to the surface of the microregion under study. Within the experimental error, we did not observe a transverse Kerr effect for $\mathbf{H} \leq 1$ kOe in a field perpendicular to the surface of the microregion. According to current views [11], the linear dependence of the magnetization on the applied magnetic field and a rectangular hysteresis loop give evidence of the magnetization reversal of samples along easy and hard axes, respectively. In this case, due to the axial symmetry of the samples, the easy axis coincides with the circumferential direction. The explanation of this experimental fact is as follows. It is known that (see, for example, [8, 9, 12]) there is no magnetocrystalline anisotropy in amorphous alloys. The magnetic anisotropy in these materials is attributed to the magnetoelasticity. The energy of magnetoelastic anisotropy depends on the magnetostriction constant λ_s and internal manufacturing-induced residual stresses in a sample. During this process, the external cladding of a wire consolidates first, while its internal part consolidates under the stresses induced by the consolidated external cladding. The resulting residual stress in cobalt-doped amorphous wires with negative magnetostriction is responsible for the circumferential orientation of the easy axis in the near-surface layer.

Figure 1 also shows that, as the length L of a microwire decreases, the slope of the hysteresis loops decreases and the saturation field H_s increases. The variation in the diameter D also leads to the variation in the magnetic properties. To illustrate this fact, we present the experimental values of H_s as a function of the ratio L/D (curve *1* in Fig. 2). An increase in H_s due to a decrease in the length of microwires and/or increase in their diameter can be attributed to the stronger effect of the macroscopic demagnetizing field $H_N = -NM_s$ on the magnetic properties of the samples under investigation. Here, N is the macroscopic demagnetizing factor, and M_s is the saturation magnetization. The calculation of H_N for microwires of various sizes was carried out with the use of the expression for the macroscopic demagnetizing factor N given in [13] (see curve *2* in Fig. 2). One can see from Fig. 2 that the behavior of curves $H_s(L/D)$ is in a qualitative agreement with the calculated data; however, the calculated values of H_s are less than the experimental ones by a factor of 4–5. This quantitative discrepancy is likely to be associated with the fact that, for samples with negative magnetostriction, which are characterized by hysteresis-free loops in an axial magnetic field, the calculation of H_N should take into account the near-surface region with circular domains. The solution of this micromagnetic problem is a sufficiently complicated problem, and we could not find such a solution in the literature.

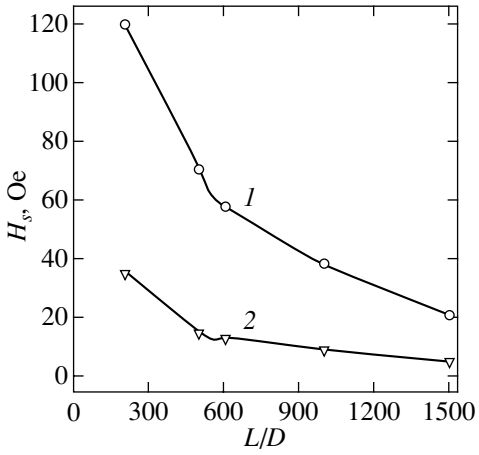


Fig. 2. Experimental (*1*) and calculated (*2*) values of the saturation field H_s versus L/D (L is the wire length and D is diameter).

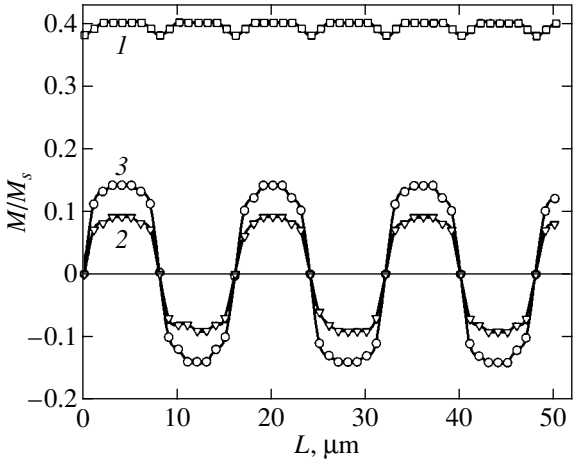


Fig. 3. Typical distributions of tangential magnetization components (*1*) parallel and (*2* and *3*) perpendicular to the external magnetic field directed along a microwire of length L. Curves *1–3* were measured in the central part of a 15-mm-long microwire 10 μm in diameter under a unipolar sinusoidal magnetic field (*1* and *2*) H = 8 and (*3*) 10 Oe.

Figure 3 presents typical distributions of tangential magnetization components that are parallel ($M_\parallel$) and perpendicular ($M_\perp$) to the external axial magnetic field; these distributions are observed in the central part of a microwire with respect to its length L. The functions $M_\parallel(L)$ were measured in the transverse configuration (the field H is parallel to L and perpendicular to the plane of incidence of light). It is known that the transverse Kerr effect is proportional to the variation (induced by the external magnetic field) in the tangential component of magnetization, which is perpendicular to the plane of incidence of light. To obtain information on the distribution of $M_\perp$ along L, we used a longitudinal configuration (H is parallel to L and to the plane of incidence of light). In this case, the TKE is proportional to the variation in the tangential component of magnetization, which is perpendicular to the field H and, hence, to the length of a microwire. Preliminary measurements showed that, in the case of a sinusoidal magnetic field, the first harmonic of the magnetooptic signal, proportional to $M_\perp$, vanishes. The dependences $M_\perp(L)$ (and, for the sake of uniformity, the distributions of the component $M_\parallel$ along L) were measured under the magnetization reversal of samples with the use of a unipolar sinusoidal magnetic field. Figure 3 shows that the component $M_\parallel$ has the same sign along L, while $M_\perp$ exhibits oscillating alternating behavior.

To explain these experimental facts, we analyzed the shape of magnetooptic signals, taking into account different mechanisms of the magnetization reversal of a sample. The results of this analysis are shown in Fig. 4. We considered both longitudinal and transverse orientations of samples with respect to the plane of incidence of light, as well as the behavior of magnetization under a sinusoidal unipolar magnetic field in different domains, which are denoted in Fig. 4 by Roman numerals I, II, III, IV, V, and VI. We found that the transverse configuration admits an alternating distribution of the component $M_\perp$ along L only when the local magnetization vector $\mathbf{M}_s$ in different microregions (domains) in the initial state is directed at angle $\pm\theta$ relative to the sample axis (state 0), and that the magnetization reversal of these microregions is due to the rotation of the vector $\mathbf{M}_s$ from state 1 to state 2 (see Fig. 4a). In this

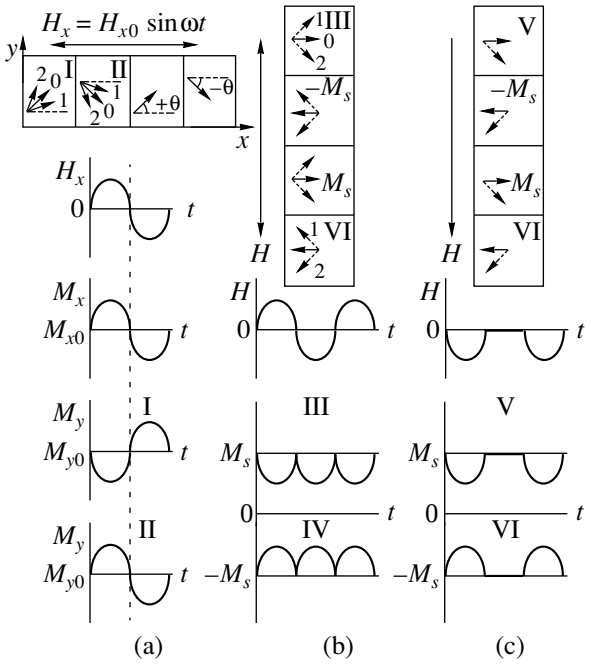


Fig. 4. The shapes of magnetooptic signals arising under the magnetization reversal of microregions of a sample due to the rotation of local magnetization vectors. (a) External sinusoidal magnetic field is parallel to the length of a sample and perpendicular to the incidence plane of light; angles $\pm\theta \neq \pm 90°$ are measured from the sample centerline and determine the orientation of magnetization in adjacent domains. (b) External sinusoidal magnetic field and the wire are parallel to the incidence plane of light; $\pm\theta = \pm 90°$. (c) Unipolar magnetic field is parallel to the wire and the incidence plane of light; $\pm\theta = \pm 90°$. The orientation of magnetization vectors M_s in domains of the types I–IV for $\mathbf{H} = 0$, $+\mathbf{H}$, and $-\mathbf{H}$ is denoted by 0, 1, and 2, respectively. The orientation of vectors M_s in domains V and VI for $\mathbf{H} \neq 0$ is shown by a dashed lines.

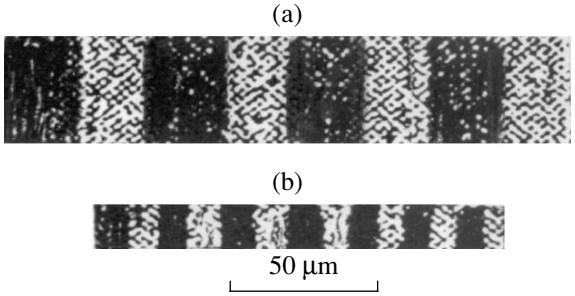


Fig. 5. Magnetic domain pattern observed by magnetooptic contrast in microwires (a) 50 and (b) 20 μm in diameter under $H = 0$. The length of the samples is 15 mm.

case, $M_\perp$ can be detected by measuring the TKE for the ±45-degree polarization of light ($\delta^{\pm 45°}$). According to our earlier calculations and experimental investigations [14, 15], there should exist a meridional intensity effect δ_{MIE} that is odd with respect to the polarization angle of light and proportional to the magnetization component parallel to the plane of incidence of light (respectively, perpendicular to the sample centerline). Thus, by measuring $\delta^{\pm 45°} = \delta_{TKE}(M_\parallel) \pm \delta_{MIE}(M_\perp)$, one can determine $\delta_{MIE}(M_\perp) = (\delta^{+45°} - \delta^{-45°})/2$ for various microregions of the wire under investigation.

The analysis of magnetooptic signals has also shown that, under the assumption that the magnetization reversal of the microregions are due to the motion of domain walls, or when θ does not change its sign, the curve of the magnetization distribution either has the same sign or vanishes. Moreover, we found out that, in the transverse and longitudinal configurations under the ±90-degree orientation of magnetization in adjacent domains relative to the sample centerline, the first harmonic of a magnetooptic signal, proportional to the magnetization component perpendicular to **H**, vanishes. The component $M_\perp$ can be measured by the TKE in the longitudinal configuration under a unipolar magnetic field (see Figs. 4b, 4c).

Thus, the experimental data obtained and the analysis of magnetooptic signals allow us to conclude that, in the microwires under investigation, there exist near-surface circular domains with a ±90-degree circumferential orientation of magnetization in adjacent domains with respect to the sample axis. Moreover, the measurements of the polar Kerr effect show that the magnetization component normal to the surface of a sample is absent. The features of the distribution of magnetization components along the centerline of microwires and the hysteresis-free loops provide evidence for the fact that the magnetization reversal of microwires in an axial magnetic field is mainly due to the rotation of the vectors of spontaneous magnetization.

The next step of our investigations consisted in the observation of the near-surface domain structure in microwires by magnetooptic contrast with the use of the meridional Kerr effect. In this case, the length L of microwires was perpendicular to the plane of incidence of light. Figure 5 presents typical domain patterns exhibited by the samples in zero field $\mathbf{H} = 0$. Actually, all the microwires exhibited distinct light and dark strips perpendicular to L. The size of the strips (or circular domains) depends on the length and diameter of the microwires. The fact that magnetooptic contrast becomes weaker and then vanishes as the orientation of a microwire is changed from the longitudinal to the transverse direction relative to the plane of incidence of light deserves special attention. This result serves as an additional confirmation of the existence of near-surface circular domains with the ±90-degree orientation of the magnetization vector in adjacent domains.

Obviously, the width W of circular domains can be determined from the distance between the zeros of the alternating curves $M_\perp(L)$ (Fig. 3) or from the domain patterns observed in the microwires (Fig. 5). Figure 6 shows the dependence of W on the diameter of microwires for a fixed sample length. The comparison of the

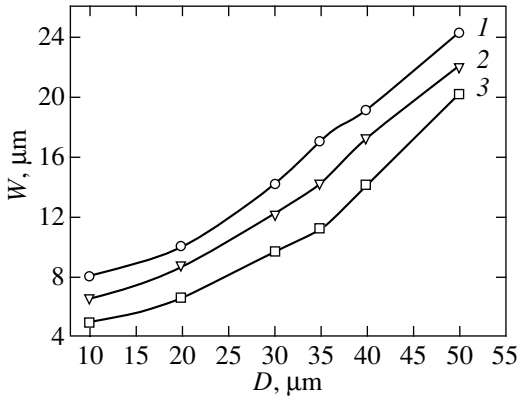


Fig. 6. The width of a circular domain as a function of the diameter of microwires of length 15 (*1*) , 10 (*2*), and 6 mm (*3*) obtained from the distribution of the magnetization component perpendicular to the axial magnetic field.

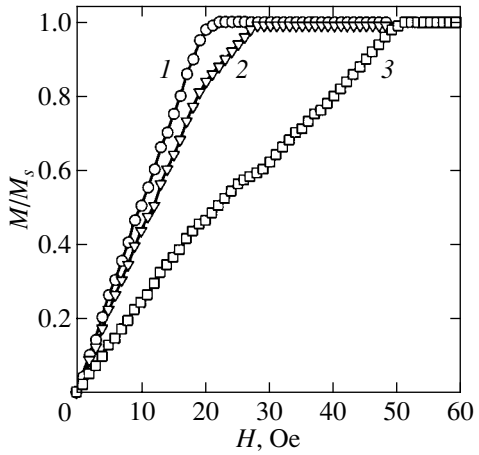


Fig. 7. Local magnetization curves $M_{||}(H)/M_s$ observed in a 15-mm-long microwire 10 μm in diameter. Curves *1*, *2*, and *3* were obtained under tensile stresses of σ = 0, 15, and 30 MPa, respectively.

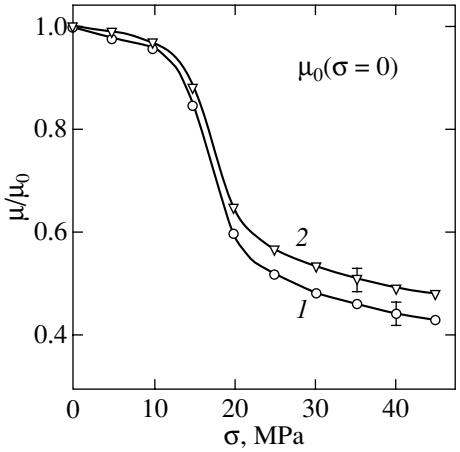


Fig. 8. Reduced initial permeability μ/μ_0 as a function of tensile stresses σ, obtained for 15-mm-long microwires 10 and 30 μm in diameter (curves *1* and *2*, respectively). Here, μ_0 is the value of μ for σ = 0.

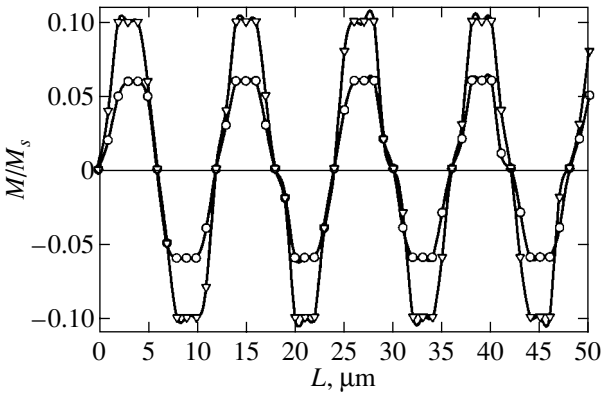


Fig. 9. Distributions of magnetization components perpendicular to the field that are obtained for the central part of a 15-mm-long microwire 10 μm in diameter in a unipolar axial field H = 8 Oe (curve *1*) and 10 Oe (curve *2*) in the presence of a tensile stress of σ = 15 MPa.

values of W obtained from the distribution of magnetization and from the domain patterns shows that the difference between these values of W is no greater than 10%.

It is well known that the micromagnetic structure of amorphous materials can substantially be changed under external effects. In this paper, we investigated the effect of tensile stresses on the local magnetic properties and the micromagnetic structure of amorphous microwires. We found that the local magnetization curves and, hence, the parameters of local hysteresis loops of the microwires are changed under stretch-induced stresses (see Fig. 7). We found that the initial permeability μ decreases and the saturation field increases as tensile stress σ increases. Figure 8 represents the reduced initial permeability μ/μ_0 as a function of stress σ for 15-mm-long microwires with diameters of 10 and 30 μm. Here, μ_0 is the value of μ for σ = 0. The observed decrease in μ/μ_0 as σ increases can be attributed to the fact that a circumferential magnetic anisotropy becomes stronger, which is characteristic of samples with negative magnetostriction. The difference between the curves $\mu/\mu_0(\sigma)$ for microwires of different diameters is due to the effect of the macroscopic demagnetizing field H_N on the local magnetic properties. As we have already mentioned above, for a sample of fixed length, H_N increases as the diameter of a microwire increases [12, 13]. As a result, other conditions being equal, the effect of tensile stresses on the local magnetic properties of microwires of large diameter decreases.

It is obvious that the tensile stresses should also affect the micromagnetic structure of microwires. Fig-

ure 9 shows the distributions $M_\perp(L)$ obtained for a 15-mm-long microwire with a diameter of 10 μm under a stress of $\sigma = 15$ MPa. The comparison of the curves in Figs. 3 and 9 shows that the width of circular domains decreases under tensile stresses ($W = 8$ and 6 μm for $\sigma = 0$ and 15 MPa, respectively). The variation in W under stress σ is also associated with the increase in the magnetic circumferential anisotropy. This result is in a good agreement with the calculations carried out in the theoretical study [16].

4. CONCLUSION

In this work, we obtained experimental results supporting the fact that scanning Kerr microscopy enables one to obtain detailed information on the equilibrium distribution of magnetization and the magnetization reversal processes in samples whose one or two dimensions lie in the micrometer range. A correct choice of methods for reversing magnetization in microwires and of magnetooptic effects allows one to compare the micromagnetic structures realized in samples in the initial state and in the presence of tensile stresses in zero and small ($H < H_s$) quasistatic magnetic fields, to analyze the effect of the sample sizes on their magnetic properties, and to determine the features of the magnetization reversal of the samples.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 01-03-32986 and 02-02-16627.

REFERENCES

1. R. S. Beach and A. E. Berkowitz, Appl. Phys. Lett. **64**, 3652 (1994).
2. L. V. Panina, K. Mohri, K. Bushida, and M. Noda, J. Appl. Phys. **76**, 6198 (1994).
3. L. V. Panina and K. Mohri, Appl. Phys. Lett. **65**, 1189 (1994).
4. F. L. A. Machado, C. S. Martins, and S. M. Rezende, Phys. Rev. B **51**, 3926 (1995).
5. M. Knobel, M. L. Sánchez, C. Gómez-Polo, *et al.*, J. Appl. Phys. **79**, 1646 (1996).
6. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8: *Electrodynamics of Continuous Media* (Nauka, Moscow, 1982; Pergamon, New York, 1975), p. 195.
7. L. V. Panina and K. Mohri, J. Magn. Magn. Mater. **157/158**, 137 (1996).
8. M. Takajo, J. Yamasaki, and F. B. Humphrey, IEEE Trans. Magn. **29**, 3484 (1993).
9. J. N. Nderu, J. Yamasaki, and F. B. Humphrey, J. Appl. Phys. **81**, 4036 (1997).
10. G. F. Taylor, Phys. Rev. **24**, 655 (1924).
11. E. C. Stoner and E. P. Wohlfarth, Philos. Trans. R. Soc. London, Ser. A **240**, 599 (1948).
12. J. Vázquez and A. P. Zhukov, J. Magn. Magn. Mater. **160**, 223 (1996).
13. A. P. Zhukov, M. Vázquez, J. Velázquez, *et al.*, J. Magn. Magn. Mater. **151**, 132 (1995).
14. G. S. Krinchik, E. E. Chepurova, and Sh. V. Égamov, Zh. Éksp. Teor. Fiz. **74**, 714 (1978) [Sov. Phys. JETP **47**, 375 (1978)].
15. E. E. Shalyguina, L. M. Bekoeva, and N. I. Tsidaeva, Sens. Actuators **81**, 216 (2000).
16. N. Usov, A. Antonov, A. Dykhne, and A. Lagar'kov, J. Magn. Magn. Mater. **174**, 127 (1997).

Translated by I. Nikitin

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 517–525.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 600–609.

SOLIDS
Electronic Properties

On the Emergence of Superconductivity and Hysteresis in a Cylindrical Type I Superconductor

G. F. Zharkov

Lebedev Physical Institute, Russian Academy of Sciences, Leninskiĭ pr. 53, Moscow, 119991 Russia
e-mail: zharkov@lpi.ru

Received November 5, 2001

Abstract—One-dimensional vortex-free solutions of the system of Ginzburg–Landau equations (the so-called precursor states) are studied. These states describe the emergence of superconductivity in a long cylindrical type I superconductor, which was initially in the supercooled normal state in a magnetic field, and are formed upon subsequent reduction of the external field. The precursor states are responsible for the magnetic hysteresis in type I superconductors (for which $\kappa < \kappa_c$, where $\kappa_c(R)$ is the critical value of the parameter κ in the Ginzburg–Landau theory, which is a function of radius). The range of fields is determined in which precursor states exist along with the Meissner state (and a hysteresis is possible) in the dependence of the cylinder radius R and parameter κ. © *2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The emergence (and degradation) of the superconducting state in a magnetic field in superconductors of various geometry was investigated on the basis of the Ginzburg–Landau macroscopic theory of superconductivity [1] by many authors [2–22]. Among other things, it was found that, for type II superconductors (with $\kappa > 1$) in the form of a long cylinder of radius R, a decrease in the external axial magnetic field H leads to the emergence of superconductivity from the normal state through a second-order phase transition in a certain field $H = H_2(m, \kappa, R)$ [4], when a small nonzero value of the modulus of the complex order parameter $\Psi = \psi e^{i\Theta}$ appears for the first time (ψ is the modulus, Θ is the phase, and m is the total number of vortices in the superconductor). In the case of large R and m, the field $H_2(m, \kappa, R)$ coincides with the field $H_{c3} = 1.69H_{c2}$ [3, 4] of emergence of surface superconductivity, where $H_{c2} = \phi_0/2\pi\xi^2$, ϕ_0 being the flux quantum and ξ being the coherence length. Recent numerical investigations [17–22] of one-dimensional (depending only on the radius) solutions of nonlinear Ginzburg–Landau equations for finite values of $\psi \sim 1$ in the case of the cylindrical geometry revealed that these solutions (with $\psi \sim 1$) exhibit a complex dependence on the parameters of the problem (m, κ, R, H). We can mention, for example, the existence of several branches of solutions in type II superconductors [18], the jumpwise rearrangement of these solutions upon a transition through the critical values of parameters [19, 21], the complex shape of the interface $S_{\text{I–II}}(\kappa, R)$ separating type I and II superconductors (this boundary or, which is the same, the critical value of $\kappa_c(R)$ depends on the cylinder radius [21] and does not coincide with the simple value of $\kappa_0 = 1/\sqrt{2}$ typical of the contact between two semi-infinite metallic superconducting (s) and normal (n) phases [1, 2]), and hysteresis phenomena in type II superconductors, associated with the existence of the "depressed" branch of solutions [22].

Among other things, it was noted in [22] that the superconducting state ($m = 0$) emerging in a supercooled (in the magnetic field) normal cylindrical sample of a type I superconductor is described by a special solution (a precursor) which preceeds complete expulsion of the field from the bulk of the sample and a jumpwise transition of the cylinder to the Meissner state. In this work, these solutions (precursors) are studied in greater detail. Since precursor states in type I superconductors (and hysteresis phenomena accompanying them) are manifested most strongly in the vicinity of the $S_{\text{I–II}}$ interface separating type I and II superconductors on the (κ, R) plane [21], we will henceforth consider the general case of arbitrary values of κ, which will allow us to describe the behavior of solutions upon a transition through the boundary $S_{\text{I–II}}$, which depends essentially on parameters R and κ.

Obviously, small-radius cylinders can contain only vortex-free states with $m = 0$. Here, we will confine our analysis to a detailed study of the properties of precisely such states. Among other things, it will be proved that the shape of the magnetization hysteresis loop for a cylinder with $m = 0$ is determined to a considerable extent by parameters R and κ. This can be used, in principle, for experimental determination of these parameters in mesoscopic superconductors. Some details of the picture, which will be obtained below, were earlier unknown and may be useful for discussing experiments with small-size (mesoscopic) superconductors [23–29].

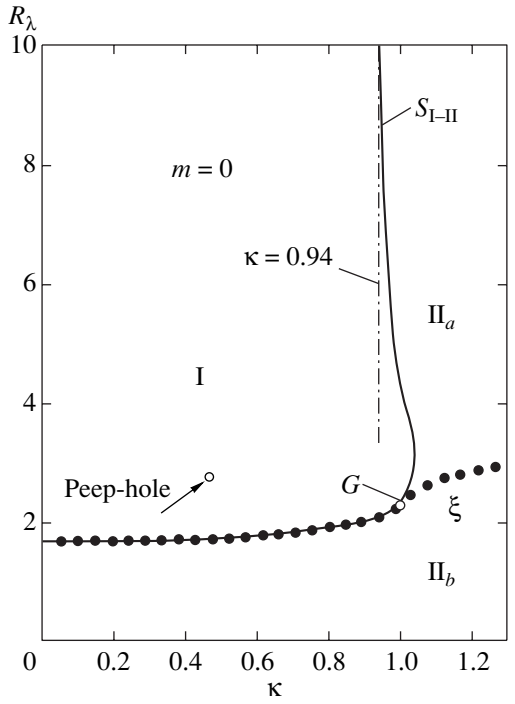


Fig. 1. Critical curves $S_{I–II}$ and ζ dividing the plane (R_λ, κ; $m = 0$) into three regions. In the field-increase regime, the degradation of the superconducting state in region I occurs through a first-order jump from the Meissner (M) state, $\psi \approx 1$, to the normal (n) state, $\psi \equiv 0$. In region II$_a$, the M state is first transformed jumpwise into the superconducting e state, which is destroyed finally through a second-order transition. In region II$_b$, superconductivity degradation (in the field-increase regime) occurs gradually (second-order phase transition), without jumps. In the field reduction regime, the restoration of the M state in region II$_a$ occurs through the e and d states and is accompanied by jumps and hysteresis. In region II$_b$, no hysteresis effects or jumps are possible. In region I (in the field reduction mode), the superconducting p state emerges from the supercooled n state through a second-order phase transition, after which a jump to the M state occurs and a hysteresis loop is formed. At point G ($k = 1$, $R_\lambda = 2.28$), the critical curves $S_{I–II}$ and ζ merge into one curve. Below the ζ curve, no hysteresis loop is formed. For $\kappa > 3.5$, curve ζ attains the constant value $R_\zeta \approx 3.6$. For $\kappa \ll 1$, we have $R_\zeta \approx 1.69$.

2. EQUATIONS

The Ginzburg–Landau macroscopic theory of superconductivity [1] leads to a system of two nonlinear equations for the order parameter Ψ and the vector potential A of a magnetic field. The equation for Ψ contains the coherence length ξ and the equation for A, the magnetic field penetration depth λ ($\lambda = \kappa\xi$, where κ is the parameter of the Ginzburg–Landau theory). These two lengths are equivalent, and any of them can be chosen as a unit of measurement. In this work, we take λ as the unit of measurements. We will study the vortex-free states ($m = 0$); in this case, we can set the phase Θ equal to zero and assume that the order parameter is a real-valued quantity, $\Psi = \psi$. In this case, the Ginzburg–Landau equations for an infinitely long cylinder in an axial external magnetic field H in the cylindrical system of coordinates (r, φ, z) can be written in the following dimensionless form:

$$\frac{d^2U}{d\rho^2} - \frac{1}{\rho}\frac{dU}{d\rho} - \psi^2 U = 0, \quad (1)$$

$$\frac{d^2\psi}{d\rho^2} + \frac{1}{\rho}\frac{d\psi}{d\rho} + \kappa^2(\psi - \psi^3) - \frac{U^2}{\rho^2}\psi = 0. \quad (2)$$

Here, $\rho = r/\lambda$ is the dimensionless radial coordinate, $U(\rho)$ is the dimensionless potential of the magnetic field,

$$A = \frac{\lambda\phi_0}{2\pi\lambda^2}\frac{U}{\rho}, \quad B = \frac{1}{r}\frac{d(rA)}{dr}, \quad b = \frac{1}{\rho}\frac{dU}{d\rho},$$

where $b = B/H_\lambda$ is the dimensionless field in the superconductor and $H_\lambda = \phi_0/2\pi\lambda^2$ is the unit of measurement of the field.

The boundary conditions to Eqs. (1) and (2) have the form

$$U|_{\rho=0} = 0, \quad \left.\frac{dU}{d\rho}\right|_{\rho=R_\lambda} = h_\lambda, \quad (3)$$

$$\left.\frac{d\psi}{d\rho}\right|_{\rho=0} = 0, \quad \left.\frac{d\psi}{d\rho}\right|_{\rho=R_\lambda} = 0. \quad (4)$$

Here, $R_\lambda = R/\lambda$ and $h_\lambda = H/H_\lambda$. The magnetic moment M of the cylinder (or the magnetization per unit volume) can be found from the formula

$$\bar{b} = h_\lambda + 4\pi M_\lambda, \quad \bar{b} = B_{av}/H_\lambda,$$

where B_{av} is the average magnetic field in the superconductor and $M_\lambda = M/H_\lambda$.

Obviously, solutions $U(\rho)$ and $\psi(\rho)$ to Eqs. (1)–(4) depend on three parameters: κ, R_λ, and h_λ. In order to find self-consistent solutions to Eqs. (1)–(4), use was made of the iterative procedure described in greater detail in [17]. This method is equivalent to the analogous numerical procedures used earlier [5–8]. However, in contrast to [5–8], where solutions were determined, as a rule, for several randomly distributed values of parameters κ, R, and h, we carried out a more detailed and systematic study of the solutions in a wide range of variation of the parameters κ, R_λ, and h_λ, which allowed us to discover some interesting features that had remained unnoticed in a less detailed analysis. Some of the results of this study will be described below.

3. NUMERICAL RESULTS

In order to obtain a compact description of the results of numerical calculations, we will first consider the plane of the variables (κ, R_λ). Each point on this

plane corresponds to a solution $\psi(\rho; h_\lambda)$ and $U(\rho; h_\lambda)$ to the system (1)–(4). A certain idea concerning the properties of this solution (for given κ and R_λ) can be grasped from an analysis of the dependence of the order parameter ψ_0 at the center of the cylinder as a function of the magnetic field, $\psi_0 = \psi(0; h_\lambda)$. Similar information on the properties of the solution at the point (κ, R_λ) is contained in the dependence of the magnetic moment $-4\pi M_\lambda$ of the system on the external field. It is convenient to imagine mentally that any point on the (κ, R_λ) plane contains a "hole" through which the dependence of ψ_0 (and $-4\pi M_\lambda$) on the field h_λ can be "seen." Comparing these dependences, we can establish that there exist three regions on the (κ, R_λ) plane with qualitatively different behavior of $\psi_0(h_\lambda)$ as well as $M_\lambda(h_\lambda)$. These regions are denoted by I, II_a, and II_b in Fig. 1.

The meaning of division of the plane (κ, R_λ) into separate regions will be clarified below. However, before making a commentary on Fig. 1, we note that the superconducting state (at temperature $T < T_c$) can be obtained in two different ways: either in zero external magnetic field at the initial instant with a subsequent increase in H (field amplification regime) or by reducing a strong magnetic field H in which the metal was initially in the normal state (field reduction regime). These two regimes generally correspond to different solutions for the same value of the field h_λ. While seeking the solutions to Eqs. (1)–(4) corresponding to the field amplification mode, the trial function for the order parameter at the beginning of the iterative procedure was specified in the form $\psi(\rho) \sim 1$. Solutions in the field reduction mode correspond to the initial trial function $\psi(\rho) \sim 0.01 \ll 1$. Figures 2–4 show examples of the dependences $\psi_0(h_\lambda)$ and $-4\pi M_\lambda(h_\lambda)$, while examples of coordinate dependences of the solutions $\psi(\rho)$ and $b(\rho)$ emerging in different regimes are presented in Fig. 5.

If a hole is made on the line $R_\lambda = 6$ at point $\kappa = 0.95$ (in region I), we can see that the solutions appearing in the amplification mode for different fields h_λ correspond to a stable Meissner state (with $\psi_0 \approx 1$). These solutions correspond to the solid line in Fig. 2a. As the field attains the value h_1, the Meissner solution becomes absolutely unstable, and the cylinder passes jumpwise to the normal state ($\psi \equiv 0$). Since the order parameter ψ_0 in the Meissner state remains finite up to the jump point h_1, the value of the critical field $h_1(\kappa, R_\lambda)$ cannot be determined from the linearized theory [4, 9], which can be applied only if the condition $\psi_0 \ll 1$ is satisfied (see [22] for details). The jump in the order parameter ψ_0 at point h_1 is denoted by δ_1. For $h > h_1$, there are no other solutions except the normal solution ($\psi \equiv 0$).

If we now seek solutions in the field reduction mode (for $h_\lambda < h_1$), the normal solution ($\psi \equiv 0$) remains stable (relative to small spatial perturbations) up to the point h_p at which a small ($\psi_0 \ll 1$) nucleus of the superconducting state appears (precursor, or p state). In the interval of fields $\Delta_n = h_1 - h_p$, there exist two stable (in the above sense) solutions: the Meissner state and the supercooled (in a magnetic field) normal state. Obviously, the supercooled normal state is metastable since the free energy is lower in the Meissner state (with $\psi \approx 1$). Since the amplitude of the emerging p state is small, the value of the critical field $H_p(\kappa, R_\lambda)$ can be found from the linearized theory [4, 9], whence it follows that $H_p = \phi_0/2\pi\xi^2 \equiv H_{c2}$ for $R_\lambda \gg 1$.

Upon a further decrease in the field ($h_\lambda < h_p$), the amplitude of the emerging superconducting p state increases (see the dashed curve in Fig. 2a) up to point h_r at which the restoration of the Meissner state occurs jumpwise (the amplitude of the jump is δ_r). In the field interval $\Delta_p = h_p - h_r$, there exist two stable (relative to small perturbations) superconducting states: the Meissner (M) state and the p state. For $h_\lambda < h_r$, only one stable Meissner state exists. The precursor state (as well as the supercooled n state) is metastable; it describes the possibility of a hysteresis in the field reduction mode (the hysteresis loop is indicated by arrows). The total width of the hysteresis loop is

$$\Delta_{pn} = \Delta_p + \Delta_n = h_1 - h_r.$$

It should be emphasized that the field h_r and amplitude ψ_r at the transition point cannot be determined with the help of the linearized theory [4, 9] (see the discussion of this problem in [22]).

A similar pattern emerges in an analysis of magnetization $(-4\pi M_\lambda)$ as a function of the field (Fig. 2b). In this case also, we have the supercooled n state (in the field interval Δ_n); the hysteresis loop, $\Delta_{pn} = \Delta_p + \Delta_n = h_1 - h_r$, associated with the existence of the n and p states; and the magnetization jump (δ_r) during the restoration of the M state. Such a pattern is typical of type I superconductors.

If we make a hole at the point $R_\lambda = 6$, $\kappa = 0.98$ lying in the II_a region in Fig.1, the emerging pattern is essentially different (Fig. 2e). Here, the Meissner state in the field amplification regime also becomes absolutely unstable in the field $h_1(\kappa, R_\lambda)$, but the jump δ_1 occurs not to the state with $\psi \equiv 0$ (as in type I superconductors), but to a special superconducting state with a suppressed order parameter, viz., the e (edge-suppressed) state typical of type II superconductors [18, 21, 22]. For the e state, the amplitude of the order parameter ψ_0 decreases smoothly upon an increase in h_λ and vanishes completely in the field $h_\lambda = h_2(\kappa, R_\lambda)$. For $R \gg 1$, the field $H_2(\kappa, R)$ coincides with $H_{c2} = \phi_0/2\pi\xi^2$. In the field interval $\Delta_e = h_2 - h_1$, the dependence $\psi_0(h_\lambda)$ has a smoothly decreasing tail corresponding to the e state.

In the field-reduction mode, the superconducting e state appears again in the field h_2 (Fig. 2e) and continues

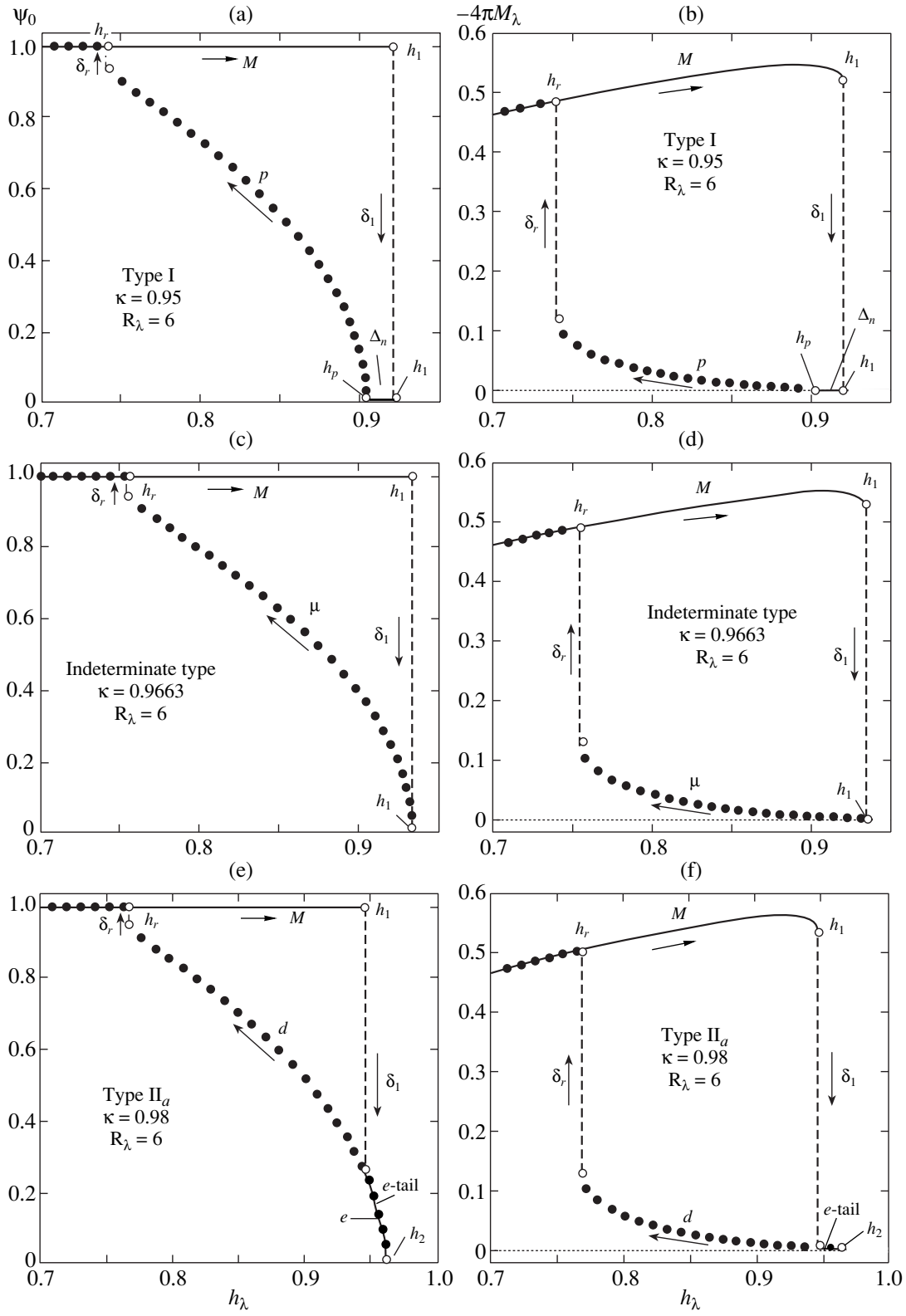


Fig. 2. Various types of superconducting states (M is the Meissner state, p is the precursor, μ is the marginal state, e is the edge-suppressed state, and d is the depressed state) existing near the $S_{\text{I–II}}$ curve in Fig. 1 for $R_\lambda = 6$. The dependences of the amplitude of states (ψ_0) and magnetization ($-4\pi M_\lambda$) on the field h_λ are plotted for different values of κ. Solid curves correspond to the field amplification mode, and dotted curves to the field reduction mode. The jumps δ_1 between states occur at points h_1 in the field amplification mode, and jumps δ_r take place at points h_r in the field reduction mode. The precursor state (p) emerges at point h_p (Figs. 2a and 2b) from the supercooled normal state ($\Delta_n = h_1 - h_p$ is the width of the supercooling region).

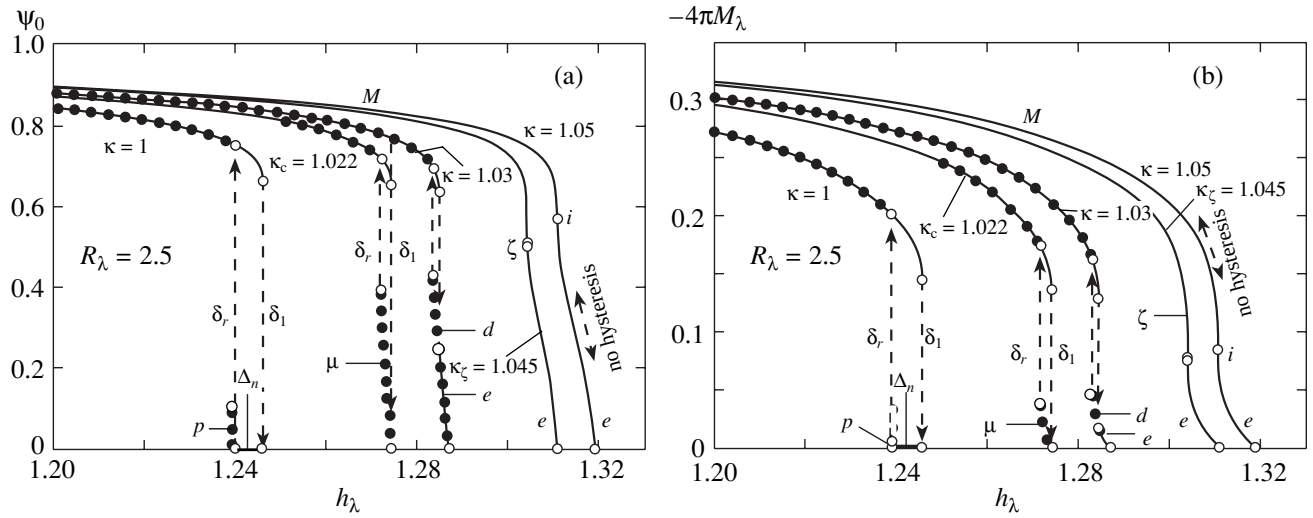


Fig. 3. Dependence of the amplitude of solutions (ψ_0) and magnetization ($-4\pi M_\lambda$) in different states (notation is the same as in Fig. 2) on field h_λ for $R_\lambda = 2.5$ and several values of κ (figures on the curves). Solid curves correspond to the field amplification mode, and dotted curve, to the field reduction mode. For $\kappa = 1$, the solutions (*M*, *p*, and *n*) belong to region I in Fig. 1. For $\kappa_c = 1.022$, the precursor state possesses the maximum amplitude $\psi_r(h_r) = 0.3855$, $h_r = 1.2716$ (μ state). Solutions with $\kappa = 1.03$ lie in region II_a (see Fig. 1); here, there are two singular points, h_1 and h_r, at which the jumps δ_1 and δ_r occur and a hysteresis loop may be formed. For $\kappa = \kappa_\zeta = 1.045$, these two singularities merge at the point $h_\zeta = 1.3037$, where $dM_\lambda/dh_\lambda = \infty$. For $\kappa > \kappa_\zeta$, the values of $dM_\lambda/dh_\lambda < \infty$, and there is no hysteresis (type *i* curves).

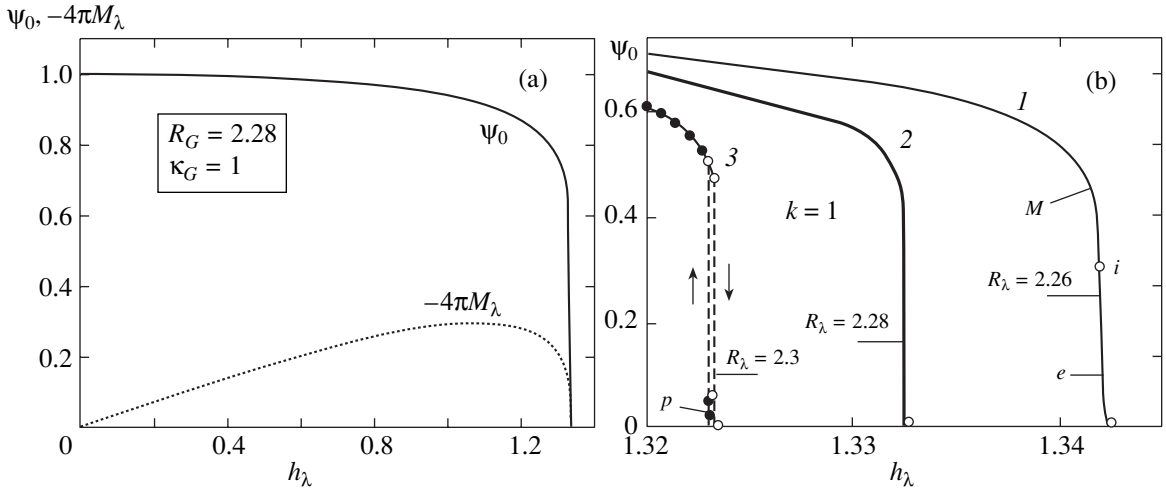


Fig. 4. (a) Dependence of the order parameter ψ_0 (solid curve) and magnetization $-4\pi M_\lambda$ (dotted curve) on the field h_λ at the critical point *G* ($\kappa_G \approx 1$, $R_G \approx 2.28$). (b) The behavior of $\psi_0(h_\lambda)$ (on a magnified scale) near the points of termination of solutions for $R_\lambda = 2.30$ ($R_\lambda > R_G$, region I) (curve *1*), 2.28 ($R_\lambda \approx R_G$) (*2*), and 2.26 ($R_\lambda < R_G$, region II_b in Fig. 1) (*3*). It can be seen that, for $R_\lambda < R_G$, the nonhysteretic curve has the point of inflection *i* with a finite derivative and with a tail of *e* states, while the *p* state and a hysteresis loop appear for $R_\lambda > R_G$.

to exist up to point h_1. In the field interval $\Delta_e = h_2 - h_1$, in both regimes (field enhancement and reduction), there exists a unique *e* solution, and no hysteresis is possible. As the field decreases further ($h_\lambda < h_1$), the *e* solution is smoothly transformed into a new *d* (depressed) state [22] which is preserved up to the point $h_r(\kappa, R_\lambda)$ at which the Meissner solution is restored jumpwise (with jump amplitude δ_r). In the field interval $\Delta_d = h_1 - h_r$, there exist two stable solutions (*M* and *d* states); for this reason a hysteresis loop of width Δ_d may be formed.

The magnetic moment $-4\pi M_\lambda(h_\lambda)$ exhibits a similar behavior (Fig. 2f). In this case, the magnetization curve also has a tail, the supercooled *n* state is absent, and there exists a hysteresis loop associated with the *d* state,

which can be treated as the supercooled *e* state. Such a pattern is typical of type II superconductors.

As the parameter κ approaches the critical line $S_{\text{I–II}}$ separating type I and II superconductors (solid curve in Fig. 1, which can be denoted by $\kappa_c(R_\lambda)$), the field intervals $\Delta_n = h_1 - h_p$ (in which the supercooled normal state is possible in type I superconductors) and $\Delta_e = h_2 - h_1$ (where the magnetization curve for type II superconductors acquires a tail) decrease and vanish exactly at the critical line. In this case (for $R_\lambda = 6$ and $\kappa_c = 0.9663$), the solution depicted in Figs. 2c and 2d appears. In this case, one cannot tell the type of superconductor (indeterminate type). The precursor solution can be termed marginal (μ) in this special case; it emerges without preliminary formation of the supercooled normal state. All marginal μ states lie on the critical curve $S_{\text{I–II}}$ in Fig. 1. The amplitude of the marginal *p* state at the jump point, $\psi_r = \psi_0(h_r, \kappa_c)$, is the largest among other *p* states existing in region I.

Figure 3 shows what happens near the critical curve $S_{\text{I–II}}$ for a smaller radius of the cylinder, $R_\lambda = 2.5$. It can be seen that (in accordance with Fig. 1) the *p* solution in region I has the maximum amplitude ψ_r on the critical curve $S_{\text{I–II}}$ ($\kappa_c = 1.022$, $\psi_r = 0.3855$, $h_r = 1.2716$; this is a marginal μ state). For $\kappa < \kappa_c$, the amplitude ψ_r decreases rapidly ($\psi_r = 0.095$ for $\kappa = 1$ and $\psi_r = 0.001$ for $\kappa = 0.9$). This allows us to state that, for $\kappa < \kappa_c(R_\lambda)$, the Meissner superconducting state (with $\psi_0 \approx 1$) is restored from the supercooled normal state (existing in the interval Δ_n), as a rule, through a "nearly first-order" phase transition (the jump δ_r occurs from the *p* state with $\psi_r \ll 1$).

For $\kappa > \kappa_c(R_\lambda)$ (in region II_a), the supercooled normal state and the corresponding hysteresis are absent, but (see the curve $\kappa = 1.03$) the *e* and *d* superconducting states are formed (*e* branch in the field amplification mode and *d* branch in the field reduction mode) and the hysteresis associated with the simultaneous existence of superconducting *d* and *M* states becomes possible. For $\kappa > 1.03$, the field interval $\Delta_d = h_1 - h_r$ in which *d* solutions exist and hysteresis is possible (see Figs. 2e and 2f) decreases and vanishes for $\kappa_\zeta = 1.045$. At the point $\kappa = \kappa_\zeta$, the jumps between the branches also disappear ($\delta_1 = \delta_r = 0$) and $dM_\lambda/dh_\lambda = \infty$ at this point (Fig. 3b). For $\kappa > \kappa_\zeta$ (in region II_b), the magnetization curve displays only the point of maximum descent (inflection point *i*) with a finite value of the derivative dM_λ/dh_λ (see the curve with $\kappa = 1.05$). In this case, the superconducting solutions (*M* and *e*) corresponding to the field reduction and enhancement modes merge into a single branch, and hysteresis is impossible; however, the magnetization curve has two distinguishable regions: in front of the point of inflection (*M* state) and behind it (*e* state).

The critical values of κ_ζ corresponding to different values of R_λ are depicted in Fig. 1 by the dotted curve ζ. Above this curve (in region II_a), hysteresis is possible (*d* solutions exist), while, below this curve (in region II_b), hysteresis is absent. At point *G* ($R_G \approx 2.28$ for $\kappa \approx 1$), the critical curves $S_{\text{I–II}}$ and ζ merge into one; for $R_\lambda < R_G$, there exists a single critical curve above which (in region I) the processes of superconductivity degradation (and restoration) are accompanied by first-order phase transitions (with jumps δ_1 and δ_2), while, below this curve (in region II_b), smooth second-order phase transitions take place. Thus, for a small radius of the cylinder, all type I superconductors (with $\kappa < \kappa_c(R_\lambda)$) become in fact type II superconductors.

Ginzburg [30], who noted that a type I superconductor with a small radius (with $\kappa \ll 1$) behaves in a magnetic field as a type II superconductor, arrived at the same conclusion (on the basis of different considerations). Consequently, point *G* can be referred to as a Ginzburg bicritical point. On the (κ, R_λ) plane, two critical curves, $S_{\text{I–II}}$ and ζ, converge at point *G* in contrast to the Landau tricritical point, at which three critical curves corresponding to supercooled, equilibrium, and superheated states converge on the plane of parameters (H, R) (see, for example, [7]). The lower part of curve ζ (lying below point *G*) determines the radius of the cylinder for which a type I superconductor becomes a nonhysteretic type II superconductor.

Figure 4a shows the dependence of the order parameter ψ_0 (and magnetization $-4\pi M_\lambda(h_\lambda)$) on field h_λ at point *G* ($\kappa = 1$, $R_\lambda \approx 2.28$). It can easily be verified that hysteretic *d* states (which are present in Fig. 3a for $R_\lambda = 2.5$) are not observed any longer (since point *G* lies on the nonhysteretic ζ curve). For the same reason, hysteretic *p* states (to be more precise, μ states since point *G* lies on the critical curve $S_{\text{I–II}}$), as well as the *e* states, must also be absent. As a result, the dependence $\psi_0(h_\lambda)$ at point *G* must have the form of a single-valued nonhysteretic curve consisting only of the *M* states, but with a vertical tangent line at the transition point, where ψ_0 vanishes.

All this is illustrated in Fig. 4b, showing on a magnified scale the dependence $\psi_0(h_\lambda)$ in the immediate vicinity of point *G* ($R_\lambda = 2.28$, curve *2*) as well as at points $R_\lambda = 2.3 > R_G$ (curve *1*, region I) and $R_\lambda = 2.26 < R_G$ (curve *3*, region II_b).

Figures 1–4 illustrate the behavior of vortex-free ($m = 0$) solutions to the Ginzburg–Landau equations as functions of parameters κ, R_λ, and h_λ. Figure 5 shows the behavior of the self-consistent solutions $\psi(x)$ and $b(x)$ as functions of the coordinate $x = r/R$. Figure 5a illustrates the form of the order parameter $\psi(x)$ for *p* solutions (precursors of transition to the Meissner state in region I in Fig. 1) for $R_\lambda = 6$. The marginal μ solution lies on the critical curve $S_{\text{I–II}}$ (with $\kappa_c = 0.9663$), where it attains the maximum amplitude ψ_r at the jump point ($h_r = 0.7548$, $\psi_r = 0.9441$), after which it becomes unstable and is transformed into the *M* solu-

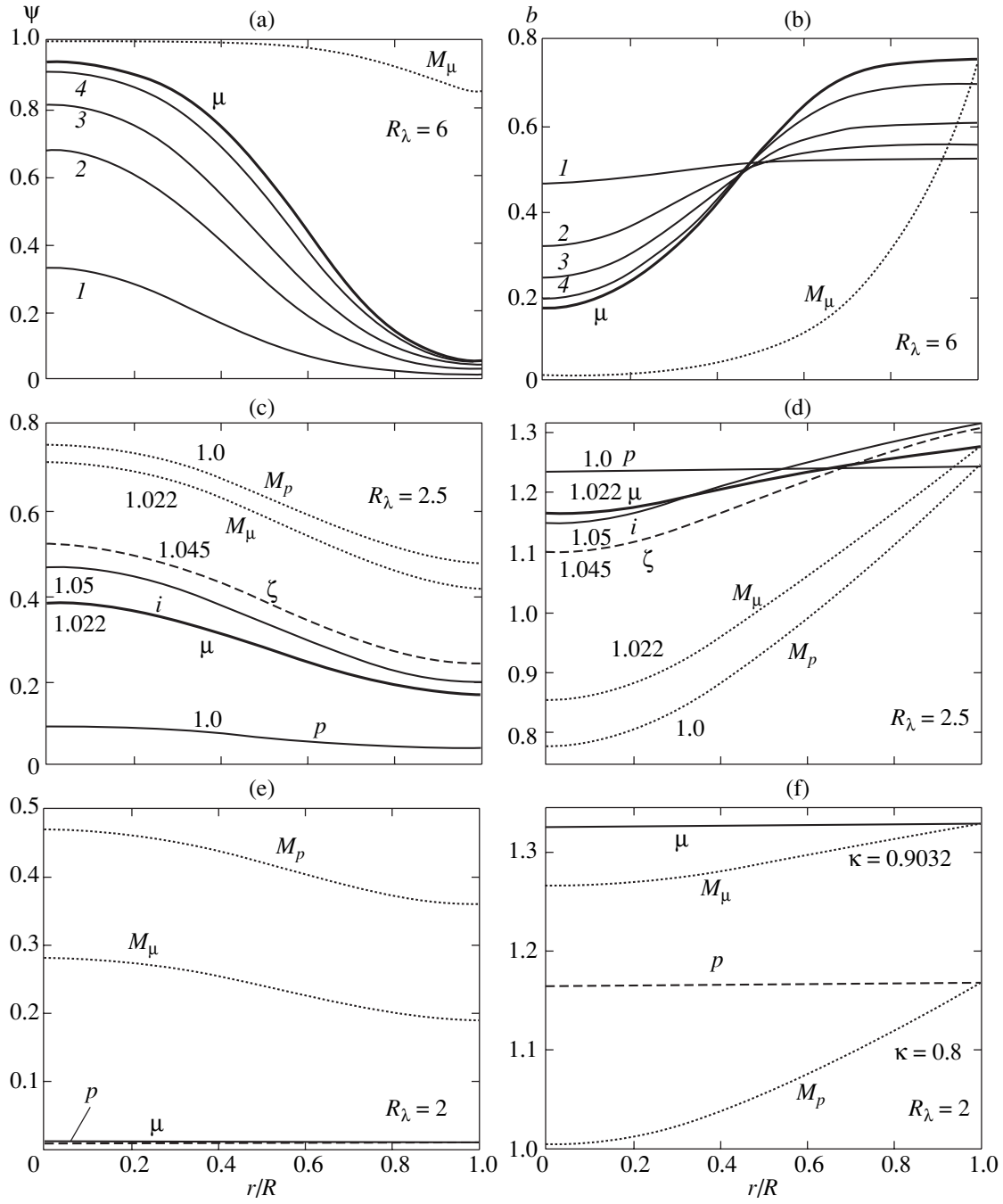


Fig. 5. Examples of the coordinate dependences of solutions $\psi(x)$ and $b(x)$ ($x = r/R$). (a, b) $R_\lambda = 6$. Precursor solutions (p) exist in region I along with the Meissner (M) solutions. Curve μ describes the marginal solution emerging for the critical value $\kappa_c = 0.9663$ (see Figs. 2c and 2d) and having the maximum amplitude $\psi_r = 0.9441$ in the field $h_r = 0.7548$. For $h_\lambda = 0.7547 < h_r$, a jump to the M state occurs. As the value of k decreases to $\kappa < \kappa_c$, the amplitude of the p solution decreases rapidly at the jump point ($\psi_r = \psi(h_r)$): $\kappa = 0.9$, $h_r = 0.6966$, $\psi_r = 0.9135$ (curve *1*); $\kappa = 0.8$, $h_r = 0.6044$, $\psi_r = 0.8150$ (curve *2*); $\kappa = 0.75$, $h_r = 0.5538$, $\psi_r = 0.6784$ (curve *3*); and $\kappa = 0.72$, $h_r = 0.5193$, $\psi_r = 0.3321$ (curve *4*); for $\kappa = 0.71$, $h_r = 0.5057$ and $\psi_r < 1 \times 10^{-4}$. Meissner solutions at the transition points h_r have the form $\psi(x) \approx 1$; these solutions are similar to M_μ and are not shown in the figure. Figures (c, d) ($R_\lambda = 2.5$) and (e, f) ($R_\lambda = 2$) are explained in the text.

tion (dotted curve). As we move from the boundary $S_{\text{I–II}}$ deeper into region I, the amplitude of the p solutions decreases rapidly (Fig. 5a shows only the p solutions at the points of their transformation into M solutions; the corresponding M solutions are not shown).

Figure 5b shows similar dependences for the magnetic field distribution over the cylinder radius, $b(x) = B(x)/H_\lambda$.

Figure 5c illustrates the behavior of the order parameter $\psi(x)$ for $R_\lambda = 2.5$. Here, the marginal μ solu-

tion corresponds to $\kappa = 1.022$ ($h_r = 1.2716$, $\psi_r = 0.3855$), the jump from the μ state occurring to the corresponding Meissner M_μ state. The p solution in region I (for $\kappa = 1.0$ and $h_r = 1.2393$) and the corresponding M_p state are also shown. In addition, the ζ solution lying (see Fig. 1) at the boundary between regions II_a and II_b (for $\kappa_\zeta = 1.045$ and $h_\lambda = 1.3037$, where the derivative $dM_\lambda/dh_\lambda \longrightarrow \infty$) is also given, as well as the i solution (see Fig. 3) from region II_b (for $\kappa = 1.05$ at the point $h_\lambda = 1.311$, where the magnetization $M_\lambda(h_\lambda)$ has a point of inflection i with a finite value of the derivative dM_λ/dh_λ (see Fig. 3)). In states of the type ζ and i, hysteresis is impossible in view of single-valuedness of these solutions (see Fig. 3). Solutions for the field $b(x)$ in the case when $R_\lambda = 2.5$ are shown in Fig. 5d.

Solutions $\psi(x)$ and $b(x)$ for $R_\lambda = 2.0$ (region I) are depicted in Figs. 5e and 5f. Here, the μ state corresponds to $\kappa_c = 0.9032$ ($h_r = 1.3301$); the p state existing for $\kappa = 0.8$ (at point $h_r = 1.1666$) is also shown. It can be seen from Figs. 5a, 5c, and 5e that the amplitude of p states decreases rapidly with increasing distance from the boundary $S_{I–II}$ (in this case, the supercooled n state is transformed onto the superconducting Meissner state through a "nearly first-order" phase transition). It can be seen, by the way, that the amplitude of the p state always attains its maximum at the midpoint of the cylinder; i.e., the superconducting state emerges in the bulk of the cylinder and not at its surface. In this connection, see [31–37], where the emergence of superconductivity in plane-parallel plates is interpreted in a different manner.

4. CONCLUSIONS

In this study, the main attention is paid to an analysis of the behavior of cylindrical type I superconductors (with small values of parameter κ) in an arbitrary magnetic field H. We have obtained self-consistent solutions to the Ginzburg–Landau equations for vortex-free states ($m = 0$) typical of cylinders with small radii R. The boundary $S_{I–II}$ separating the regions of behavior of the order parameter and magnetization of the cylinder, typical of type I and II superconductors, has been determined. It is shown that the field dependence of magnetization $-4\pi M_\lambda(h_\lambda)$ in type I superconductors exhibits a hysteresis loop associated with the existence of a supercooled normal phase and with the emergence (in the field-reduction regime) of special p states (precursors, see Fig. 3) preceding the complete expulsion of the field from the sample and a transition of the cylinder to the Meissner state in the field $h_r(\kappa, R_\lambda)$. This Meissner state is destroyed through a first-order jump (in the field amplification regime) in the field $h_1(\kappa, R_\lambda)$ and is restored also jumpwise (in the field reduction regime) in the field $h_r(\kappa, R_\lambda)$. In the vicinity of the boundary $S_{I–II}$, the amplitude of the emerging superconducting p state may become large ($\psi \approx 1$), which can in principle be detected in experiments. With increasing distance from the boundary $S_{I–II}$ to the bulk of region I, the amplitude of the p state decreases rapidly; for this reason, the restoration of the Meissner state ($\psi \approx 1$) for most type I superconductors (with $\kappa < \kappa_c(R)$) occurs through a "nearly first-order" phase transition (jumpwise from the p state with $\psi_r \ll 1$). The shape of the magnetization hysteresis loop, its location, and the sites and magnitudes of the jumps (see Figs. 2–4) strongly depend on κ (for given R and temperature T), which can be used, in principle, for the experimental determination of the parameters κ, R, and T (see also [20] in connection with the inclusion of temperature dependence).

Here, we confine our analysis to these qualitative remarks concerning the possible relation to experiments since, to our knowledge, no direct observations of hysteretic (and other) phenomena in long mesoscopic cylinders have been reported. Such experiments were mainly conducted on superconducting mesoscopic discs of various shapes [23–29]; the discussion of the results in the framework of the Ginzburg–Landau theory can be found in [10–16]. It turns out, however, that many theoretical results weakly depend on the choice of the sample geometry. Consequently the predictions obtained in the special case of cylindrical geometry may be of a more general significance and should be borne in mind in discussing the results of specific experiments.

Finally, let us clarify why the boundary $S_{I–II}$ between type I and II superconductors does not coincide with the generally accepted criterion $\kappa_0 = 1/\sqrt{2}$ [1]. This discrepancy can be explained by several factors. First, in [1], an unbounded superconductor is considered, while we are dealing with a superconducting cylinder of a finite radius. Second, in our case, the superconductor borders on a vacuum, while in [1] the contact of two semi-infinite s and n metals is considered. Third, we distinguish between two types of superconductors from the form of the dependence $-4\pi M_\lambda(h_\lambda)$ (i.e., from the presence or absence of a tail on the magnetization curve), while superconductors in [1] are divided into two groups according to a different criterion, i.e., according to the sign of the free energy $\sigma(\kappa)$ on the interface between the s and n phases (in this case, $\sigma = 0$ for $\kappa_0 = 1/\sqrt{2}$ [1]). Thus, the noncoincidence of our boundary $S_{I–II}$ with the value $\kappa_0 = 1/\sqrt{2}$ is due to the difference in the formulation of the problem.

ACKNOWLEDGMENTS

The author is grateful to V.L. Ginzburg for his interest in this work and for valuable remarks and also to V.G. Zharkov and A.Yu. Tsvetkov for fruitful discussions.

This study was supported by the Russian Foundation for Basic Research (project no. 02-02-16285).

REFERENCES

1. V. L. Ginzburg and L. D. Landau, Zh. Éksp. Teor. Fiz. **20**, 1064 (1950).
2. A. A. Abrikosov, Zh. Éksp. Teor. Fiz. **32**, 1442 (1957) [Sov. Phys. JETP **5**, 1174 (1957)].
3. D. Saint-James and P. de Gennes, Phys. Lett. A **7**, 306 (1963).
4. D. Saint-James, Phys. Lett. A **15**, 13 (1965).
5. H. J. Fink and A. G. Presson, Phys. Rev. **151**, 219 (1966); Phys. Rev. **168**, 399 (1968).
6. F. de la Cruz, H. J. Fink, and J. Luzuriaga, Phys. Rev. B **20**, 1947 (1979).
7. H. J. Fink, D. S. McLachlan, and B. Rothberg-Bibby, in *Progress in Low Temperature Physics*, Ed. by D. F. Brewer (North-Holland, Amsterdam, 1978), Vol. VIIb, p. 435.
8. R. Doll and P. Graf, Z. Phys. **197**, 172 (1966); Z. Phys. **204**, 205 (1967).
9. Yu. N. Ovchinnikov, Zh. Éksp. Teor. Fiz. **79**, 1496 (1980) [Sov. Phys. JETP **52**, 755 (1980)]; Zh. Éksp. Teor. Fiz. **79**, 1825 (1980) [Sov. Phys. JETP **52**, 923 (1980)].
10. V. V. Moshchalkov, X. G. Qiu, and V. Bruindoncx, Phys. Rev. B **55**, 11 793 (1997).
11. J. J. Palacios, Phys. Rev. B **58**, R5948 (1998); Physica B (Amsterdam) **256–258**, 610 (1998); Phys. Rev. Lett. **84**, 1796 (2000).
12. P. Deo, V. Shweigert, F. Peeters, and A. K. Geim, Phys. Rev. Lett. **79**, 4653 (1997).
13. V. Schweigert, F. Peeters, and P. Deo, Phys. Rev. Lett. **81**, 2783 (1998).
14. V. Schweigert and F. Peeters, Phys. Rev. B **57**, 13 817 (1998); Phys. Rev. B **59**, 6039 (1999).
15. F. M. Peeters, V. A. Schweigert, B. J. Baelus, and P. S. Deo, Physica C (Amsterdam) **332**, 255 (2000).
16. V. A. Schweigert and F. M. Peeters, Physica C (Amsterdam) **332**, 266 (2000); Physica C **332**, 426 (2000).
17. G. F. Zharkov and V. G. Zharkov, Phys. Scr. **57**, 664 (1998).
18. G. F. Zharkov, V. G. Zharkov, and A. Yu. Zvetkov, Phys. Rev. B **61**, 12 293 (2000).
19. G. F. Zharkov, V. G. Zharkov, and A. Yu. Zvetkov, cond-mat/0008217; G. F. Zharkov, V. G. Zharkov, and A. Yu. Tsvetkov, Kratk. Soobshch. Fiz., No. 11, 35 (2001); No. 12, 31 (2001).
20. G. F. Zharkov, Phys. Rev. B **63**, 214 502 (2001).
21. G. F. Zharkov, Phys. Rev. B **63**, 224 513 (2001).
22. G. F. Zharkov, cond-mat/0109451; J. Low Temp. Phys. **128** (3/4), 87 (2002).
23. G. Dolan, J. Low Temp. Phys. **15**, 133 (1974).
24. O. Buisson, P. Gandit, R. Rammal, *et al.*, Phys. Lett. A **150**, 36 (1990).
25. W. Braunish, N. Knauf, G. Bauer, *et al.*, Phys. Rev. Lett. **68**, 1908 (1992); Phys. Rev. B **48**, 4030 (1993).
26. A. Geim, S. Dubonos, I. Grigorieva, *et al.*, Nature **390**, 259 (1997); **396**, 144 (1998); **407**, 55 (2000); Phys. Rev. Lett. **85**, 1528 (2000).
27. L. Pust, L. Wenger, and M. Koblishka, Phys. Rev. B **58**, 14191 (1998).
28. C. Bolle, V. Aksyuk, F. Pardo, *et al.*, Nature **399**, 43 (1999).
29. F. Müller-Alinger and A. Motta, Phys. Rev. Lett. **84**, 3161 (2000).
30. V. L. Ginzburg, Zh. Éksp. Teor. Fiz. **34**, 113 (1958) [Sov. Phys. JETP **7**, 78 (1958)].
31. P. M. Markus, Rev. Mod. Phys. **36**, 294 (1964).
32. A. A. Abrikosov, Zh. Éksp. Teor. Fiz. **47**, 720 (1964) [Sov. Phys. JETP **20**, 480 (1964)].
33. H. J. Fink, Phys. Rev. Lett. **14**, 309 (1965).
34. L. J. Barnes and H. J. Fink, Phys. Lett. A **20**, 583 (1966).
35. J. Matricon and D. Saint-James, Phys. Lett. A **24**, 241 (1967).
36. P. V. Cristiansen and H. Smith, Phys. Rev. **171**, 445 (1968).
37. J. Feder and D. McLachlan, Phys. Rev. **177**, 763 (1969).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 526–537.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 610–623.

SOLIDS
Electronic Properties

Optical Conductivity in a 2D Model of the Pseudogap State

M. V. Sadovskiĭ* and N. A. Strigina**

Institute of Electrophysics, Ural Division, Russian Academy of Sciences,
ul Komsomol'skaya 34, Yekaterinburg, 620016 Russia
**e-mail: sadovski@iep.uran.ru*
***e-mail: strigina@iep.uran.ru*
Received March 28, 2002

Abstract—A 2D model of the pseudogap state is considered on the basis of the scenario of strong electron scattering by short-range-order fluctuations of the "dielectric" (antiferromagnetic or charge density wave) type. A system of recurrence relations is constructed for a one-particle Green's function and the vertex part, describing the interaction of electrons with an external field. This system takes into account all Feynman diagrams for electron scattering at short-range-order fluctuations. The results of detailed calculations of optical conductivity are given for various geometries (topologies) of the Fermi surface, demonstrating both the effects of pseudogap formation in the electron spectrum and the localization effects. The obtained results are in qualitative agreement with experimental data for underdoped HTSC cuprates. © *2002 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

One of the central problems in the physics of high-temperature copper-oxide superconductors (HTSC) is the description of the nature of the so-called pseudogap state [1, 2] existing in a wide region of the phase diagram. In our opinion [2], the preferable scenario for the pseudogap formation in HTSC oxides is based on the existence of strong scattering of charge carriers in this region at short-range-order fluctuations of the "dielectric" type (antiferromagnetic (AFM) fluctuations or charge-density wave (CDW) type fluctuations). This scattering is strong in the vicinity of the characteristic vector $\mathbf{Q} = (\pi/a, \pi/a)$ (a is the 2D lattice constant), corresponding to doubling of the period (antiferromagnetism vector) and is a precursor of the spectral rearrangement due to the establishment of the long-range AFM order. Accordingly, an essentially non-Fermi-liquid rearrangement of the electron spectrum occurs in this pretransition region of the phase diagram in certain regions of the momentum space in the vicinity of so-called hot spots on the Fermi surface [2], where its effective destruction takes place. A direct experimental verification of such a pattern of formation of a pseudogap was obtained in recent ARPES experiments on the system $Nd_{1.85}Ce_{0.15}CuO_4$ [3], in which the above-mentioned spectral rearrangement could be studied in the vicinity of hot spots.

In the framework of the above scenario of the pseudogap state formation, it is possible to construct a simplified "almost exactly" solvable model describing the main features of this state [2] and taking into account the contribution of all Feynman diagrams in the perturbation theory on the scattering by short-range-order (Gaussian) fluctuations with characteristic scattering momentum from the vicinity of $\mathbf{Q}$, determined by the corresponding correlation length ζ [4, 5]. This model is based on a generalization of the model of formation of a pseudogap in a 1D system due to developed short-range-order fluctuations of the CDW type (which was proposed earlier by one of the authors [6, 7]) to the 2D case. A simplified version of this 2D model (the model of hot patches) was used in [8–11] for describing the main properties of superconducting state formed against the background of a dielectric pseudogap.

In [4, 5], one-particle properties of the model under investigation (such as spectral density and the density of states) were mainly analyzed. A remarkable feature of this model is the possibility of summation of the entire series of Feynman diagrams also in the two-particle problem of calculation of the vertex part describing the response of the system to external perturbation (e.g., electromagnetic field) [6, 12, 13]. In the simplified version of the model of "hot patches" on the Fermi surface, the required calculations of optical conductivity in the 2D case were made in [14]. Here, we aim both at a detailed analysis of theoretical aspects of the calculation of two-particle properties in the framework of the general model [4, 5] and at the calculation of optical conductivity for various geometries (topologies) of the Fermi surface, emerging when a realistic form of the free electron spectrum is used.

2. MODEL OF HOT SPOTS

2.1. Description of the Model and "Almost Exact" Solution for One-Particle Green's Function

In the model of a "nearly antiferromagnetic" Fermi liquid, which is actively used for explaining the micro-

scopic mechanism of HTSC [15, 16], the effective interaction of electrons with short-range-order AFM spin fluctuations is introduced. This interaction is described by the dynamic spin susceptibility $\chi_{\mathbf{q}}(\omega)$ whose shape is determined from fitting to NMR data [16]:

$$V_{\text{eff}}(\mathbf{q}, \omega) = g^2\chi_{\mathbf{q}}(\omega) \approx \frac{g^2\xi^2}{1 + \xi^2(\mathbf{q} - \mathbf{Q})^2 - i\omega/\omega_{\text{sf}}}, \quad (1)$$

where g is the coupling constant, ξ is the correlation length of spin fluctuations, $\mathbf{Q} = (\pi/a, \pi/a)$ is the antiferromagnetic ordering vector in the dielectric phase, and ω_{sf} is the characteristic frequency of spin fluctuations. The dynamic susceptibility and, hence, the effective interaction (1) have peaks in the region $\mathbf{q} \sim \mathbf{Q}$; accordingly, two types of quasiparticles emerge in the system, i.e., hot particles, whose momenta lie in the vicinity of hot spots on the Fermi surface (Fig. 1), and cold particles, whose momenta lie in the vicinity of the regions on the Fermi surface surrounding the diagonals of the Brillouin zone [4]. As a matter of fact, quasiparticles from the regions of hot spots are strongly scattered with the momentum transfer of the order of $\mathbf{Q}$ due to their interaction with the spin fluctuations (1), while the same interaction for particles with momenta away from hot spots is quite weak.

Considering the region of rather high temperatures $\pi T \gg \omega_{sf}$, we can neglect the spin dynamics [4], confining our analysis of relation (1) to the static approximation. A considerable simplification of calculations, which makes it possible to analyze higher-order contribution of perturbation theory, can be obtained if we go over in relation (1) to a model interaction of the form [5]

$$V_{\text{eff}}(\mathbf{q}) = \Delta^2 \frac{2\xi^{-1}}{\xi^{-2} + (q_x - Q_x)^2} \frac{2\xi^{-1}}{\xi^{-2} + (q_y - Q_y)^2}, \quad (2)$$

where Δ is an effective parameter having the dimensions of energy. Following [4, 5], in the subsequent analysis we will treat Δ and ξ as phenomenological parameters (that can be determined experimentally). Expression (2) is qualitatively similar to the static limit (1) and differs from it quantitatively only slightly in the most interesting region $|\mathbf{q} - \mathbf{Q}| < \xi^{-1}$ determining scattering in the vicinity of hot spots.

We will take the spectrum of the "bare" (free) quasiparticles in the form [4]

$$\begin{aligned}\xi_{\mathbf{p}} &= -2t(\cos p_x a + \cos p_y a) \\ &\quad - 4t' \cos p_x a \cos p_y a - \mu,\end{aligned} \quad (3)$$

where t is the integral of transfer between the nearest neighbors, t' is the same for next-to-nearest neighbors in the square lattice, and μ is the chemical potential. This expression provides a satisfactory approximation to the results of band calculations for real HTSC systems. For example, for $YBa_2Cu_3O_{6+\delta}$, we have t = 0.25 eV and $t' = -0.45t$ [4]. The chemical potential μ is fixed by charge carrier concentrations. In this work, we consider various characteristic relations between parameters t and t' leading to different geometries (topologies) of the Fermi surface, aiming at an analysis of the general pattern, which is not necessarily associated with known specific systems.

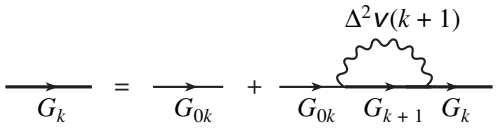


Fig. 1. Diagrammatic representation of recurrence relation for the Green's function.

In [5], a detailed analysis of contributions of all diagrams was carried out for the self-energy part $\Sigma(\varepsilon_n\mathbf{p})$ of an electron. It turns out that, in the case when the signs of the velocity components $v^x_{\mathbf{p}}$ and $v^x_{\mathbf{p}+\mathbf{Q}}$ (as well as of $v^y_{\mathbf{p}}$ and $v^y_{\mathbf{p}+\mathbf{Q}}$) coincide in hot spots on the Fermi surface, the Feynman integrals in a diagram of any order are determined only by the contributions from the poles of the Lorentzians in relation (2) and can easily be evaluated.[1] In this case, the contribution of an arbitrary diagram for the self-energy component of the Nth order in the interaction with fluctuations (2) has the form ($\varepsilon_n = (2n + 1)\pi T$)

$$\Sigma^{(N)}(\varepsilon_n\mathbf{p}) = \Delta^{2N} \prod_{j=1}^{2N-1} \frac{1}{i\varepsilon_n - \xi_j(\mathbf{p}) + in_j v_j \kappa}, \quad (4)$$

where $\xi_j(\mathbf{p}) = \xi_{\mathbf{p}+\mathbf{Q}}$ and $v_j = |v^x_{\mathbf{p}+\mathbf{Q}}| + |v^y_{\mathbf{p}+\mathbf{Q}}|$ for odd j, $\xi_j(\mathbf{p}) = \xi_{\mathbf{p}}$ and $v_j = |v^x_{\mathbf{p}}| + |v^y_{\mathbf{p}}|$ for even j, and $\kappa = \xi^{-1}$. Here, n_j is the number of interaction lines embracing the jth Green's function in the given diagram; for the sake of definiteness, we assume that $\varepsilon_n > 0$.

The conditions under which the above constraints are imposed on the velocities at the points on the Fermi surface connected by vector $\mathbf{Q}$ (hot spots) are analyzed in detail in [5], where examples of corresponding geometries of the Fermi surfaces realized for certain relations between parameters t and t' in Eq. (3) are considered. In these cases, expression (4) is virtually exact. In all remaining cases (for other relations between t and t'), expression (4) is used as a successful ansatz for an arbitrary-order contribution obtained by simple continuation of the spectrum in parameters t and t' to the region of interest. Even in the most unfavorable 1D case [7] corresponding to a square Fermi surface emerging from Eq. (3) for $t' = 0$ and $\mu = 0$, the use of this ansatz leads to results (e.g., for the density of states) very close quantitatively [17] to the results of the exact numerical

[1] A similar situation also emerges in the case when the velocities in the hot spots connected by vector $\mathbf{Q}$ are exactly perpendicular [4].

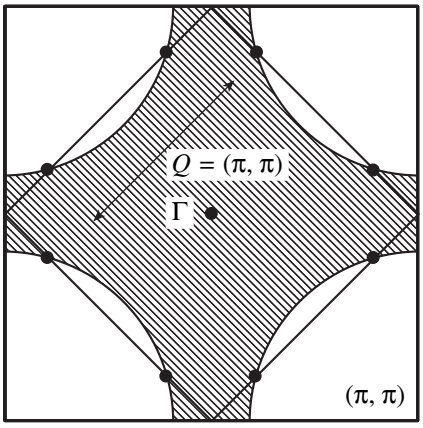


Fig. 2. Fermi surface with hot spots connected through the scattering vector of the order of $\mathbf{Q} = (\pi/a, \pi/a)$.

simulation of this problem [18]. In this sense, we are using the term "almost exact" solution.

When ansatz (4) is used, it is found that the contribution of any diagram with crossing interaction lines is equal to the contribution of a diagram of the same order without intersection of these lines [7]. For this reason, we can, in fact, take into account the contributions from diagrams without intersection of the interaction lines, taking into account the contribution from diagrams with intersection, with the help of additional combinatorial factors compared to the "initial" interaction vertices (or lines) [7]. As a result, we obtain the following recurrence relation (representation in the form of a continued fraction [7]) for a one-electron Green's function, which gives an effective logarithm for subsequent numerical calculations [5]:

$$G_k(\varepsilon_n\xi_\mathbf{p}) = \frac{1}{i\varepsilon_n - \xi_k(\mathbf{p}) + ikv_k\kappa - \Sigma_{k+1}(\varepsilon_n\xi_\mathbf{p})} \equiv \{G_{0k}^{-1}(\varepsilon_n\xi_\mathbf{p}) - \Sigma_{k+1}(\varepsilon_n\xi_\mathbf{p})\}^{-1}, \quad (5)$$

$$\Sigma_k(\varepsilon_n\xi_\mathbf{p}) = \Delta^2\frac{v(k)}{i\varepsilon_n - \xi_k(\mathbf{p}) + ikv_k\kappa - \Sigma_{k+1}(\varepsilon_n\xi_\mathbf{p})}. \quad (6)$$

Figure 2 is a graphical representation of this recurrence relation. The physical Green's function we are interested in is $G(\varepsilon_n\xi_\mathbf{p}) = G_{k=0}(\varepsilon_n\xi_\mathbf{p})$. In relation (5), we have also introduced the following auxiliary notation:

$$G_{0k}(\varepsilon_n\xi_\mathbf{p}) = \frac{1}{i\varepsilon_n - \xi_k(\mathbf{p}) + ikv_k\kappa}. \quad (7)$$

In the case of commensurate fluctuations with $\mathbf{Q} = (\pi/a, \pi/a)$ [7] under investigation, the combinatorial factor is given by

$$v(k) = k \quad (8)$$

if we disregard their spin structure (CDW-type fluctuations). If the spin structure of interactions is taken into account in the model of a nearly antiferromagnetic Fermi liquid (spin–fermion model [4]), the combinatorics of the diagrams becomes more complicated. In particular, the scattering with spin conservation gives a formally commensurate combinatorics, while scattering with spin flip is described by the diagrams for the incommensurate case ("charged" random field in the terminology used in [4]). As a result, the recurrence relation for the Green's function, as before, has the form (6), but the combinatorial factor $v(k)$ has the form [4]

$$v(k) = \begin{cases} \dfrac{k+2}{3} & \text{for odd } k \\ \dfrac{k}{3} & \text{for even } k. \end{cases} \quad (9)$$

In the subsequent analysis, we confine ourselves to cases (8) and (9); the details corresponding to incommensurate fluctuations of the CDW type can be found in [5–7].

The obtained solution for a one-particle Green's function is exact in the limit $\xi \longrightarrow \infty$, when a solution can be found in analytic form [4, 6]. This solution is exact in the trivial limit $\xi \longrightarrow 0$, when interaction (2) just vanishes for a fixed value of Δ. For all intermediate values of ξ, it gives a very good interpolation (see above) since it is virtually exact for certain geometries of the Fermi surface emerging for specific ranges of variation of the parameters of spectrum (3) [5].

Using relation (5), we can easily carry out numerical calculations of the one-electron spectral density and density of states:

$$A(E\mathbf{p}) = -\frac{1}{\pi}\mathrm{Im}G^R(E\mathbf{p}), \qquad N(E) = \sum_\mathbf{p} A(E\mathbf{p}). \quad (10)$$

In these relations, $G^R(E\mathbf{p})$ is the retarded Green's function obtained by the conventional analytical continuation of Eq. (5) from the Matsubara frequencies to the real axis E. The details of corresponding calculations and the discussion of the obtained results for the 2D model under investigation can be found in the publications [4, 5] mentioned above.

2.2. Recurrence Equations for the Vertex Part and Conductivity

In order to calculate the optical conductivity, we must calculate the vertex part describing the electromagnetic response of the system. This apex can be determined by the method proposed for an analogous one-dimensional model in [12, 13]. Any diagram for an irreducible vertex component can be obtained by inserting the external field lines into the corresponding diagram for the self-energy component [6]. Since our model can take into account only the diagrams for the self-energy component without intersection of the

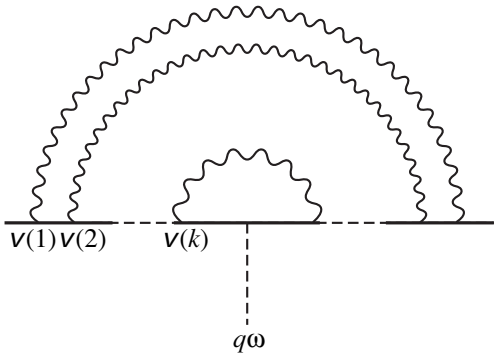


Fig. 3. General form of the higher order correction for the vertex part.

interaction lines with additional combinatorial factors $v(k)$ at initial vertices, it is sufficient to consider only diagrams of the type shown in Fig. 3 for calculating vertex corrections. This immediately gives a system of recurrence equations for the vertex parts, presented graphically in Fig. 4. In order to obtain the corresponding analytic expressions, we consider the simplest vertex correction shown in Fig. 5a. Carrying out calculations for $T = 0$ in the *RA* channel, we can easily obtain the corresponding contribution in the form

$$\begin{aligned}
&\mathcal{J}_1^{(1)RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q})\\
&= \sum_{\mathbf{K}} V_{\text{eff}}(\mathbf{k}) G_{00}^A(\varepsilon\xi_{\mathbf{p}-\mathbf{K}}) G_{00}^R(\varepsilon+\omega\xi_{\mathbf{p}-\mathbf{K}+\mathbf{q}})\\
&= \Delta^2\{G_{00}^A(\varepsilon,\ \xi_1(\mathbf{p})+iv_1\kappa)\\
&- G_{00}^R(\varepsilon+\omega,\ \xi_1(\mathbf{p}+\mathbf{q})-iv_1\kappa)\}\\
&\times\frac{1}{\omega+\xi_1(\mathbf{p})-\xi_1(\mathbf{p}+\mathbf{q})} \qquad (11)\\
&= \Delta^2 G_{00}^A(\varepsilon,\ \xi_1(\mathbf{p})+iv_1\kappa) G_{00}^R(\varepsilon+\omega,\ \xi_1(\mathbf{p}+\mathbf{q})-iv_1\kappa)\\
&\times\left\{1+\frac{2iv_1\kappa}{\omega+\xi_1(\mathbf{p})-\xi_1(\mathbf{p}+\mathbf{q})}\right\}\\
&\equiv \Delta^2 G_{01}^A(\varepsilon,\ \xi_{\mathbf{p}}) G_{01}^R(\varepsilon+\omega,\ \xi_{\mathbf{p}+\mathbf{q}})\\
&\times\left\{1+\frac{2iv_1\kappa}{\omega+\xi_1(\mathbf{p})-\xi_1(\mathbf{p}+\mathbf{q})}\right\},
\end{aligned}$$

where we have evaluated the integrals using the following identity valid for free-electron Green's functions:

$$\begin{aligned}
&G_{00}^A(\varepsilon\xi_p) G_{00}^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}})\\
&= \{G_{00}^A(\varepsilon\xi_{\mathbf{p}}) - G_{00}^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}})\}\frac{1}{\omega-\xi_{\mathbf{p}+\mathbf{q}}+\xi_{\mathbf{p}}}. \qquad (12)
\end{aligned}$$

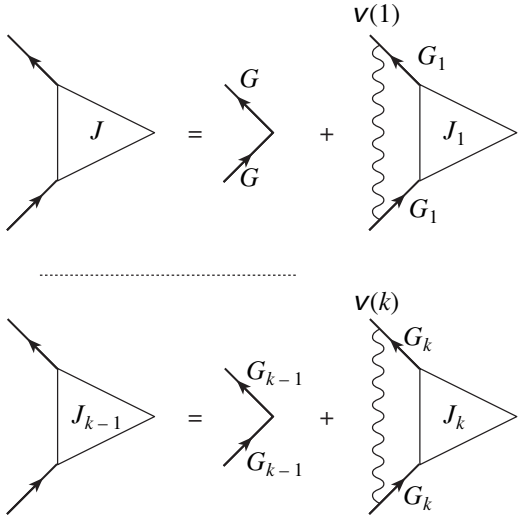


Fig. 4. Recurrence equations for the vertex part.

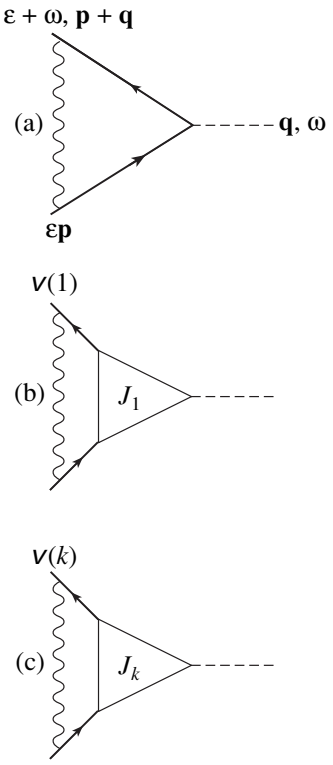


Fig. 5. Simplest corrections to vertex parts.

"Dressing" internal electron lines, we pass to the diagram in Fig. 5b. Using the identity

$$\begin{aligned}
&G^A(\varepsilon\xi_{\mathbf{p}}) G^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}}) = \{G^A(\varepsilon\xi_{\mathbf{p}}) - G^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}})\}\\
&\times\frac{1}{\omega-\xi_{\mathbf{p}+\mathbf{q}}+\xi_{\mathbf{p}}-\Sigma_1^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}})+\Sigma_1^A(\varepsilon\xi_{\mathbf{p}})}, \qquad (13)
\end{aligned}$$

which is valid for exact Green functions, we can write the contribution of this diagram in the form

$$\mathcal{J}_1^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q})$$
$$= \Delta^2 v(1) G_1^A(\varepsilon, \xi_{\mathbf{p}}) G_1^R(\varepsilon+\omega, \xi_{\mathbf{p}+\mathbf{q}}) \left\{ 1 + \frac{2iv_1\kappa}{\omega - \xi_1(\mathbf{p}+\mathbf{q}) + \xi_1(\mathbf{p}) - \Sigma_2^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}}) + \Sigma_2^A(\varepsilon\xi_{\mathbf{p}})} \right\} \quad (14)$$
$$\times J_1^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q}).$$

Here, we have assumed that the line of interaction on the diagram for the vertex correction in Fig. 5b "transforms" the self-energy component $\Sigma_1^{R,A}$ of internal electron lines into $\Sigma_2^{R,A}$ in accordance with the approximation used above for the self-energy component (see Fig. 2).[2]

We can now easily write a similar expression for a general diagram shown in Fig. 5c:

$$\mathcal{J}_k^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q}) = \Delta^2 v(k) G_k^A(\varepsilon, \xi_{\mathbf{p}}) G_k^R(\varepsilon+\omega, \xi_{\mathbf{p}+\mathbf{q}})$$
$$\times \left\{ 1 + \frac{2iv_k\kappa k}{\omega - \xi_k(\mathbf{p}+\mathbf{q}) + \xi_k(\mathbf{p}) - \Sigma_{k+1}^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}}) + \Sigma_{k+1}^A(\varepsilon\xi_{\mathbf{p}})} \right\} J_k^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q}). \quad (15)$$

Accordingly, the fundamental recurrence relation for the vertex part in Fig. 4 can be written in the form

$$J_{k-1}^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q}) = 1 + \Delta^2 v(k) G_k^A(\varepsilon, \xi_{\mathbf{p}}) G_k^R(\varepsilon+\omega, \xi_{\mathbf{p}+\mathbf{q}})$$
$$\times \left\{ 1 + \frac{2iv_k\kappa k}{\omega - \xi_k(\mathbf{p}+\mathbf{q}) + \xi_k(\mathbf{p}) - \Sigma_{k+1}^R(\varepsilon+\omega\xi_{\mathbf{p}+\mathbf{q}}) + \Sigma_{k+1}^A(\varepsilon\xi_{\mathbf{p}})} \right\} J_k^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q}). \quad (16)$$

The physical apex $J^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q})$ is defined as $J_{k=0}^{RA}(\varepsilon\mathbf{p};\ \varepsilon+\omega,\ \mathbf{p}+\mathbf{q})$. The recurrence procedure (16) takes into account all diagrams in perturbation theory for the vertex component. As $\kappa \longrightarrow 0$ ($\xi \longrightarrow \infty$), Eq. (16) is reduced to the series studied in [6] (see also [4]), which can be summed exactly in analytic form. In our scheme of analysis, the standard ladder approximation corresponds to the case when all combinatorial factors $v(k)$ in Eq. (16) are assumed to be equal to unity [13].

The conductivity of the system can be expressed [19] in terms of the retarded density–density response function $\chi^R(q, \omega)$:

$$\sigma(\omega) = e^2 \lim_{q\to 0}\left(-\frac{i\omega}{q^2}\right)\chi^R(q, \omega), \quad (17)$$

where e is the electron charge and

$$\chi^R(q, \omega) = \omega\{\Phi^{RA}(0, q, \omega) - \Phi^{RA}(0, 0, \omega)\}, \quad (18)$$

while the two-particle Green's function $\Phi^{RA}(\varepsilon, q, \omega)$ is determined by the loop graph shown in Fig. 6.

Direct numerical calculations confirm that the recurrence procedure (16) satisfies the exact relation following (for $\omega \longrightarrow 0$) from the Ward identity [19]:

$$\Phi^{RA}(0, 0, \omega) = -N(E_F)/\omega, \quad (19)$$

where $N(E_F)$ is the density of states at the Fermi level $E_F = \mu$. This is the main argument in favor of the ansatz used in the derivation of Eqs. (14)–(16).

Ultimately, we can write conductivity in the symmetrized form convenient for numerical calculations:

$$\sigma(\omega) = \frac{e^2\omega^2}{\pi}\lim_{q\to 0}\frac{1}{q^2}\sum_{\mathbf{p}}\left\{ G^R\left(\frac{\omega}{2}, \mathbf{p}+\frac{\mathbf{q}}{2}\right)\right.$$
$$\times J^{RA}\left(\frac{\omega}{2}, \mathbf{p}+\frac{\mathbf{q}}{2}; -\frac{\omega}{2}, \mathbf{p}-\frac{\mathbf{q}}{2}\right) G^A\left(-\frac{\omega}{2}, \mathbf{p}-\frac{\mathbf{q}}{2}\right) \quad (20)$$
$$\left. - G^R\left(\frac{\omega}{2}, \mathbf{p}\right) J^{RA}\left(\frac{\omega}{2}, \mathbf{p}; -\frac{\omega}{2}, \mathbf{p}\right) G^A\left(-\frac{\omega}{2}, \mathbf{p}\right)\right\},$$

[2] A motivation for this notation is that it ensures the fulfillment of the Ward identity which will be discussed below.

where we have also taken into account the additional factor 2 associated with the summation over spin.

Numerical calculations were carried out directly by using formulas (20), (16), and (5), the recurrence procedure being terminated at a high "level" k, where all Σ_k and J_k were set equal to zero. Integration of Eq. (20) was carried out over the entire 2D Brillouin zone. The "bare" electron spectrum was taken in the form (3). Integration momenta are naturally reduced to dimensionless form with the help of lattice constant a, and all energies will be henceforth given in units of the transfer integral t. In this case, conductivity is measured in units of universal conductivity $\sigma_0 = e^2/\hbar = 2.5 \times 10^{-4}\ \Omega^{-1}$ of a 2D system, and the density of states is measured in units of $1/ta^2$.

3. RESULTS AND DISCUSSION

Optical conductivity and other parameters of the model under investigation were calculated for various values of parameters determining the spectrum (3) of free quasiparticles and for $\Delta = t$. Let us first consider the case when the Fermi surfaces are in the vicinity of half-filled band with $\mu = 0$ and $t' = 0$, which are presented in Fig. 7a for the first quadrant of the Brillouin zone. It is well known that, for $\mu = 0$ and $t' = 0$, the Fermi surface has the form of a square (complete nesting), so that the situation is equivalent to a certain extent to the 1D case considered in [6, 12, 13]. The results of calculations for the real part of optical conductivity in the 2D problem under investigation for the case of spin–fermion combinatorics of the diagrams and for various values of correlation length of the short-range AFM order (parameter $\kappa = \xi^{-1}$, where ξ is measured in units of the lattice constant a) are presented in Fig. 8. The form of conductivity is qualitatively quite similar to that obtained in [12, 13] in the 1D model (for the case of incommensurate CDW-type fluctuations). It is characterized by the presence of a well-defined peak due to pseudogap absorption (the corresponding curves for the density of states, demonstrating the presence of a pseudogap near the Fermi level, are shown in the inset to Fig. 8) for $\omega \sim 2\Delta$ and the presence of a maximum in the low-frequency region, which is associated with the localization of charge carriers in the static random field of AFM fluctuations. The localization nature of this maximum is confirmed by its conversion into the characteristic Drude peak (with a maximum at $\omega = 0$) for calculations in the ladder approximation, when the combinatorial factors $V(k) = 1$, which corresponds to the exclusion of the contribution from diagrams with crossed interaction lines which directly lead to 2D Anderson localization [19, 20]. The qualitative form of conductivity in this case is also quite similar to that obtained in [13]. The narrowing of the localization peak upon a decrease in the correlation length of fluctuations can be explained, according to [13], by a decrease in the effective interaction (2) upon a decrease in ξ (for a fixed value of Δ),

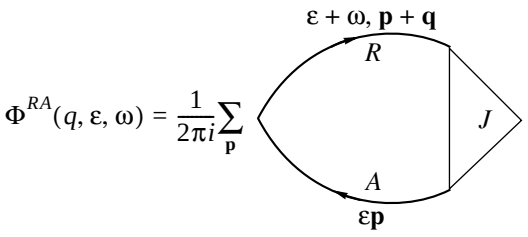


Fig. 6. Diagrammatic representation for the two-particle response function $\Phi^{RA}(q, \varepsilon, \omega)$.

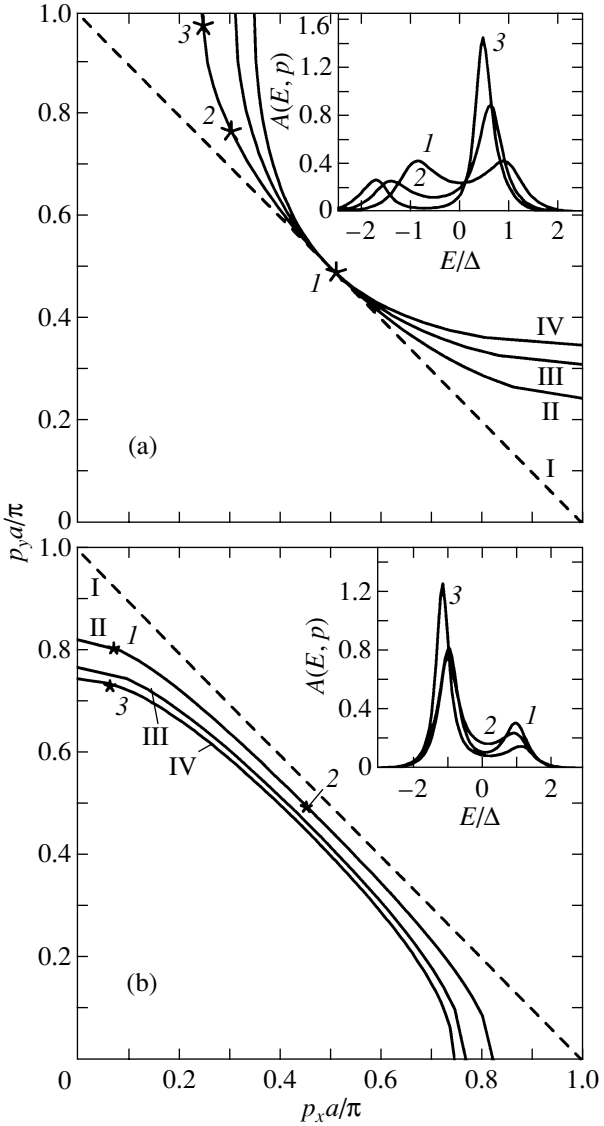


Fig. 7. Fermi surfaces for different values of parameter t' and chemical potential μ: (a) $\mu = 0$ and $t'/t = 0$ (I), –0.2 (II), –0.4 (III), and –0.6 (IV); (b) $t' = 0$ and $\mu/t = 0$ (I), –0.3 (II), –0.5 (III), and –0.6 (IV). The insets show the energy dependences of spectral density for the spin–fermion model for $\kappa a = 0.1$ at the points of the momentum space marked by asterisks.

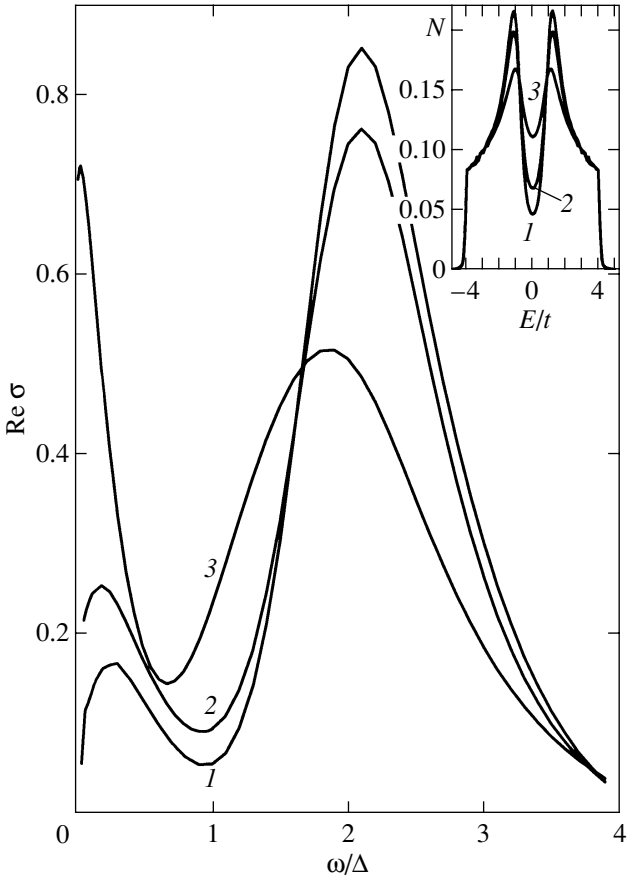


Fig. 8. Real part of optical conductivity in the spin–fermion model for a square Fermi surface ($\mu = 0$, $t' = 0$) for different values of the inverse short-range-order correlation length: $\kappa a = 0.1$ (*1*), 0.2 (*2*), and 0.5 (*3*). The inset shows the corresponding densities of states.

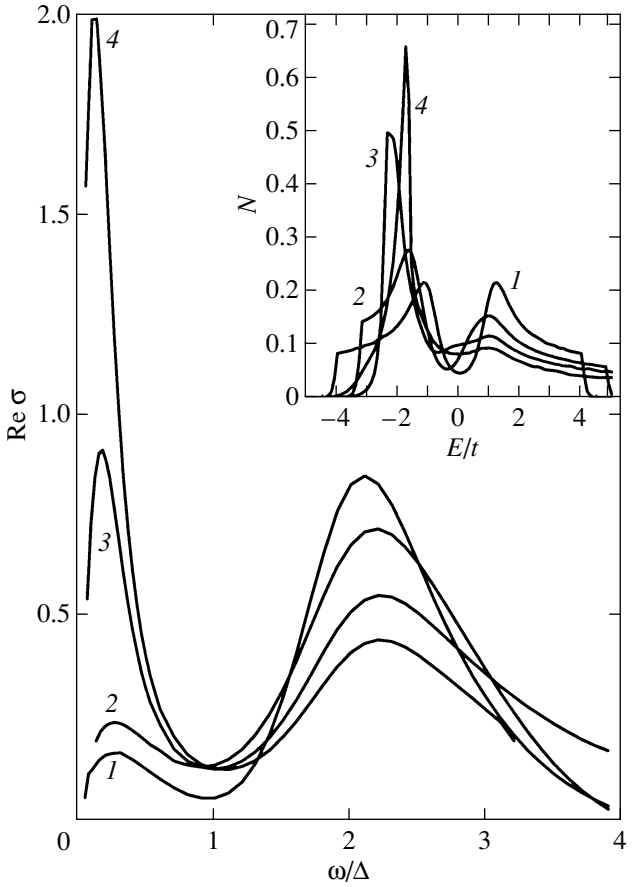


Fig. 9. Real part of optical conductivity in the spin–fermion model for $\mu = 0$ and $\kappa a = 0.1$ for various shapes of the Fermi surface obtained from the square surface taking into account the transfer integral: $t'/t = 0$ (*1*), –0.2 (*2*), –0.4 (*3*), and –0.6 (*4*). The inset shows the corresponding densities of states.

leading to a general decrease in scattering rate (including that at the cold part of the Fermi surface). It should be noted that the behavior of the density of states and optical conductivity determined here is in complete qualitative agreement with the results obtained for an analogous 2D model of the Peierls transition with the help of the quantum Monte Carlo method in a recent publication [21].

If we now include the transfer integral t' between the next-to-nearest neighbors in Eq. (3), assuming, as before, that $\mu = 0$, we arrive at shapes of the Fermi surface differing from a square and depicted in Fig. 7a. The inset to this figure shows the energy dependence of the spectral density (10) at several characteristic points on these Fermi surfaces. It can be seen that it displays a characteristic non-Fermi-liquid behavior of the type of that studied in [4, 5] practically at all points on the Fermi surface as long as the shape of this surface differs from a square not very strongly, in spite of the fact that a hot spot in the case under investigation lies strictly at the intersection of the Fermi surface with the diagonal of the Brillouin zone. The corresponding curves for the real part of optical conductivity are shown in Fig. 9; the inset to this figure depicts the shape of the corresponding densities of states. It can be seen that, as the situation differs more and more strongly from complete nesting, the pseudogap absorption peak decreases, while the localization peak increases in conformity with the general summation rule for conductivity. It should be noted, however, that the pseudogap absorption peak remains quite noticeable even when the pseudogap in the density of states is virtually imperceptible (curves *4* in Fig. 9).

Let us return to the case when $t' = 0$, but the value of μ is varied, so that we pass to the Fermi surfaces whose shape is quite close to the square shown in Fig. 7b. Strictly speaking, hot spots on the Fermi surface are absent altogether, but the spectral density shown in the inset to Fig. 7b preserves a typical pseudogap form. The corresponding dependences for the real part of optical conductivity are presented in Fig. 10.

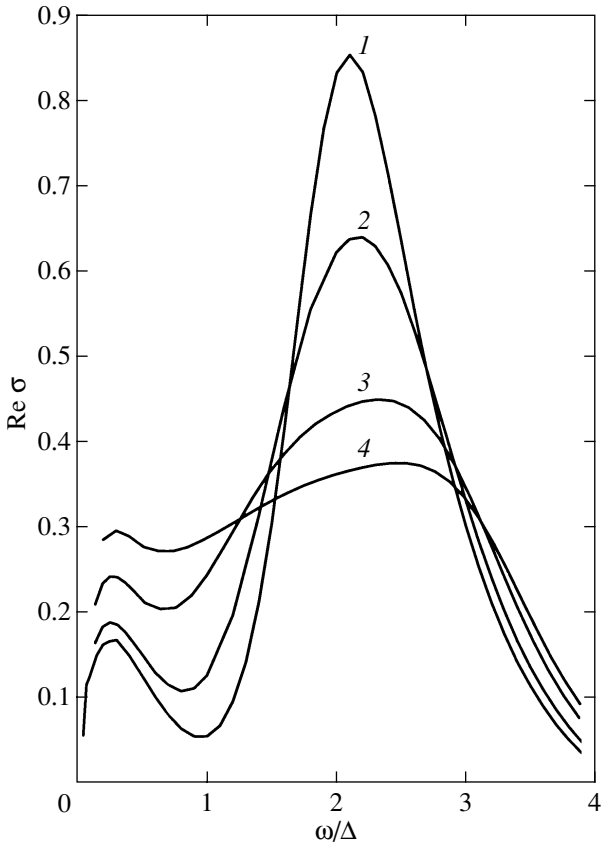


Fig. 10. Real part of optical conductivity in the spin–fermion model for different values of parameter t' and $\kappa a = 0.1$ for various shapes of the Fermi surface obtained from the square surface as a result of departure from half-filled band. The chemical potential corresponds to values of $\mu/t = 0$ (*1*), –0.3 (*2*), –0.5 (*3*), and –0.6 (*4*).

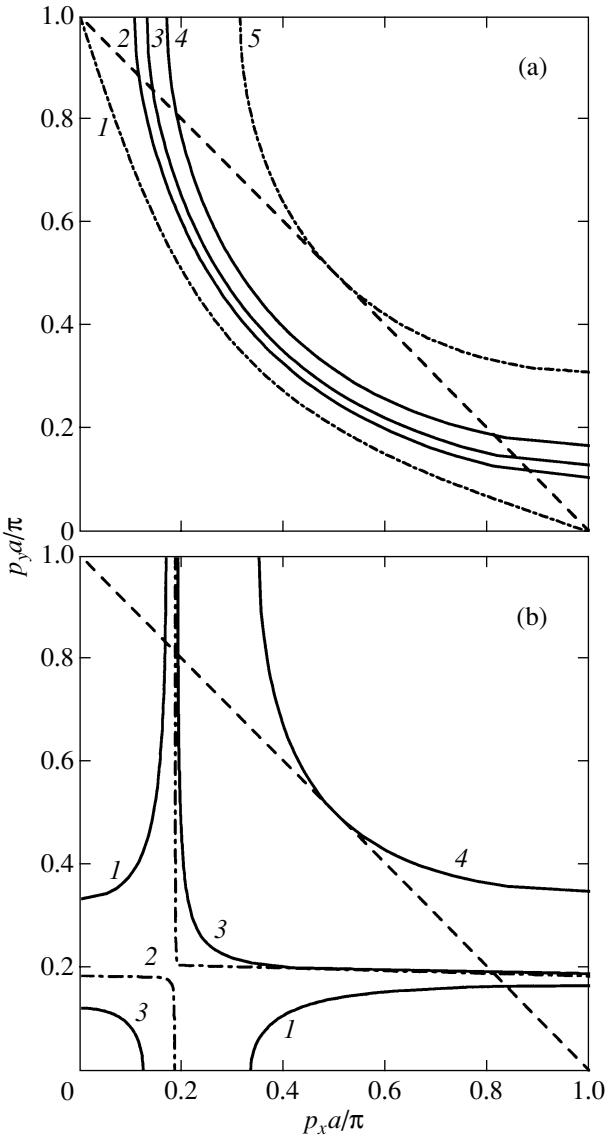


Fig. 11. Fermi surfaces for different values of parameter t' and chemical potential μ: (a) $t'/t = -0.4$ (which is typical of HTSC cuprates) and $\mu/t = -1.6$ (*1*), –1.4 (*2*), –1.3 (*3*), –1.1 (*4*), and 0 (*5*); hot spots exist for $-1.6 < \mu/t < 0$; (b) $t'/t = -0.6$ and $\mu/t = -1.8$ (*1*), –1.666 (*2*), –1.63 (*3*), and 0 (*4*). Hot spots exist for $\mu < 0$. Dashed lines mark the boundary of the magnetic Brillouin zone.

Let us now consider various geometries of the Fermi surface with hot spots shown in Fig. 11. Figures 12 and 13 depict the real part of optical conductivity, calculated (for different combinatorics of the diagrams) for two characteristic values $t' = -0.4t$ and $t' = -0.6t$ for the chemical potential $\mu = 0$, when hot spots are on the diagonal of the Brillouin zone (curve *5* in Fig. 11a and curve *4* in Fig. 11b). It can be seen that the pseudogap behavior of the conductivity persists even in the case when there is practically no pseudogap in the density of states (shown in the insets to Figs. 12, 13). The dashed curve in Fig. 12 shows the results of the ladder approximation, demonstrating the typical disappearance of 2D localization. Figure 13 illustrates the smearing of the pseudogap maximum of conductivity upon a decrease in the short-range-order correlation length.

For most high-temperature copper-oxides superconductors, the characteristic geometry of the Fermi surface is described by the case $t' = -0.4t$ and $\mu = -1.3t$ [4] (curve *3* in Fig. 11a). The results of calculation of optical conductivity for this case for different values of the inverse correlation length κ are presented in Fig. 14 (for the case of the spin–fermion combinatorics of diagrams). We have introduced additional weak scattering due to inelastic processes through the standard substitution $\omega \longrightarrow \omega + i\gamma$ [22], which leads to the emergence of a narrow Drude peak in the frequency range $\omega < \gamma$ (violation of 2D localization due to dephasing). It can easily be verified that, as the rate γ of inelastic scattering increases, the localization peak is smeared and is transformed into a conventional Drude peak in the low-frequency region. The pseudogap absorption peak becomes more pronounced upon an increase in correla-

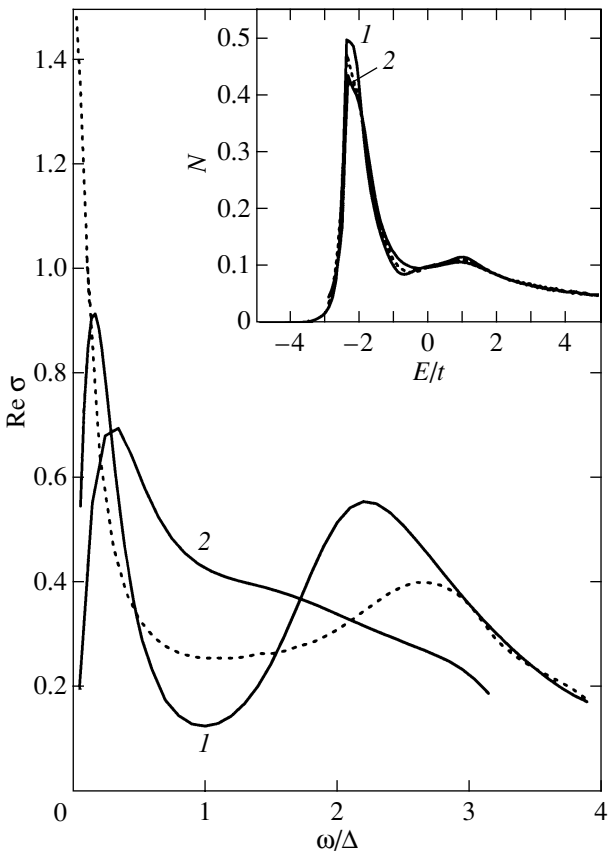


Fig. 12. Real part of optical conductivity for $t'/t = -0.4$, $\mu = 0$, and $\kappa a = 0.1$ for different combinatorics of diagrams: spin–fermion model (*1*) and commensurate case (*2*). The dotted curve corresponds to the ladder combinatorics. The inset shows the corresponding densities of states.

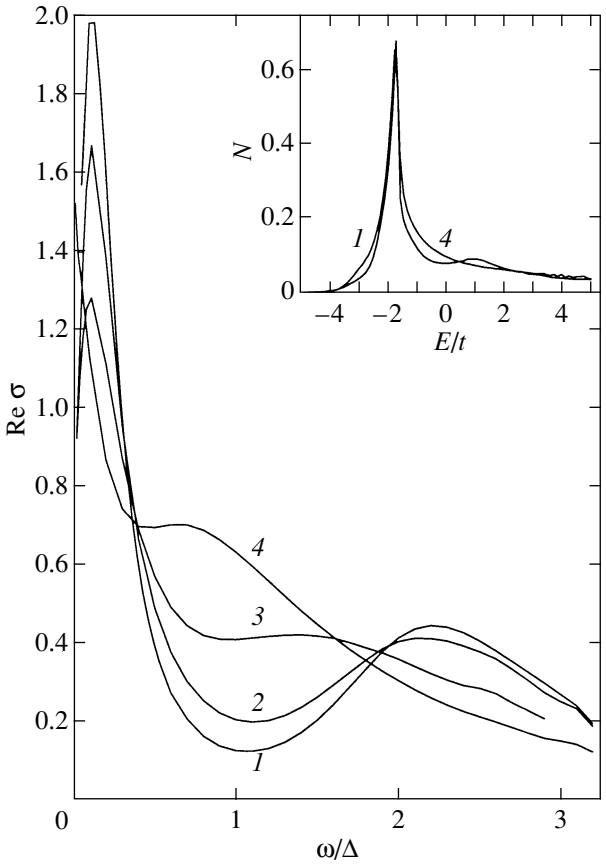


Fig. 13. Real part of optical conductivity in the spin–fermion model for $t'/t = -0.6$ and $\mu = 0$ for the values of inverse correlation length $\kappa a = 0.1$ (*1*), 0.2 (*2*), 0.5 (*3*) and 1.0 (*4*). The inset shows the densities of states corresponding to curves *1* and *4*.

tion length ξ (a decrease in parameter κ). Figure 15 shows the frequency dependences of the effective scattering rate $1/\tau(\omega)$ and effective mass $m^*(\omega)$, determined from the results of our calculations with the help of the generalized Drude formula, which is often used for experimental data fitting [1]:

$$\frac{1}{\tau(\omega)} = \frac{\omega_p^2}{4\pi}\mathrm{Re}\left(\frac{1}{\sigma(\omega)}\right), \tag{21}$$

$$\frac{m^*(\omega)}{m} = -\frac{1}{\omega}\frac{\omega_p^2}{4\pi}\mathrm{Im}\left(\frac{1}{\sigma(\omega)}\right). \tag{22}$$

Here, ω_p is the plasma frequency, and m is the free electron mass. It can be seen from Fig. 15 that the quantity $1/\tau(\omega)$ (which is expressed in units of $\omega_p^2\hbar/4\pi e^2$ in this figure) demonstrates a typical pseudogap behavior in the frequency range $\omega < 2\Delta$. It should be noted that the density of states in this case exhibits only a weakly pronounced pseudogap [5] (see the inset to Fig. 12). Figure 16 presents similar results for the same case (typical of HTSC oxides) obtained for a model with diagram combinatorics corresponding to commensurate fluctuations of the CDW type. It can be seen that the pseudogap absorption peak is virtually unnoticeable in this case.

It can be seen from Fig. 11b that, as the chemical potential changes from $\mu = 0$ to $\mu = -1.666t$, the Fermi surface acquires flat regions of increasing size and is transformed into a virtually cross-shaped surface for $\mu \approx 1.666t$. Such a Fermi surface was observed in ARPES experiments on the system $La_{1.28}Nd_{0.6}Sr_{0.12}CuO_4$ [23, 24]. In this case, the components of velocities at hot spots connected by the vector $\mathbf{Q} = (\pi/a, \pi/a)$ become orthogonal. For $\mu/t = -1.666\ldots$, the topology of the Fermi surface changes (Fig. 11b), and these components have the same sign in the entire region $\mu/t < -1.666\ldots$, which ensures exact fulfillment of our fundamental ansatz (4) for the contributions of higher order diagrams [5]. It is interesting to consider the results of calculations of optical conductivity in this region of variation of μ also. The corresponding results in the case of a commensurate (CDW) combinatorics

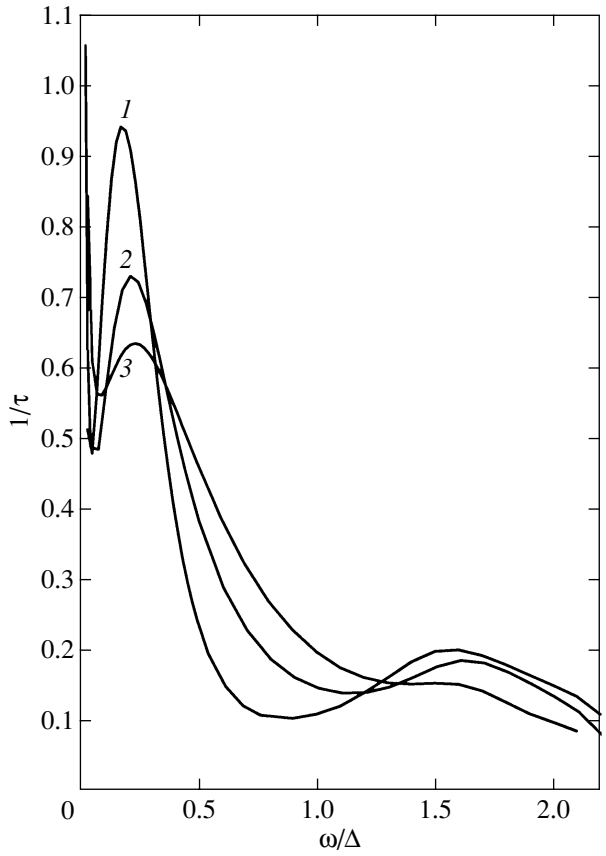


Fig. 14. Real part of optical conductivity in the spin–fermion model for $t'/t = -0.4$ and $\mu/t = -1.3$ for the values of inverse correlation length $\kappa a = 0.05$ (*1*), 0.1 (*2*), and 0.2 (*3*). The dephasing rate $\gamma/t = 0.005$.

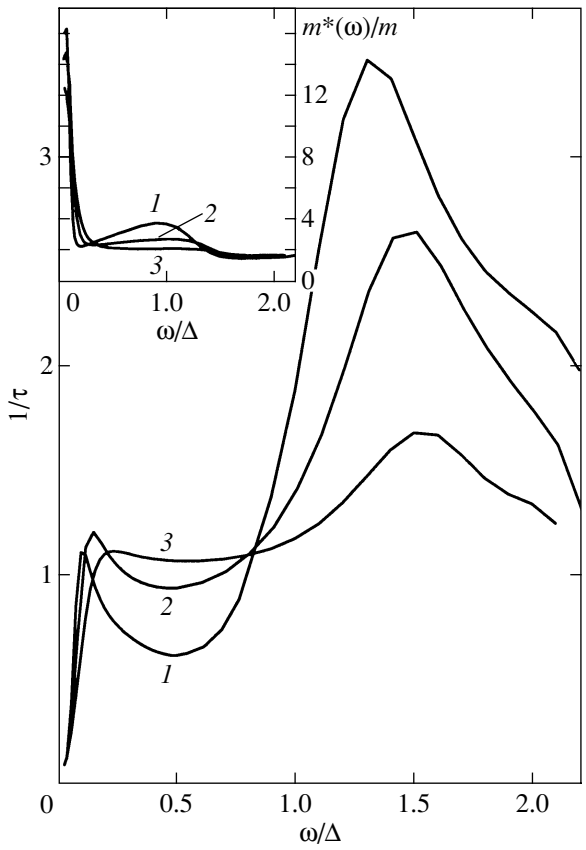


Fig. 15. Generalized scattering rate and effective mass for the case $t'/t = -0.4$ and $\mu/t = -1.3$ typical of high-temperature superconductors. The parameters of the generalized Drude model are obtained in the spin–fermion model for the values of inverse correlation length $\kappa a = 0.05$ (*1*), 0.1 (*2*), and 0.2 (*3*). The dephasing rate $\gamma/t = 0.005$.

are presented in Fig. 17, where the variation of the localization conductivity peak during the transition of chemical potential through the topological transition region can be traced. A low-intensity pseudogap absorption peak virtually remains unchanged. The inset to Fig. 17 shows the evolution of the localization peak taking into account inelastic scattering (parameter γ) for $\mu = -1.8t$. It can clearly be seen how a transition from the localization to the Drude behavior occurs due to dephasing processes. The obtained results show that the change in the Fermi surface topology itself does not lead to strong qualitative changes in optical conductivity in the framework of the model under investigation.

4. CONCLUSIONS

The above analysis demonstrates the variety of the results that can be obtained in the model under investigation for different geometries and topologies of the Fermi surface, emerging upon a change in the parameters of the "bare" quasiparticle spectrum (3). It is interesting to compare these results with those obtained earlier in the simplified model of hot patches on the Fermi surface [14]. Since the pseudogap anomalies in the hot-patches model are mainly determined by strong scattering precisely in these (flat) regions on the Fermi surface and by their relative size, the localization conductivity peak was virtually unnoticeable in this model, and the dominating role was played by the Drude peak associated with scattering from cold regions, which is determined by an auxiliary scattering rate γ (whose meaning is similar to the inelastic scattering rate introduced above). The above analysis of a more realistic model shows that the contribution of the localization peak may be quite noticeable and that it is this peak that can be transformed into a narrow Drude peak when dephasing processes are taken into account.

The main drawback of the model considered above is probably the disregard of the dynamics of short-range-order fluctuations. This approximation is justified, according to [4, 5], only at high temperatures, but the processes of inelastic scattering responsible for the dephasing and violation of localization become more significant just at such temperatures. Another draw-

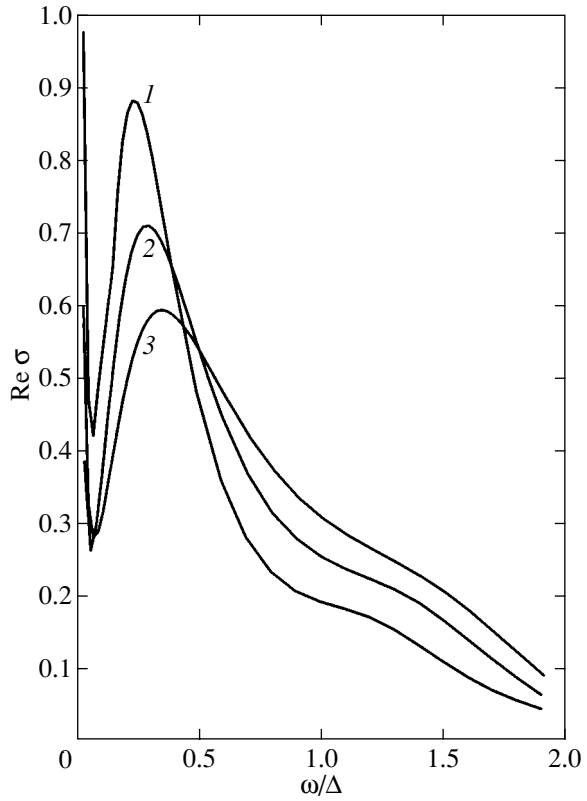


Fig. 16. Real part of optical conductivity in the commensurate case for $t'/t = -0.4$ and $\mu/t = -1.3$ for the values of inverse correlation length $\kappa a = 0.05$ (*1*), 0.1 (*2*), and 0.2 (*3*). The dephasing rate $\gamma/t = 0.005$.

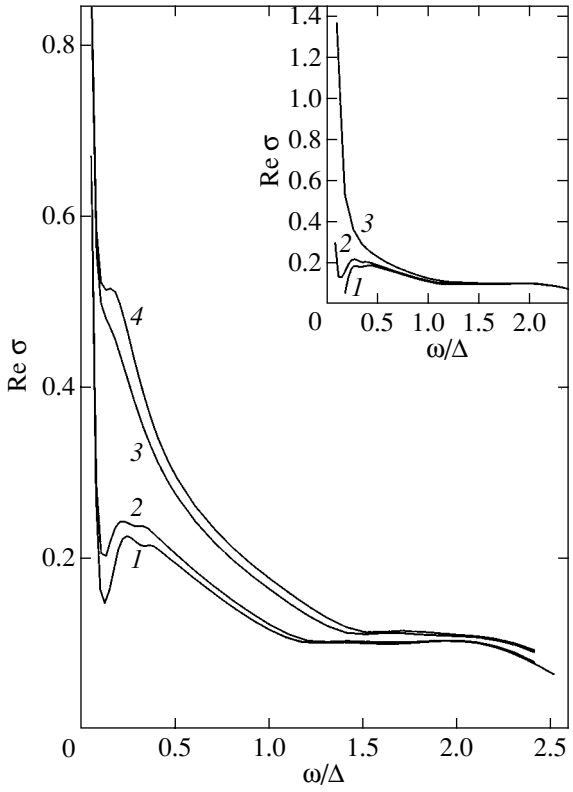


Fig. 17. Evolution of the real part of optical conductivity in the commensurate case for $t'/t = -0.6$ and $\kappa a = 0.2$ with a change in the chemical potential in the topological transition region. The curves correspond to values of $\mu/t = -1.79$ (*1*), -1.77 (*2*), -1.66 (*3*), and -1.63 (*4*). The dephasing rate $\gamma/t = 0.01$. The inset shows the real part of optical conductivity in the case when $\mu/t = -1.8$ for values of $\gamma/t = 0$ (*1*), 0.01 (*2*), and 0.05 (*3*).

back, as was noted repeatedly in [5, 7], is the confinement to the Gaussian approximation for fluctuation statistics, which can also be justified only for the region of high temperatures.

While considering a possible relation between the results obtained above and real experiments on HTSC cuprates, it should be borne in mind that no localization peak was observed in most of such experiments [1, 2], which can apparently be attributed to a noticeable role of inelastic processes (dephasing) at the high temperatures used in these experiments. Optical conductivity peaks in the low-frequency region, attributed to localization, were observed in disordered samples of the YBaCuO system in [25, 26]. Recent experiments on the NdCeCuO system [27, 28], in which such a peak was observed especially clearly, are worth mentioning. In particular, the qualitative behavior of optical conductivity observed in [28] for a series of NdCeCuO samples of various compositions (from underdoped to optimally doped) is in complete agreement with the behavior depicted in Fig. 14, which may be typical of HTSC cuprates (see above). Thus, in our opinion, the model of hot spots may claim at a realistic description of anomalies in the optical conductivity of high-temperature superconductors.

ACKNOWLEDGMENTS

The authors are grateful to É.Z. Kuchinskiĭ for numerous discussions.

This study was partly financed by the Russian Foundation for Basic Research (project no. 02-02-16031), CRDF no. REC-005, the Program of Fundamental Studies “Quantum Macrophysics” of the Presidium of the Russian Academy of Sciences, and under the project “Investigation of Collective and Quantum Effects in Condensed Media” of the Ministry of Industry and Science of the Russian Federation.

REFERENCES

1. T. Timusk and B. Statt, Rep. Prog. Phys. **62**, 61 (1999).
2. M. V. Sadovskiĭ, Usp. Fiz. Nauk **171**, 539 (2001).
3. N. P. Armitage, D. H. Lu, C. Kim, *et al.*, Phys. Rev. Lett. **87**, 147 003 (2001).
4. J. Schmalian, D. Pines, and B. Stojkovic, Phys. Rev. Lett. **80**, 3839 (1998); Phys. Rev. B **60**, 667 (1999).
5. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **115**, 1765 (1999) [JETP **88**, 968 (1999)].

6. M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **66**, 1720 (1974) [Sov. Phys. JETP **39**, 845 (1974)]; Fiz. Tverd. Tela (Leningrad) **16**, 2504 (1974) [Sov. Phys. Solid State **16**, 1632 (1974)].
7. M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **77**, 2070 (1979) [Sov. Phys. JETP **50**, 989 (1979)].
8. A. I. Posazhennikova and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **115**, 632 (1999) [JETP **88**, 347 (1999)].
9. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **117**, 613 (2000) [JETP **90**, 535 (2000)].
10. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **119**, 553 (2001) [JETP **92**, 480 (2001)].
11. É. Z. Kuchinskiĭ and M. V. Sadovskiĭ, Zh. Éksp. Teor. Fiz. **121**, 758 (2002) [JETP **94**, 654 (2002)].
12. M. V. Sadovskiĭ and A. A. Timofeev, Sverkhprovodimost: Fiz., Khim., Tekh. **4**, 11 (1991).
13. M. V. Sadovskii and A. A. Timofeev, J. Mosc. Phys. Soc. **1**, 391 (1991).
14. M. V. Sadovskiĭ, Pis'ma Zh. Éksp. Teor. Fiz. **69**, 447 (1999) [JETP Lett. **69**, 483 (1999)]; Physica C (Amsterdam) **341–348**, 939 (2000).
15. P. Monthoux, A. Balatsky, and D. Pines, Phys. Rev. B **46**, 14 803 (1992).
16. P. Monthoux and A. Balatsky, Phys. Rev. B **47**, 6069 (1993); Phys. Rev. B **48**, 4261 (1994).
17. M. V. Sadovskii, Physica C (Amsterdam) **341–348**, 811 (2000).
18. L. Bartosch and P. Kopietz, Phys. Rev. B **60**, 15 488 (1999).
19. D. Vollhardt and P. Wolfle, Phys. Rev. B **22**, 4666 (1980).
20. L. P. Gor'kov, A. I. Larkin, and D. E. Khmel'nitskiĭ, Pis'ma Zh. Éksp. Teor. Fiz. **30**, 248 (1979) [JETP Lett. **30**, 228 (1979)].
21. P. Monthoux and D. J. Scalapino, Phys. Rev. B **65**, 235 104 (2002).
22. A. A. Gogolin and G. T. Zimanyi, Solid State Commun. **46**, 469 (1983).
23. X. J. Zhou, P. Bogdanov, S. A. Kellar, *et al.*, Science **286**, 268 (1999).
24. A. Damascelli, D. H. Lu, and Z.-X. Shen, cond-mat/0107042.
25. D. N. Basov, A. V. Puchkov, R. A. Hughes, *et al.*, Phys. Rev. B **49**, 12 165 (1994).
26. D. N. Basov, B. Dabrowski, and T. Timusk, Phys. Rev. Lett. **81**, 2132 (1998).
27. E. J. Singley, D. N. Basov, K. Kurahashi, *et al.*, cond-mat/0103480.
28. Y. Onose, Y. Taguchi, K. Ishizaka, and Y. Tokura, Phys. Rev. Lett. **87**, 217 001 (2001).

Translated by N. Wadhwa

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 538–549.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 624–635.

SOLIDS
Electronic Properties

One-Dimensional Anisotropic Heisenberg Model in the Transverse Magnetic Field¶

D. V. Dmitriev[a, b, *], V. Ya. Krivnov[a, b], A. A. Ovchinnikov[a, b], and A. Langari[b, c]

[a] *Joint Institute of Chemical Physics, Russian Academy of Sciences, Moscow, 117977 Russia*
**e-mail: dmitriev@deom.chph.ras.ru*
[b] *Max-Planck-Institut für Physik Komplexer Systeme 01187, Dresden, Germany*
[c] *Institute for Advanced Studies in Basic Sciences, Zanjan 45195-159, Iran*

Received March 29, 2002

Abstract—The one-dimensional spin-1/2 *XXZ* model in a transverse magnetic field is studied. It is shown that the field induces a gap in the spectrum of the model with the easy-plane anisotropy. Using conformal invariance, the field dependence of the gap is found at small fields. The ground state phase diagram is obtained. It contains four phases with the long-range order of different types and a disordered phase. These phases are separated by critical lines, where the gap and the long-range order vanish. Using scaling estimates, the mean-field approach, and numerical calculations in the vicinity of all critical lines, we find the critical exponents of the gap and the long-range order. It is shown that the transition line between the ordered and disordered phases belongs to the universality class of the transverse Ising model. © *2002 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

The effect of the magnetic field on an antiferromagnetic chain has been attracting much interest from theoretical and experimental standpoints. In particular, a strong dependence of the properties of quasi-one-dimensional anisotropic antiferromagnets on the field orientation was observed experimentally [1]. It is therefore interesting to study the dependence of properties of the one-dimensional antiferromagnet on the direction of the applied field. The simplest model of the one-dimensional anisotropic antiferromagnet is the spin-1/2 *XXZ* model. This model in a uniform longitudinal magnetic field (along the *z* axis) was studied in great detail [2]. Because the longitudinal field commutes with the *XXZ* Hamiltonian, the model can be exactly solved by the Bethe ansatz. This is not the case if the symmetry-breaking transverse magnetic field is applied and the exact integrability is lost. Because of its mathematical complexity, this model has not been studied much. From this standpoint, it is of particular interest to study the ground state properties of the 1*D XXZ* model in the transverse magnetic field. The Hamiltonian of this model is given by

$$H = \sum_{n=1}^{N} (S_n^x S_{n+1}^x + S_n^y S_{n+1}^y + \Delta S_n^z S_{n+1}^z) + h \sum_{n=1}^{N} S_n^x \quad (1)$$

with periodic boundary conditions and even *N*.

The spectrum of the *XXZ* model is gapless for $-1 < \Delta \le 1$. In the longitudinal field, the spectrum remains gapless if the field does not exceed the saturation value $(1 + \Delta)$. On the other hand, a gap in the excitation spectrum seems to open up when the transverse magnetic field is applied. It is supposed [3] that this effect can explain the peculiarity of the low-temperature specific heat in Yb_4As_3 [1]. The magnetic properties of this compound are described by the *XXZ* Hamiltonian with $\Delta \approx 0.98$; it was shown that the magnetic field in the easy plane induces a gap in the spectrum resulting in a dramatic decrease of the linear term in the specific heat [3].

First of all, what do we know about model (1)?

The first part of the Hamiltonian is the well-known *XXZ* model, whose exact solution is given by the Bethe ansatz. In the Ising-like region $\Delta > 1$, the ground state of the *XXZ* model has a Néel long-range order along the *z* axis and there is a gap in the excitation spectrum. In the region $-1 < \Delta \le 1$, the system is in the so-called spin-liquid phase with a power-law decay of correlations and a linear spectrum. Finally, for $\Delta < -1$, the classical ferromagnetic state is the ground state of the *XXZ* model with a gap over the ferromagnetic state.

In the transverse magnetic field, the total spin projection S^z is not a good quantum number and the model is essentially complicated, because the transverse field breaks rotational symmetry in the *xy* plane and destroys the integrability of the *XXZ* model, except at some special points. In particular, the exact diagonalization study of this model is difficult for finite systems because of a nonmonotonic behavior of energy levels.

The first special case of model (1) is the limit as $\Delta \longrightarrow \pm\infty$. In this case, the model reduces to the 1*D* Ising model in a transverse field (ITF), which can be exactly solved by transforming it to the system of noninteracting fermions. In both limits, the system has the

¶This article was submitted by the authors in English.

phase transition point $h_c = |\Delta|/2$, where the gap closes and the long-range order in the z direction vanishes.

It is suggested [4] that the phase transition of the ITF type occurs for any $\Delta > 0$ at some critical value $h = h_c(\Delta)$. It can also be expected that such a transition exists for any Δ and the transition line connects two limiting points $h_c = |\Delta|/2$, $\Delta \longrightarrow \pm\infty$.

Similarly to these limiting cases, for any $|\Delta| > 1$ and $h < h_c(\Delta)$, the system has a long-range order in the z direction (the Néel order for $\Delta > 1$ and the ferromagnetic order for $\Delta < -1$). But for $|\Delta| < 1$ and $h < h_c(\Delta)$, the ground state changes, and instead of the long-range order in the z direction, a staggered magnetization along the y axis appears at $h < h_c(\Delta)$.

This assumption is confirmed on the "classical" line $h_{\mathrm{cl}} = \sqrt{2(1+\Delta)}$ ($h_{\mathrm{cl}} < h_c(\Delta)$), where the quantum fluctuations of the *XXZ* model are compensated by the transverse field and the exact ground state of (1) at $h = h_{\mathrm{cl}}$ is a classical one [5]. The excited states on the classical line are generally unknown, although it is assumed that the spectrum is gapped.

The second case where model (1) remains integrable is the isotropic antiferromagnetic case $\Delta = 1$. In this case, the direction of the magnetic field is not important and the ground state of the system remains the spin-liquid one up to the point $h = 2$, where a phase transition of the Pokrovsky–Talapov type occurs and the ground state becomes a completely ordered ferromagnetic state.

The last special case is $\Delta = -1$. Model (1) then reduces to the isotropic ferromagnetic model in a staggered magnetic field. This model is nonintegrable, but as shown [6], the system remains gapless up to some critical value $h = h_0$, where a phase transition of the Kosterlitz–Thouless type occurs.

Summarizing, we expect that the phase diagram of model (1) (in the (Δ, h) plane) has the form shown in Fig. 1. The phase diagram contains four regions that correspond to different phases and are separated by transition lines. Each phase is characterized by a long-range order of its own type the Néel order along the z axis in region (1); the ferromagnetic order along the z axis in region (2); the Néel order along the y axis in region (3); in region (4), there is no long-range order except the magnetization along the field direction x (which certainly exists in all the above regions). By the long-range order, we hereafter understand the one of the type corresponding to a given region.

In this paper, we investigate the behavior of the gap and the long-range order near the transition (critical) lines. In Section 2, devoted to the classical line, we review the exact ground state and construct three exact excitations. In Section 3, we study the transition line $h_c(\Delta)$ using the mean-field approach and the exact diagonalization of finite systems. In Section 4, we find the critical exponents in the vicinity of the line $h = 0$. The properties of the model near the critical lines $\Delta = \pm 1$ and in the vicinity of the points ($\Delta = \pm 1$, $h = 0$), in particular, are studied in Sections 5 and 6.

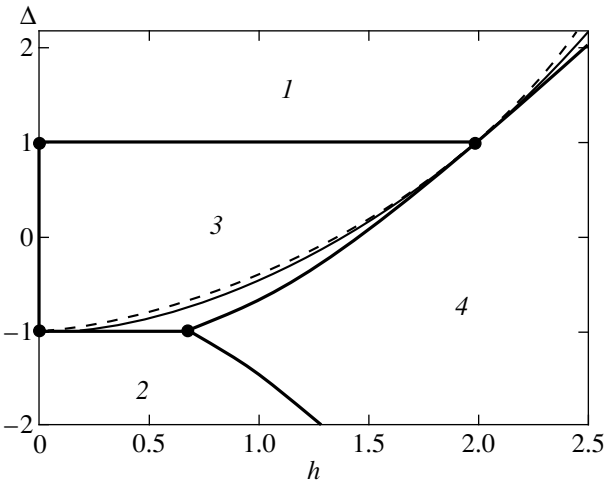


Fig. 1. Phase diagram of model (1). The thick solid lines are the critical lines, the thin solid line is the "classical" line, and the dashed line is the line $h_1(\Delta)$.

2. THE CLASSICAL LINE

We first we consider the classical line

$$h_{\mathrm{cl}} = \sqrt{2(1+\Delta)}, \quad \Delta > -1,$$

because we often refer to it in what follows. It is remarkable in the sense that the ground state is identical to the classical one on this line and quantum fluctuations are missing. It was shown in [5] that the ground state of (1) is twofold degenerate on this line and the ground state wave functions with the momentum $k = 0$ and $k = \pi$ are given by

$$\Psi_{1,2} = \frac{1}{\sqrt{2}}(\Phi_1 \pm \Phi_2),$$

where $\Phi_{1(2)}$ are direct products of single-site functions,

$$|\Phi_1\rangle = |\alpha_1\bar{\alpha}_2\alpha_3\bar{\alpha}_4\ldots\rangle,$$

$$|\Phi_2\rangle = |\bar{\alpha}_1\alpha_2\bar{\alpha}_3\alpha_4\ldots\rangle.$$

Here, $|\alpha_i\rangle$ is the state of the ith spin lying in the xy plane for $|\Delta| < 1$ (or in the xz plane for $\Delta > 1$) at the angle φ with the x axis. These states can be written as

$$|\alpha_i\rangle = (e^{i\varphi}S_i^+ - 1)|\downarrow\rangle, \quad |\Delta| < 1,$$

$$|\alpha_i\rangle = (e^{\varphi}S_i^+ - 1)|\downarrow\rangle, \quad \Delta > 1$$

with

$$\cos\varphi = h_{\mathrm{cl}}/2, \quad |\Delta| < 1$$

and

$$\cosh\varphi = h_{cl}/2, \quad \Delta > 1.$$

The state $|\bar{\alpha}_i\rangle$ is obtained by rotation of the ith spin by π about the magnetic field axis x,

$$|\bar{\alpha}_i\rangle = e^{i\pi S_i^x}|\alpha_i\rangle.$$

The ground state has a two-sublattice structure and is characterized by the presence of the long-range order in the y ($|\Delta| < 1$) or in the z ($\Delta > 1$) directions. In particular, for $|\Delta| < 1$, the staggered magnetization $\langle S_n^y\rangle$ is

$$\langle S_n^y\rangle = \frac{(-1)^n}{2}\sqrt{1-\frac{h_{cl}^2}{4}}.$$

In general, the excited states of (1) on the classical line are nontrivial. Some of them can nevertheless be found exactly. For this, it is convenient to introduce the operator overturning the ith spin,

$$R_i = e^{i\pi S_i^z}, \quad |\Delta| < 1,$$

$$R_i = e^{i\pi S_i^y}, \quad \Delta > 1,$$

such that the states of the "overturned" ith spin $|\beta_i\rangle = R_i|\alpha_i\rangle$ and $|\bar{\beta}_i\rangle = R_i|\bar{\alpha}_i\rangle$ are orthogonal to $|\alpha_i\rangle$ and $|\bar{\alpha}_i\rangle$,

$$\langle\alpha_i|\beta_i\rangle = \langle\bar{\alpha}_i|\bar{\beta}_i\rangle = 0.$$

The exact excited states are then written as

$$|\psi^1_{1(2)}\rangle = \sum_m R_m|\Phi_{1(2)}\rangle,$$

$$|\psi^2_{1(2)}\rangle = \sum_n (-1)^n R_n R_{n+1}|\Phi_{1(2)}\rangle,$$

$$|\psi^3_{1(2)}\rangle = \sum_{n,m} (-1)^n R_n R_{n+1} R_m|\Phi_{1(2)}\rangle,$$

and therefore, each of the three exact excitations is also twofold degenerate. This degeneracy is in fact a consequence of the Z_2 symmetry describing the rotation of all spins by π about the magnetic field axis x.

To show that these states are indeed the exact ones, it is convenient to rotate the coordinate system such that in one of the ground states, for example, Φ_1, all spins point down. In the case where $|\Delta| < 1$, this transformation is the rotation of the spins at even (odd) sites by an angle φ ($-\varphi$) around the z axis followed by the rotation by $\pi/2$ around the y axis,

$$\begin{aligned} S_n^x &= \sigma_n^z\cos\varphi + (-1)^n\sigma_n^y\sin\varphi, \\ S_n^y &= (-1)^n\sigma_n^z\sin\varphi - \sigma_n^y\cos\varphi, \\ S_n^z &= -\sigma_n^x. \end{aligned} \tag{2}$$

In the case where $\Delta > 1$, the transformation of the spin operators is defined by

$$\begin{aligned} S_n^x &= \sigma_n^z\cos\varphi + (-1)^n\sigma_n^x\sin\varphi, \\ S_n^y &= \sigma_n^y, \\ S_n^z &= -(-1)^n\sigma_n^z\sin\varphi + \sigma_n^x\cos\varphi. \end{aligned} \tag{3}$$

On the classical line, Hamiltonian (1) then becomes

$$\begin{aligned} H_1 &= \Delta\sum_n \sigma_n\sigma_{n+1} + (1+\Delta)\sum_n \sigma_n^z \\ &+ h_{cl}\sqrt{1-\frac{h_{cl}^2}{4}}\sum_n (-1)^n\sigma_n^y(\sigma_{n+1}^z + \sigma_{n-1}^z + 1) \end{aligned} \tag{4}$$

for $\Delta < 1$ and

$$\begin{aligned} H_2 &= \sum_n \sigma_n\sigma_{n+1} - (\Delta-1)\sum_n \sigma_n^z\sigma_{n+1}^z \\ &+ 2\sum_n \sigma_n^z + \sqrt{h_{cl}^2-4}\sum_n (-1)^n\sigma_n^x(\sigma_{n+1}^z + \sigma_{n-1}^z + 1) \end{aligned} \tag{5}$$

for $\Delta > 1$.

The ground state of both Hamiltonians and of (1) is twofold degenerate. Obviously, in one of the ground states, all spins σ_n point down,

$$\Phi_1 = |0\rangle \equiv |\downarrow\downarrow\downarrow\ldots\rangle.$$

The energy of this state is

$$E_0 = -\frac{N}{2} - N\frac{\Delta}{4}. \tag{6}$$

In this representation, the second ground state Φ_2 has a more complicated form,

$$\tilde{\Phi}_2 = \prod_n (\cos\varphi + (-1)^n\sigma_n^+\sin\varphi)|0\rangle.$$

It is now easy to see that the following three excited states are exact:

$$\begin{aligned} |\psi_1^{(1)}\rangle &= \sum_n \sigma_n^+|0\rangle, \quad E_1 - E_0 = 1+\Delta, \\ |\psi_1^{(2)}\rangle &= \sum_n (-1)^n\sigma_n^+\sigma_{n+1}^+|0\rangle, \\ E_2 - E_0 &= 2+\Delta, \\ |\psi_1^{(3)}\rangle &= \sum_{n,m} (-1)^n\sigma_n^+\sigma_{n+1}^+\sigma_m^+|0\rangle, \\ E_3 - E_0 &= 3+2\Delta. \end{aligned} \tag{7}$$

It can be verified that the last terms in (4) and (5) annihilate these three functions and are therefore the exact excited states of (1) for any even N. Similarly to

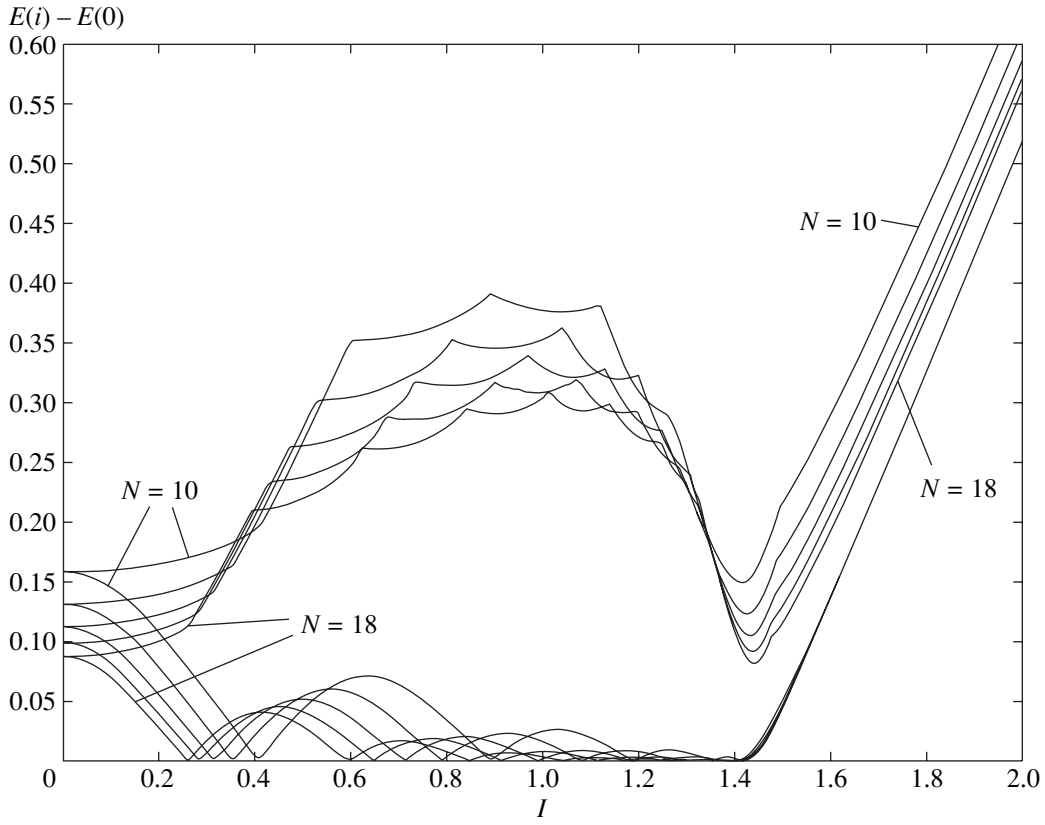


Fig. 2. The dependence of the difference between the energy of the two lowest levels $E(1)$, $E(2)$ and the ground state energy $E(0)$ on magnetic field h for finite chains with $N = 10$–18.

the ground state, excited states (7) are degenerate with the states $|\psi_2^k\rangle$. These states $|\psi_2^k\rangle$ can be represented in the same form (7), but in the coordinate system where the function Φ_2 describes all spins pointing down.

The states $|\psi_{1(2)}^1\rangle$ are especially interesting because they define the gap of model (1) on the classical line at small values of h_{cl}. Our numerical calculations of finite systems show that, as $h_{cl} \longrightarrow 0$ ($\Delta \longrightarrow -1$), the lowest branch of the excitations has a minimum at $k = 0$ and the corresponding excitation energy is $(1 + \Delta)$ (of course, because of the Z_2 symmetry, there is another branch with the minimum at $k = \pi$ and the same minimum energy, but we consider one branch only). The excitation energy at $k = \pi$ obtained by the extrapolation of numerical calculations as $N \longrightarrow \infty$ is $2(1 + \Delta)$. As h_{cl} increases, the excitation energies at $k = 0$ and $k = \pi$ are drawn together and become equal to each other at some $\tilde{h}_{cl}$. Our numerical results give

$$\tilde{h}_{cl} \approx 0.76 \quad (\Delta \approx -0.79).$$

On the classical line, the gap is therefore $(1 + \Delta)$ for $-1 < \Delta < -0.79$.

3. THE TRANSITION LINE $h = h_c(\Delta)$

The existence of the transition line $h_c(\Delta)$ passing through the entire phase diagram is quite natural, because all types of long-range order except the long-range order along the field must vanish at some value of the magnetic field. The transition line connects two obvious limits as $\Delta \longrightarrow \pm\infty$, where model (1) reduces to the ITF model. The line passes through the exactly solvable point ($\Delta = 1$, $h = 2$) and the point ($\Delta = -1$, $h = h_0$) studied in [6]. We suppose that the entire line $h_c(\Delta)$ is of the ITF type with algebraically decaying correlations with the corresponding critical exponents [7].

The transition line can also be observed from the numerical calculations of finite systems. As an example, the dependences of the excitation energies of the three lowest levels on h are shown in Fig. 2 for $\Delta = 0$ and for $N = 10$–18. From this figure, it can be seen that the two lowest states cross each other $N/2$ times and the last crossing occurs at the classical point $h_{cl} = \sqrt{2}$. These two states form a twofold degenerate ground state in the thermodynamic limit. They have different momenta $k = 0$ and π and different quantum numbers describing the Z_2 symmetry that remains in the system after applying the field. As for the first excitation above the degenerate ground state, we also see numerous level crossings in Fig. 2. These level crossings lead to incom-

mensurate effects that manifest themselves in the oscillatory behavior of the spin correlation functions. The correlation functions at $n \gg 1$ are given by

$$\langle S_1^\alpha S_n^\alpha \rangle - \langle S^\alpha \rangle^2 = f(n) e^{-\kappa n}, \tag{8}$$

where $\langle S^\alpha \rangle$ ($\alpha = x, y, z$) is the corresponding magnetization (the long-range order) and is the oscillatory function of n with the oscillation period depending on h and Δ. All crossings disappear at $h > h_{cl}(\Delta)$ and the correlation functions do not contain oscillatory terms in this region of the phase diagram.

The energy of the first excitation near h_{cl} decreases rapidly, and after extrapolation we found that, for $\Delta = 0$, the gap vanishes at the magnetic field $h_c \approx 1.456(6) > h_{cl}$. Inside the region $h_{cl} < h < h_c$, the ground state remains twofold degenerate, although there are no level crossings. At $h > h_c$, the mass gap appears again; for a large field, the gap is proportional to h.

To determine the transition line $h_c(\Delta)$ and to study the model in the vicinity of $h_c(\Delta)$, we use the Fermi representation of (1). This representation gives the exact solution in the limits as $\Delta \longrightarrow \pm\infty$ and in addition yields the exact ground state on the classical line.

First, it is convenient to perform a rotation of the spins around the y axis by $\pi/2$ in (1) such that the magnetic field is directed along the z axis,

$$H = \sum_n (\Delta S_n^x S_{n+1}^x + S_n^y S_{n+1}^y + S_n^z S_{n+1}^z) + h \sum_n S_n^z. \tag{9}$$

After the Jordan–Wigner transformation to Fermi operators a_n^+ and a_n,

$$\begin{aligned} S_n^+ &= e^{i\pi \Sigma a_j^+ a_j} a_n, \\ S_n^z &= a_n^+ a_n - \frac{1}{2}, \end{aligned} \tag{10}$$

Hamiltonian (9) becomes

$$\begin{aligned} H_f = &-\frac{hN}{2} + \frac{N}{4} + \sum_k \left(h - 1 - \frac{1+\Delta}{2}\cos k\right) a_k^+ a_k \\ &+ \frac{1-\Delta}{4} \sum_k \sin k (a_k^+ a_{-k}^+ + a_{-k} a_k) + \sum_n a_n^+ a_n a_{n+1}^+ a_{n+1}. \end{aligned} \tag{11}$$

Treating the Hamiltonian H_f in the mean-field approximation, we find the ground state energy E_0 and the one-particle excitation spectrum $\varepsilon(k)$,

$$\begin{aligned} \frac{E_0}{N} = &(h-1)\left(\gamma_1 - \frac{1}{2}\right) + \frac{1}{4} - \left(1 - \frac{g}{2}\right)\gamma_2 \\ &+ \frac{g}{2}\gamma_3 + \gamma_1^2 - \gamma_2^2 + \gamma_3^2, \end{aligned} \tag{12}$$

$$\varepsilon(k) = \sqrt{a^2(k) + b^2(k)}, \tag{13}$$

where $g = 1 - \Delta$ and

$$\begin{aligned} a(k) &= (h-1) - \left(1 - \frac{g}{2}\right)\cos k + 2\gamma_1 - 2\gamma_2 \cos k, \\ b(k) &= \left(\frac{g}{2} + 2\gamma_3\right)\sin k. \end{aligned} \tag{14}$$

The quantities γ_1, γ_2, and γ_3 are the ground state averages determined by the self-consistent equations

$$\begin{aligned} \gamma_1 &= \langle a_n^+ a_n \rangle = \sum_{k>0} \left(1 - \frac{a(k)}{\varepsilon(k)}\right), \\ \gamma_2 &= \langle a_n^+ a_{n+1} \rangle = -\sum_{k>0} \frac{a(k)}{\varepsilon(k)} \cos k, \\ \gamma_3 &= \langle a_n^+ a_{n+1}^+ \rangle = -\sum_{k>0} \frac{b(k)}{2\varepsilon(k)} \sin k. \end{aligned} \tag{15}$$

The magnetization $S = \langle S_n^x \rangle$ of model (1) is given by

$$S = \frac{1}{2} - \gamma_1. \tag{16}$$

The numerical solution of Eq. (15) shows that the function $\varepsilon(k)$ has a minimum at k_{min}, which changes from $\pi/2$ at $h = 0$ to zero at $h = h_1(\Delta)$ and $k_{min} = 0$ for $h > h_1(\Delta)$. The gap in the spectrum $\varepsilon(k)$ vanishes at $h_c(\Delta)$ ($h_c > h_1$) and is given by $m = |h - h_c|$ for $h > h_1$. The functions $h_1(\Delta)$ and $h_c(\Delta)$ are shown in Fig. 1. We note that the Hamiltonian H_f differs from the domain-wall fermionic Hamiltonian that is mapped from (1) in [4]. The transition line obtained in [4] is a linear function of Δ in contrast to $h_c(\Delta)$ in Fig. 1.

It is interesting to note that the mean-field approximation gives the exact ground state on the classical line $h_{cl} = \sqrt{2(1+\Delta)}$. On this line, the solution of Eqs. (15) has the simple form

$$\begin{aligned} &\gamma_1 = \frac{1}{2} - \frac{h_{cl}}{4}, \quad \gamma_2 = -\gamma_3 = \frac{4 - h_{cl}^2}{16}, \quad |\Delta| < 1, \\ &\gamma_1 = \frac{1}{2} - \frac{1}{h_{cl}}, \quad \gamma_2 = \gamma_3 = \frac{h_{cl}^2 - 4}{4h_{cl}^2}, \quad \Delta > 1, \end{aligned} \tag{17}$$

and the energy is given by

$$\frac{E_0}{N} = -\frac{1}{2} - \frac{\Delta}{4}.$$

On the classical line in the mean-field approximation, the gap is

$$m = \frac{1}{4}(2-h_{\rm cl})^2, \quad |\Delta|<1,$$
$$m = \frac{h_{\rm cl}^2-2}{2h_{\rm cl}^2}(h_{\rm cl}-2)^2, \quad \Delta>1. \tag{18}$$

We compared (18) with the results of the extrapolation of finite systems in the classical line. The coincidence is sufficiently good for $\Delta > 0.5$. Equation (18) gives a satisfactory estimate for the gap up to $\Delta \approx -0.5$. For example, at $\Delta = 0$ ($h_{\rm cl} = \sqrt{2}$), it follows that $m = 0.086$ from Eq. (18), while the extrapolated gap is $m \approx 0.076(4)$.

The smaller the fermion density, the better the mean-field approximation works. It becomes worse as the magnetization $S \longrightarrow 0$. This is the reason for incorrect behavior of the gap as $h_{\rm cl} \longrightarrow 0$ ($\Delta \longrightarrow -1$). It follows from (18) that $m = 1$, while m vanishes in this limit as $m = (1+\Delta)$ (7).

In the mean-field approximation, the Hamiltonian H_f is similar to the well-known bilinear Fermi Hamiltonian describing the anisotropic XY model or the ITF model. Using results in [7], the following facts related to the model under consideration can be established.

1. There is a staggered magnetization $\langle S_n^y\rangle$ along the y axis for $|\Delta| < 1$ or $\langle S_n^z\rangle$ along the z axis for $|\Delta| > 1$, and they vanish as $(h_c - h)^{1/8}$ for $h \longrightarrow h_c$.

2. The magnetization S has a logarithmic singularity as $h \longrightarrow h_c$.

3. The spin correlation function decays exponentially (excluding the transition line) as $n \longrightarrow \infty$,

$$G^\alpha(n) = \langle S_1^\alpha S_n^\alpha\rangle - \langle S^\alpha\rangle^2 = f(n)e^{-\kappa n}. \tag{19}$$

The function $f(n)$ has an oscillatory behavior for $0 < h < h_{\rm cl}$ and is monotonic for $h > h_{\rm cl}$; $f(n) = 0$ at $h = h_{\rm cl}$, and

$$f(n) \approx \frac{\cos\omega n}{n^2}, \quad \omega = \sqrt{2\frac{h_{\rm cl}-h}{h_{\rm cl}}}$$

for $h_{\rm cl} - h \ll 1$. The classical line therefore determines the boundary on the phase diagram where the spin correlation functions show the incommensurate behavior.

On the transition line $h = h_c(\Delta)$, the spin correlation functions have a power-law decay,

$$G^x(n)\propto 1/n^2, \quad G^y(n)\propto 1/n^{1/4},$$
$$G^z(n)\propto 1/n^{9/4}, \quad |\Delta|<1,$$
$$G^x(n)\propto 1/n^2, \quad G^y(n)\propto 1/n^{9/4},$$
$$G^z(n)\propto 1/n^{1/4}, \quad |\Delta|>1. \tag{20}$$

These results show that the transition at $h = h_c(\Delta)$ belongs to the universality class of the ITF model.

In the vicinity of the point $h = 2$, $\Delta = 1$, the fermion density is small ($S \approx 1/2$) and the mean-field approximation of the four-fermion term gives an accuracy up to g^3 or $(2-h)^4$ at least. In this case, we give the corresponding expressions (for $g \ll 1$):

$$h_c = 2 - \frac{g}{2} - \frac{g^2}{32}, \quad h_1 = h_c - \frac{g^2}{16},$$
$$m = \begin{cases} |h-h_c|, & h > h_1 \\ \dfrac{g}{2\sqrt{2}}\sqrt{h_c - h - \dfrac{g^2}{32}}, & h < h_1. \end{cases} \tag{21}$$

The magnetization S is

$$S = \begin{cases} \dfrac{1}{2} - \dfrac{\sqrt{2}}{\pi}\sqrt{h_c-h} - \dfrac{g}{8\pi}, & g \ll \sqrt{h_c-h} \\ \dfrac{1}{2} - \dfrac{g}{4\pi} - \dfrac{2(h_c-h)}{\pi g}\ln\left(\dfrac{g^2}{h_c-h}\right), & g \ll \sqrt{h_c-h}. \end{cases} \tag{22}$$

The susceptibility $\chi(h) = dS/dh$ is

$$\chi(h) = \begin{cases} \dfrac{2}{\pi g}\ln\left(\dfrac{g^2}{h_c-h}\right), & g \gg \sqrt{h_c-h} \\ \dfrac{1}{\sqrt{2}\pi}\dfrac{1}{\sqrt{h_c-h}}, & g \ll \sqrt{h_c-h}. \end{cases} \tag{23}$$

It follows from (23) that there is a crossover from the square root to the logarithmic divergence of χ as the parameter $g^2/(h_c - h)$ varies from 0 to ∞.

4. THE LINE $h = 0$, $|\Delta| < 1$

4.1. Scaling Estimates

The XXZ model is integrable and its low-energy properties are described by a free massless boson field theory with the Hamiltonian

$$H_0 = \frac{v}{2}\int dx[\Pi^2 + (\partial_x\Phi)^2], \tag{24}$$

where $\Pi(x)$ is the momentum conjugate to the boson field $\Phi(x)$, which can be separated into the left and right moving terms,

$$\Phi = \Phi_L + \Phi_R.$$

The dual field $\tilde{\Phi}$ is defined as the difference

$$\tilde{\Phi} = \Phi_L - \Phi_R.$$

The spin-density operators are represented as

$$S_n^z \approx \frac{1}{2\pi R}\partial_x\Phi + \text{const}(-1)^n\cos\frac{\Phi}{R},$$
$$S_n^x \approx \cos(2\pi R\tilde{\Phi})\left[C(-1)^n + \text{const}\cdot\cos\frac{\Phi}{R}\right] \quad (25)$$

with the constant C found in [8]. The compactification radius R is known from the exact solution

$$2\pi R^2 = \theta = 1 - \frac{\arccos\Delta}{\pi}.$$

The nonoscillating part of the operator S^x in Eq. (25) has the scaling dimension

$$d = \frac{\theta}{2} + \frac{1}{2\theta}$$

and conformal spin $S = 1$. A nonzero conformal spin of the perturbation operator S^x can lead to incommensurability in the system [9], which agrees with Eq. (19). As shown in [10], the general formula for the mass gap

$$m \sim h^\nu, \quad \nu = \frac{1}{2-d} = \frac{2}{4-\theta-1/\theta}, \quad (26)$$

is not applicable in the entire region $|\Delta| < 1$. Because of a nonzero conformal spin of the nonoscillating part of the operator S^x, higher order effects in h must be considered. The analysis shows [10] that the original perturbation with a nonzero conformal spin generates another perturbation with zero conformal spin,

$$V = h^2\cos(4\pi R\tilde{\Phi}). \quad (27)$$

This perturbation gives the critical exponent for the mass gap

$$m \sim h^\gamma, \quad \gamma = \frac{1}{1-\theta}. \quad (28)$$

Comparing Eqs. (26) and (28), we see that perturbation (27) becomes more relevant in the region

$$\Delta < \cos(\pi\sqrt{2}) \approx -0.266.$$

It turns out that the oscillating part of the operator S^x gives another, more relevant, index for the gap at $\Delta < 0$. We now reproduce the standard "conformal" chain of arguments for this oscillating part. The perturbed action of the model is given by

$$S = S_0 + h\int dt\,dx\,S^x(x, t), \quad (29)$$

where S_0 is the Gaussian action of the *XXZ* model. The time-dependent correlation functions of the *XXZ* chain show the power-law decay at $|\Delta| < 1$ and have the asymptotic form [11]

$$\langle S^x(x,\tau)S^x(0,0)\rangle \sim \frac{(-1)^x A_1}{(x^2+v^2\tau^2)^{\theta/2}} - \frac{A_2}{(x^2+v^2\tau^2)^{\theta/2+1/2\theta}}, \quad (30)$$

where A_1 and A_2 are known constants [8] and $\tau = it$ is the imaginary time. We can therefore estimate the large-distance contribution to the action of the oscillating part of the operator $S^x(x, \tau)$ as

$$h\int d\tau dx S^x(x,\tau) \sim h\int d\tau\sum_n \frac{(-1)^n}{(n^2+v^2\tau^2)^{\theta/4}}$$
$$\sim h\int d\tau\sum_{\text{even }n}\frac{\theta n}{(n^2+v^2\tau^2)^{\theta/4}} \sim h\int d\tau dx\frac{\theta x}{(x^2+v^2\tau^2)^{\theta/4+1}}.$$

The relevant field $S^x(x, \tau)$ leads to a finite correlation length ξ. This correlation length is such that the contribution of the field $S^x(x, \tau)$ to the action is of the order of unity. That is,

$$h\int_0^{\xi/v} d\tau\int_0^{\xi} dx\frac{\theta x}{(x^2+v^2\tau^2)^{\theta/4+1}} \sim \frac{\theta h\xi^{1-\theta/2}}{v} \sim 1$$

which gives the mass gap

$$m \sim \frac{v}{\xi} \sim h^\mu, \quad \mu = \frac{1}{1-\theta/2}. \quad (31)$$

In fact, the oscillating factor $(-1)^n$ in the correlator in some sense eliminates one singular integration over x, and the general conformal formula

$$m \propto h^{1/(D-d)},$$

where D is the dimension of space and d is the scaling dimension of the perturbation operator, must be taken with $D = 1$ instead of conventional $D = 2$.

The comparison of Eqs. (26), (28), and (31) shows that, for $0 < \Delta < 1$, the leading term is given by Eq. (26) and, for $-1 < \Delta < 0$, by Eq. (31). We therefore have

$$m \sim h^\nu, \quad 0 < \Delta < 1,$$
$$m \sim h^\mu, \quad -1 < \Delta < 0. \quad (32)$$

The functions $\nu(\Delta)$, $\mu(\Delta)$, and $\gamma(\Delta)$ are shown in Fig. 3. In this respect, model (1) is different from the *XXZ* model in the staggered transverse field, for which

$$m \propto h^{2/(4-\theta)}$$

for all $|\Delta| < 1$ [12].

The staggered magnetization (long-range order) along the *y* axis behaves as

$$\langle S_n^y \rangle \sim \frac{(-1)^n}{\xi^{\theta/2}} \sim (-1)^n m^{\theta/2}. \quad (33)$$

Hence, the long-range order also has two different critical exponents,

$$\begin{aligned} \langle |S^y| \rangle &\sim h^{\theta/(4-\theta-1/\theta)}, \quad 0 < \Delta < 1, \\ \langle |S^y| \rangle &\sim h^{\theta/(2-\theta)}, \quad -1 < \Delta < 0. \end{aligned} \quad (34)$$

4.2. Perturbation Series

The critical exponents ν and μ can also be derived from the analysis of infrared divergences of the perturbation theory in *h*. Obviously, only even orders in *h* give contributions. We now estimate the large-distance behavior of the operator

$$U = \frac{1}{E_0 - H_0} V \frac{1}{E_0 - H_0} V \quad (35)$$

determining the perturbation theory order, where $V = h\sum S_i^x$ and H_0 is the Hamiltonian of the *XXZ* model. The perturbation series for the ground state energy is given by

$$\delta E \sim V \frac{1}{E_0 - H_0} V (1 + U + U^2 + \ldots). \quad (36)$$

We consider a large but finite system of the length *N*. We keep the powers of *N* and *h* only, omitting all other factors. We first consider the nonoscillating part of correlator (30). Taking only low-lying excitations of the spectrum of the *XXZ* model into account (these excitations give the most divergent part) and estimating the large-distance behavior of the nonoscillating part of correlator (30), we arrive at

$$U \sim h^2 \frac{\sum_{i,j} \langle S_i^x S_j^x \rangle}{(1/N)^2} \sim h^2 N^2 \frac{N^2}{N^{\theta + 1/\theta}} = h^2 N^{4-\theta-1/\theta}. \quad (37)$$

It follows that if $4 - \theta - 1/\theta > 0$, then each next order in perturbation theory (36) diverges more and more strongly. To absorb these infrared divergences, we must introduce the scaling parameter $y = Nh^\nu$ and assume that the series $(1 + U + U^2 + \ldots)$ in (36) forms some function of the scaling parameter *y*. In our case,

$$\nu = \frac{2}{4 - \theta - 1/\theta}$$

(see Eq. (26)) and

$$U \propto y^{2/\nu}.$$

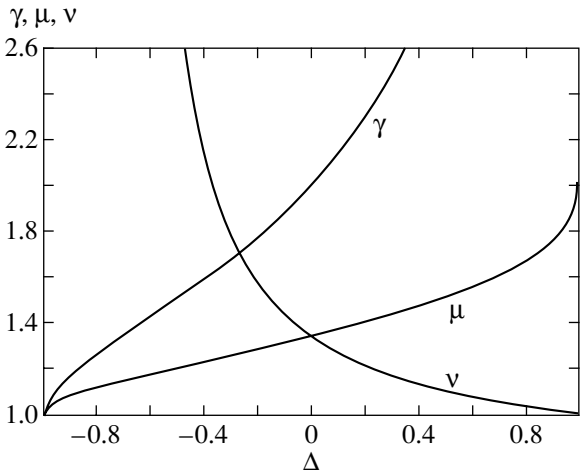


Fig. 3. The dependence of the critical exponents ν, μ, and γ on Δ. The smallest exponent gives the perturbation of the most relevant type and defines the index for the mass gap.

The leading second-order divergence of the ground state energy can be found similarly to (37),

$$\delta E^{(2)} = V \frac{1}{E_0 - H_0} V \sim h^2 N^{3-\theta-1/\theta}. \quad (38)$$

Combining Eqs. (37) and (38), we can write

$$\delta E \sim N h^{2\nu} f(y)$$

with some unknown function *f*(*y*) whose small-*y* expansion is given by

$$f(y) = \frac{1}{y^2} \sum_{n=1}^{\infty} c_n y^{2n/\nu}.$$

In the thermodynamic limit as $N \longrightarrow \infty$, the scaling parameter $y = Nh^\nu$ also tends to infinity, $y \longrightarrow \infty$. Because the energy is proportional to *N*, the function *f*(*y*) has a finite limit $f(\infty) = a$. In the thermodynamic limit for the correction to the ground state energy, we therefore have

$$\delta E \sim aNh^{2\nu}. \quad (39)$$

For the first excited state, the perturbation theory divergences have the same form as in (37) and (38). For the gap, we therefore find the same scaling parameter $y = Nh^\nu$ and

$$m \sim Nh^{2\nu} g(y).$$

In the thermodynamic limit, the mass gap is of the order of unity (in terms of *N*), and therefore, the function $g(y) \propto 1/y$ as $y \longrightarrow \infty$. Thus, finally we arrive at Eq. (26).

We now consider the more subtle oscillating part of correlator (30). For the oscillating part at large distances, we can write

$$\sum_{i,j}\langle S_i^x S_j^x\rangle \sim N\sum_r \frac{(-1)^r}{r^\theta} \sim N\sum_r \frac{1}{r^{\theta+1}} \sim N\frac{1}{N^\theta}.$$

The oscillating part of the perturbation operator V connects the low-lying gapless states with finite-energy states. That is, each second level in all orders of the perturbation series is separated from the ground state by a finite gap. For the operator U, we therefore have

$$U \sim h^2 \frac{\sum_{i,j}\langle S_i^x S_j^x\rangle}{(1/N)} \sim h^2 N^{2-\theta}.$$

Because θ is always less than 2, the divergences grow with the order of the perturbation theory. To eliminate these divergences, we introduce the scaling parameter $y = Nh^\mu$, with μ defined in Eq. (31), such that $U \sim y^{2/\mu}$.

The second-order correction to the ground state energy is then given by

$$\delta E^{(2)} \sim h^2 \frac{\sum_{i,j}\langle S_i^x S_j^x\rangle}{1} \sim h^2 N^{1-\theta}$$

and the total correction to the ground state energy is

$$\delta E \sim N h^{2\mu} f(y),$$

where $f(y)$ is an unknown function with a finite limit $f(\infty) = b$.

In the thermodynamic limit, the ground state energy therefore behaves as

$$\delta E \sim bNh^{2\mu}.$$

The mass gap is found similarly,

$$m \sim Nh^{2\mu} g(y)$$

with the function $g(y) \propto 1/y$ as $y \longrightarrow \infty$. We thus reproduce Eq. (31) in the thermodynamic limit.

We note that we have estimated only the long-wavelength divergent part of the perturbation theory. In addition, the regular part of the perturbation theory gives the leading term of the order h^2. Combining all the above facts, we thus arrive at

$$\frac{\delta E}{N} = -\frac{\chi}{2}h^2 + ah^{2\nu} + bh^{2\mu}. \tag{40}$$

As can be seen from Eq. (40), δE consists of a regular term h^2 and two singular terms. Because $\nu > 1$ and $\mu > 1$, the susceptibility χ is finite for any Δ, in contrast to the model with the staggered transverse field [12], where the singular term is h^η with $\eta = 4/(4-\theta) < 2$.

It follows from Eqs. (26) and (31) that $\nu \longrightarrow 1$ as $\Delta \longrightarrow 1$ and $\mu \longrightarrow 1$ as $\Delta \longrightarrow -1$. In both limits, one of the singular terms therefore becomes proportional to h^2 and, hence, contributes to the susceptibility. This implies that the susceptibility has a jump at the symmetric points $\Delta = \pm 1$.

5. THE LINE $\Delta = 1$

In the vicinity of the line $\Delta = 1$, it is convenient to rewrite Hamiltonian (1) as

$$H = H_0 + V,$$

$$H_0 = \sum_n (\mathbf{S}_n \cdot \mathbf{S}_{n+1}) + h\sum_n S_n^x, \tag{41}$$

$$V = -g\sum_n S_n^z S_{n+1}^z,$$

where the parameter $g = 1 - \Delta \ll 1$ is small. On the isotropic line $\Delta = 1$, model (1) is exactly solvable by the Bethe ansatz. The properties of the system remain critical up to the transition point $h_c = 2$, where the ground state becomes ferromagnetic. Therefore, for $h < 2$ and small perturbation V, we can use a conformal estimates.

The asymptotic form of the correlation function on this line is given by

$$\langle S_i^z S_{i+n}^z\rangle \sim \frac{(-1)^n}{n^{\alpha(h)}}, \tag{42}$$

where $\alpha(h)$ is a known function obtained from the Bethe ansatz [13]. It has the asymptotic forms

$$\alpha(h) \sim \begin{cases} 1 - \dfrac{1}{2\ln(1/h)}, & h \longrightarrow 0 \\ \dfrac{1}{2}, & h \longrightarrow 2. \end{cases} \tag{43}$$

The scaling dimension of the operator S^z is $d_z = \alpha(h)/2$, and the scaling dimension of $S_i^z S_{i+1}^z$ is four times greater, $d_{zz} = 4d_z = 2\alpha(h)$. Because $\alpha(h) < 1$, the perturbation V is relevant and leads to the mass gap and the staggered magnetization given by

$$\begin{aligned} m &\sim |g|^{1/(2-d_{zz})} = |g|^{1/(2-2\alpha)}, \\ \langle|S^y|\rangle &\sim |g|^{\alpha/(4-4\alpha)}, \quad \Delta < 1, \\ \langle|S^z|\rangle &\sim |g|^{\alpha/(4-4\alpha)}, \quad \Delta > 1. \end{aligned} \tag{44}$$

From the general expressions for the mass gap in Eq. (44), we find that $m \sim g$ in the limit as $h \longrightarrow 2$, which agrees with the result of the mean-field approximation in Eq. (21).

In the vicinity of the point $\Delta = 1$, $h = 2$, the long-range order vanishes on both lines: at $\Delta = 1$ as $g^{1/4}$ (see Eqs. (44)) and at $h = h_c$ as $|h_c - h|^{1/8}$. We also have the

exact expression for the long-range order on the classical line,

$$\langle |S^y| \rangle_{\rm cl} = \frac{\sqrt{g}}{2\sqrt{2}}. \quad (45)$$

Combining all these facts, we arrive at the formula

$$\langle |S^y| \rangle = 2^{-7/8} g^{1/4} |h_c - h|^{1/8}. \quad (46)$$

The behavior of the system near the point $\Delta = 1$, $h = 0$ is more subtle. As follows from Eq. (32), the mass gap is $m \sim h$ for very small h; on the other hand, Eq. (44) implies a different scaling $m \sim g^{\ln(1/h)}$. Therefore, there are two regions near this point with different behaviors of the mass gap. The boundary between these two regions can be found as follows. We rewrite the perturbation in Hamiltonian (41) as

$$V = V_1 + V_2,$$

$$V_1 = -\frac{g}{2}\sum_n (S_n^y S_{n+1}^y + S_n^z S_{n+1}^z),$$

$$V_2 = \frac{g}{2}\sum_n (S_n^y S_{n+1}^y - S_n^z S_{n+1}^z).$$

The part $H_0 + V_1$ of the Hamiltonian corresponds to the *XXZ* model in the longitudinal magnetic field, which is gapless for the magnetic field

$$h > \exp\left(-\frac{\pi^2}{2\sqrt{g}}\right).$$

Therefore, in the region of very small magnetic field

$$h < \exp\left(-\frac{\pi^2}{2\sqrt{g}}\right),$$

the perturbation V_1 is relevant, leading to the mass gap $m \sim h$. The two-cutoff scaling procedure [9, 10] leads to the mass gap

$$m \approx h \exp\left(-\frac{\pi^2}{2\sqrt{g}}\right)$$

for

$$h > \exp\left(-\frac{\pi^2}{2\sqrt{g}}\right).$$

Finally, when g is much less than h, the scaling dimension of the operator V_2 defines the exponent for the gap in Eq. (44). Summarizing, the mass gap in the vicinity of the isotropic point $\Delta = 1$, $h = 0$ is given by

$$\begin{aligned} &m \sim h, \quad \ln h \ll -\frac{1}{\sqrt{g}}, \\ &m \sim h e^{-\pi^2/2\sqrt{g}}, \quad \frac{1}{\sqrt{g}\ln g} \gg \ln h \gg -\frac{1}{\sqrt{g}}, \\ &m \sim g^{-\ln h}, \quad \ln h \gg \frac{1}{\sqrt{g}\ln g}. \end{aligned} \quad (47)$$

6. THE LINE $\Delta = -1$

In this section, we consider model (1) in the vicinity of the line $\Delta = -1$, where

$$1 + \Delta = \delta \ll 1$$

is a small parameter. It is convenient to rotate spins on each odd site by π around the z axis, such that model (1) becomes

$$H = -\sum_n (\mathbf{S}_n \cdot \mathbf{S}_{n+1}) + \delta \sum_n S_n^z S_{n+1}^z - h\sum_n (-1)^n S_n^x. \quad (48)$$

At $\delta = 0$ and $h = 0$, the ground state of (48) is the ferromagnetic state with zero momentum degenerate with respect to total S^z. The states that can be reached from the ground state by means of the transition operator

$$\sum_n (-1)^n S_n^x$$

are the states with $q = \pi$ and a finite gap over the ground state. For $\delta \ll 1$, the transition operator connects the low-energy states and the states with the energies $\varepsilon_s \approx 2$ The second-order correction to low-energy states is given by

$$\delta E_l^{(2)} = h^2 \sum_{s,n,m} \frac{\langle l|(-1)^n S_n^x|s\rangle\langle s|(-1)^m S_m^x|l\rangle}{E_l - E_s}, \quad (49)$$

where $|l\rangle$ is the low-energy state and $|s\rangle$ is a state with the high energy $E_s - E_l \approx 2$. For $\delta \ll 1$, Eq. (49) can therefore be rewritten as

$$\begin{aligned} \delta E_l^{(2)} &= -\frac{h^2}{2}\sum_{n,m} \langle l|(-1)^{n-m} S_n^x S_m^x|l\rangle \\ &= -\frac{h^2 N}{8} - h^2 \sum_{n<m} \langle l|(-1)^{n-m} S_n^x S_m^x|l\rangle. \end{aligned} \quad (50)$$

The spin correlation function $\langle l|S_n^x S_m^x|l\rangle$ is a slowly varying function of $|m - n|$ for $\delta \ll 1$. Therefore,

$$\sum_{n<m} \langle l|(-1)^{n-m} S_n^x S_m^x|l\rangle \approx -\frac{1}{2}\sum_n \langle l|S_n^x S_{n+1}^x|l\rangle. \quad (51)$$

In accordance with Eqs. (49)–(51), the low-lying states of (48) are therefore described for $|\delta| \ll 1$ and $h \ll 1$ by the *XYZ* Hamiltonian

$$\begin{aligned} H &= -\frac{h^2 N}{8} \\ &- \sum_n \left[\left(1 - \frac{h^2}{2}\right) S_n^x S_{n+1}^x + S_n^y S_{n+1}^y - \Delta S_n^z S_{n+1}^z\right]. \end{aligned} \quad (52)$$

The coincidence of the low-energy spectra of (48) and (52) in the vicinity of the ferromagnetic point $\Delta = -1$, $h = 0$ has been checked numerically for finite sys-

tems. The spectrum of low-lying excitations of the $s = 1/2$ *XYZ* model in Eq. (52) and of the original model in Eq. (1) near the ferromagnetic point $\Delta = -1$, $h = 0$ can be asymptotically exactly described by the spin-wave theory, which gives

$$\begin{aligned} m &= h\sqrt{(1+\Delta)/2}, \quad \Delta > -1, \\ m &= \sqrt{(1+\Delta)(1+\Delta+h^2/2)}, \quad \Delta < -1. \end{aligned} \tag{53}$$

It can be verified that Eq. (53) yields the exact gap of the *XYZ* model [14] for $|\delta|, h \ll 1$. The validity of the spin-wave approximation is quite natural because the number of magnons forming the ground state is small in the vicinity of the ferromagnetic point $\Delta = -1$, $h = 0$.

We also note that the gap in Eq. (53) for $\Delta \geq -1$ agrees with the conformal theory result (32) and gives the preexponential factor for the gap. On the classical line

$$h_{\mathrm{cl}} = \sqrt{2(1+\Delta)},$$

Eq. (53) yields the gap $m = 1 + \Delta$, which confirms that the function $\psi_1^{(1)}$ in Eq. (7) gives the exact gap.

A similar mapping of model (1) with an arbitrary spin s to the *XYZ* model can be performed for $\Delta \approx -1$, $h \ll 1$. Taking into account that $\varepsilon_s = 4s$, the corresponding *XYZ* Hamiltonian is

$$\begin{aligned} H = -\sum_n \left[\left(1 - \frac{h^2}{2}\right) S_n^x S_{n+1}^x + S_n^y S_{n+1}^y - \Delta S_n^z S_{n+1}^z\right] \\ - \frac{h^2}{4s}\sum_n (S_n^x)^2, \end{aligned} \tag{54}$$

where S_n^α are spin-s operators.

The leading term of the gap of model (1) with an arbitrary spin s in the vicinity of the point $\Delta = -1$, $h = 0$ is exactly given by the spin-wave theory,

$$\begin{aligned} m &= h\sqrt{(1+\Delta)/2}, \quad \Delta > -1, \\ m &= 2s\sqrt{(1+\Delta)(1+\Delta+h^2/8s^2)}, \quad \Delta < -1. \end{aligned} \tag{55}$$

On the classical line h_{cl}, Eq. (55) gives the correct result $m = 2s\delta$.

Strictly on the line $\Delta = -1$, model (1) reduces to the isotropic ferromagnet in the staggered magnetic field. This model is nonintegrable, but it was suggested in [6] that the system is governed by a $c = 1$ conformal field theory up to some critical value $h = h_0$, where the phase transition of the Kosterlitz–Thouless type occurs.

For $h \ll 1$, where the mapping of (48) to *XYZ* model (52) is valid, the line $\Delta = -1$ is described by the *XXZ* model and the correlation functions have a power-law decay,

$$\begin{aligned} \langle S_i^z S_{i+n}^z \rangle &= \langle S_i^y S_{i+n}^y \rangle \sim \frac{(-1)^n}{n^{1/\beta(h)}}, \\ \langle S_i^x S_{i+n}^x \rangle &\sim \frac{(-1)^n}{n^{\beta(h)}}. \end{aligned} \tag{56}$$

We believe that the relation between the indices of x and y, z correlators on the line $\Delta = -1$ is given by (56) for $0 < h < h_0$. The scaling dimensions of the operators S_i^x, S_i^y, and S_i^z on this line are therefore given by $d_x = \beta/2$ and $d_y = d_z = 1/2\beta$.

On the line $\Delta = -1$, model (1) is gapless for $h < h_0$. This implies that the magnetic field term is irrelevant for $h < h_0$ ($\beta(h) > 4$) and becomes marginal at $h = h_0$, where $d_x = 2$ and $\beta(h_0) = 4$. Therefore, at the point $h = h_0$, the transition is of the Kosterlitz–Thouless type, and for $h > h_0$, the mass gap is exponentially small.

In the vicinity of the line $\Delta = -1$, the term

$$\delta \sum_n S_n^z S_{n+1}^z$$

in (48) can be considered as a perturbation and the scaling dimension of the perturbation operator $S_n^z S_{n+1}^z$ is

$$d_{zz} = 4d_z = 2/\beta(h).$$

Because $\beta(h) \geq 4$ for $h < h_0$, the perturbation is relevant and leads to the mass gap and the long-range order

$$\begin{aligned} m &\sim |\delta|^{1/(2-2\beta)}, \\ \langle |S^y| \rangle &\sim \delta^{1/4(\beta-1)}, \quad \delta > 0, \\ \langle |S^z| \rangle &\sim |\delta|^{1/4(\beta-1)}, \quad \delta < 0. \end{aligned} \tag{57}$$

In particular, $m \propto |\delta|^{2/3}$ and $\langle |S^y| \rangle \sim |\delta|^{1/12}$ as $h \longrightarrow h_0$.

The function $\beta(h)$ is generally unknown, except in the where case $h \ll 1$, the mapping to the *XXZ* model is valid, and

$$\beta(h) = \left[1 - \frac{1}{\pi}\arccos\left(\frac{h^2}{2} - 1\right)\right]^{-1} \approx \frac{\pi}{h}.$$

But because model (1) is conformally invariant at $\Delta = -1$ and $h < h_0$, we can use a finite-size scaling analysis to determine the exponent $\beta(h)$ and the value of h_0. According to the standard scaling approach [15],

$$\beta(h) = \frac{2\pi v}{A},$$

where v is the speed of sound and A/N is the difference between the two lowest energies of the system. We calculated $\beta(h)$ for finite systems. The extrapolated function $\beta(h)$ agrees well with the dependence π/h at $h \ll 1$ and $\beta = 4$ at $h_0 \approx 0.52$. This estimate is close to our

direct numerical estimates $h_0 \approx 0.549$. On the other hand, the mean-field approach gives a rather crude value

$$h_0 = h_c(-1) = 0.69.$$

7. CONCLUSIONS

In summary, we have studied the effect of the symmetry-breaking transverse magnetic field on the $s = 1/2$ *XXZ* chain. Unlike the longitudinal field, the transverse field generates the staggered magnetization in the *y* direction and the gap in the spectrum of the model with the easy-plane anisotropy. Using conformal invariance, we have found the critical exponents of the field dependence of the gap and the long-range order. We have shown that the spectrum of the model is gapped on the entire $h\Delta$ plane except at several critical lines, where the gap and the long-range order vanish. The behavior of the gap and the long-range order in the vicinity of the critical lines $\Delta = \pm 1$ is considered on the basis of the conformal field theory. We note that, in the vicinity of the points ($\Delta = 1$, $h = 0$) and ($\Delta = 1$, $h = 2$), there is a crossover between different regimes of the behavior of the system. We have shown that, near the point ($\Delta = -1$, $h = 0$), the original model can be mapped to the effective exactly solvable 1*D XYZ* model and has the spin-wave spectrum. The transition line $h_c(\Delta)$ between the ordered phases and the disordered one is studied in the mean-field approximation. This study shows that this transition is similar to that in the Ising model in the transverse field. But the behavior of the model on the transition line near the Kosterlitz–Thouless point ($\Delta = -1$, $h = h_0$) is not so clear. The mean-field approximation worsens as $\Delta \longrightarrow -1$, and a more sophisticated theory is needed.

ACKNOWLEDGMENTS

We thank Prof. P. Fulde for many useful discussions. We are grateful to Max-Planck-Institut für Physik Komplexer Systeme for kind hospitality. This work is supported by the Russian Foundation for Basic Research (project nos. 00-03-32981 and 00-15-97334) and ISTC (grant no. 2207).

REFERENCES

1. R. Helfrich, M. Koppen, M. Lang, *et al.*, J. Magn. Magn. Mater. **177**, 309 (1998).
2. C. N. Yang and C. P. Yang, Phys. Rev. B **150**, 321 (1966); Phys. Rev. B **150**, 327 (1966).
3. G. Uimin, Y. Kudasov, P. Fulde, and A. A. Ovchinnikov, Eur. Phys. J. B **16**, 241 (2000).
4. S. Mori, J.-J. Kim, and I. Harada, J. Phys. Soc. Jpn. **64**, 3409 (1995); Y. Hieida, K. Okunishi, and Y. Akutsu, Phys. Rev. B **64**, 224 422 (2001).
5. J. Kurmann, H. Tomas, and G. Muller, Physica A (Amsterdam) **112**, 235 (1982); G. Muller and R. E. Shrock, Phys. Rev. B **32**, 5845 (1985).
6. F. C. Alcaraz and A. L. Malvezzi, J. Phys. A **28**, 1521 (1995); M. Tsukano and K. Nomura, J. Phys. Soc. Jpn. **67**, 302 (1998).
7. E. Barouch and B. M. McCoy, Phys. Rev. A **3**, 786 (1971).
8. S. Lukyanov and A. Zamolodchikov, Nucl. Phys. B **493**, 571 (1997); T. Hikihara and A. Furusaki, Phys. Rev. B **58**, R583 (1998).
9. A. O. Gogolin, A. A. Nersesyan, and A. M. Tsvelik, *Bosonization and Strongly Correlated Systems* (Cambridge Univ. Press, Cambridge, 1998).
10. A. A. Nersesyan, A. Luther, and F. V. Kusmartsev, Phys. Lett. A **176**, 363 (1993).
11. A. Luther and I. Peschel, Phys. Rev. B **12**, 3908 (1975).
12. I. Affleck and M. Oshikawa, Phys. Rev. B **60**, 1038 (1999).
13. N. M. Bogoliubov, A. G. Izergin, and V. E. Korepin, Nucl. Phys. B **275**, 687 (1986).
14. J. D. Johnson, S. Krinsky, and B. M. McCoy, Phys. Rev. A **8**, 2526 (1973); I. M. Babich and A. M. Kosevich, Zh. Éksp. Teor. Fiz. **82**, 1277 (1982) [Sov. Phys. JETP **55**, 743 (1982)].
15. J. L. Cardy, in *Phase Transitions and Critical Phenomena* (Academic, New York, 1986), Vol. XI.

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 550–559.
Translated from Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 636–646.

SOLIDS
Electronic Properties

Critical Behavior of Disordered Systems with Replica Symmetry Breaking

V. V. Prudnikov* and P. V. Prudnikov

Omsk State University, Omsk, 644077 Russia

**e-mail: prudnikv@univer.omsk.su*

Received April 17, 2002

Abstract—A field-theoretic description of the critical behavior of weakly disordered systems with a p-component order parameter is given. For systems of an arbitrary dimension in the range from three to four, a renormalization group analysis of the effective replica Hamiltonian of the model with an interaction potential without replica symmetry is given in the two-loop approximation. For the case of the one-step replica symmetry breaking, fixed points of the renormalization group equations are found using the Padé–Borel summing technique. For every value p, the threshold dimensions of the system that separate the regions of different types of critical behavior are found by analyzing those fixed points. Specific features of the critical behavior determined by the replica symmetry breaking are described. The results are compared with those obtained by the ε expansion, and the scope of the method applicability is determined. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

When the renormalization group approach is applied to describe the critical behavior of disordered systems with quenched disorder, the method of replicas [1–3] is used to restore the translation symmetry of the effective Hamiltonian describing the interaction of fluctuations. However, in the studies [4–6], it was conjectured that the replica symmetry could be broken in systems with quenched disorder. In [4, 5], the physical concept of the occurrence of numerous local energy minima in disordered systems with random transition temperature was used to give a renormalization group description of the ϕ^4 model with the interaction potential characterized by the broken replica symmetry. For this purpose, the ϵ-expansion technique was used in the lower order of the theory. For systems in which the number of components of the order parameter p is less than four, it was discovered that the breaking of replica symmetry is the crucial factor in the critical behavior. It was shown that, for p in the range from one through four, two modes of system behavior are possible. The first one determines a nonuniversal critical behavior, which depends on seed values of the model parameters and, ultimately, on the concentration of impurities in the system. The second mode is characterized by the absence of a stable critical behavior, as also is the most interesting case of Ising (p =1) systems. Even though the implications of these studies are very interesting, the results of a field-theoretic description of certain homogeneous and disordered systems in the two-loop and higher order approximations based on the asymptotic series summation techniques [7] showed that the stability analysis of various types of critical behavior that uses the first-order terms of the ϵ expansion can be considered only as a coarse estimate, especially for multivertex statistical models [8]. For this reason, the results of investigation of the replica symmetry-breaking (RSB) effects obtained in [4–6] require revaluation from the viewpoint of a more accurate approach.

To this end, we proposed in [9, 10], in the framework of the field-theoretic approach, a renormalization group description of the model of weakly disordered three- and two-dimensional systems with the fourth-order interaction potential with respect to the order parameter fluctuations, which determines the replica symmetry breaking. An analysis of solutions to the renormalization group equations carried out in the two-loop approximation with the sequential application of the summation technique for Padé–Borel series showed that the critical behavior of three- and two-dimensional systems is stable with respect to the relative influence of the RSB effects, and the former scenario of the quenched disorder influence on the critical behavior is realized [11].

However, the scope of the results obtained in [4, 5] remains unclear. In particular, it is interesting to establish the threshold dimensions of the disordered system, $d_c(p)$, that separate the domain of influence of RSB effects from the critical behavior domains in which these effects are insignificant. It is also interesting to apply the renormalization group approach to analyze the behavior of systems with replica symmetry breaking in which no stable critical behavior exists and the strong coupling mode occurs (see [4–6]). A theoretical analysis of this phenomenon is especially important from the viewpoint of the possible manifestation of RSB effects in strongly disordered systems and their observation in computer models of the critical behavior

under an impurity concentration exceeding the impurity percolation threshold when extended impurity structures are formed in the system [12].

This paper is devoted to the consideration of the above-mentioned problems. For weakly disordered systems of an arbitrary dimension in the range from three to four, an analysis of the critical behavior of the model with an RSB potential is carried out based on the renormalization group approach in the two-loop approximation with the use of summation techniques. Our analysis does not rely on the ϵ-expansion technique.

2. DEFINITION OF THE MODEL AND THE CALCULATION PROCEDURE

The model Ginzburg–Landau Hamiltonian, which describes the behavior of a p-component spin system with weakly quenched disorder near the critical point has the form

$$H = \int d^d x \left\{ \frac{1}{2}\sum_{i=1}^{p} [\nabla \phi_i(x)]^2 + \frac{1}{2}[\tau - \delta\tau(x)]\sum_{i=1}^{p}\phi_i^2(x) + \frac{1}{4}g\sum_{i,j=1}^{p}\phi_i^2(x)\phi_j^2(x) \right\}, \quad (1)$$

where the random phase transition temperature has the Gaussian distribution $\delta\tau(x)$ with the variance $\langle\langle(\delta\tau(x))^2\rangle\rangle \sim u$, which is determined by a positive constant u and is proportional to the concentration of the structure defects. The application of the conventional replica method (see, for example, [6]) makes it possible to average over the temperature fluctuations $\delta\tau(x)$ and reduce the problem of the statistical description of a weakly disordered system to the problem of the statistical description of a homogeneous system with the effective Hamiltonian

$$H_n = \int d^d x \left\{ \frac{1}{2}\sum_{i=1}^{p}\sum_{a=1}^{n} [\nabla \phi_i^a(x)]^2 + \frac{1}{2}\tau\sum_{i=1}^{p}\sum_{a=1}^{n} [\phi_i^a(x)]^2 + \frac{1}{4}\sum_{i,j=1}^{p}\sum_{a,b=1}^{n} g_{ab}[\phi_i^a(x)]^2[\phi_j^b(x)]^2 \right\}. \quad (2)$$

Here, the index a enumerates replicas (images) of the original homogeneous component in Hamiltonian (1); and the additional vertex u, which occurs in the interaction matrix $g_{ab} = g\delta_{ab} - u$, specifies the effective interaction of fluctuations of the $(n \times p)$-component order parameter through the defect field. This statistical model is thermodynamically equivalent to the original disordered model in the limit $n \longrightarrow 0$. The subsequent renormalization group procedure, which statistically takes into account the contribution of long-wave fluctuations of the order parameter relative to the ground state of the system with the configuration $\phi(x) = 0$ (at $T \geq T_c$), is performed at the scale of the correlation length, which turns to infinity at the transition temperature T_c. This procedure makes it possible to analyze possible types of critical behavior and conditions of their realization and calculate the critical indexes.

However, it was shown in [4–6] that a macroscopically large number of spatial regions with $\phi(x) \neq 0$ appears in the system due to fluctuations of the random transition temperature at $[\tau - \delta\tau(x)] < 0$. These regions are separated from the ground state by potential barriers. To describe the statistical properties of systems with multiple local energy minima, the replica symmetry-breaking formalism (suggested by Parisi) was used in [4–6] by analogy with spin glasses [9]. According to the reasoning presented in [4–6], the statistical calculation of the contribution of nonperturbation degrees of freedom associated with the order parameter fluctuations relative to the configurations of the field $\phi(x)$ at the local energy minima results (when the replica procedure is applied for the weak disorder) in the appearance of additional interactions of the form $\sum_{a,b} g_{ab}\phi_a^2\phi_b^2$ in the effective replica Hamiltonian. Here, the final matrix g_{ab} is no longer replica-symmetric with $g_{ab} = g\delta_{ab} - u$, but rather has the RSB Parisi replica structure [13]. According to [4–6, 13], the matrix g_{ab} with the RSB structure is parameterized (in the limit $n \longrightarrow 0$) in terms of its diagonal elements $\tilde{g}$ and the off-diagonal function $g(x)$, which is defined on the interval $0 < x < 1$: $g_{ab} \longrightarrow (\tilde{g}, g(x))$. Here, operations with the matrices g_{ab} are defined by the rules

$$g_{ab}^k \longrightarrow (\tilde{g}^k; g^k(x)),$$
$$(\hat{g}^2)_{ab} = \sum_{c=1}^{n} g_{ac}g_{cb} \longrightarrow (\tilde{c}; c(x)), \quad (3)$$
$$(\hat{g}^3)_{ab} = \sum_{c,d=1}^{n} g_{ac}g_{cd}g_{db} \longrightarrow (\tilde{d}; d(x)),$$

where

$$\tilde{c} = \tilde{g}^2 - \int_0^1 dx g^2(x),$$
$$c(x) = 2\left[\tilde{g} - \int_0^1 dy g(y)\right]g(x) - \int_0^x dy[g(x) - g(y)]^2,$$
$$\tilde{d} = \tilde{c}\tilde{g} - \int_0^1 dx c(x)g(x), \quad (4)$$

$$d(x) = \left[\tilde{g} - \int_0^1 dy g(y)\right] c(x) + \left[\tilde{c} - \int_0^1 dy c(y)\right] g(x) - \int_0^x dy [g(x) - g(y)][c(x) - c(y)].$$

The replica-symmetric situation corresponds to the function $g(x) = \text{const}$, which is independent of x.

The renormalization group description of the model specified by the replica Hamiltonian (2) was carried out in the framework of the field-theoretic approach in the two-loop approximation for systems of an arbitrary dimension in the range from three to four. Possible types of critical behavior and their stability in the fluctuation domain are determined by the renormalization group equation for the coefficients of the matrix g_{ab}. They were determined by the conventional method based on the Feynman diagram technique for the vertex parts of the irreducible Green's functions and the renormalization procedure. For example, in the two-loop approximation, the two-point vertex function $\Gamma^{(2)}$, the four-point vertex functions $\Gamma^{(4)}_{ab}$, and the two-point function $\Gamma^{(2,1)}_{aa}$ with the insertion $(\phi^a_i)^2$ have the form

$$\left.\frac{\partial\Gamma^{(2)}}{\partial k^2}\right|_{k^2=0} = 1 + 4fg^2_{aa} + 2pf\sum_{c=1}^{n} g_{ac}g_{ca}, \tag{5}$$

$$\begin{aligned}\Gamma^{(4)}_{ab}\big|_{k_i=0} &= g_{ab} - p\sum_{c=1}^{n} g_{ac}g_{cb} - 4g_{aa}g_{ab} - 4g^2_{ab} \\ &+ (8+16h)g^3_{ab} + (24+8h)g^2_{aa}g_{ab} + 48hg_{aa}g^2_{ab} \\ &+ 4g_{aa}g_{bb}g_{ab} + 8ph\sum_{c=1}^{n} g_{ac}g^2_{cb} + 8phg_{ab}\sum_{c=1}^{n} g_{ac}g_{cb} \\ &+ 4phg_{ab}\sum_{c=1}^{n} g^2_{ac} + 2p\sum_{c=1}^{n} g_{ac}g_{cc}g_{cb} \\ &+ 4pg_{aa}\sum_{c=1}^{n} g_{ac}g_{cb} + p^2\sum_{c,d=1}^{n} g_{ac}g_{cd}g_{db},\end{aligned} \tag{6}$$

$$\begin{aligned}\Gamma^{(2,1)}_{aa}\big|_{k_i=0} &= 1 - p\sum_{c=1}^{n} g_{ca} - 2g_{aa} \\ &+ 2pg_{aa}\sum_{c=1}^{n} g_{ca} + (4+12h)g^2_{aa} \\ &+ 6ph\sum_{c=1}^{n} g^2_{ca} + p\sum_{c=1}^{n} g_{cc}g_{ca} + p^2\sum_{c,d=1}^{n} g_{dc}g_{ca},\end{aligned} \tag{7}$$

where the notation

$$\begin{aligned}f(d) &= -\frac{1}{J^2}\frac{\partial}{\partial k^2} \\ &\times \int \frac{d^dk_1 d^dk_2}{(k_1^2+1)(k_2^2+1)((k_1+k_2+k)^2+1)}\bigg|_{k^2=0}, \\ h(d) &= \frac{1}{J^2}\int \frac{d^dk_1 d^dk_2}{(k_1^2+1)^2(k_2^2+1)((k_1+k_2)^2+1)}, \\ J &= \int \frac{d^dk}{(k^2+1)^2}\end{aligned} \tag{8}$$

is used, and the redefinition $g_{ab} \longrightarrow g_{ab}/J$ is carried out. The diagram representation of the corresponding contributions to $\Gamma^{(2)}$, $\Gamma^{(4)}_{ab}$, and $\Gamma^{(2,1)}_{aa}$ is shown in Fig. 1.

However, the subsequent renormalization procedure for the vertex functions and the calculation of the β and γ functions, which determine the renormalization group transformations for the interaction constants, are difficult due to the complicated structure of relations (3) and (4) defining operations with matrices g_{ab}. The steplike structure of the function $g(x)$ established in [4–6] makes it possible to implement the renormalization procedure. In this paper, we restrict ourselves to the consideration of the one-step function $g(x)$:

$$g(x) = \begin{cases} g_0, & 0 \le x < x_0 \\ g_1, & x_0 < x \le 1, \end{cases} \tag{9}$$

where the coordinate of the step $0 \le x_0 \le 1$ is an arbitrary parameter that does not evolve under scale transformations and remains the same as in the seed function $g_0(x)$. As a result, the renormalization group transformations of the replica Hamiltonian with RSB are determined by the three parameters $\tilde{g}$, g_0, and g_1.

The critical properties of the model can be revealed by analyzing the coefficients $\beta_i(\tilde{g}, g_0, g_1)$ $(i = 1, 2, 3)$, $\gamma_\phi(\tilde{g}, g_0, g_1)$, and $\gamma_{\phi^2}(\tilde{g}, g_0, g_1)$ of the renormalization group Callan–Symanzik equation [14]. We obtained the following expressions for the β and γ functions in the two-loop approximation in the form of series in the renormalized parameters $\tilde{g}$, g_0, and g_1:

$$\begin{aligned}\beta_1 &= -\tilde{g} + (8+p)\tilde{g}^2 - px_0g_0^2 - p(1-x_0)g_1^2 \\ &+ [(8f - 40h + 20)p + 16f - 176h + 88]\tilde{g}^3 \\ &+ (24h - 8f - 12)x_0p\tilde{g}g_0^2 \\ &+ (24h - 8f - 12)(1-x_0)p\tilde{g}g_1^2 \\ &- (16h-8)x_0pg_0^3 - (16h-8)(1-x_0)pg_1^3,\end{aligned}$$

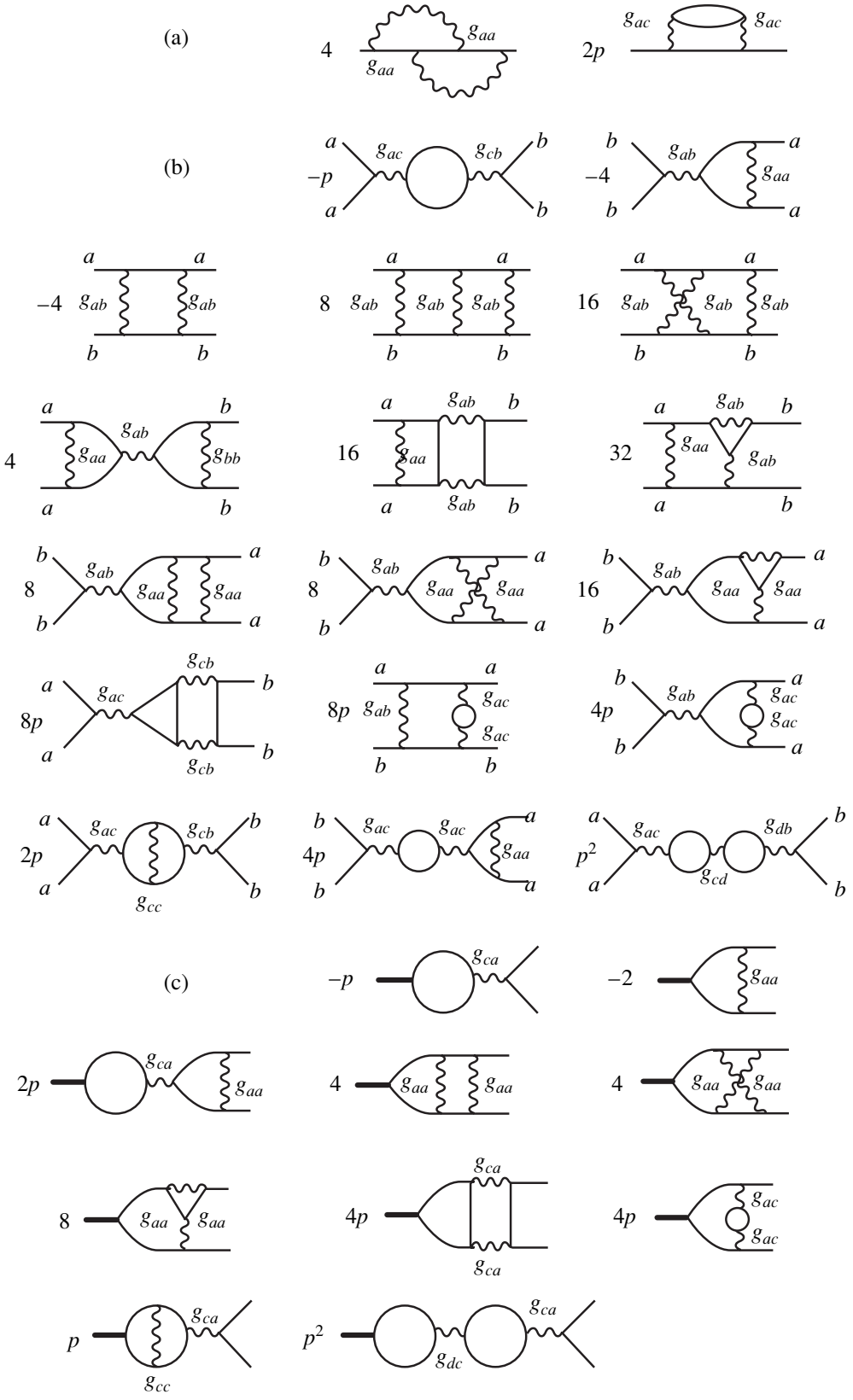


Fig. 1. The diagram representation of contributions to the two-point $\Gamma^{(2)}$ (a), four-point $\Gamma^{(4)}_{ab}$ (b), and two-point $\Gamma^{(2,1)}_{aa}$ with the inclusion $(\phi^a_i)^2$ (c) vertex functions in the one- and two-loop approximations with the corresponding weighting coefficients.

$$\begin{aligned}
\beta_2 &= -g_0 + (4+2p)\tilde{g}g_0 + (2px_0 - 4)g_0^2 \\
&\quad + 2(1-x_0)pg_0g_1 \\
&\quad + [(8f - 48h + 28)p + 16f - 48h + 24]\tilde{g}^2 g_0 \\
&\quad - [((32h-16)x_0 + 8 - 32h)p + 48 - 96h]\tilde{g}g_0^2 \\
&\quad - (32h-16)(1-x_0)p\tilde{g}g_0g_1 \\
&\quad + [(48h - 8f - 20)x_0 p - 32h + 16]g_0^3 \\
&\quad + (32h-8)(1-x_0)pg_0^2g_1 \\
&\quad + (16h - 12 - 8f)(1-x_0)pg_0g_1^2, \\
\beta_3 &= -g_1 + px_0g_0^2 - [p(x_0-2)+4]g_1^2 + (4+2p)\tilde{g}g_1 \\
&\quad + [(8f - 48h + 28)p + 16f - 48h + 24]g_1\tilde{g}^2 \\
&\quad - (16h-8)x_0p\tilde{g}g_0^2 \\
&\quad - [((8-16h)x_0 - 8)p + 48 - 96h]u_0g_1^2 \\
&\quad + (16h-8)x_0pg_0^3 + (8h - 8f - 4)x_0pg_1g_0^2 \\
&\quad + [(8f - 24h + 12)x_0p \\
&\quad + (48h - 8f - 20)p + 16 - 32h]g_1^3, \\
\gamma_\phi &= 4(4-d)f(d) \\
&\quad \times [(p+2)\tilde{g}^2 - px_0g_0^2 - p(1-x_0)g_1^2], \\
\gamma_{\phi^2} &= -(4-d)[(p+2)\tilde{g} + px_0g_0 + p(1-x_0)g_1 \\
&\quad - 2(6h - 2f - 3)((p+2)\tilde{g}^2 - px_0g_0^2 - p(1-x_0)g_1^2)].
\end{aligned} \tag{10}$$

In order to compare the results of this paper with those obtained in [4–6], we reversed, by analogy with [4–6], the signs of the off-diagonal elements in the matrix: $g_{a\neq b} \longrightarrow -g_{a\neq b}$. As a result, g_0 and g_1 became positive definite. The integrals $f(d)$ and $h(d)$ were calculated numerically for $3 \le d < 4$.

It is well known that the series used in perturbation theory are asymptotic, and vertices of the interaction of the order parameter fluctuations in the fluctuation domain $\tau \longrightarrow 0$ are sufficiently large to directly apply expressions (10). For this reason, in order to extract the physical information from those expressions, we use the generalized (for the three-parameter case) Padé–Borel method for summing asymptotic series. The direct and inverse Borel transformations have the form

$$\begin{aligned}
f(\tilde{g}, g_0, g_1) &= \sum_{i,j,k} c_{ijk}\tilde{g}^i g_0^j g_1^k = \int_0^\infty e^{-t}F(\tilde{g}t, g_0t, g_1t)dt, \\
F(\tilde{g}, g_0, g_1) &= \sum_{i,j,k} \frac{c_{ijk}}{(i+j+k)!}\tilde{g}^i g_0^j g_1^k.
\end{aligned} \tag{11}$$

In order to find the analytic continuation of the Borel image of a function, we use the following series in the auxiliary variable θ:

$$\begin{aligned}
&\tilde{F}(\tilde{g}, g_0, g_1, \theta) \\
&= \sum_{k=0}^{\infty}\theta^k\sum_{i=0}^{k}\sum_{j=0}^{k-i}\frac{c_{i,j,k-i-j}}{k!}\tilde{g}^i g_0^j g_1^{k-i-j}.
\end{aligned} \tag{12}$$

The Padé approximation $[L/M]$ is applied to this series at the point $\theta = 1$. This technique was successfully used in [8] to describe the critical behavior of certain systems characterized by several interaction vertices of the order parameter fluctuations. The symmetry conservation property of a system when applying a Padé approximant in the variable θ becomes essential in the description of multivertex models. In this paper, we used the [2/1] approximant for the calculation of β functions in the two-loop approximation.

3. CALCULATION RESULTS

It is well known that the nature of the critical behavior is determined by the existence of a stable fixed point satisfying the system of equations

$$\beta_i(\tilde{g}^*, g_0^*, g_1^*) = 0, \quad i = 1, 2, 3. \tag{13}$$

By numerically solving system (13) for β functions found by the Padé–Borel summation technique (for p = 1, 2, and 3), three types of nontrivial fixed points were found in the physically interesting domain of parameters $\tilde{g}^*, g_0^*, g_1^* \ge 0$ (see Tables 1–3). The first-type fixed point with $\tilde{g}^* \neq 0$ and $g_0^* = g_1^* = 0$ corresponds to the critical behavior of the homogeneous system; the second-type fixed point with $\tilde{g}^* \neq 0$ and $g_0^* = g_1^* \neq 0$ corresponds to the critical behavior of the disordered system with replica symmetry; and the third-type fixed point with $\tilde{g}^* \neq 0$, $g_0^* = 0$, and $g_1^* \neq 0$ corresponds to the critical behavior of the disordered system with RSB. The values of $\tilde{g}^*$ and g_1^* at the fixed point with RSB depend on the coordinate of the step x_0. Tables 1–3 present the values of $\tilde{g}^*$ and g_1^* for $0 \le x_0 \le 1$ with the step $\Delta x_0 = 0.1$.

The possibility of realizing one or another type of critical behavior for each p depends on the stability of the corresponding fixed point. The stability criterion of a fixed point reduces to a condition that the eigenvalues λ_i of the matrix

$$B_{ij} = \frac{\partial\beta_i(\tilde{g}^*, g_0^*, g_1^*)}{\partial g_j} \tag{14}$$

belong to the complex right half-plane. An analysis of λ_i for every type of fixed point (see Tables 1–3) provides the following conclusions.

Table 1. The values of fixed points and eigenvalues at $p = 1$

d	Type	x_0	$\tilde{g}^*$	g_0^*	g_1^*	λ_1	λ_2	λ_3
	1		0.1774	0	0	0.6536	−0.1692	−0.1692
	2		**0.1844**	**0.0812**	**0.0812**	**0.5253 ± 0.0893*i***		**0.2112**
		0.0	0.1844	0	0.0812	0.5253 ± 0.0893*i*		−0.0392
		0.1	0.1840	0	0.0829	0.5352 ± 0.0983*i*		−0.0492
		0.2	0.1835	0	0.0846	0.5471 ± 0.1067*i*		−0.0599
		0.3	0.1830	0	0.0863	0.5607 ± 0.1133*i*		−0.0712
(a) $d = 3.0$		0.4	0.1824	0	0.0880	0.5765 ± 0.1180*i*		−0.0832
	3	0.5	0.1817	0	0.0895	0.5951 ± 0.1203*i*		−0.0959
		0.6	0.1810	0	0.0910	0.6172 ± 0.1189*i*		−0.1093
		0.7	0.1802	0	0.0924	0.6439 ± 0.1114*i*		−0.1234
		0.8	0.1793	0	0.0936	0.6760 ± 0.0921*i*		−0.1381
		0.9	0.1784	0	0.0947	0.7135 ± 0.0353*i*		−0.1534
		1.0	0.1774	0	0.0957	0.8573	0.6536	−0.1692
	1		0.0917	0	0	0.6315	−0.4163	−0.4163
(b) $d = 3.985$	2		**0.1231**	**0.1090**	**0.1090**	**0.6986 ± 0.1311*i***		**0.0022**
	3	0.0	0.1231	0	0.1090	0.7047 ± 0.1069*i*		−0.0363
	1		0.0916	0	0	0.6318	−0.4165	−0.4165
(c) $d = 3.986$	2		0.1230	0.1092	0.1092	0.6895 ± 0.1453*i*		−0.0076
	3	0.0	0.1230	0	0.1092	0.7018 ± 0.0935*i*		−0.0359

(i) For the three-dimensional Ising model ($p = 1$), the second-type fixed point is stable (Table 1, (a)). The complex values λ_1 and λ_2, for positive $|\lambda_1|$, $|\lambda_2|$, and λ_3, show that the second-type fixed point, in contrast to the third-type one, is a stable focus in the parametric space ($\tilde{g}$, g_0, g_1), and the renormalization group flows approach the second-type fixed point in a spiral trajectory. At the threshold dimension $d_c = 3.986$ (see Table 1, blocks (a) and (b)), the second-type fixed point looses stability (λ_3 changes sign). Since all other fixed points remain unstable in the entire range of the dimension variation ($3 \le d < 4$), no critical behavior is realized in the system at $3.986 \le d$ due to the replica symmetry breaking. The analysis of the behavior of renormalization group flows at $3.986 \le d$ provides the following results.

(ii) For the three-dimensional *XY* model ($p = 2$), small values of λ_i (see Table 2, (a)) indicate that the second-type replica symmetric fixed point is weakly stable. However, already for the dimension $d_c = 3.1$ (see Table 2, (b) and (c)), the third-type fixed point with RSB effects becomes stable. However, the critical behavior determined by this point is nonuniversal and depends on the parameter x_0 and, therefore, on the concentration of impurities. A stability analysis of the third-type fixed point reveals that it is stable only in the interval $0 \le x_0 \le x_c(d)$, where x_c is a threshold value of the parameter, which depends on the dimension of the system. For example, for $d = 3.1$, $x_c = 0.1$; and, for $d = 3.999$, $x_c = 0.3$. In the interval $x_c(d) < x_0 < 1$, all fixed points are unstable.

(iii) For the isotropic three-dimensional Heisenberg model ($p = 3$), the first-type fixed point becomes stable (Table 3, (a)), while at the other fixed points the constants g_0^* and g_1^* take physically senseless negative values. Only at $d_c = 3.999$ do g_0^* and g_1^* take physically meaningful values for the third-type point, and this point becomes stable in the interval $0 \le x_0 \le 0.4$ (Table 3, (b) and (c)). In the interval $0.4 < x_0 < 1$, all fixed points are unstable.

Note that, although the calculations indicate the stability of the impurity replica-symmetric second-type fixed point for the three-dimensional *XY* model ($p = 2$), there is reason to believe that, in the higher order approximations (as is the case for disordered systems considered without taking into account RSB effects [11]), the first-type fixed point, which corresponds to the critical behavior of the homogeneous system, will become stable. On the one hand, this is indicated by the very weak stability ($\lambda_3 = 0.000004$) of the second-type fixed point and by the fact that the threshold value of the order parameter $p_c = 2.0114$ found in the two-loop approximation, which separates the critical behavior domains determined by the first- ($p > p_c$) and second-type ($p < p_c$) fixed points, is very close to $p = 2$. This

Table 2. The values of fixed points and eigenvalues at $p = 2$

d	Type	x_0	$\tilde{g}^*$	g_0^*	g_1^*	λ_1	λ_2	λ_3
	1		0.155830	0	0	0.667315	−0.001672	−0.001672
	2		**0.155831**	**0.000584**	**0.000584**	**0.667312**	**0.001682**	**0.000004**
		0.0	0.155831	0	0.000584	0.667313	0.001683	−0.000001
		0.1	0.155831	0	0.000614	0.667313	0.001684	−0.000088
		0.2	0.155831	0	0.000648	0.667313	0.001685	−0.000186
		0.3	0.155831	0	0.000686	0.667313	0.001686	−0.000296
(a) $d = 3.0$		0.4	0.155831	0	0.000729	0.667313	0.001687	−0.000419
	3	0.5	0.155831	0	0.000778	0.667313	0.001687	−0.000559
		0.6	0.155831	0	0.000833	0.667313	0.001688	−0.000717
		0.7	0.155831	0	0.000896	0.667314	0.001690	−0.000901
		0.8	0.155831	0	0.000971	0.667314	0.001692	−0.001116
		0.9	0.155831	0	0.001058	0.667315	0.001694	−0.001369
		1.0	0.155830	0	0.001163	0.667316	0.001696	−0.001672
	1		0.1499955	0	0	0.689608	−0.009539	−0.009539
	2		0.1500170	0.00325	0.00325	0.689535	0.009887	−0.000003
		0.0	**0.1500170**	**0**	**0.00325**	**0.689535**	**0.009887**	**0.000109**
(b) $d = 3.10$	3	0.1	0.1500169	0	0.00341	0.689535	0.009899	−0.000401
		0.2	0.1500167	0	0.00360	0.689536	0.009926	−0.000961
	1		0.089762	0	0	1.119442	−0.133591	−0.133591
	2		0.092307	0.036991	0.036991	1.103421	0.227335	−0.025378
		0.0	**0.092307**	**0**	**0.036991**	**1.103421**	**0.227335**	**0.030783**
		0.1	**0.092270**	**0**	**0.038723**	**1.102142**	**0.235506**	**0.021563**
(c) $d = 3.999$	3	0.2	**0.092205**	**0**	**0.040559**	**1.100913**	**0.244667**	**0.011135**
		0.3	0.092108	0	0.042500	1.099845	0.254810	−0.000648
		0.4	0.091970	0	0.044547	1.099106	0.265820	−0.013939

explains a very slow variation of the eigenvalues λ_i of the stability matrix for the disordered *XY* model with the variation of the system dimension (Table 2). On the other hand, the negative value of the critical heat capacity coefficient α of the *XY* model also suggests, according to the Harris criterion, that the critical behavior of the model is stable with respect to the influence of the quenched disorder and, therefore, it can be expected that $p_c < 2$ in the higher order approximations. For example, the value $p_c = 1.912(4)$ was found in [15] on the basis of the six-loop approximation with the use of the pseudo ε expansion and the Padé–Borel–Leroy summation technique with a thoroughly chosen fitting parameter.

Because p_c is very close to $p = 2$ for the *XY* model, one can expect that the calculations based on higher order approximations will substantially change the threshold dimension $d_c(p = 2)$, although, for the Ising and Heisenberg models, the changes of $d_c(p)$ should be small. This assumption is supported by the calculation of critical indexes for three-dimensional homogeneous models with $p = 1, 2, 3$ and the disordered Ising model. We performed these calculations in the two-loop approximation with the use of the Padé–Borel summation technique (Table 4). The comparison of these results with the corresponding indexes reported in [16, 17], where all-time accurate calculations for three-dimensional models were performed in the six-loop approximation, shows that the difference in the values of critical indexes does not exceed 0.02.

The values of the threshold dimensions $d_c(p)$, which separate the domain of critical behavior with RSB effects $d_c(p) < p < 4$ from the domain where these effects are inessential, can be considered as threshold dimensions that restrict the scope of the ε-expansion method as applied to the three-vertex model of the weakly disordered system and the corresponding results reported in [4–6]. Our analysis also shows that the results of application of the ε-expansion technique to multivertex statistical models are unreliable independently of the approximation order. This is explained by the competition between different types of fixed points

Table 3. The values of fixed points and eigenvalues at $p = 3$

d	Type	x_0	$\tilde{g}^*$	g_0^*	g_1^*	λ_1	λ_2	λ_3
(a) $d = 3.0$	1		**0.1383**	**0**	**0**	**0.6814**	**0.1315**	**0.1315**
	2		0.1419	−0.0359	−0.0359	0.6727	−0.0891	−0.1420
	3	0.0	0.1419	0	−0.0359	0.6727	−0.0891	−0.0058
		0.1	0.1420	0	−0.0382	0.6727	−0.0865	0.0011
		0.2	0.1420	0	−0.0408	0.6728	−0.0836	0.0088
(b) $d = 3.998$	1		**0.090189**	**0**	**0**	**1.008004**	**0.024111**	**0.024111**
	2		0.090269	−0.005167	−0.005167	−3.346714	−0.829868	−0.861435
	3	0.0	0.090269	0	−0.005167	1.007806	−0.022461	−0.005642
		0.1	0.090271	0	−0.005519	1.007803	−0.022334	−0.004451
		0.2	0.090273	0	−0.005922	1.007801	−0.022185	−0.003093
(c) $d = 3.999$	1		0.081989	0	0	1.113633	−0.000820	−0.000820
	2		0.081989	0.000171	0.000171	1.113633	0.000822	−0.000228
	3	0.0	**0.081989**	**0**	**0.000171**	**1.113633**	**0.000822**	**0.000228**
		0.1	**0.081989**	**0**	**0.000183**	**1.113633**	**0.000822**	**0.000188**
		0.2	**0.081989**	**0**	**0.000196**	**1.113633**	**0.000823**	**0.000142**
		0.3	**0.081989**	**0**	**0.000212**	**1.113633**	**0.000823**	**0.000088**
		0.4	**0.081989**	**0**	**0.000230**	**1.113633**	**0.000823**	**0.000025**
		0.5	0.081989	0	0.000251	1.113633	0.000823	−0.000050
		0.6	0.081989	0	0.000277	1.113633	0.000824	−0.000140
		0.7	0.081989	0	0.000309	1.113633	0.000824	−0.000251
		0.8	0.081989	0	0.000350	1.113633	0.000825	−0.000391
		0.9	0.081989	0	0.000402	1.113633	0.000826	−0.000574
		1.0	0.081989	0	0.000473	1.113633	0.000828	−0.000820

Table 4. The values of critical indexes for three-dimensional models at replica-symmetric fixed points (FP)

Model	FP	η	ν	γ	β	α
Ising's	FP1	0.0280	0.637	1.256	0.327	0.088
	[16]	0.031(4)	0.630(2)	1.241(2)	0.325(2)	0.110(5)
	FP2	0.0283	0.679	1.339	0.349	−0.037
	[17]	0.030(3)	0.678(10)	1.330(17)	0.349(5)	−0.034(30)
XY	FP1	0.0288	0.674	1.328	0.347	−0.022
	[16]	0.034(3)	0.669(1)	1.316(1)	0.346(1)	−0.007(6)
Heisenberg's	FP1	0.0283	0.706	1.392	0.363	−0.118
	[16]	0.034(3)	0.705(1)	1.387(1)	0.364(1)	−0.115(9)

in the parametric space of multivertex models, which usually does not allow one to pass to the limit as $\varepsilon \longrightarrow 1$ without crossing the marginal dimensions $3 \le d_c < 4$ separating the stability domains of different fixed points.

To reveal the character of the behavior of a disordered system with RSB effects in the domain without stable critical states, we analyzed the phase portrait of the model based on the system of equations

$$r\frac{\partial g_i}{\partial r} = \beta_i(\tilde{g}, g_0, g_1), \qquad (15)$$

which specifies phase trajectories in the space of vertices $(\tilde{g}, g_0, g_1)$. An analysis shows (see Fig. 2) that, for

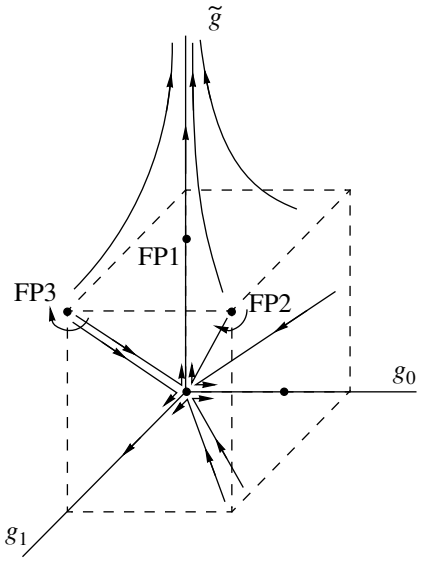


Fig. 2. The picture of renormalization group flows in the parametric space ($\tilde{g}$, g_0, g_1) for the Ising model with the system dimension $d = 3.99$.

the Ising model with $d_c = 3.986$ at $d \geq 3.986$, where none of the fixed points is stable, the strong coupling regime with the renormalization group flows determined by $(\tilde{g}, g_0, g_1) \longrightarrow (\infty, 0, 0)$ is realized if $\tilde{g} > \tilde{g}^*$. At the same time, at $\tilde{g} < \tilde{g}^*$, the flows with $(\tilde{g}, g_0, g_1) \longrightarrow (0, 0, 0)$ are realized, which asymptotically approach the Gauss fixed point (0, 0, 0) and then also tend to infinity along the axes $\tilde{g}$, g_0, and g_1. Such a behavior of the flows at $\tilde{g} < \tilde{g}^*$ is caused by the closeness of the system dimension d to four when the effect of fluctuations is negligibly small and the Gauss fixed point becomes an attractor.

4. CONCLUSIONS

The renormalization group analysis of weakly disordered systems of an arbitrary dimension in the range from three to four conducted in the two-loop approximation showed that the critical behavior of three-dimensional systems is stable with respect to the effect of the replica symmetry breaking. In systems with a one-component order parameter, the critical behavior determined by the structural disorder with a replica-symmetric fixed point is realized. The presence of weak disorder does not affect the critical behavior of multicomponent systems, although the proof of this fact for systems with $p = 2$ requires calculations with higher order approximations.

Effects of the replica symmetry breaking manifest themselves only in disordered systems with the dimension greater than three, and the threshold dimensions d_c depend on the number of components of the order parameter p and the value of the parameter x_0. The predicted picture of the influence of replica symmetry breaking on the critical behavior of disordered systems with a dimension $d > d_c$ qualitatively agrees with the results reported in [4–6]. The latter results were obtained on the basis of the ϵ-expansion technique. For systems with $p = 1$, RSB effects destroy the stable critical behavior, and the strong coupling regime is realized; for systems with $p = 2$ and 3, a domain of nonuniversal critical behavior occurs at $0 \leq x_0 \leq x_c(d)$. For x_0 outside this interval, the system exhibits no critical behavior, as is the case at $p = 1$.

The values of threshold dimensions $d_c(p)$—$d_c(p = 1) = 3.986$, $d_c(p = 2) = 3.10$, and $d_c(p = 3) = 3.999$—which separate the domain of critical behavior with RSB effects $d_c(p) < d < 4$ from the domain where these effects are insignificant, simultaneously specify the lower bound of the domain where the results obtained by ε expansion are applicable to the description of the model of weakly disordered systems with RSB effects [4–6]. It is noted that calculations carried out in higher order approximations of the theory can significantly change the threshold dimension d_c for the XY model. On the other hand, changes in $d_c(p)$ for the Ising and Heisenberg models are expected to be small, which leaves the scope of the results obtained by the ε-expansion technique close to dimension four.

As the concentration of defects increases, one can expect a decrease in the threshold values d_c down to $d_c \leq 3$ beginning with a certain threshold concentration. In this case, the influence of replica symmetry-breaking effects can be significant. Due to specific features of the manifestation of RSB effects, the concentration n_s corresponding to the spin percolation threshold can play the role of the threshold concentration of defects for the Ising model, so that no stable critical behavior is observed at $n > n_s$. For the XY and Heisenberg models, this role can be played by the concentration of defects corresponding to the impurity percolation threshold $n_{\text{imp}} = 1 - n_s$ with a nonuniversal critical behavior for $n_{\text{imp}} < n < n_s$ and the absence of a stable critical behavior at $n > n_s$.

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research (project nos. 00-02-16455 and 02-02-06181) and by the Ministry of Education of the Russian Federation (project nos. E00-3.2-43 and UR.01.01.052).

REFERENCES

1. S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).
2. J. Emery, Phys. Rev. B **11**, 239 (1975).

3. G. Grinstein and A. Luther, Phys. Rev. B **13**, 1329 (1976).
4. Vik. S. Dotsenko, A. B. Harris, D. Sherrington, and R. B. Stinchcombe, J. Phys. A **28**, 3093 (1995).
5. Vik. S. Dotsenko and D. E. Feldman, J. Phys. A **28**, 5183 (1995).
6. Vik. S. Dotsenko, Usp. Fiz. Nauk **165**, 481 (1995) [Phys. Usp. **38**, 457 (1995)].
7. V. V. Prudnikov, A. V. Ivanov, and A. A. Fedorenko, Pis'ma Zh. Éksp. Teor. Fiz. **66**, 793 (1997) [JETP Lett. **66**, 835 (1997)]; V. V. Prudnikov, S. V. Belim, A. V. Ivanov, *et al.*, Zh. Éksp. Teor. Fiz. **114**, 972 (1998) [JETP **87**, 527 (1998)]; V. V. Prudnikov, P. V. Prudnikov, and A. A. Fedorenko, Zh. Éksp. Teor. Fiz. **116**, 611 (1999) [JETP **89**, 325 (1999)]; V. V. Prudnikov, P. V. Prudnikov, and A. A. Fedorenko, Phys. Rev. B **62**, 8777 (2000).
8. K. B. Varnashev and A. I. Sokolov, Fiz. Tverd. Tela (St. Petersburg) **38**, 3665 (1996) [Phys. Solid State **38**, 1996 (1996)]; A. I. Sokolov, K. B. Varnashev, and A. I. Mudrov, Int. J. Mod. Phys. B **12**, 1365 (1998); A. I. Sokolov and K. B. Varnashev, Phys. Rev. B **59**, 8363 (1999).
9. V. V. Prudnikov, P. V. Prudnikov, and A. A. Fedorenko, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 153 (2001) [JETP Lett. **73**, 135 (2001)].
10. V. V. Prudnikov, P. V. Prudnikov, and A. A. Fedorenko, Phys. Rev. B **63**, 184 201 (2001).
11. A. Pelissetto and E. Vicari, Phys. Rev. B **62**, 6393 (2000).
12. V. V. Prudnikov and A. N. Vakilov, Zh. Éksp. Teor. Fiz. **103**, 962 (1993) [JETP **76**, 469 (1993)].
13. G. Parisi, J. Phys. A **13**, 1101 (1980); G. Parisi, J. Phys. A **13**, L115 (1980); G. Parisi, J. Phys. A **13**, 1887 (1980); M. Mezard, G. Parisi, and M. Virasoro, *Spin-Glass Theory and Beyond* (World Sci., Singapore, 1987); Vik. S. Dotsenko, Usp. Fiz. Nauk **163** (6), 1 (1993) [Phys. Usp. **36**, 455 (1993)].
14. J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Clarendon, Oxford, 1996).
15. M. Dudka, Yu. Holovatch, and T. Yavorskii, J. Phys. Stud. **5**, 233 (2001).
16. J. C. LeGuillou and J. Zinn-Justin, Phys. Rev. B **21**, 3976 (1980).
17. A. Pelissetto and E. Vicari, cond-mat/0002402.

Translated by A. Klimontovich

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 560–571.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 647–659.

NONLINEAR PHYSICS

Fractal Diffusion in Smooth Dynamical Systems with Virtual Invariant Curves¶

B. V. Chirikov* and V. V. Vecheslavov**
Budker Institute of Nuclear Physics, Novosibirsk, 630090 Russia
**e-mail: chirikov@inp.nsk.su*
***e-mail: vecheslavov@inp.nsk.su*
Received April 5, 2002

Abstract—Preliminary results of extensive numerical experiments with a family of simple models specified by the smooth canonical strongly chaotic 2D map with global virtual invariant curves are presented. We focus on the statistics of the diffusion rate D of individual trajectories for various fixed values of the model perturbation parameters *K* and *d*. Our previous conjecture on the fractal statistics determined by the critical structure of both the phase space and the motion is confirmed and studied in some detail. In particular, we find additional characteristics of what we earlier termed the virtual invariant curve diffusion suppression, which is related to a new very specific type of critical structure. A surprising example of ergodic motion with a "hidden" critical structure strongly affecting the diffusion rate was also encountered. At a weak perturbation ($K \ll 1$), we discovered a very peculiar diffusion regime with the diffusion rate $D = K^2/3$ as in the opposite limit of a strong ($K \gg 1$) uncorrelated perturbation, but in contrast to the latter, the new regime involves strong correlations and exists for a very short time only. We have no definite explanation of such a controversial behavior. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION: VIRTUAL INVARIANT CURVES

In a two-dimensional map (2.1) that we study here, the diffusion crucially depends on the global invariant curves (GICs) that cut the 2D phase space of the motion (a cylinder, see the next section). Even a single such curve is sufficient to completely block the global diffusion in the action variable along the cylinder. As is well known by now, the existence of GICs depends not only on the perturbation strength but also on its smoothness. It is convenient to characterize the latter by the temporal Fourier spectrum of the perturbation. For an analytical perturbation, the Fourier amplitudes decay exponentially fast. In this case, the global diffusion sets up if the perturbation $\epsilon \gtrsim \epsilon_{cr}$ exceeds some critical value. Otherwise, the chaos remains localized within relatively narrow chaotic layers of nonlinear resonances. As a result, either the global diffusion is completely blocked by GICs or the rate of the diffusion and the measure of its domain decay exponentially in the parameter $1/\epsilon$ as $\epsilon \longrightarrow 0$ (the so-called Arnold diffusion; see, e.g., [1–3] for a general review).

By definition, the Hamiltonian of a smooth system has the power-law Fourier spectrum with a certain exponent $\beta + 1$ (see, e.g., [4] and references therein). In this case, the global diffusion is always blocked for some sufficiently small perturbation strength $\epsilon < \epsilon_{cr}(\beta)$ if the smoothness parameter $\beta > \beta_{cr}$ exceeds the critical value. This is similar to the case of an analytical Hamiltonian except that the critical perturbation now depends on the Hamiltonian smoothness ($\epsilon_{cr}(\beta) \longrightarrow 0$ as $\beta \longrightarrow \beta_{cr}$).

To the best of our knowledge, the strongest rigorous result is that $\beta_{cr} < 4$ for a 2D map as in this paper (see [5]). But a simple physical consideration [4] leads to an even smaller value $\beta_{cr} = 3$, which is still to be confirmed somehow, theoretically or numerically. In any event, the smoothness $\beta = 2$ of our model here is even less.

Until recently, the behavior of dynamical systems in the opposite case $\beta < \beta_{cr}$ of a poor smoothness remained rather vague. Even though most of the numerical data seemed to confirm the simplest behavior of some universal global diffusion (see, e.g., [6]), several counterexamples were also observed (see, e.g., [7, 8]).

In these counterexamples, some trajectories remained within a certain restricted part of the phase space for a sufficiently long computation time. No clear explanation of these strange events has yet been given.

Meanwhile, about 20 years ago (!) a number of mathematical studies revealed various possibilities for the existence of GICs in smooth systems with $\beta < \beta_{cr}$ (see, e.g., [8–10]). To us, the most comprehensive analysis of this problem was given by Bullett [9], who rigorously proved a strange survival of infinitely many GICs amid a strong local chaos. Surprisingly, all these interesting results remain essentially unknown, at least to physicists. Apparently, this is because the above mathematical papers were restricted (perforce!) to what could be done rigorously, that is, to the invariant curves

¶This article was submitted by the authors in English.

only, without any attempt to analyze very interesting and important transport processes such as diffusion. This is still within reach of physical analysis and numerical (or laboratory) experiments only. As a result, only after the recent accidental rediscovery of GICs in chaos by Ovsyannikov [11] (which is still unpublished; see [12, 13] for the full text of Ovsyannikov's theorem) have intense physical studies of this interesting phenomenon begun [12–16].

Interestingly, the authors of both [9] and [11] used exactly the same model, in which a strange locked-in trajectory was observed much earlier [7]. Apparently, this is because this model (a particular case of our model with the parameter $d = 1/2$; see Section 2) is the simplest one possessing those curious GICs (see [15] for discussion). Perhaps the main surprise was that the GICs include the separatrices of nonlinear resonances, which have always been considered as ones destroyed first by almost any perturbation. The principal difference is that the invariant curves, separatrices including, now exist for special values of the system parameters only (e.g., $K = K_m$).

Although there are infinitely many such special values of the parameter and infinitely many GICs such that a single GIG completely blocks the global diffusion for each of the parameter values, the probability of global diffusion (that is, the measure of such K values) is apparently zero. Therefore, a principal question to be answered is: What would be the behavior of that system for an arbitrary value of K? In [16], we conjectured that, even though the set of K_m is not everywhere dense [9] in general, the density of this set is rather high, and we can therefore expect some change (presumably suppression) of the diffusion for every K value compared to the "usual" (familiar) dynamical system. In other words, we hypothesized that the structure of the phase space and of the motion therein can be changed by the formation of GIG at a close K value even if no GICs occur for almost all K. This is why we now call such a neighbor-K invariant curve the virtual one (VIC) with respect to any K [16].

Preliminary numerical experiments presented in [16] did confirm our conjecture. These experiments were done by the prompt computation of the average diffusion rate $D(K)$ as a function of the parameter K in the domain with GICs, real or virtual ones. The experiments revealed a very strong suppression of the diffusion, up to many orders of magnitude, restricted only by the computation time. But even more interestingly, a very complicated (apparently fractal) structure of the dependence $D(K)$ was revealed. This seems to be a result of a very complicated structure of the model phase space itself. Preliminarily, it looks like the so-called critical structure (see, e.g., [4]), but a rather specific one due to a forest of VICs.

In the present paper, we begin the study of this seemingly new type of the critical structure. Specifically, we start with the investigation of the statistical properties of diffusion as one of the characteristic processes in chaotic motion.

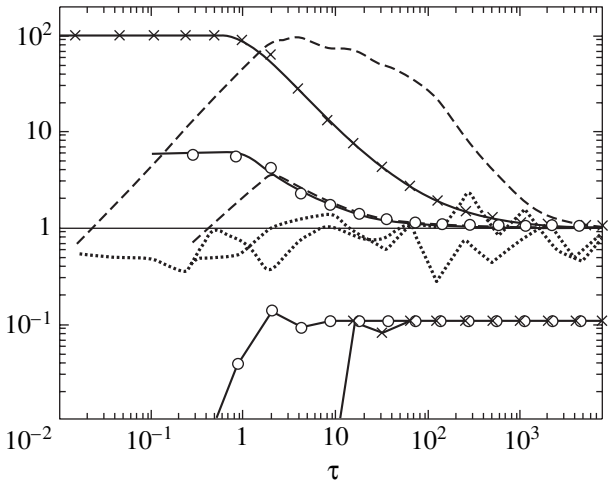


Fig. 1. The diffusion relaxation $D^*(\tau) = D(\tau)/D_\infty \longrightarrow 1$ in model (2.2) with the parameter $d = 0$ (without invariant curves) is presented as a function of the dimensionless time τ [Eq. (3.5)] for the two values $K = 0.01$ (circles) and $K = 3 \times 10^{-5}$ (crosses). Two smooth solid lines show empirical relation (3.5) with two fitting parameters $c = 1$ and $\gamma = 4$. Dashed lines are variances the $V_M(\tau)$ in Eq. (3.2), and dotted lines show the variances $V_N(\tau)$ in Eq. (3.8). In the lower part, the scaling in Eq. (3.6) is presented reduced by the factor 10 to avoid overlapping with other data. The full volume of empirical data is $J = M \times N = 10^4 \times 10 = 10^5$.

2. MODEL: THE SAME AGAIN

For the reader's convenience, we here repeat the description of the model in [15, 16]. In the canonical variables given by the action (momentum) p and the phase x, the model is specified by the map

$$\bar{p} = p + K f(x), \quad \bar{x} = x + \bar{p} \mod 1, \qquad (2.1)$$

where $K = \varepsilon > 0$ is the perturbation strength (not necessarily weak) and the "force" $f(x)$ is the antisymmetric piecewise linear "saw" of period 1 ($f(-y) = -f(y)$, $y = x - 1/2$). The phase space of the model is the cylinder $0 < x < 1$, $-\infty < p < +\infty$.

As in [15, 16], we actually consider a family of maps with another parameter d (see Fig. 1 in [15]) and the force

$$f(x) = \begin{cases} \dfrac{2x}{1-d}, & |x| \le \dfrac{1-d}{2} \\ -\dfrac{2y}{d}, & |y| \le \dfrac{d}{2}, \end{cases} \qquad (2.2)$$

where $y = x - 1/2$ and the second parameter d ($0 \le d \le 1$) is the distance between the two "teeth" of the saw $|f(x)| = 1$ at the points $y = y_\pm = \pm d/2$. The most studied particular case of the family corresponds to $d = 1/2$,

where the saw $f(x)$ with two teeth is symmetric. In the limit $d = 0$, the two teeth merge into one and all the invariant curves are destroyed. This was observed and explained in [15] for $K > 0$. In the opposite case, where $K < 0$ (which is equivalent to $K > 0$, $d = 1$), the dynamics of the model is completely different, and we do not consider it in this paper (see [15] for a brief discussion). In our 2D map (2.1), the GIC supports rotation of the phase x around the cylinder, which bars any motion in p over GICs. In contrast to this, a local invariant curve (LIC) surrounding, e.g., the domain of regular motion (see [4] and Section 5 below) corresponds to oscillation in the phase x, which allows other trajectories to bypass that obstacle.

The GICs, separatrices including, exist in the entire interval $0 < d < 1$, but for special K values only [9, 15, 16]. In particular, the invariant curves are completely absent [9] for sufficiently large parameter values

$$K > K_B(d) = \frac{2d^2}{1+d}, \quad 0 < d < 1. \tag{2.3}$$

If $K \gg K_B$ (see below), the physical quantity of main interest to us, the diffusion rate D, can be approximately calculated from the Fourier expansion of force (2.2) (see [16] for details)

$$f(x) = \sum_{n \geq 1} \frac{f_n}{n^\beta} \sin(2\pi n x), \tag{2.4}$$

where

$$f_n = -\frac{2}{\pi^2} \frac{\cos(n\pi)\sin(n\pi d)}{d(1-d)}, \quad \beta = 2. \tag{2.5}$$

In particular, in the limit $d = 0$,

$$f_n = -\frac{2}{\pi}\cos(n\pi), \quad \beta = 1, \tag{2.6}$$

the smoothness parameter β becomes less by one but both values are less than the critical one $\beta_{cr} = 3$.

The diffusion rate and other quantities are calculated using the standard analysis of nonlinear resonances and their interaction (overlap) (see, e.g., [1–3, 16]). The calculation is especially simple if we neglect the variation of the coefficients $|f_n| \approx \text{const}$ in (2.4). This simplification is exact for $d = 0$ [see (2.6)] and remains reasonably accurate [16] for

$$K \gtrsim 3K_B = \frac{6d^2}{1+d}. \tag{2.7}$$

The diffusion rate is then approximately given by a very simple standard relation

$$D(K) = \frac{\overline{(\Delta p)_t^2}}{t} \approx \frac{256}{\pi^5} K^{5/2} \approx 0.57 K^{5/2}, \tag{2.8}$$

where t is the motion time in map iterations and the parameter $K \ll 1$ is assumed to be sufficiently small. The latter expression in (2.8), which we use below, is the result of extensive numerical experiments in [6], also confirmed in [16] for $K \lesssim 0.1$ (see [16] and Section 3).

We note that the dependence $D(K) \propto K^{5/2}$ is different from the usual, or, better to say, the simplest, one $D(K) \propto K^2$. This is explained by the dynamical correlation of motion that is determined by the frequency of the phase oscillation on nonlinear resonances,

$$\Omega_n = \sqrt{\frac{2\pi K f_n}{n^{\beta-1}}} \approx 2\sqrt{K} \approx \Lambda_n(K) \ll 1, \tag{2.9}$$

where Λ_n stands for the Lyapunov exponent charaterizing the local exponential instability of the motion, which is the main criterion for dynamical chaos. We note that for $\beta = 1$, both Ω_n and Λ_n are independent of the Fourier harmonic number n. The exact value of the Lyapunov exponent in the limit $d = 0$ is given by

$$\Lambda = \ln(1 + K + \sqrt{2K + K^2}) \approx \sqrt{2K} \ll 1. \tag{2.10}$$

The latter expression is the approximation for small K [cf. Eq. (2.9)], which is sufficiently good within the region of Eq. (2.8) ($K \lesssim 0.1$) with the accuracy of ~1%. Because the time is discrete in our model (the number of the map iterations), both correlation characteristics, Eqs. (2.9) and (2.10), must be small, which implies the above restriction on the parameter K.

In the opposite limit $K \gg 1$, the correlation between successive x values is negligible, and we arrive at the "usual" relation for the diffusion rate,

$$D(K) = K^2 \int_0^1 f^2(x)dx = \frac{K^2}{3}, \tag{2.11}$$

which is independent of the parameter d. In the intermediate region ($K \sim 1$), the correlation causes a decaying oscillation (see [6]), which is beyond the scope of the present paper.

3. DIFFUSION WITHOUT ANY INVARIANT CURVES: AVERAGES AND MOMENTS

As mentioned above, there are no invariant curves for $d = 0$. Moreover, the motion is ergodic, which implies the simplest structure of the phase space (cf. Section 4 below). Therefore, this particular case is not of the main interest to us by itself. It is nevertheless a good introduction to our central problem considered in Section 6 below. A similar approach was taken in our previous paper [16].

We first consider the time dependence of the diffusion rate $D(K; t)$. The semicolon instead of the usual comma is intended to emphasize that this time dependence is not a real physical contribution to the diffusion but rather a combination of two different processes: the proper diffusion via accumulation of random perturba-

tion effects and a stationary regular oscillation of the diffusing variable (p in our case), which is a certain type of background for the diffusion. This phenomenon can be roughly represented by the simple relation

$$D(K; t) \sim D_\infty(K) + \frac{B(K)}{t}, \tag{3.1}$$

where $B(K)$ is some function of the perturbation (see, e.g., [16] and Eq. (3.5) below). In other words, in many cases, the present studies including, the nondiffusing stationary part can be separated from the diffusing part, thereby considerably simplifying the analysis of this complicated process. All this can be described, of course, via the standard method of the correlation of perturbation. But this would lead to much more intricate theoretical relations and, in addition, to much less information on the diffusion dynamics (see, e.g., [6]).

An example of the diffusion kinetics is presented in Fig. 1. The computation was done as follows. The number of trajectories $M \gg 1$ with random initial conditions homogeneously distributed within the unit area of the phase cylinder ($0 \le x_0 < 1, 0 \le p_0 < 1$) were run for a sufficiently long time with successive outputs at certain intermediate moments of time t as shown in Fig. 1. We recall that t is measured in the number of the map iterations. Each output includes the diffusion rate (D), averaged over all M trajectories and the dimensionless variance

$$V_M = \frac{\langle D^2\rangle - \langle D\rangle^2}{2\langle D\rangle^2}. \tag{3.2}$$

For the Gaussian distribution of the action p, this variance must be equal to unity. This is indeed the case for a sufficiently long motion time when the measured diffusion rate reaches its asymptotic value D_∞ in Eq. (3.1). A quite different dependence $V_M(t)$ for the previous smaller time is not surprising (nor is it very interesting) because $D(t)$ then depends on a completely different physical process that must be passed over.

A real surprise was the very beginning of the diffusion, the plateau in Fig. 1. This looks like a real diffusion unlike the following part of the stationary oscillation. Moreover, the diffusion rate $D_0 = K^2/3$ on the plateau is the maximum one, Eq. (2.11), as for $K \gg 1$. Another interesting observation is the duration of this strange diffusion,

$$t_0 \approx \frac{1}{\Lambda} \approx \frac{1}{\sqrt{2K}}, \tag{3.3}$$

which is close to the inverse Lyapunov exponent, the rise time of the local exponential instability of the underlying chaotic motion. The last but not the least curious property is the fast increase in variance (3.2),

$$V_M(t) \approx \frac{t}{3}, \quad 2 \le t \lesssim t_0, \tag{3.4}$$

as shown in Fig. 1. This is qualitatively different from the behavior of the same diffusion rate for $K \gg 1$ with the usual variance $V_M \approx 1$. The dynamical mechanism of this strange transitional diffusion is not completely clear and requires further studies. Apparently, it is somehow related to the main correlation (2.9) on dynamical scale (3.3). Although the initial "diffusion" is relatively fast, it lasts for only a short time, and the relative change of the initial distribution of trajectories

$$\frac{|\Delta p|}{|\Delta p|_0} \sim \sqrt{\frac{D_0}{\Lambda}} \sim K^{3/4} \ll 1$$

is therefore negligible for $K \ll 1$ unless the initial distribution $|\Delta p|_0 \lesssim K^{3/4}$ is very narrow. But in the latter case, the dependence $D(t)$ is very sensitive to the form of the initial distribution in p, as several of our preliminary numerical experiments reveal. The variance of $D(t)$ is especially strong for small $t \sim t_0$ in the region of that mysterious plateau but eventually decays as $t \longrightarrow \infty$, with the diffusion approaching its limit value D_∞. Apparently, this is related to a complicated fine structure of the phase space and/or of the motion correlations. This interesting question certainly deserves further studies, but in the present paper, we consider the simplest, homogeneous, distribution of the trajectory initial conditions on the phase cylinder.

In this particular case, a very simple and surprisingly accurate empirical relation for the diffusion time dependence has been found starting from the qualitative picture in (3.1). It is given by

$$D(t) \approx \frac{D_0 + \tau D_\infty}{(1+\tau^\gamma)^{1/\gamma}}, \quad \tau = c\Lambda t, \tag{3.5}$$

where τ is the dimensionless time with an empirical fitting parameter c that is very close to one. The second empirical parameter $\gamma \approx 4$ is less definite, but it affects the turn of the dependence $D(t)$ at $\tau \approx 1$ only. This relaxation of the diffusion rate has two time scales: the plateau

$$\tau_{\rm pl} = 1 \ \text{ or } \ t_{\rm pl} = 1/c\Lambda \approx 1/\sqrt{2K} \gg 1$$

and the relaxation

$$\tau_R = D_0/D_\infty \sim 1/\sqrt{K} \gg 1 \ \text{ or } \ t_R \sim 1/K,$$

which is much longer. Interestingly, the usual diffusion spreading of a very narrow initial p distribution on the relaxation time scale

$$|\Delta p|_R^2 = D_\infty t_R = \frac{D_\infty(D_0/D_\infty)}{c\Lambda} = \frac{D_0}{c\Lambda} = |\Delta p|_{\rm pl}^2$$

is exactly equal to the spreading on the plateau. Hence, the full relaxation spreading is twice as large, which is also directly seen from empirical relation (3.5),

$$|\Delta p|_R^2 = D(\tau_R)\frac{\tau_R}{c\Lambda} \approx \frac{D_0 + \tau_R D_\infty}{(1+\tau_R^\gamma)^{1/\gamma}}\frac{\tau_R}{c\Lambda} \sim K^{3/2} \ll 1,$$

and which is still much less than the unit p period.

In Fig. 1, empirical relation (3.5) is presented and compared with the numerical data in the dimensionless variables τ and $D^* = D/D_\infty$, where D_∞ is the asymptotic ("true") diffusion rate (2.8). In these variables, the curves with various K values are similar and converge in the limit as $\tau \longrightarrow \infty$.

Another interesting scaling can be done as follows. We calculate the diffusion rate $D_\infty(D(\tau)) = D_{\text{th}}$ from Eq. (3.5) and plot its ratio to the true rate in Eq. (2.8),

$$\frac{D_{\text{th}}}{D_\infty} \approx \frac{D(\tau)(1+\tau^\gamma)^{1/\gamma} - D_0}{\tau D_\infty} \approx 1. \tag{3.6}$$

Then, within the accuracy of scaling (3.5) and of fluctuations, this ratio must always be close to unity. This is indeed the case except on the plateau ($t \lesssim t_0$), where the rate $D(\tau)$ is almost independent of τ (see Fig. 1). The next important statistical property consists in fluctuations of the diffusion rate. One characteristic of these fluctuations is the dispersion of trajectories, which is characterized by the variance in Eq. (3.2). If all the trajectories were statistically independent, the dispersion of the mean diffusion rate would be

$$\left(\frac{\Delta\langle D\rangle}{\langle D\rangle}\right)^2 = \frac{2V_M}{M-1}. \tag{3.7}$$

By construction, the trajectories are indeed independent with respect to their initial conditions but not necessarily with respect to the corresponding diffusion rate. To verify this, we repeated the computation of the diffusion N times with new and independent initial conditions and then calculated the second (new) dimensionless variance for the average diffusion rate,

$$V_N = \left(\frac{\langle\langle D\rangle^2\rangle_N}{\langle\langle D\rangle\rangle_N^2} - 1\right)\frac{M-1}{2V_M} \approx 1. \tag{3.8}$$

Again, if Eq. (3.7) is valid, the variance V_N must be close to one.

The time dependence of both variances, $V_M(t)$ and $V_N(t)$, is shown in Fig. 1. Remarkably, their behavior is qualitatively different. The first variance $V_M(t)$ depends on the distribution function of p in the ensemble of trajectories, while the second variance $V_N(t)$ is affected by the statistical dependence (or independence) among trajectories for any distribution function. The results of our numerical experiments presented in Fig. 1 clearly demonstrate that the distribution in p quickly deviates from the Gaussian one during the diffusion on the plateau and returns only in the limit as $t \longrightarrow \infty$, when the diffusion rate $D \longrightarrow D_\infty$ approaches the asymptotic value without any nondiffusing part. Unlike this, the trajectories remain statistically independent during the entire process of the diffusion relaxation. We return to this interesting point in Section 7.

We now consider the most informative statistical characteristic, the distribution function $f(D)$ of the diffusion rate.

4. DIFFUSION WITHOUT INVARIANT CURVES: THE DISTRIBUTION FUNCTION

In the main part of our paper (Section 6), we are primarily interested in the distribution tail $D \longrightarrow 0$ of a very low diffusion rate. The shape of this tail is known to be an important characteristic of the critical structure of the motion (see, e.g., [4]). The first indications of such a structure in the presence of virtual invariant curves were observed in [16]. Here, we continue these studies.

Because the statistics of the far tail are always rather poor, we follow [16] in using a special version of the integral distribution

$$F(D) = \int_0^D f(D')dD' \approx \frac{j}{J}, \tag{4.1}$$

the so-called "rank-ordering statistics of extreme events" (see, e.g., [17]). The following simple ordering of the $D(j)$ values (events) of the diffusion rate is sufficient for this: $D(j+1) > D(j)$, $j = 1, 2, \ldots, J$. The integral probability is then approximately given by the ratio j/J, as shown in Eq. (4.1).

In computation, we typically ran M trajectories N times (see Section 3), and the maximum number of events therefore reached $J = M \times N = 10^4 \times 10 = 10^5$. To obtain the lowest possible D values and simultaneously minimize a rather big output, we ordered all the computed events but printed only J_0 of those, with $J_0 \ll J$, such that some (the smallest) D_j were obtained first, while the rest were printed on a logarithmic scale. An example of such a distribution is presented in Fig. 2 for $K = 0.001$ in the variables $D^* = D/\langle D\rangle$ and $F(D^*) = j/J$, where $\langle D\rangle$ is some average diffusion rate (see below). The upper distribution corresponds to a rather long motion time $t = 10^4 \gg 1/K$, with the mean diffusion rate already very close to the limit D_∞. For the lower distribution, $t = 10$ is very short and corresponds to the plateau.

At least in the former case, where the p distribution is Gaussian (see Section 3), the distribution

$$f(D) = \frac{\alpha^\lambda}{\Gamma(\lambda)} D^{\lambda-1} e^{-\alpha D} \tag{4.2}$$

is the so-called Pearson Γ distribution with the two moments

$$\langle D\rangle = \frac{\lambda}{\alpha}, \quad (\Delta D)^2 = \langle D^2\rangle - \langle D\rangle^2 = \frac{\lambda}{\alpha^2}, \tag{4.3}$$

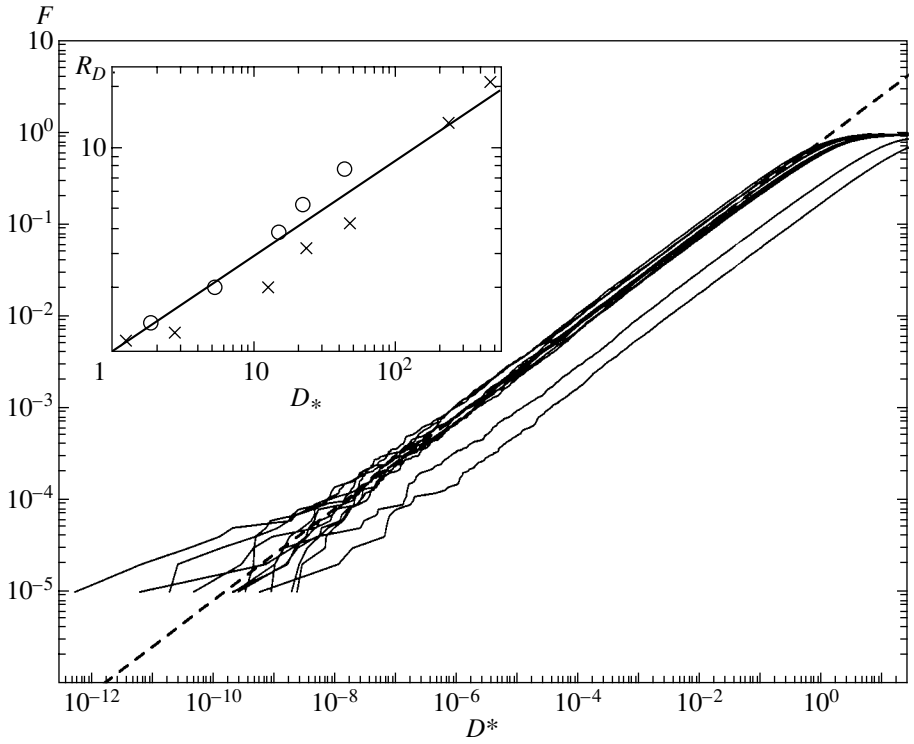


Fig. 2. The distribution function F [Eq. (4.1)] of the reduced diffusion rate D^* [Eq. (4.5)] in model (2.2) without invariant curves ($d = 0$). The thick dashed straight line represents asymptotic behavior (4.6) of the integrated D distribution (4.2) for the Gaussian p statistics. Two lower wiggly lines correspond to large deviations from the Gaussian statistics: $D_* = 42$ ($K = 10^{-3}$) and 461 ($K = 3 \times 10^{-5}$) (see insert). A group of ten D distributions in a large interval ($10 \le D_* \le 461$) are brought together using empirical relation (4.9). Insert: the shift factor R_D vs. the deviation D_* [Eq. (4.8)] for $K = 10^{-3}$ (circles) and 3×10^{-5} (crosses); the straight line is empirical relation (4.7).

which are the mean and the variance, respectively. For the Gaussian p distribution, the reduced variance in Eq. (3.2) becomes $V_M = 1$, and therefore,

$$\left(\frac{\Delta D}{\langle D\rangle}\right)^2 = \frac{1}{\lambda} = 2 \tag{4.4}$$

and $\lambda = 1/2$ is independent of α. Moreover, if we introduce the dimensionless diffusion rate

$$D \longrightarrow D^* = \frac{D}{D_\infty} \tag{4.5}$$

with the average $\langle D^*\rangle = 1$, we also obtain from Eq. (4.3) that $\alpha = \lambda = 1/2$. The new distribution then becomes

$$f(D^*) = \frac{(D^*)^{-1/2}\exp(-D^*/2)}{\sqrt{2\pi}}$$

and

$$F(D^*) = \int_0^{D^*} f(D')dD' \longrightarrow \sqrt{\frac{2}{\pi}D^*}, \tag{4.6}$$

where the latter expression gives the asymptotic behavior as $D^* \longrightarrow 0$ that we need. This asymptotic form is in very good agreement with the empirical data in Fig. 2 even at $D^* \approx 0.1$ (!). For very small D^*, the accuracy of the agreement is limited by the fluctuations caused by several remaining points. The smallest value $D^* = 8.3 \times 10^{-11}$ corresponds to the estimate $D^*_{\min} \sim 1/J^2 = 10^{-10}$.

Because the distribution $f(D^*)$ in (4.6) is also Gaussian in $\sqrt{D^*}$, the integral $F(D^*)$ admits a very simple approximation found in [18],

$$F(D^*) = \begin{cases} 1 - \dfrac{\exp(-D^*/2)}{\sqrt{D^*}+1}, & D^* > 1/2 \\ \sqrt{\dfrac{2D^*}{\pi}}, & D^* < 1/2. \end{cases} \tag{4.6a}$$

The relative accuracy $|\Delta F/F| < 0.05$ of this approximation is better than 5% in the entire range of F. Actually, the accuracy is even much better except in a narrow interval at $D^* \sim 1/2$.

Thus, the upper distribution in Fig. 2, which describes the real diffusion at a sufficiently long motion

time, is in a good agreement with the available theory. This is no longer the case for the lower distribution on the plateau. In itself, this is not a surprise, because, contrary to the previous case, the measured diffusion rate is mainly determined by nondiffusive processes. But a very interesting feature of this nondiffusive distribution is that the exponent of the power-law tail remains exactly the same as if the p distribution were a Gaussian one. The simplest explanation, quite plausible to us, is that the far tail still represents a distribution that is a part of the entire distribution according to our original picture expressed by estimate (3.1). One immediate inference is then the decrease in the tail probability if we use the same variable $D^* = D/D_\infty$. This is indeed the case according to the data in Fig. 2!

A more difficult problem is the quantitative estimate of the distribution shift for the motion time $t \lesssim 1/K$ with the ratio $\langle D^* \rangle = \langle D(t) \rangle / D_\infty > 1$. This shift can be characterized either via the probability decrease by R_F times for a fixed D^* or via the increase in D^* itself by R_D times for a fixed probability. We note that $R_D = R_F^2$ on the tail because of the square-root dependence in Eq. (4.6). The characteristic R_D seems more preferable to us because it describes the shift not only of the tail but also (qualitatively) of the entire distribution $F(D^*)$.

Having analyzed the data, we found the empirical relation for the tail shift,

$$R_D(D_*) \approx D_*^a, \tag{4.7}$$

where the new diffusion ratio is

$$D_*(\tau) \approx \frac{D_0}{\tau D_\infty} + 1 \tag{4.8}$$

and the fitted exponent is $a = 0.45$.

The philosophy behind this relation is as follows. We start with our original picture of a combined diffusive/nondiffusive process described by Eq. (3.1), which is almost our final choice (4.8). But at the beginning, we seemed to improve the original relation by including our surprising discovery, the plateau. Specifically, we tried to use Eq. (3.5), which is in good agreement with the empirical data, for the dependence $D(t)$ (see Fig. 1). We also found that it partly describes the distribution $F(D)$, except on that mysterious plateau! Our final step was then to return from (3.5) to a version of (3.1) in form (4.8).

Although it may have seemed strange, this did work with a reasonable accuracy, as the inset in Fig. 2 demonstrates. The question "why?" is still to be answered in further studies. This is actually a serious general problem of the dynamical mechanism underlying the plateau formation and statistics.

Our empirical relation (4.7) can be represented differently. Namely, instead of describing the actual distribution tail shifted with respect to the asymptotic form in Eq. (4.6), we can introduce the scaled diffusion rate

$$D \longrightarrow \frac{D}{R_D},$$

which implies that

$$D^* \longrightarrow \frac{D^*}{R_D}. \tag{4.9}$$

The result is shown in Fig. 2 as a beam of ten scaled distributions scattered around asymptotic line (4.6).

5. DIFFUSION AMID VIRTUAL INVARIANT CURVES: THE LYAPUNOV EXPONENTS

In the previous sections, we considered a very particular and simplest limiting case of our model (2.2) with the parameter $d = 0$. In this case, the motion is ergodic [6], which greatly simplifies the problem under consideration. Nevertheless, we obtained a number of new results that form a firm foundation for further studies.

The most important new feature of the motion for $d > 0$ is the so-called divided phase space of the system, that is, a mixture of both chaotic and regular components of the motion. This is a typical structure of dynamical systems with several degrees of freedom (see, e.g., [4]).

First of all, we must eliminate the regular trajectories from further analysis of the diffusion statistics. The standard well-known method to achieve this consists in simultaneously computing for each trajectory the so-called Lyapunov exponent Λ, which is the rate of the local exponential instability of the motion (see, e.g., [1–3] and references therein). A two-dimensional canonical (Hamiltonian) map such as our model (2.2) involves two Lyapunov exponents whose sum is always zero, $\Lambda_1 + \Lambda_2 = 0$. For a chaotic trajectory, one exponent, e.g., $\Lambda_1 = \Lambda_+ > 0$, is positive and the other is negative, $\Lambda_2 = \Lambda_- < 0$. As a result, in accordance with the standard definition of the Lyapunov exponent in the limit as $t \longrightarrow \infty$, any tangent vector (dx, dp) of the linearized motion approaches the eigenvector corresponding to $\Lambda_+ > 0$.

A simple well-known procedure for computing Λ_+ that we also use in the present work is as follows. For each of M trajectories with random initial conditions x_0 and p_0, we chose the tangent vector (dx, dp) of a random direction and the unit modulus, $d\rho^2 = dx^2 + dp^2 = 1$. Both maps, the main one and the one linearized with respect to the main reference trajectory $x(t, x_0, p_0)$, $p(t, x_0, p_0)$, were then run simultaneously during some time t. The current $\Lambda(t)$ was finally calculated from the standard relation

$$\Lambda(t) = \frac{\langle \ln \rho(t) \rangle}{t}, \tag{5.1}$$

where the brackets denote averaging over M trajectories. In contrast to the formal mathematical definition of Λ in the limit as $t \longrightarrow \infty$, the Lyapunov exponent $\Lambda(t)$ is always time-dependent, perforce, in numerical experiments.

In Fig. 3, several typical examples of the Λ distribution are depicted for the number of events in (4.1) $J = M$ equal to that of trajectories and with a smaller number of printed points $J_0 = M' \le M$ except in the case where $d = 0$. The simplest distribution is for the ergodic motion ($d = 0$). It has the form of an almost vertical step, whose derivative $dF/d\Lambda \sim 10^4$ is a very narrow δ function. We note that the regular chain of points along the F axis has no special physical meaning but simply reflects a particular type of distribution accepted, $F(\Lambda_j) = j/J$ with integer j [see (4.1)]. The mean value of Λ depends only on K [see Eq. (2.10)] but not on the initial conditions. This example in Fig. 3 shows the empirical/theoretical ratio, which is very close to unity, as expected.

The other two examples correspond to the same values $K = 0.45$ and $M = 10^4$ but different motion times $t = 10^4$ and 10^5 iterations. Both distributions have the same step at the largest Λ, which corresponds to diffusive components (not necessarily a single one) of the motion, similarly to the ergodic case. But the most interesting part is the rest of the distribution, which represents a rich motion structure, contrary to a dull one in the ergodic motion.

The largest (but not the most interesting) part of this structure is related to the steep distribution cutoff at small Λ. Comparison of the two distributions for different motion times $t = 10^4$ and 10^5 shows that, in this region, the Λ values of the trajectories decrease with increasing time approximately as $\Lambda \sim 1/t$. This means that all these trajectories are regular [see Eq. (5.1)] because the tangent vector ρ does not grow. The relative number of such trajectories gives the total area of regular motion on the phase cylinder of the system. In the example under consideration, it is given by $A_{\text{reg}} = 3177/10\,000 \approx 0.318$ ($t = 10^5$). Generally, this value depends on a particular choice of the cutoff border (see the arrow in Fig. 3). This delicate experimental problem is considerably mitigated by a fortunate feature of the Λ distribution in our model, namely, the occurrence of a relatively wide plateau of $F(\Lambda)$ immediately above the cutoff with only several trajectories on it. But the statistical accuracy

$$\frac{\Delta A_{\text{reg}}}{A_{\text{reg}}} \approx (MA_{\text{reg}})^{-1/2} \tag{5.2}$$

is typically much worse and can be improved by increasing the number of trajectories (and the computation time) only.

Another interesting feature of the Λ distribution in our model is a characteristic "fork" shape of the cutoff. This is a result of negative Λ for many regular trajectories. Such a peculiar representation is obtained by ordering $\Lambda(t)$ values with their signs but plotting the moduli $|\Lambda(t)|$ only. The lower prong of the fork therefore corresponds to $\Lambda(t) < 0$, while $\Lambda(t) > 0$ on the upper one. This is because of the complex-conjugate Lyapunov exponents, resulting in a strictly bounded oscillation of the tangent vector (dx, dp) in this case. However, the area $A_{\pm}$ [see Eq. (5.4) in what follows] is noticeably smaller than the total area of regular domains A_{reg}, $A_{\pm} \approx 0.20 < A_{\text{reg}} \approx 0.318$. The rest is filled with trajectories that are also regular but linearly unstable.

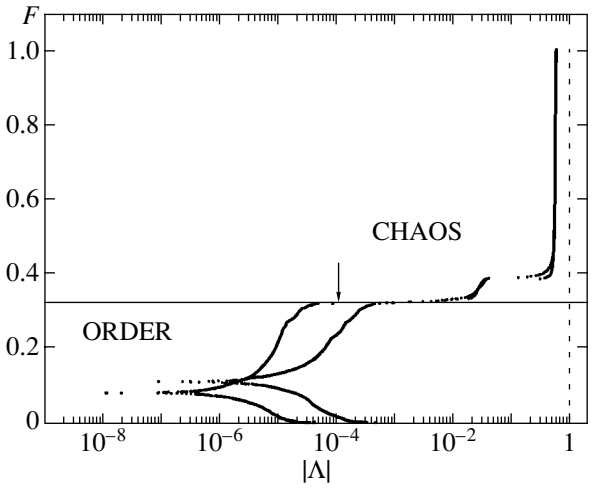


Fig. 3. Examples of the distribution function $F(\Lambda)$ of type (4.1) with the Lyapunov exponent in model (2.2) for $d = 0$, $M = M' = 80$, $t = 10^4$ (the rightmost step F(Λ), ergodic motion) and for $d = 1/2$, $M = 10^4$, $M' = 1000$, $t = 10^4$, 10^5 (nonergodic motion); in all cases, $K = 0.45$. The horizontal line indicates the total share $A_{\text{reg}} \approx 0.318$ of the motion regular components. The arrow at $\Lambda = 10^{-4}$ shows the lower border of chaotic trajectories chosen for further analysis (for $t = 10^5$).

This implies the linear growth of the tangent vector in time, $\rho(t) \sim t$, such that $\Lambda(t) \longrightarrow 0$ remains positive but vanishes in the limit as $t \longrightarrow \infty$. This is the so-called marginal local instability with both $\Lambda_{\pm} = 0$ equal zero (see [19] for a discussion). A curious point is that this seemingly exceptional case becomes the typical one in a nonlinear oscillator system because oscillation frequencies depend on the trajectory initial conditions. In fact, the bounded ρ oscillation producing negative $\Lambda(t)$ is the exceptional case. The origin of this peculiarity is in a piecewise linear force in our model (2.2). As a result, the motion in the main (and, for large K, the biggest) regular domain around the fixed point $x = 1/2$, $p = 0$ is precisely the harmonic oscillation with the frequency (for $K < d$)

$$\Omega = \arccos\left(1 - \frac{K}{d}\right) \approx 1.47, \tag{5.3}$$

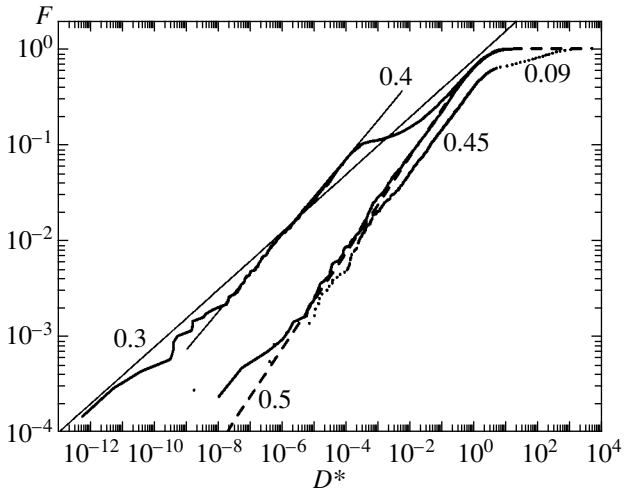


Fig. 4. Three characteristic examples of the diffusion statistics in the critical structure including virtual invariant curves ($d = 1/2$). Shown are the integral distributions F [Eq. (4.1)] of the normalized diffusion rate $D^* = D/D_{\text{norm}}$. The numbers at the curves are the critical diffusion exponents c_m. The largest one $c_0 = 0.5$ corresponds to the ergodic motion ($d = 0$) without any critical structure (the dashed curve). Two straight lines show the averaged ($c_1 = 0.3$) and local ($c_1' = 0.4$) critical exponents for $K = 0.45$ (the solid line connecting 500 values of $F(D^*)$). The distribution for $K = 0.335$ with two local critical exponents ($c_2 = 0.09$ and $c_2' = 0.45$) is represented by 300 points shifted to the right to avoid overlapping with the other two distributions. The third distribution (a solid line through 1000 points, $K = 0.3294$) is surprisingly close to that in the ergodic case (dashed line). In all three examples, $M = 10^4$, $t = 10^5$.

which remains the same in the entire regular domain of the area

$$A_{\pm} = \frac{2\pi K}{d} y_{\pm}^2 \left(1 - \frac{K}{2d}\right) \approx 0.20. \qquad (5.4)$$

Here, $y_{\pm} = x_{\pm} - 0.5 = \pm d/2$ is the position of two singularities of the force (see Eq. (2.2) and below) that restrict the size of the regular domain surrounded by the limiting ellipse to which both lines of the singularity $y_{\pm} = d/2 = 0.25$ are tangent. This ellipse is determined by the initial conditions

$$p_0 = 0, \quad x_0 = 0.5 + y_{\pm}\left(1 - \frac{K}{2d}\right) \approx 0.5 \pm 0.185. \qquad (5.5)$$

All the numerical values above correspond to $K = 0.45$ and $d = 1/2$. Within the ellipse, the motion of the tangent vector obeys the same equation as the main motion, the only difference being an arbitrary length ρ of the tangent vector (for details, see [3] and references therein).

Returning to Fig. 3, we note that the measured area $A_{\pm}$ decreases as the motion time increases. This is explained by the penetration of trajectories into a very complicated critical structure at the chaos border surrounding each regular domain (for details, see, e.g., [4]). For the same reason, the direct measurement of the entire regular region $A_{\text{reg}} \approx 0.40$ by a single chaotic trajectory for 10^9 iterations gives a noticeably larger value compared to $A_{\text{reg}} \approx 0.318$ obtained from 10^4 trajectories with 10^5 iterations each.

With all the curiosity of the $\Lambda(t)$ distribution being in regular components of the motion, our main interest in the present study is in the intermediate region between the regular cutoff at smallest $\Lambda(t) \longrightarrow 0$ and the chaotic step at maximum Λ independent of t. In this region, the distribution is also independent of the motion time and characterizes the proper critical structure of the chaotic motion. In the example in Fig. 3, this structure is represented by a relatively small probability step $\Delta F \approx 0.06$ at $\Lambda \approx 0.03$. Several other examples are also considered in the next section.

6. DIFFUSION AMID VIRTUAL INVARIANT CURVES: THE CRITICAL STATISTICS

In Fig. 4, we present three characteristic examples of the effect of the critical structure on the diffusion statistics. The dashed curve shows the "unperturbed" distribution $F(D^*)$ of the normalized diffusion rate $D^* = D/D_{\text{norm}}$ [see Eq. (4.1)], with the normalizing rate D_{norm} to be chosen in each particular case (see below). The term "unperturbed" refers to the ergodic case $d = 0$ without any invariant curves and critical structure (see Section 4; the problem of the critical structure in this case is not as simple as it may seem; see below and Section 7). The normalizing rate $D_{\text{norm}} = D_\infty$ is then the true asymptotic diffusion rate (4.5).

We are now interested in the effect of the critical structure that typically arises in a nonergodic motion with its barriers for the chaos, or chaos borders. The latter are a particular, and a very important, case of an invariant curve transformed into itself under the dynamics of the system. As discussed in Section 1, there are several different types of invariant curves.

One is the well-studied and rather familiar chaos border surrounding any domain with regular motion. In this paper, we call it the local invariant curve (LIC); it does not block the global diffusion around such a domain. An important property of a LIC is the robustness, which means that a small change of the system, e.g., of the parameter K or d, cannot destroy the LIC but can only deform it slightly. This implies that LICs are always present in any divided phase space.

Here, we are mainly interested in invariant curves of a different type, the global invariant curves. Each GIG cuts the entire phase-space cylinder (x mod 1) of our model and therefore completely prevents global diffusion in p. Such invariant curves are less known, especially the most surprising of them, the separatrix of a nonlinear resonance. But those GICs are not robust in the model under consideration (see [9]), being

destroyed by almost any arbitrarily small perturbation of the system, in particular by a change of even a single one of its parameters. In other words, such GICs exist only for special values, e.g., $K = K_m$. Although there are typically infinitely many such special values, the probability of finding a GIG in a randomly chosen system is zero. This is why we are interested in a more generic situation where our model has no GICs at all. But the effect of those still persists in a certain domain around each K_m! For this reason, we call such GICs virtual invariant curves (VICs) in analogy with other virtual quantities in physics, e.g., virtual energy levels in quantum mechanics. We note that unlike a GIG, a VIC is robust and, hence, generic.

Both LICs and GICs produce the so-called critical structure of motion (see, e.g., [4]), which is typically characterized by a power-law distribution of principal quantities. The corresponding exponents c_n are called critical exponents. Their values are shown in Fig. 4 at the related distributions. We note that the opposite is generally not true; that is, a particular power law does not necessarily indicate any critical structure. In our model, this is the case for the ergodic motion where the diffusion rate distribution is also characterized by an asymptotic ($D \longrightarrow 0$) power law with the exponent $c_0 = 0.5$ (see above and Section 7). An important difference between ergodic and nonergodic dynamics, however, is that, in the latter case, all the critical exponents $c_n < c_0$ are less than the (generally noncritical) ergodic exponent c_0. This is the main physical result of our preliminary numerical experiments that we can present and already discuss now (see Fig. 4).

We start with the distribution for $K = 0.45$ (the upper solid line), which is far in the region without VICs (the border of this region is at $K_B(d = 1/2) = 1/3$; see Eq. (2.3) above and [16]). But the regular trajectories ($A_{\text{reg}} \approx 0.318$) together with LICs and the related critical structure are present. As a result, the distribution (with $D_{\text{norm}} = D_\infty$) deviates considerably from the unperturbed one for the ergodic motion with $d = 0$. The critical structure of this type in a relatively narrow layer around a LIC is well studied by now (see, e.g., [4]), including the case where the typical distribution deviates from a pure power law. The latter would imply the exact scale invariance of the underlying critical structure in both the system phase space and its motion time.

The critical structure is described by the so-called renormalization group, or renormgroup for brevity. On the other hand, the equations of motion also form a certain (dynamical) group for any dynamical system. Such a fundamental similarity allows interpreting the critical structure as a certain dynamics, which was called the renormdynamics [4, 20]. In this picture, the exact scale invariance with a pure power-law distribution corresponds to the simplest, periodic renormdynamics, even though the original dynamics may be the most complicated chaotic motion. The resolution of this apparent paradox is that the complexity of the original dynamics is "transferred" to the dynamical infinite-dimensional space of the renormdynamics, leaving behind the simplest renormdynamics itself (sometimes!).

This case is best studied only because it is the simplest one. But the generic case is just the opposite—a typical renormchaos is also chaotic [20, 21]. This implies a certain chaotic oscillation of the characteristic distribution around some average power law. This is precisely the case for the upper distribution in Fig. 4. It is characterized by the average critical exponent $c_1 = 0.3$ with fluctuations of the order $c_1' - c_1 = 0.1$. Such an interpretation of the critical structure in question is known to be typical but not necessarily unique (see below). The truly unique property of this critical structure is the infinite power law, with or without fluctuations. The term "infinite" here corresponds to the range of a renormdynamical variable $\ln D \longrightarrow -\infty$ with an unrestricted variation, even though the diffusion rate itself $D > 0$ is strictly bounded from below.

This is no longer the case for the critical structure of a new type that we have encountered in our problem and which is produced by VICs (=robust GICs) rather than by robust LICs. As explained above, the principal difference between the two is that the VIC is not an invariant curve at all. In terms of renormdynamics, this implies that a VIC can mimic a GIC for relatively large $\ln D$ only. This is clearly seen in Fig. 4 in the upper part of the distribution with the local critical exponent $c_2 = 0.09$ and the parameter $K = 0.335$ (points). Here, we have taken $D_{\text{norm}} = 10^{-6} < D_\infty \approx 2 \times 10^{-5}$ much smaller than the true diffusion rate D_∞. This shifts the entire distribution to the right in order to avoid overlapping with other distributions. This value is slightly above the border $K_B(1/2) = 1/3$ [see Eq. (2.3)], where there are many VICs without any GIC. As a result, the range of the characteristic critical exponent c_2, $\Delta \ln D^* \approx 5$ is very short compared to the total available range ≈25. The rest of the distribution remains sufficiently close to the unperturbed one. This implies the absence of the critical structure or its sharp change at $\ln D \lesssim 2$ at least. With this interpretation, the renorm-motion stops in the specified region.

This in turn implies a "dissipative" rather than "Hamiltonian" renormdynamics. We note that the main part of the distribution is close but not identical to the unperturbed one because of a slight difference in the characteristic exponent. Whether this implies a certain very slow renorm-motion remains a very interesting open question. Interestingly, the larger critical exponent $c_2' = 0.45$ is also close to the local critical exponent $c_1' = 0.4$ in the region without VICs or GICs; above, it was interpreted as a random fluctuation in renormchaos. Whether this is indeed true remains unclear.

Finally, the third distribution in Fig. 4 (the lower solid line) actually coincides with the unperturbed distribution ($D_{\text{norm}} \approx D_\infty$), even though it corresponds to the

region with many VICs and a strong suppression of the diffusion ($K = 0.3294$—see Fig. 3 in [16]). A deviation for very small D^* is due to poor statistics at this end. We note that the coincidence of both distributions is not only asympotic (as $F \longrightarrow 0$), but also complete, including the opposite limit as $F \longrightarrow 1$. This occurs in spite of a rather large regular region $A_{reg} \approx 0.581$. The origin of this peculiarity for a particular K value remains unclear. One possibility is that the area of the critical structure at the chaos border around this regular domain is unusually small for some reason. Examples of such a peculiarity are known in different models (see [22]) where the critical structure was found to be unusually large but hidden. In other words, the motion was ergodic but with strong correlations [cf. the unusual diffusion rate in Eq. (2.8) for $K \ll 1$ in the ergodic system at $d = 0$]. Returning to this case in Fig. 4, we conclude that our "unperturbed" power-law distribution with the exponent $c_0 = 0.5$ (dashed line) may well represent a peculiar critical structure related to the strong hidden temporal correlations rather than to a purely spatial geometry of the phase space. If this is true, the correlation decay may indeed be not a power-law one, as is the case in the model in [22], where such a hidden decay is purely exponential (see Fig. 6 in [22]).

We finally mention another peculiarity of the critical structure in question: all the critical exponents found so far are smaller, albeit by a small amount, than the "unperturbed" or "hidden" one $c_0 = 0.5$. The physical meaning of this universal inequality is that the critical structure under consideration always increases the probability of a very low diffusion rate $D \longrightarrow 0$. The general mechanism of this effect is known (see, e.g., [4]) and is explained by the trajectory "sticking" within a complicated critical structure, which slows down the diffusion. Interestingly, the sign of the sticking effect can be opposite when the sticking accelerates the diffusion up to the absolute maximum $D(t) \propto t$ of the homogeneous diffusion rate [23, 24].

To summarize, we see that our "simple" model considered in this paper reveals a great variety of critical structures still to be further studied and understood.

7. CONCLUSION: A HIDDEN CRITICAL STRUCTURE?

In this paper, we present some preliminary results of the numerical experiments with a family of simple models specified by the smooth canonical 2D map (2.1) with global virtual invariant curves. As in [16], we use here the same strongly chaotic model and again focus on the statistics of the diffusion rate D, which proves to be of a very complicated (apparently fractal) type determined by the so-called critical structure of both the phase space and the motion (see, e.g., [4]). In [16], we studied the statistics of the mean diffusion rate $\langle D(K)\rangle$ averaged over the ensemble of trajectories with random initial conditions. Our main result there was the observation of very big and irregular fluctuations of the dependence $\langle D(K)\rangle$ and a long and very slowly decaying tail of the $\langle D\rangle$ distribution as $\langle D\rangle \longrightarrow 0$. We termed the latter effect the VIC diffusion suppression.

In the present paper, we continue studying this interesting phenomenon in more detail. For this, we pass from the statistics of averages $\langle D(K)\rangle$ as functions of the model parameter K to the statistics of individual trajectories for a given K. In principle, this approach provides the deepest insight into the statistical problem. As the main statistical characteristic, we have chosen the integral distribution $F(D)$ in form (4.1) for poor statistics as $\langle D\rangle \longrightarrow 0$. Preliminary results of our extensive numerical experiments presented in Fig. 4 confirm our earlier conjecture on a critical structure underlying the fractal dependence $\langle D(K)\rangle$ in [16], the true sign of such a structure being various power-law distributions found. Moreover, in addition to the familiar well-known critical structure exemplified in Fig. 4 by the case with the parameter $K = 0.45$, we observed many cases of a rather different structure, as the one with $K = 0.335$. The principal difference of the latter is its finite size in the structure variable $\Delta \ln D \lesssim 5$. A natural explanation of this difference is as follows. First, the VIC is not a true invariant curve like a GIC. The latter completely blocks the global diffusion, while the former can at most inhibit the diffusion only. The mechanism of inhibition is known to be the trajectory sticking inside a very complicated critical structure. The sticking is the stronger (longer), the smaller the spatial and/or the longer the temporal scale of the critical structure. But for the VIC structure, both scales are strictly limited. On the other hand, this restriction is the weaker, the higher the VIC density. In the system under consideration, the VIC density is rather large, and hence, the restriction leaves enough freedom for a strong suppression of the global diffusion for almost any K. Moreover, because the critical exponent of the VIC structure is typically very small (for example, $c_2 = 0.09$ in Fig. 4), the probability of large suppression is high even for a short critical structure (cf. [16] for a different characteristic of this phenomenon). This slowly decaying suppression probability is well ascertained in our numerical experiments, but we have no theoretical explanation of such behavior.

We now come to possibly the most interesting result of our current studies. Strange although it may seem, this brings us to the apparently simplest case of our model with $d = 0$, when the motion is ergodic. The problem is whether it can still reveal any structure on the grounds that the distribution $F(D)$ is also a power law (Fig. 4). This is certainly not the case if in addition $K \gg 1$ and the diffusion rate has standard form (2.11), $D \propto K^2$. But if $K \ll 1$, the diffusion rate becomes qualitatively different at least, $D \propto K^{5/2}$. This does not imply anything in general. But in the particular case under consideration, this dependence $D(K)$ can be, and actually was, derived [16] from the resonance structure of

motion. If the system were not ergodic (with a divided phase space), this structure would be clearly seen in the phase space. The question is what happens for the ergodic motion with the same dependence $D(K)$. In [16], we conjectured that some structure would persist in the form of correlations that determine the diffusion rate, which is in some "hidden" form and cannot be directly seen in the picture of the motion in phase space. An example of such a hidden critical structure was found in [22] (see Section 6). But in that case, a particular distribution function was exponential rather than a power-law one(?). Hence, the question is whether this qualitative difference can depend on a particular characteristic of the critical structure. Another question arises from a very strange temporal behavior of the diffusion rate in the same "simple" case of the ergodic motion for $d = 0$—a "mysterious" plateau at the very beginning of diffusion under a weak perturbation ($K \ll 1$; see Fig. 1). In this case, the dependence $D(K) = K^2/3$ is the same as in the opposite limit of strong ($K \gg 1$) uncorrelated perturbation(?) but for a very short time only, the shorter the stronger the perturbation(?!). Moreover, the correlations on the plateau not only are very large as in the weak-perturbation limit $K \longrightarrow 0$ but also increase during the entire plateau regime [see Fig. 1, dashed lines for the variances $V_M(\tau)$ in Eq. (3.2)]. At present, we have no definite explanation for this controversial behavior. A discreet current conjecture is as follows. The duration of the plateau is $t_{pl} \approx 1$, or $\tau_{pl} \approx 1/\Lambda \approx 1/\Omega$ [see Eq. (2.9)]. But the latter expression gives the phase oscillation period on the critical nonlinear resonance that determines the diffusion rate [16]. One can then imagine that this period characterizes not only the correlation decay, as usual, but also the correlation uprise. But, the invariable diffusion rate over the entire plateau region is yet to be explained.

ACKNOWLEDGMENTS

The authors are grateful to Ms. L.F. Hailo for her permanent and very important assistance in computer experiments. This work was partly supported by the Russian Foundation for Basic Research (project no. 01-02-16836) and by the complex program "Nonlinear Dynamics and Solitons" of the Russian Academy of Science.

REFERENCES

1. B. V. Chirikov, Phys. Rep. **52**, 263 (1979).
2. G. M. Zaslavsky and R. Z. Sagdeev, *Introduction to Nonlinear Physics* (Nauka, Moscow, 1988).
3. A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics* (Springer-Verlag, New York).
4. B. V. Chirikov, Chaos, Solitons, and Fractals **1**, 79 (1991).
5. J. Moser, *Stable and Random Motion in Dynamical Systems* (Princeton Univ. Press, Princeton, 1973).
6. I. Dana, N. Murray, and I. Percival, Phys. Rev. Lett. **62**, 233 (1989).
7. B. V. Chirikov, E. Keil, and A. Sessler, J. Stat. Phys. **3**, 307 (1971).
8. M. Hénon and J. Wisdom, Physica D (Amsterdam) **8**, 157 (1983).
9. S. Bullett, Commun. Math. Phys. **107**, 241 (1986).
10. M. Wojtkowski, Commun. Math. Phys. **80**, 453 (1981); Ergodic Theory Dyn. Syst. **2**, 525 (1982).
11. L. V. Ovsyannikov, private communication (1999).
12. V. V. Vecheslavov, nlin.CD/0005048.
13. V. V. Vecheslavov, Zh. Éksp. Teor. Fiz. **119**, 853 (2001) [JETP **92**, 744 (2001)].
14. V. V. Vecheslavov, Preprint No. 99-69 (Budker Institute of Nuclear Physics, Novosibirsk, 1999).
15. V. V. Vecheslavov and B. V. Chirikov, Zh. Éksp. Teor. Fiz. **120**, 740 (2001) [JETP **93**, 649 (2001)].
16. V. V. Vecheslavov and B. V. Chirikov, Zh. Éksp. Teor. Fiz. **122**, 175 (2002) [JETP **95**, 154 (2002)].
17. D. Sornette, L. Knopoff, Y. Kagan, and C. Vanneste, J. Geophys. Res. **101**, 13 883 (1996).
18. B. V. Chirikov and O. V. Zhirov, nlin.CD/0102028.
19. G. Casati, B. V. Chirikov, and J. Ford, Phys. Lett. A **77**, 91 (1980).
20. B. V. Chirikov and D. L. Shepelyansky, Physica D (Amsterdam) **13**, 395 (1984).
21. S. Ostlund, D. Rand, J. Sethna, *et al.*, Physica D (Amsterdam) **8**, 303 (1983).
22. B. V. Chirikov, Preprint 1999-7 (Budker Institute of Nuclear Physics, Novosibirsk, 1999).
23. B. V. Chirikov and D. L. Shepelyansky, Phys. Rev. Lett. **82**, 528 (1999).
24. B. V. Chirikov, Zh. Éksp. Teor. Fiz. **119**, 205 (2001) [JETP **92**, 179 (2001)].

Journal of Experimental and Theoretical Physics, Vol. 95, No. 3, 2002, pp. 572–580.
From Zhurnal Éksperimental'noĭ i Teoreticheskoĭ Fiziki, Vol. 122, No. 3, 2002, pp. 660–669.

NONLINEAR PHYSICS

Approximate Analytical Solutions of the Baby Skyrme Model¶

T. A. Ioannidou[a, *], V. B. Kopeliovich[b, **], and W. J. Zakrzewski[c, *]**

[a]*Institute of Mathematics, University of Kent, Canterbury, CT2 7NF, UK*
**e-mail: T.Ioannidou@ukc.ac.uk*
[b]*Institute for Nuclear Research, Russian Academy of Sciences, Moscow, 117312 Russia*
***e-mail: kopelio@al20.inr.troitsk.ru, kopelio@cpc.inr.ac.ru*
[c]*Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, UK*
****e-mail: W.J.Zakrzewski@durham.ac.uk*

Received March 1, 2002

Abstract—We show that many properties of the baby skyrmions, which have been determined numerically, can be understood in terms of an analytic approximation. In particular, we show that the approximation captures properties of the multiskyrmion solutions (derived numerically) such as their stability towards decay into various channels, and that it is more accurate for the "new baby Skyrme model" describing anisotropic physical systems in terms of multiskyrmion fields with axial symmetry. Some universal characteristics of configurations of this kind are demonstrated that are independent of their topological number. *© 2002 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

It is known that the two-dimensional O(3) σ model [1] possesses metastable states that can shrink or spread out under perturbation because of the conformal (scale) invariance of the model [2–4]. This implies that the metastable states can be of any size, and therefore, a term of the fourth order in derivatives, the so-called Skyrme term, must be added to break the scale invariance of the model. But the resulting energy functional has no minima, and a further extra term is needed to stabilize the size of the corresponding solitons; this term contains no derivatives of the field and is often called the potential (or mass) term. The field can then be viewed as the magnetization vector of a two-dimensional ferromagnetic substance [1], and the potential term describes the coupling of the magnetization vector to a constant external magnetic field. Because the extra terms contribute to the masses of the solitons, their dependence deviates from a simple law in which the skyrmion mass is proportional to the skyrmion (topological) number and the two-skyrmion configuration becomes stable, showing that the model possesses bound states [5].

In this paper, we demonstrate that the simple analytical method used for the description of the three-dimensional Skyrme model presented in [6] can also be used to study various properties of the low-energy states of the corresponding two-dimensional σ model when the parameters that determine the contributions of the Skyrme and the potential terms are not large. More precisely, it was possible to describe the basic properties of the three-dimensional skyrmions for large baryon numbers analytically [6], and it is therefore worthwhile to derive such a description for the two-dimensional O(3) σ model as well. In general, such analytical discussions of soliton models are useful because they lead to a better understanding of the soliton properties. The two-dimensional O(3) σ model is widely used to describe ferromagnetic systems, high-temperature superconductivity, etc., and the results obtained here can therefore be useful for the understanding of these phenomena.

Our method is based on the ansatz introduced in [6] and is accurate for the so-called "new baby Skyrme model" [7] that describes anisotropic physical systems. Its accuracy actually increases as the skyrmion number n increases, and this method allows predicting some universal properties of the ringlike configurations for large n, independently of its particular value. Although such models are not integrable, the new baby Skyrme model appears to have the properties of an integrable system in the case where n is large.

2. NEAR THE NONLINEAR O(3) σ MODEL

The Lagrangian density of the O(3) σ model with the additional terms introduced and discussed in [5, 7, 8] is[1]

$$\mathcal{L} = \frac{g^2}{2}(\partial_\alpha \mathbf{n})^2 - \frac{1}{4e^2}[\partial_\alpha \mathbf{n}, \partial_\beta \mathbf{n}]^2 - g^2 V. \qquad (1)$$

¶This article was submitted by the authors in English.

[1] The first several paragraphs of this section follow [5, 8] very closely and are included to make the paper more self-contained.

Here, $\partial_\alpha = \partial/\partial x^\alpha$; x^α, $\alpha = 0, 1, 2$, refer to both time and spatial components of t, x, y); and the field $\mathbf{n}$ is a scalar field with three components n_a, $a = 1, 2, 3$, satisfying the condition

$$\mathbf{n}^2 = n_1^2 + n_2^2 + n_3^2 = 1.$$

The constants g and e are free parameters, with g^2 having the dimension of energy. It is useful to think of g^2 and $1/ge$ as natural units of energy and length, respectively. The first term in (1) is familiar from σ models, the second term, which is of the fourth order in derivatives, is the analogue of the Skyrme term, and the last term is the potential term. The respective potentials for the old baby Skyrme model (OBM) and the new baby Skyrme model (NBM) describing anisotropic systems are given by

$$\begin{aligned} V_{OBM} &= \mu^2(1-n_3), \\ V_{NBM} &= \frac{1}{2}\mu^2(1-n_3^2), \end{aligned} \tag{2}$$

where μ has the dimension of energy, and $1/\mu$ therefore determines a second length scale in our model. Evidently, $V_{NBM} \le V_{OBM}$ at a fixed value of μ.

In three spatial dimensions, the Skyrme term is necessary for the existence of soliton solutions, but the inclusion of a potential is optional from the mathematical standpoint. Physically, however, a potential of a certain form is required in order to give the pions a mass [9]. By contrast, in two dimensions, a potential term must be included in the above Lagrangian in order for soliton solutions to exist. As shown in [10], the different potential terms give quite different properties to the multiskyrmion configurations when the skyrmion number is large. Our analytical treatment here supports this conclusion, as shown in Sections 3–5.

We are only interested in configurations with finite energy, and we therefore define the configuration space to be the space of all maps $\mathbf{n}$: $R^2 \longrightarrow S^2$ that tend to the constant field (0, 0, 1) (the so-called vacuum) at spatial infinity,

$$\lim_{|x|\to\infty} \mathbf{n}(\mathbf{x}) = (0, 0, 1). \tag{3}$$

Every configuration $\mathbf{n}$ can thus be regarded as a representative of a homotopy class in $\pi_2(S^2) = \mathbf{Z}$ and has the corresponding integer degree given by

$$\deg[\mathbf{n}] = \frac{1}{8\pi}\int d^2x \epsilon^{bc}\mathbf{n}(\partial_b\mathbf{n}\times\partial_c\mathbf{n}). \tag{4}$$

The vacuum field is invariant under the symmetry group

$$G = E_2 \times SO(2)_{\text{iso}} \times P,$$

where E_2 is the Euclidean group of two-dimensional translations and rotations, acting on fields via pullback. The $SO(2)_{\text{iso}}$ subgroup of the three-dimensional rotation group acting on S^2 is the subgroup that leaves the vacuum invariant (we call its elements isorotations to distinguish them from rotations in physical space). Finally, P is a combined reflection in both space and the target space S^2.

We are interested in stationary points of $\deg[\mathbf{n}] \neq 0$; the maximal subgroups of G under which such fields can be invariant are labeled by a nonzero integer n and consist of spatial rotations by some angle $\alpha \in [0, 2\pi]$ and simultaneous isorotation by $-n\alpha$. Fields that are invariant under such a group are of the form

$$\begin{aligned} n_1 = \sin f(\tilde{r})\cos(n\phi), \quad n_2 = \sin f(\tilde{r})\sin(n\phi), \\ n_3 = \cos f(\tilde{r}), \end{aligned} \tag{5}$$

where $(\tilde{r}, \phi)$ are polar coordinates and $f(\tilde{r})$ is the profile function. Such fields are the analogues and generalizations of the hedgehog fields in the Skyrme model. In this parametrization, which involves azimuthal symmetry of the fields, it is assumed that all the skyrmions sit on top of each other in forming the multiskyrmion configuration.

It is easy to show that the degree of field (5),

$$\deg[\mathbf{n}] = n, \tag{6}$$

is equal to the azimuthal winding number n.

The respective static energy functionals related to Lagrangian (1) for the OBM and the NBM are given by

$$\begin{aligned} E_{\text{cl}}(n)_{OBM} = \frac{g^2}{2}\int r dr\Bigg(f'^2 + \frac{n^2\sin^2 f}{r^2} \\ + a\left[\frac{n^2 f'^2 \sin f^2}{r^2} + 2(1-\cos f)\right]\Bigg), \end{aligned} \tag{7}$$

$$\begin{aligned} E_{\text{cl}}(n)_{NBM} = \frac{g^2}{2}\int r dr\Bigg(f'^2 + \frac{n^2\sin^2 f}{r^2} \\ + a\left[\frac{n^2 f'^2 \sin f^2}{r^2} + (1-\cos^2 f)\right]\Bigg). \end{aligned} \tag{8}$$

In (7) and (8), the length $(\sqrt{ge\mu})^{-1}$ is absorbed such that the scale size of the localized structures is a function of the dimensionless spatial coordinate $r = \sqrt{ge\mu}\tilde{r}$ and the dimensionless parameter $a = \mu/ge$ becomes the only nontrivial parameter of the model. Finiteness of the energy functional requires that the profile function must satisfy the boundary conditions $f(0) = \pi$ and $f(\infty) = 0$.

Setting $\phi = \cos f$ in (7), we rewrite the energy functional as

$$E_{\rm cl}(n)_{OBM} = \frac{g^2}{2}\int r dr \times\left(\frac{\phi'^2}{1-\phi^2} + \frac{n^2(1-\phi^2)}{r^2} + a\left[\frac{n^2\phi'^2}{r^2} + 2(1-\phi)\right]\right) \quad (9)$$

and similarly for $E_{\rm cl}(n)_{NBM}$. We next parametrize the field ϕ using the ansatz introduced in [6] for the description of the three-dimensional skyrmions,

$$\phi = \cos f = \frac{(r/r_n)^p - 1}{(r/r_n)^p + 1}, \quad \phi' = \frac{p}{2r}(1-\phi^2). \quad (10)$$

After the integration with respect to r, this leads to the analytic expressions for the energy

$$E_{\rm cl}(n)_{OBM} = \pi g^2\left(\frac{4n^2}{p} + p + \frac{4a\pi}{p\sin(2\pi/p)}\left[\frac{n^2(p^2-4)}{3r_n^2 p} + r_n^2\right]\right), \quad (11)$$

$$E_{\rm cl}(n)_{NBM} = \pi g^2\left(\frac{4n^2}{p} + p + \frac{4a\pi}{p\sin(2\pi/p)}\left[\frac{n^2(p^2-4)}{3r_n^2 p} + \frac{2}{p}r_n^2\right]\right). \quad (12)$$

Here, p and r_n are parameters that still must be determined by minimizing the energy. In fact, r_n corresponds to the radius of the n-soliton configuration. We remark that in deriving (11) and (12) we used the Euler-type integrals (see also [6])

$$\int_0^\infty \frac{2r dr}{1+(r/r_n)^p} = \frac{2\pi r_n^2}{p\sin(2\pi/p)}, \quad p>2,$$

$$\int_0^\infty \frac{dr(r/r_n)^p}{r[1+(r/r_n)^p]^2} = \frac{1}{p}, \quad p>0,$$

$$\int_0^\infty \frac{dr(r/r_n)^{2p}}{r^3[1+(r/r_n)^p]^4} = \frac{\pi(p^2-4)}{3r_n^2 p^4\sin(2\pi/p)}, \quad p>1, \quad (13)$$

$$\int_0^\infty \frac{2r dr}{[1+(r/r_n)^p]^2} = \left(1-\frac{2}{p}\right)\frac{2\pi r_n^2}{p\sin(2\pi/p)}, \quad p>1.$$

It can be easily proved that the minimization of the energies in Eqs. (11) and (12) implies that

$$(r_n^{\min})^2_{OBM} = \frac{n}{\sqrt{3}}\sqrt{\frac{p^2-4}{p}}, \quad (r_n^{\min})^2_{NBM} = n\sqrt{\frac{p^2-4}{6}}, \quad (14)$$

i.e., $((r_n^{\min})^2_{NBM} = \sqrt{p/2}\,(r_n^{\min})^2_{OBM}$, and the minimum energy values are therefore equal to

$$E_{\rm cl}(n)_{OBM} = 4\pi g^2\times\left[\frac{n^2}{p} + \frac{p}{4} + \frac{2an\pi}{\sqrt{3}p\sin(2\pi/p)}\frac{\sqrt{p^2-4}}{\sqrt{p}}\right], \quad (15)$$

$$E_{\rm cl}(n)_{NBM} = 4\pi g^2\times\left[\frac{n^2}{p} + \frac{p}{4} + \frac{2\sqrt{2}an\pi}{\sqrt{3}\sin(2\pi/p)}\frac{\sqrt{p^2-4}}{p^2}\right]. \quad (16)$$

It is obvious that the energy contributions of the Skyrme and the potential terms are equal due to (14), which is in agreement with the result obtained from Derrick's theorem. Equations (15) and (16) provide an upper bound for the energies of baby skyrmions for any value of p. To obtain the lowest upper bound, we must minimize the right-hand sides of (15) and (16) with respect to the parameter p. In what follows, we investigate various cases that correspond to different values of the only nontrivial parameter of the model, a.

We first consider the case where $a \ll 1$, i.e., the model parameter is very small. We observe that, for $a = 0$, ansatz (10) is a solution of the model for $p = 2n$, which implies that $p \longrightarrow 2n$ as $a \longrightarrow 0$. In accordance with (14), the radius of the multiskyrmion configuration then increases with n,

$$(r_n^{\min})^2_{OBM} \propto n^{3/2}, \quad (r_n^{\min})^2_{NBM} \propto n^2.$$

Moreover, the configuration consists of a ring of the thickness $\delta \approx 4r_n/p$, and therefore

$$\delta_{OBM} \propto 2n^{-1/4}, \quad \delta_{NBM} \propto \text{const}.$$

We remark that the ring thickness is determined as the difference of the values of ϕ inside (which is equal to -1) and outside (which is equal to $+1$) the ring (i.e., $d\phi = 2$) divided by its derivative at $r = r_n$, where $\phi(r_n) = 0$; as a consequence of (10); $\phi'(r_n) = p/2r_n$.

Magnetic solitons of this type have been observed in [11, 12] as solutions of the Landau–Lifshitz equations defining the dynamics of ferromagnets. (We note that the static solutions of the baby Skyrme model and the Landau–Lifshitz equations are related.) In general, $\phi(r)$ given by (10) for $p = 2n$ is a low-energy approxi-

mation of multiskyrmion configurations (for $n > 1$), because the corresponding energies given by (15) and (16) are infinite for $n = 1$. Indeed, it is a matter of simple algebra to show that

$$\begin{aligned} E_{\rm cl}(n=2)_{OBM} &= 4\pi g^2(2+a\pi),\\ E_{\rm cl}(n=2)_{NBM} &= 4\pi g^2\left(2+\frac{a\pi}{\sqrt{2}}\right),\\ E_{\rm cl}(n=3)_{OBM} &= 4\pi g^2\left(3+a\pi\frac{8}{3\sqrt{3}}\right),\\ E_{\rm cl}(n=3)_{NBM} &= 4\pi g^2\left(3+a\pi\frac{8}{9}\right),\\ E_{\rm cl}(n=4)_{OBM} &= 4\pi g^2(4+a\pi\sqrt{5}),\\ E_{\rm cl}(n=4)_{NBM} &= 4\pi g^2\left(4+a\pi\frac{\sqrt{5}}{2}\right). \end{aligned} \tag{17}$$

For large n, the energies take the asymptotic values

$$\begin{aligned} E_{\rm cl}(n)_{OBM} &= 4\pi n g^2\left(1+\sqrt{\frac{2n}{3}}a\right),\\ E_{\rm cl}(n)_{NBM} &= 4\pi n g^2\left(1+\sqrt{\frac{2}{3}}a\right). \end{aligned} \tag{18}$$

We note that the energy of the OBM per unit skyrmion number increases as n increases, while the energy of the NBM per skyrmion decreases as n increases and becomes constant for large n. In fact, the energies given by (17) are the upper bounds of the multiskyrmion energies because the exact profile function corresponding to the minimum of the energy differs from that given by (10).

3. PERTURBATION THEORY FOR THE MODEL PARAMETER

In this section, we obtain energy corrections up to the second or higher orders with respect to the model parameter a. The corresponding energies for the OBM and NBM can be written as

$$E_{\rm cl}(n) = 4\pi g^2[f(p)+ah(p)], \tag{19}$$

where $f(p)$ and $h(p)$ can be evaluated from (15) and (16), respectively. Letting $p = 2n + \epsilon$ and expanding energies (15) and (16) up to the second order in ϵ, we obtain $f(p) = n + \epsilon^2/8n$, $h(p) = h_0 + \epsilon h_1$, where

$$h_1 = (2n)^{-1}\beta h_0.$$

In fact, the corresponding functions for the OBM and the NBM are given by

$$\begin{aligned} \frac{h_{0_{OBM}}}{n} &= \sqrt{\frac{2n}{3}}\frac{\pi}{n\sin(\pi/n)}\sqrt{1-1/n^2},\\ \beta_{OBM} &= \frac{\pi}{n}\cot(\pi/n)-\frac{1}{2}+\frac{1}{n^2-1}, \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{h_{0_{NBM}}}{n} &= \sqrt{\frac{2}{3}}\frac{\pi}{n\sin(\pi/n)}\sqrt{1-1/n^2},\\ \beta_{NBM} &= \frac{\pi}{n}\cot(\pi/n)-1+\frac{1}{n^2-1}. \end{aligned} \tag{21}$$

Minimization of (19) with respect to ϵ implies that

$$\epsilon^{\min} = -4anh_1 = -2a\beta h_0.$$

At large values of n, the parameters ϵ and $p = 2n + \epsilon$ take the values

$$\begin{aligned} \epsilon(n)_{OBM} &\approx -an\sqrt{\frac{2n}{3}},\\ \epsilon(n)_{NBM} &\approx 2a\sqrt{\frac{2}{3}}\frac{\pi^2/3-1}{n}, \end{aligned} \tag{22}$$

$$\begin{aligned} p(n)_{OBM} &\approx 2n-an\sqrt{\frac{2n}{3}},\\ p(n)_{NBM} &\approx 2n+2a\sqrt{\frac{2}{3}}\frac{\pi^2/3-1}{n}. \end{aligned} \tag{23}$$

For any a, the effective power $p(n)_{OBM}$ becomes negative as n increases and the approach based on the assumption that ϵ_{OBM} is small is not self-consistent (also see the next section). On the contrary, for the NBM, $p(n)_{NBM} \approx 2n$ as n increases, which implies that our consideration is self-consistent in this case. In terms of (19)–(21), the energy per skyrmion of the n-skyrmion configuration takes the value

$$\frac{E_{\rm cl}(n)}{4\pi g^2 n} = 1+a\frac{h_0}{n}-a^2\frac{h_0^2\beta^2}{2n^2}, \tag{24}$$

which gives

$$\begin{aligned} \frac{E_{\rm cl}(2)_{OBM}}{4\pi g^2 2} &= 1+1.5708a-0.034a^2,\\ \frac{E_{\rm cl}(2)_{NBM}}{4\pi g^2 2} &= 1+1.1107a-0.2741a^2,\\ \frac{E_{\rm cl}(3)_{OBM}}{4\pi g^2 3} &= 1+1.6120a-0.068a^2,\\ \frac{E_{\rm cl}(3)_{NBM}}{4\pi g^2 3} &= 1+0.9308a-0.0317a^2, \end{aligned}$$

Table 1. Energy per unit skyrmion number (in $4\pi g^2$) for different values of the parameter a for the OBM with second-order corrections in a taken into account

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 8$
$a = 0.001$	1.0063	1.00157	1.0016	1.0017	1.0019	1.0021	1.0023
$a = 0.01$	1.0384	1.0157	1.0161	1.0176	1.0191	1.0206	1.0234
$a = 0.0316$	1.0933	1.0496	1.0508	1.0553	1.0601	1.0649	1.0737
$a = 0.1$	1.2227	1.1567	1.1605	1.1737	1.1882	1.2025	1.2291
$a = 0.316$	1.5113	1.4930	1.5026	1.5358	1.5638	1.6126	1.6835
$a_{\text{hed}} = 0.316$ (num)	1.5647	1.4681	1.4901	1.5284	1.5692	1.6092	1.6832
$a = 0.316$ (num)	1.564	1.468	1.460	1.450	1.456	1.449	–

Note: The last two lines contain the exact results obtained from the numerical simulations of the respective multiskyrmions with ringlike shapes ($n \geq 2$) and with shapes other than ringlike ($n \geq 3$) [10]. In the first case, we have numerically solved the equations using the hedgehog ansatz (5).

Table 2. Energy per unit skyrmion number for different values of the parameter a for the NBM

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 8$	$n = 12$	$n = 16$
$a = 0.01$	1.0363	1.0111	1.0093	1.0088	1.0085	1.0084	1.0083	1.0082	1.0082
$a = 0.0316$	1.0851	1.0348	1.0294	1.0277	1.0270	1.0266	1.0262	1.0260	1.0259
$a = 0.1$	1.1887	1.1083	1.0928	1.0877	1.0855	1.0843	1.0831	1.0823	1.0820
$a = 0.316$	1.3814	1.3238	1.2912	1.2768	1.2699	1.2662	1.2626	1.2602	1.2593
$a = 0.4213$	1.44	1.4193	1.3865	1.3684	1.3597	1.3549	1.3501	1.3467	1.3455
$a = 0.4213$ (num)	1.564	1.405	1.371	1.358	1.352	1.349	1.3447	1.3407	1.3385

Note: The last line contains the exact results determined by the numerical simulations [10] of multiskyrmions with ringlike shapes for $a = 0.4213$, which coincide with ours for $n \leq 6$.

$$\begin{aligned}
\frac{E_{\text{cl}}(4)_{OBM}}{4\pi g^2 4} &= 1 + 1.7562a - 0.191a^2,\\
\frac{E_{\text{cl}}(4)_{NBM}}{4\pi g^2 4} &= 1 + 0.8781a - 0.0084a^2,\\
\frac{E_{\text{cl}}(5)_{OBM}}{4\pi g^2 5} &= 1 + 1.9122a - 0.302a^2,\\
\frac{E_{\text{cl}}(5)_{NBM}}{4\pi g^2 5} &= 1 + 0.8552a - 0.0032a^2,\\
\frac{E_{\text{cl}}(6)_{OBM}}{4\pi g^2 6} &= 1 + 2.0649a - 0.404a^2,\\
\frac{E_{\text{cl}}(6)_{NBM}}{4\pi g^2 6} &= 1 + 0.8430a - 0.0015a^2.
\end{aligned} \tag{25}$$

For large n, the energies in Eq. (24) take the asymptotic values

$$\begin{aligned}
\frac{E_{\text{cl}}(n)_{OBM}}{4\pi g^2 n} &= \left(1 + a\sqrt{\frac{2n}{3}} - a^2\frac{n}{12}\right),\\
\frac{E_{\text{cl}}(n)_{NBM}}{4\pi g^2 n} &= \left(1 + a\sqrt{\frac{2}{3}} - a^2\frac{(\pi^2/3-1)^2}{3n^4}\right).
\end{aligned} \tag{26}$$

We note that the energies of the two models behave differently when we consider terms of the second order in the model parameter, i.e., the terms $\sim a^2$ Indeed, for the OBM, the contribution to the energy is linearly proportional to the skyrmion number n, while, for the NBM, the contribution decreases rapidly as the skyrmion number increases. This implies that the linear approximation in a is accurate for the NBM because the quadratic term becomes negligible for large n. Numerical results obtained for different values of a for the OBM and NBM are presented in Tables 1 and 2, respectively.

As we have noted previously, our method cannot describe the one-skyrmion configuration because the corresponding energies become infinite. But setting $p = 2 + \varepsilon$ in (15) and (16) and expanding all terms up to the third order in $\varepsilon \ll 1$, we obtain

$$\begin{aligned}
&E_{\text{cl}}(n = 1)\\
&= 4\pi g^2\left(1 + \frac{\varepsilon^2}{8} - \frac{\varepsilon^3}{16} + 2a\sqrt{\frac{2}{3\varepsilon}}(1 - \gamma\varepsilon)\right),
\end{aligned} \tag{27}$$

where γ takes a different value for each of the two models,

$$\gamma_{OBM} = \frac{1}{8}, \quad \gamma_{NBM} = \frac{3}{8}. \tag{28}$$

We note that, with the terms of only up to the second order in ε considered, the corresponding energy in Eq. (27) simplifies to

$$E_{\rm cl} = 4\pi g^2\left(1 + \frac{\varepsilon^2}{8} + 2a\sqrt{\frac{2}{3\varepsilon}}\right),$$

and the minimum occurs at

$$\varepsilon_1 = 2\left(\frac{a}{\sqrt{3}}\right)^{2/5}.$$

Finally, the minimum of (27) occurs at

$$\varepsilon^{\rm min} = 2\left(\frac{a}{\sqrt{3}}\right)^{2/5}\left[1 + \frac{4}{5}\left(\frac{a}{\sqrt{3}}\right)^{2/5}\left(\gamma + \frac{3}{4}\right)\right] \qquad (29)$$

and corresponds to a shift of ε_1 because higher order corrections in ε were considered in (27). The energy of the one-skyrmion configuration is

$$\frac{E_{\rm cl}(n=1)}{4\pi g^2}$$

$$= \left\{1 + \frac{5}{2}\left(\frac{a}{\sqrt{3}}\right)^{4/5}\left[1 - \frac{1}{5}\left(\frac{a}{\sqrt{3}}\right)^{2/5}(8\gamma + 1)\right]\right\} \qquad (30)$$

$$\approx [1 + 1.611a^{4/5}(1 - 0.1605a^{2/5}(8\gamma + 1))].$$

Equation (30) implies that, for a single skyrmion, the energy expansion in a is proportional to a power of a instead of being linearly proportional to a (which is the case for the multiskyrmion configurations with $n \geq 2$), while its convergence is worse than for multiskyrmions, especially for the NBM. In fact, for $a = 0.4213$, the first two terms in (30) are equal to 1.807, and the next-order term decreases this value to 1.44, which gives an error of 7% compared to the exact value 1.564 obtained from numerical simulations. We note that our one-skyrmion parameterization gives the same energy for both models if only the expansions up to the lowest order in a are considered: the difference appears only in the term $\sim a\gamma\sqrt{\varepsilon}$ in (27).

It is clear from the results in Tables 1 and 2 that our approximate method gives energy values that are very close to the exact values obtained by numerical simulations, especially for the NBM. In particular, the difference between the exact and the approximate energies for $a = 0.4213$ is less than 0.5% for $n \geq 6$. For smaller values of a, the agreement between analytical and numerical results is even better. In evident agreement with (2), the energies of the NBM skyrmions given in Table 2 are smaller than those of the OBM skyrmions (see Table 1) at the same values of the model parameters.

We note that, for the OBM (when a is small), the energy per skyrmion of a multiskyrmion configuration with $n \geq 2$ is smaller compared to the single skyrmion energy, and therefore, these configurations are bound states, stable with respect to the decay into n individual skyrmions. On the contrary, the ringlike OBM multiskyrmions with even n (where $n \geq 4$) are unstable with respect to the decay into two-skyrmion configurations, while configurations with odd n (where $n \geq 5$) are unstable with respect to the breakup into two- and three-skyrmion configurations. In addition, Table 1 and Eq. (30) show that, for any $n \neq 1$, there is an upper limit for the model parameter, $a \leq a_{\rm cr}(n)$, above which the ringlike n-skyrmion configuration can decay into n individual skyrmions.

We now consider the case where $n = 3$ in more detail. As can be observed from the energies in Eqs. (17) and (30), the ringlike three-skyrmion configuration is stable with respect to the decay into a single and a two-skyrmion configuration for $a \leq 0.77$ because

$$E_1 + E_2 - E_3 \approx 1.611a^{4/5} - a\pi\left(\frac{8}{3\sqrt{3}} - 1\right), \qquad (31)$$

and this difference becomes positive if and only if

$$a \leq \left(\frac{3\sqrt{3}1.611}{\pi(8 - 3\sqrt{3})}\right)^5 \approx 0.77. \qquad (32)$$

For the skyrmion configurations with $n = 1, 2, 3$, corrections to the energy of the higher order in a lead to smaller critical values $a_{\rm cr}(n)$.

Because our fields with axial symmetry (5) and (10) correspond to ringlike solutions of the Euler–Lagrange equations [1] for $a = 0$, they must also be solutions of the corresponding equations as $a \longrightarrow 0$, i.e., when a takes values in a small region close to zero. (In fact, this region actually becomes narrower as n increases because the expansion in a becomes less convergent in this limit.) On the other hand, the latticelike configurations (tripole for $n = 3$, quadrupole for $n = 4$, etc.) are solutions of the equations when $a \geq a_{\rm cr}(n)$ for given n [5, 10, 13]. But the transition from the ringlike configuration to any other minimum energy configuration is a phenomenon that has not been studied in much detail and deserves further investigation.

Finally, it should be stressed that, in contrast to the linear approximation, the quadratic approximation given by (25) does not provide an upper bound for the energy.

4. AWAY FROM THE NONLINEAR O(3) σ MODEL

In the general case, for arbitrary values of the parameter a and the skyrmion number n, soliton solutions can be obtained by numerically minimizing the energy in Eqs. (15) and (16) with respect to the variable p. This leads to an upper bound for the corresponding energies because the profile function is given by (10).

For large a at fixed n (or for large n at fixed a), expansion (20) is not self-consistent for the OBM. But

some analytical results can also be obtained in this case because, for large a, Eq. (15) can be approximated by

$$E_{\rm cl}(n)_{OBM} \approx 4\pi g^2 \frac{2an\pi}{\sqrt{3}p\sin(2\pi/n)}\frac{\sqrt{p^2-4}}{\sqrt{p}}. \qquad (33)$$

The expansion of (33) up to second-order terms with respect to p gives

$$E_{\rm cl}(n)_{OBM} \approx 4\pi g^2 \frac{an\sqrt{p}}{\sqrt{3}}\left(1+\frac{c_2}{p^2}\right), \qquad (34)$$

where

$$c_2 = 2(\pi^2/3-1);$$

its minimization implies that

$$p_{\min} \approx \sqrt{3c_2} = 3.71$$

and the corresponding energy is therefore given by

$$\frac{E_{\rm cl}(n)_{OBM}}{4\pi g^2} \approx \frac{4}{3}an\left(\frac{c_2}{3}\right)^{1/4} = 1.48an. \qquad (35)$$

We note that, in contrast with the results obtained near the nonlinear σ model, the parameter p is independent of the skyrmion number n for large a. For $a \gg n$, the skyrmion radius is proportional to the square root of the skyrmion number, $r_n \propto n^{1/2}$; the skyrmion thickness is given by

$$\delta \propto r_n/p \propto n^{1/2},$$

and therefore, the ringlike structure of the configuration is not very pronounced. Direct numerical minimization of (33) with respect to p gives $p_{\min} = 4.5$, and the corresponding value of the energy is

$$\frac{E_{\rm cl}(n)_{OBM}}{4\pi g^2} = 1.55an. \qquad (36)$$

The energy obtained by solving the Euler–Lagrange equation numerically is [8]

$$\frac{E_{\rm cl}(n)}{4\pi g^2} = 1.333an.$$

The profile function corresponding to this solution is given by

$$\cos f = \frac{r^2}{8n^2}(r_n^2-r^2)+2\frac{r^2}{r_n^2}-1 \quad \text{for} \quad r \le r_n$$

and

$$f = 0 \quad \text{for} \quad r > r_n.$$

This solution is quite different from our parameterization (10), and the 16% difference between the exact and the approximate solutions is therefore understandable.

To conclude, we recall that, for the NBM, parameterization (10) works well for arbitrarily large n and its accuracy increases with increasing n, as illustrated in Table 2.

5. PROPERTIES OF THE SKYRMIONS: MEAN SQUARE RADII, ENERGY DENSITY, AND MOMENT OF INERTIA

Many properties of multiskyrmions can be determined using ansatz (10). For example, the mean square radius of the n-skyrmion configuration takes the simple form

$$\langle r^2\rangle_n = \frac{1}{2}\int dr r^2\phi' = \frac{2\pi r_n^2}{p\sin(2\pi/p)}, \qquad (37)$$

where r_n is given by (14) for the OBM and NBM. For small a, it was shown in Section 2 that $p = 2n$, implying that the mean square radius becomes

$$\langle r^2\rangle_{n_{OBM}} \approx \frac{\pi\sqrt{2(n^2-1)}}{\sin(\pi/n)\sqrt{3n}},$$
$$\langle r^2\rangle_{n_{NBM}} \approx \frac{\pi\sqrt{2(n^2-1)}}{n\sin(\pi/n)\sqrt{3}}, \qquad (38)$$

which takes the respective values π, $8\pi/3\sqrt{3}$, $\pi\sqrt{5}$, … and $\sqrt{2}\pi$, $8\pi/3$, $2\pi\sqrt{5}$, … for $n = 2, 3, 4, \ldots$.

For the NBM, even for a sufficiently large value of the parameter a, analytical formula (14) with the power p taken from (23) gives the values of $\langle r^2\rangle_{n_{NBM}}$, in remarkably good agreement with those obtained in numerical calculations. For example, the analytical result for $n = 3$, $a = 0.4213$ is $\sqrt{\langle r^2\rangle_3} = 2.987$ (in natural units of the model, $1/ge\mu$), to be compared with 2.872 obtained numerically. This agreement improves with increasing n, and we have $\sqrt{\langle r^2\rangle_{12}} \sim 10.92$ for $n = 12$, to be compared with 10.85 determined numerically. A similar agreement between analytical and numerical results takes place for the mean square radius of the energy distribution of multiskyrmions (the 3D case was considered in detail in [6]).

We note that the one-skyrmion configuration is (still) a singular case because (37) is undefined for $n = 1$. But as we have shown earlier, by expressing $p = 2 + \varepsilon$ and expanding (14) in ε, we obtain $r_{n=1}^2 = \sqrt{2\varepsilon/3}$, which leads to

$$\langle r^2\rangle_1 = 2\sqrt{\frac{2}{3\varepsilon^{\min}}} \qquad (39)$$

for $\varepsilon^{\min}$ given by (29). Our approximate method therefore shows that, as the model parameter tends to zero,

the mean square radius of the one-skyrmion field tends to infinity because

$$\langle r^2\rangle_1 \sim a^{-1/5};$$

on the other hand, because

$$\langle r^2\rangle_{NBM}(n) = \sqrt{n}\langle r^2\rangle_{OBM}(n),$$

the mean square radius is given by (39) for both models in this case.

The average energy density per unit surface element is defined as

$$\rho_E = \frac{E_{\rm cl}(n)}{2\pi r_n \delta} \tag{40}$$

with $\delta \approx 2r_n/n$ [see discussion after (16)]. For the NBM, when n is large, (40) takes the constant value

$$\rho_{E_{NBM}} \approx e\mu g^3\left(\sqrt{\frac{3}{2}} + a\right), \tag{41}$$

i.e., is independent of n. Equation (41) therefore represents the fundamental property of multiskyrmions of this type. On the contrary, for the OBM with ringlike configurations (which do not correspond to the minimum of the energy [5, 10]) taken into account, the energy density increases with n as $\sqrt{n}$ for small values of a.

Another quantity of physical significance determining the quantum corrections to the energy of skyrmions is the moment of inertia; it has been considered for two-dimensional models in [13]. To obtain the energy quantum correction of the soliton, due to its rotation around the axis perpendicular to the plane in which the soliton is located, we must take the t-dependent ansatz of the form

$$\begin{aligned} n_1 &= \sin f(\tilde r)\cos[n(\phi - \omega t)],\\ n_2 &= \sin f(\tilde r)\sin[n(\phi - \omega t)],\\ n_3 &= \cos f(\tilde r). \end{aligned} \tag{42}$$

The ω dependence of the energy is then given by the simple formula

$$E^{\rm rot} = \frac{\Theta_J}{2}\omega^2, \tag{43}$$

where Θ_J, the so-called moment of inertia, is given by [13]

$$\Theta_J(n) = g^2 n^2 \int d^2 r \sin^2 f(1 + a f'^2). \tag{44}$$

Using (10) and the relations

$$\frac{1}{4}\int(1-\phi^2)r\,dr = \int_0^\infty \frac{(r/r_n)^p r\,dr}{[1+(r/r_n)^p]^2} = \frac{2\pi r_n^2}{p^2\sin(2\pi/p)}, \quad p > 2, \tag{45}$$

$$\frac{1}{16}\int(1-\phi^2)^2\frac{dr}{r} = \int_0^\infty \frac{(r/r_n)^{2p}dr}{[1+(r/r_n)^p]^4 r} = \frac{1}{6p}, \quad p > 0,$$

we find that, at large values of n, the moment of inertia simplifies to

$$\Theta_J(n) \approx 4\pi g^2 n r_n^2\left(\frac{2n}{p} + \frac{anp}{3r_n^2}\right), \tag{46}$$

which holds for any multiskyrmion configuration described by ansatz (10), for both models. For small values of a, letting $p = 2n$ and taking r_n^2 given by (14), we find that

$$\begin{aligned} \Theta_J(n)_{OBM} &\approx 4\pi g^2 n r_n^2\left(1 + a\sqrt{\frac{2n}{3}}\right),\\ \Theta_J(n)_{NBM} &\approx 4\pi g^2 n r_n^2\left(1 + a\sqrt{\frac{2}{3}}\right), \end{aligned} \tag{47}$$

which implies that, for large n, the moment of inertia is

$$\Theta_J(n) \approx E_{\rm cl}(n) r_n^2, \tag{48}$$

in agreement with simple semiclassical arguments for the thin massive ring. Similar semiclassical formulas have been obtained for the three-dimensional skyrmions (see, e.g., [6, 14]), and the moment of inertia was shown to be given by

$$\Theta_J = 2M_B r_B^2/3$$

for large baryon numbers; this expression is valid for a classical spherical bubble with the mass concentrated in its shell.

6. CONCLUSIONS

We have presented an analytical approach for deriving approximate expressions of skyrmion solutions in the two-dimensional O(3) σ model. These approximations are very accurate for small values of the parameter a that determines the weight of the Skyrme term and the potential term in the Lagrangian. For other values of the model parameter, we have performed some numerical calculations and then combined them with further analytical work to investigate the binding and other properties of multiskyrmion states.

Two models have been studied: the old baby Skyrme model and the new baby Skyrme model, which differ from each other in the form of potentials (2). For both models, the a dependence of the energy of a single skyrmion differs from the cases where topological

number $n \geq 2$. For the OBM, when a is small, the $n = 3$ skyrmion configuration is stable with respect to the decay into a single skyrmion and a two-skyrmion configuration, while the ringlike multiskyrmion configurations with $n \geq 4$ are unstable with respect to the breakup into two- and three-skyrmion configurations. For the NBM, on the other hand, the hedgehog multiskyrmion configurations considered in [10] and here describe bound states, because the energy per skyrmion decreases as the skyrmion number increases. We note that the results obtained for the NBM are similar to the ones obtained for the three-dimensional model studied in [6]. In both cases, the energy per skyrmion decreases as the skyrmion number increases. The three-dimensional skyrmions obtained using the rational map ansatz [15] for large n have the form of a bubble with the energy and the baryon number concentrated in the shell. The thickness and the energy density of the shell (which is analogous to the thickness of the ring in the two-dimensional case) are independent of the skyrmion number [6]. Similarly, in this paper we have shown that, for large n, the two-dimensional baby skyrmions of the NBM correspond to ringlike configurations with a constant thickness and a constant energy density per unit surface of the ring. The building material for these objects is a band of matter with a constant thickness and the average energy density per unit surface. The baby skyrmions can therefore be obtained as dimensional reductions of the three-dimensional skyrmions at large n; the three-dimensional skyrmions can be derived from the two-dimensional baby skyrmions as dimensional extensions.

It was concluded in [8] that the Casimir energy, or quantum loop corrections, can destroy the binding properties of the two-skyrmion bound states. The validity of this argument for the two- and three-skyrmion bound states of the NBM would be worth investigating. Another interesting problem is to determine to what extent the region of sufficiently small a is of importance from the standpoint of physics. For large a, the method overestimates the skyrmion masses for the OBM but is accurate for the NBM, especially for large n.

The existence of bound states of the three-dimensional skyrmions has rich phenomenological consequences in elementary particles and nuclear physics. It suggests possible existence of multibaryons with nontrivial flavor, strangeness, charm, or beauty; more details are given in [14] and references therein. Similarly, the existence of bound states of two-dimensional baby skyrmions with universal properties in the NBM, which describes anisotropic systems, can also have some consequences for the condensed state physics, which would be worth investigating in detail.

ACKNOWLEDGMENTS

V.B.K. is indebted to G. Holzwarth for drawing his attention to the paper [8]; his work is supported by the Russian Foundation for Basic Research (project no. 01-02-16615). T.I. thanks the Nuffield Foundation for a newly appointed lecturer award.

REFERENCES

1. A. A. Belavin and A. M. Polyakov, Pis'ma Zh. Éksp. Teor. Fiz. **22**, 503 (1975) [JETP Lett. **22**, 245 (1975)].
2. R. A. Leese, M. Peyrard, and W. J. Zakrzewski, Nonlinearity **3**, 387 (1990).
3. B. M. A. G. Piette and W. J. Zakrzewski, Nonlinearity **9**, 897 (1996).
4. T. A. Ioannidou, Nonlinearity **10**, 1357 (1997).
5. B. M. A. G. Piette, B. Schroers, and W. J. Zakrzewski, Z. Phys. C **65**, 165 (1995).
6. V. B. Kopeliovich, Pis'ma Zh. Éksp. Teor. Fiz. **73**, 667 (2001) [JETP Lett. **73**, 587 (2001)]; hep-ph/0109229; J. Phys. G **28**, 103 (2002).
7. A. E. Kudryavtsev, B. M. A. G. Piette, and W. J. Zakrzewski, Nonlinearity **11**, 783 (1998).
8. H. Walliser and G. Holzwarth, hep-ph/9907492.
9. G. S. Adkins and C. R. Nappi, Nucl. Phys. B **233**, 109 (1984).
10. T. Weidig, hep-th/9811238; Nonlinearity **12**, 1489 (1999); T. Weidig, hep-th/9911056.
11. A. M. Kosevich, B. A. Ivanov, and A. S. Kovalev, Phys. Rep. **194**, 117 (1990).
12. B. M. A. G. Piette and W. J. Zakrzewski, Physica D (Amsterdam) **119**, 314 (1998).
13. B. M. A. G. Piette, B. Schroers, and W. J. Zakrzewski, Nucl. Phys. B **439**, 205 (1995).
14. V. B. Kopeliovich, Zh. Éksp. Teor. Fiz. **120**, 499 (2001) [JETP **93**, 435 (2001)].
15. C. J. Houghton, N. S. Manton, and P. M. Sutcliffe, Nucl. Phys. B **510**, 507 (1998).